A local global infrastructure for Autonomous Vehicle Development

Gilles TOURPE, gtourpe@amazon.com, HPC Business Development Executive
Agenda

- Autonomous Vehicle (AV) Landscape & Challenges
- Cloud for AV Development
- Case Studies & References
Autonomous Systems & Machine Learning at Amazon

Twenty years of innovation

- Delivery Robots
- Fulfillment Automation & Inventory Management
- Drones
- Voice-driven Interactions
- Inventing New Customer Experiences
## AV Industry Overview

<table>
<thead>
<tr>
<th>CASE</th>
<th>Connected</th>
<th>Autonomous</th>
<th>Shared</th>
<th>Electrified</th>
</tr>
</thead>
<tbody>
<tr>
<td>% new cars 2015</td>
<td>8%</td>
<td>0%</td>
<td>2%</td>
<td>0.1%</td>
</tr>
<tr>
<td>% new cars 2030</td>
<td>100%</td>
<td>35%</td>
<td>25%</td>
<td>20%</td>
</tr>
</tbody>
</table>

- **Connected**: 8% in 2015, 100% in 2030
- **Autonomous**: 0% in 2015, 35% in 2030
- **Shared**: 2% in 2015, 25% in 2030
- **Electrified**: 0.1% in 2015, 20% in 2030
Autonomous Driving: Challenges & Pain Points

50 Car Fleet, Driving 6 Hours/Day, Generates 2PB+ Each Day

Ingestion
2PB+/day needs to be transported, encoded, stored

Labeling
Manage 1000+ workers with 50+ projects. Ensure quality every frame.

Curation
Billions of frames. Find the 5-10% that are useful

Training
20+ models. 100s Engineers, Optimize each model w/ 50+ parallel experiments.

Replay
Test against 10,000s hours of sensor data. Repeat Daily

Simulation
Drive hundreds of millions of miles. Find the most critical scenarios to test.
What are Autonomous Driving challenges?

**Challenge # 1**
**DATA**
TBs of data to collect, ingest and store every day translates into PB scale data processing, storage and transfer problem

**Challenge # 2**
**SPEED**
Increased competition and need to simulate millions of miles to shorten TTM and optimizing engineering time requires significant acceleration

**Challenge # 3**
**COST**
PB scale data storage costs, managing fleet operations, significant capex if on-prem compute, lack of AV expertise requires significant human investment

**EXPERTISE**
Lack of Cloud expertise to benefit from scale

**GLOBAL**
Global fleet requires managed service for complex operations, attain data and security compliance across the globe
Edge-to-Cloud Continuum

Localized machine learning

- AWS IoT Greengrass
- AWS RoboMaker
- AWS Snowball Edge

On-premises facility

- Smart Home
- AWS Wavelength
- AWS Local Zone

End-to-end:
- Security
- Development
- Deployment
- Management
A Global Infrastructure

We add the equivalent of an entire Fortune 500 company’s compute capacity every day.

Global Infrastructure:
Redundant 100Gbps network and private capacity between all regions except China

Direct Connect:
90+ locations; customers can reach every AWS Region from their local Direct Connect PoP
AV Development Workflow

Data Management, Processing & Analytics

DevOps, MLOps, Orchestration

Labeling

Model & algorithm development

Simulation

Verification & Validation

Intelligent Storage

AI/ML Frameworks

Partners

CI/CD Pipelines

Specialized Compute
### Autonomous vehicle Ingest

**Offline transfer options**

<table>
<thead>
<tr>
<th></th>
<th>Snowball Edge</th>
<th>Snowcone</th>
<th>Logger removable media</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>THROUGHPUT CAPACITY</strong></td>
<td>&lt;10 Gbps</td>
<td>&lt;2 Gbps</td>
<td>5-50Gbps</td>
</tr>
<tr>
<td></td>
<td>&lt;100TB</td>
<td>&lt;10TB</td>
<td>&lt;120TB</td>
</tr>
<tr>
<td><strong>INTERFACES</strong></td>
<td>1/10/40 GE</td>
<td>1/10GE</td>
<td>PCIe / SATA</td>
</tr>
<tr>
<td></td>
<td>NFS/S3</td>
<td>NFS/S3</td>
<td></td>
</tr>
<tr>
<td><strong>POWER</strong></td>
<td>250W additional</td>
<td>45W additional</td>
<td>Included in data logger</td>
</tr>
<tr>
<td><strong>LOGISTICS</strong></td>
<td>AWS shipping partner direct to AWS (3-5 days for data on Amazon S3)</td>
<td>AWS shipping partner direct to AWS (2-3 days for data on Amazon S3)</td>
<td>Managed services OR customer managed with copy station</td>
</tr>
<tr>
<td><strong>COSTS</strong></td>
<td>$30 per day + shipping costs (&lt;$100)</td>
<td>$8 per day + shipping costs (&lt;$50)</td>
<td>~$15,000–$30,000 one-time costs + shipping</td>
</tr>
</tbody>
</table>
Autonomous Vehicle Data Lake

**Build data lakes quickly**
- Identify, crawl, and catalog sources
- Ingest and clean data
- Transform into optimal formats

**Simplify security management**
- Enforce encryption
- Define access policies for data sharing/access
- Implement audit login

**Enable self-service and combined analytics**
- Analysts/Developers can search all data available for analysis from a single/multiple data catalogs
- Use multiple analytics tools for search/visualization

---

**AWS Services**
- Amazon Kinesis Data Streams
- Amazon Athena
- Amazon Redshift
- Amazon EMR
- AI services
- Amazon QuickSight
- Amazon Greengrass / IOT Core
- AWS Lake Formation
- Amazon S3
- AWS Snowball Edge/Snowcone
- Removeable media
- Amazon Direct Connect
- Amazon Kinesis Data Streams
- IAM
- KMS

---

Ingestion and cleaning
Security
Processing, Analytics, and Visualization

Amazon Privileged and Confidential | © 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Choosing the right AV data lake storage class
Select storage class by data pipeline stage

Raw drive data
- Small log files
- Overwrites if synced
- Short lived
- Moved & deleted
- Batched & archived

ETL
- Data churn
- Small intermediates
- Multiple transforms
- Deletes <30 days
- Output to data lake

AV data lake
- Optimized sizes (MBs)
- Many users
- Unpredictable access
- Long-lived assets
- Hot to cool

Online cool data
- Replicated DR data
- Infrequently accessed
- Infrequent queries
- ML model training

Historical data
- Historical assets
- ML model training
- Compliance/Audit
- Data protection
- Planned restores

Optimize costs for all stages of data lake workflows
AV Development Workflow

Data Management, Processing & Analytics

Labeling

Model & algorithm development

Verification & Validation

Simulation

HD Maps Model Deployment

Offline & Online Data Ingestion

Data Acquisition

SNOWCONE

SNOW FAMILY DIRECT CONNECT OUTPOST

AD/ADAS Data Lake

SIMPLE STORAGE SERVICE (S3)

EC2 C5/M5/R5/G4

EC2 SPOT

FSX Lustre

EC2 P3/P4/G4 INSTANCES

EFA, FSX Lustre

SAGEMAKER DL AMI

SAGEMAKER GROUND TRUTH

SAGEMAKER

GLUE

EMR

DYNAMODB

AURORA

ELASTICSEARCH

LAMBDA

SIMPLE STORAGE SERVICE (S3)

EC2 C5/M5/R5/G4

EC2 SPOT

FSX Lustre

Batch Parallel Cluster

EKS

ECS

DevOps, MLOps, Orchestration

Intelligent Storage

AI/ML Frameworks

Partners

CI/CD Pipelines

Specialized Compute
Autonomous driving system modeling challenges

1. Iterate over large volumes of annotated heterogeneous data
2. Tightly coupled compute infrastructure to support distributed model building over millions of miles of acquired and simulated data in a data-parallel pattern
3. Reduce model training time with distributed GPU compute
4. Integrate model building and simulation infrastructure to enable learning on a virtual environment
The usual Autonomous driving software stack is comprised of many modeling steps:

1. **Sensing**
2. **Perception**
3. **Localization and mapping**
4. **Planning**
5. **Control**
6. **Actuation**

Each step might require different supporting infrastructure, i.e.,:

- **Perception**: latest GPU technology with large memory (g4dn, p3dn) to support Deep Learning training, over TB scale distributed file systems
- **Control**: mix of general purpose GPU (p2, g3) and latest technology to support Deep Reinforcement Learning over on and off-line simulated data
AV sample model stack

Main tasks are to locate the vehicle and, identify and classify elements of the environment

- **Main inputs:** GPS, Inertia Measurement Unit (IMU), vehicle odometry, camera images, Lidar point clouds and radar maps
- **Outputs:** Ego vehicle pose, objects segments and classes, dynamic objects state

Heavy compute workload—object detection and localization:

- ML/AI applications of Computer Vision models.
- Large Semantic Segmentation tasks running over multiple cameras and point clouds
- Real time 2D and 3D object detection and tracking
Main task is to locate environment elements around the ego
Heavily integrated with Perception stack for Simultaneous Localization and Mapping (SLAM)
- **Main inputs:** GPS, Lidar, Perception localization and tracking
- **Outputs:** Occupancy grids, localization maps and road segments

Example heavy compute tasks:
- DL base sensor fusion for object detection
- Camera and Lidar based pose regressions
AV sample model stack

1. Sensing
2. Perception
3. Localization and mapping
4. Planning
5. Control
6. Actuation

Defines drive path and execution based on Localization and Perception
Planning can comprise layers: mission, behavior, and local planning
Control agent development based on Reinforcement or Supervised Learning

- **Main inputs:** Environment representations from mapping, objects, and tracking from perception
- **Outputs:** Path plan and drive profiles, longitudinal and lateral controls

Example heavy compute tasks:
- Deep Reinforcement learning for longitudinal and lateral control
- Domain adaptation for Simulation-to-Real deployments
- Simulated Driving Environments

From: Design Space of Behaviour Planning for Autonomous Driving
From: Interpretable End-to-end Urban Autonomous Driving with Latent Deep Reinforcement Learning
Sample Model Building Architecture

AWS Cloud

Third-party labeling providers
Compressed LIDAR/Camera Image Blobs
Metadata (Timestamp/GPS)

Smart garage
Camera
LIDAR/RADAR
Vehicle

4G/5G Vehicle Telemetry

Input Labeling Staging
LIDAR SageMaker GT Labeling Job
Labeled Data Staging

4K Camera SageMaker GT Labeling Job

Labeled Data Staging

Labeled Data Staging

Production DL Development w/ SageMaker
Deep Learning Containers

4K Camera SageMaker GT Labeling Job

Production DL Development Workflow using EKS

ML Model

MLOps Training/Sim/Inference

ML Inferencing Pod

Simulations and ML Model Validation

ML Inferencing Pod

Re-sim and Log Analysis

Simulations Container Registry
Simulations Container Registry

ML Model

MLOps Containers
MLOps Deployment

MLOps Training/Sim/Inference

ML Model
AV Development Workflow

Data Management, Processing & Analytics

DevOps, MLOps, Orchestration

Labeling

Model & algorithm development

Map Development

Simulation

Verification & Validation

Office & Online Data Ingestion

HD Maps Model Deployment

SNOW CONE
Two kinds of simulations in autonomous driving

Log Replay/Re-Simulation
Replay recorded sensor data to the driving stack and evaluate how it reacts.

Synthetic simulation
Evaluate scenarios and variants in a simulated world. Sensor data is sent to the driving stack which carries driving commands.
# AV Simulations Typical Requirements

## Driving/Synthetic Simulation

<table>
<thead>
<tr>
<th>Compute</th>
<th>1-4+ vCPUs (C5, M5, R5) and/or 1 GPU (P3, P4d, G4dn)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory</td>
<td>2+GB / vCPU</td>
</tr>
<tr>
<td>Storage</td>
<td>S3, Local scratch</td>
</tr>
<tr>
<td>Runtime</td>
<td>1min to 1h+ per simulation</td>
</tr>
<tr>
<td>Sims / day</td>
<td>100s-1M+</td>
</tr>
</tbody>
</table>

## Log Replay Simulation

<table>
<thead>
<tr>
<th>Compute</th>
<th>S3, Local scratch, FSx</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory</td>
<td></td>
</tr>
<tr>
<td>Storage</td>
<td></td>
</tr>
<tr>
<td>Runtime</td>
<td></td>
</tr>
<tr>
<td>Sims / day</td>
<td>10k+</td>
</tr>
</tbody>
</table>

- **Large scale compute**
- **Data intensive and GPU**
1.6M vCPUs: all on AWS EC2
Spare Capacity
TuSimple Built an Autonomous Level 4 Truck Driving System with the World’s Longest Perceptual Range Using AWS

Challenge
TuSimple needed a platform on which to develop and test its artificial intelligence decision-making system that guides vehicles along a safe and fuel-efficient route.

Solution
TuSimple uses AWS Snowball Edge to collect data, Amazon EC2 P3 instances, and Machine Learning to train deep learning algorithms, and AWS infrastructure for its simulation environment to test algorithms.

Benefits
• On-demand access to the latest GPU instances and integrated deep learning frameworks reduces training time from days to hours.
• Global collaboration between test and development sites.

“AWS is very important to us. It provides the most comprehensive suite that we can use on the cloud without reinventing the wheel for ourselves again.”

Xiaodi Hou, President & CTO

Company: TuSimple
Country: USA/China
Employees: 400
Website: TuSimple.com

About TuSimple
TuSimple is a level 4 autonomous commercial trucking company that uses deep learning and artificial intelligence. Using an array of cameras, TuSimple’s platform scans the surrounding environment to navigate heavy freight trucks.
Lyft Increases Simulation Capacity, Lowers Costs Using Amazon EC2 Spot Instances

**Challenge**

Rideshare company Lyft runs millions of compute-intensive simulations each year to improve the performance and safety of its self-driving system and needed lots of computing power that could scale up and down at an affordable price.

**Solution**

The company significantly increased its AV simulation testing while reducing the corresponding computing costs by two-thirds with Amazon EC2 Spot Instances and Amazon EKS.

**Benefits**

- Reduced compute costs by two-thirds
- Scaled up computing capacity significantly
- Increased velocity of development for AVs

“About 77% of our computing fleet is now on Amazon EC2 Spot Instances. We were able to scale up our computing capacity significantly while reducing the overall cost of operation.”

—Timothy Perrett, Level 5 Senior Staff Engineer, Lyft

**About Lyft**

Lyft, one of the largest transportation networks in the United States and Canada, is on a mission: improve people’s lives with the world’s best transportation. It provides shared rides, electric scooters, bikeshare systems, and public transit partnerships.

Company: Lyft

Industry: Transportation & Logistics

Country: United States

Website: [lyft.com](http://lyft.com)
Mobileye runs large scale simulations on AWS Batch and EC2 Spot Instances

Mobileye reaches a daily peak of 500k concurrent vCPUs and typically runs between 200k to 300k concurrent vCPUs to run not only their simulation workloads but also their analytics and machine learning workloads. EC2 Spot instances are spare compute capacity that are interruptible but offer up to 90% discount over on demand instances.
Challenge & Solution

Vehicles with self-driving technology can bring many benefits to society. One of the top priorities at Toyota Research Institute (TRI) is to apply the latest advancements in artificial intelligence (AI) to help Toyota produce cars that are safer, more accessible, and more environmentally friendly. To help TRI achieve their goals, they turned to deep learning on AWS.

Using deep learning on Amazon EC2 P3 instances, Amazon S3, Amazon SQS, and AWS networking services, TRI built a scalable solution to enable their development teams to make rapid progress and deliver on their grand vision of applying AI to help Toyota produce cars that are safer, and get closer to realizing a future without traffic injuries or fatalities.

“Using the AWS Cloud and specifically Amazon EC2 P3 instances, we’re able to build a scalable and highly performant applications stack to efficiently handle and process the huge amount of data that we collect.”

Mike Garrison, Technical Lead, Infrastructure Engineering

Benefits

- Using Amazon EC2 P3 instances, TRI is seeing a 4x increase in time-to-train, reducing training time from days to hours.
- Lower operating costs with performance improvements in P3 instances and the AWS pay-as-you-go model.

About TRI

Toyota Research Institute is a wholly owned subsidiary of Toyota Motor North America under the direction of Dr. Gill Pratt. The company, established in 2015, aims to strengthen Toyota’s research structure and has four initial mandates: 1) enhance the safety of automobiles, 2) increase access to cars to those who otherwise cannot drive, 3) translate Toyota’s expertise in creating products for outdoor mobility into products for indoor mobility, and 4) accelerate scientific discovery by applying techniques from artificial intelligence and machine learning.
WeRide deployed its machine learning and simulation platform on AWS

**Challenge & Solution**

WeRide deployed its machine learning and simulation platform on AWS. WeRide was able to reduce its model training time from weeks to hours, while also reducing total cost of ownership by a third, and improving maintenance efficiency by 50%.

**Company:** WeRide  
**Country:** China/USA  
**Employees:** 300+  
**Website:** www.weride.ai

**About WeRide**

WeRide is a Chinese/American smart mobility company, established in 2017, with leading Society of Automotive Engineers (SAE) autonomy Level 4, Advanced Driving (AD) technology. WeRide currently operates an exploratory robotaxi program in Guangzhou covering nearly 145 KMs of Operational Design Domain (ODD) where their vehicles help locals with their daily commutes.
The Latest Compute Technology in the Cloud

P3/P4 GPU compute instance
- P4: Up to 2.5 PetaFLOP of compute with 8x NVIDIA A100 GPUs
- P3: Up to 1 PetaFLOP of compute with 8x NVIDIA V100 GPUs
- Up to 320 GB of GPU memory and up to 400 Gbps of networking on p4d.
- Designed for HPC and to handle large distributed machine learning training jobs

G4 GPU compute instance
- Up to 520 TeraFLOPs of compute with 8x NVIDIA T4 GPUs
- Up to 1.8 TB of Local NVMe storage and up to 100 Gbps of networking throughput
- Designed for cost-effective machine learning inference and graphics intensive applications
- Simulation driven workloads, Reinforcement-learning

New Architectures
- Trainium: Instances will offer the most TFLOPS of any compute instance in the cloud
- Habana Gaudi: Instances will offer 40% better price performance compared to existing, GPU-based EC2 instances.
- Graviton2: Arm Neoverse-based CPU architectures offering up to 40% better price/performance versus comparable x86-based EC2 instances.

Inf1 Inferentia instance
- Up to 2000 TOPs with 16x AWS-designed Inferentia accelerators
- Featuring AWS Inferentia, the first custom ML chip designed by AWS
- Designed for high throughput and low latency machine learning inference
AWS Graviton2: ARM-based instances

Up to **40% better price-performance** over comparable current generation x86-based instances.

**M6g**
General purpose workloads

**C6g**
Compute-intensive workloads

**R6g**
Memory-intensive workloads

Available Now!

Local NVMe-based SSD storage options are also available:
general purpose (**M6gd**), compute-optimized (**C6gd**), and memory-optimized (**R6gd**)

Every instance type also has a bare-metal option:
(*M6g.metal, M6gd.metal, C6g.metal, C6gd.metal, R6g.metal, R6gd.metal*)
Largest Pool of Spare ("Spot") Compute Capacity

- **Scale using Spot for flexible, fault-tolerant workloads**
- **Scale using On-Demand for new or stateful SPIKY workloads**
- **Use RIs and Savings Plans for known/ steady-state workloads**

**AWS Services make this easy and efficient**

- Amazon EC2 Auto Scaling
- EC2 Fleet
- Amazon Elastic Container Service (Amazon ECS)
- Amazon Elastic Kubernetes Service (Amazon EKS)
- AWS Thinkbox
- AWS Batch
- Amazon EMR
- AWS CloudFormation