THE INTERNATIONAL MEETING
CONFERENCES | WORKSHOPS | EXHIBITION

22-23-24 JUNE
DIGITAL EVENT

Forum Teratec 2021

Unlock the future !

SIMULATION |
HPC | HPDA
AI | QUANTUM

PLATINUM SPONSORS: Atos, ddn, GRAPHCORE, Hewlett Packard Enterprise, intel., VAST

GOLD SPONSORS: ATEMPO, cea, doitnow, exaion EDF GROUP, Lenovo, UCIT

SILVER SPONSORS: arm, aws, GENCI, NVIDIA, ORACLE, rescale, XILINX

PARTENAIRE EUROPA VILLAGE: Inria

A local global infrastructure for Autonomous Vehicle Development

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

aws

Agenda

- *Autonomous Vehicle (AV) Landscape & Challenges*
- *Cloud for AV Development*
- *Case Studies & References*

Autonomous Systems & Machine Learning at Amazon

Twenty years of innovation

robotics

Delivery Robots

prime

Fulfillment Automation
& Inventory Management

amazon Prime Air

Drones

amazon alexa

Voice-driven
Interactions

amazon go

Inventing New
Customer Experiences

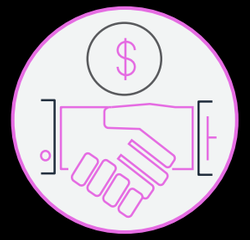
aws

AV Industry Overview

CASE

Connected

Autonomous



Shared

Electrified

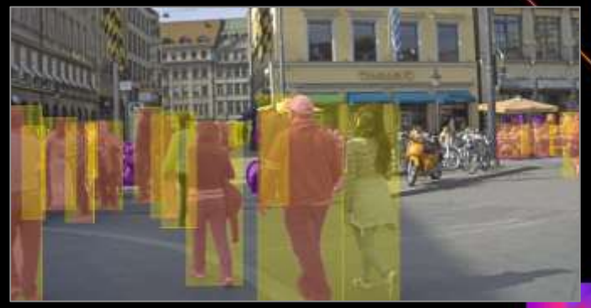
% new cars 2015	8%	0%	2%	0.1%
% new cars 2030	100%	35%	25%	20%

Autonomous Driving : Challenges & Pain Points

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

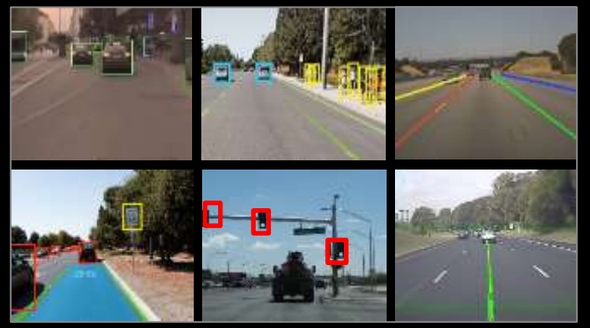
Ingestion
2PB+/day needs to be transported, encoded, stored

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



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

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

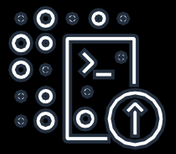


Replay
Test against 10,000s hours of sensor data.
Repeat Daily

Simulation
Drive hundreds of millions of miles.
Find the most critical scenarios to test.

AWS Autonomous Vehicle Customer References

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

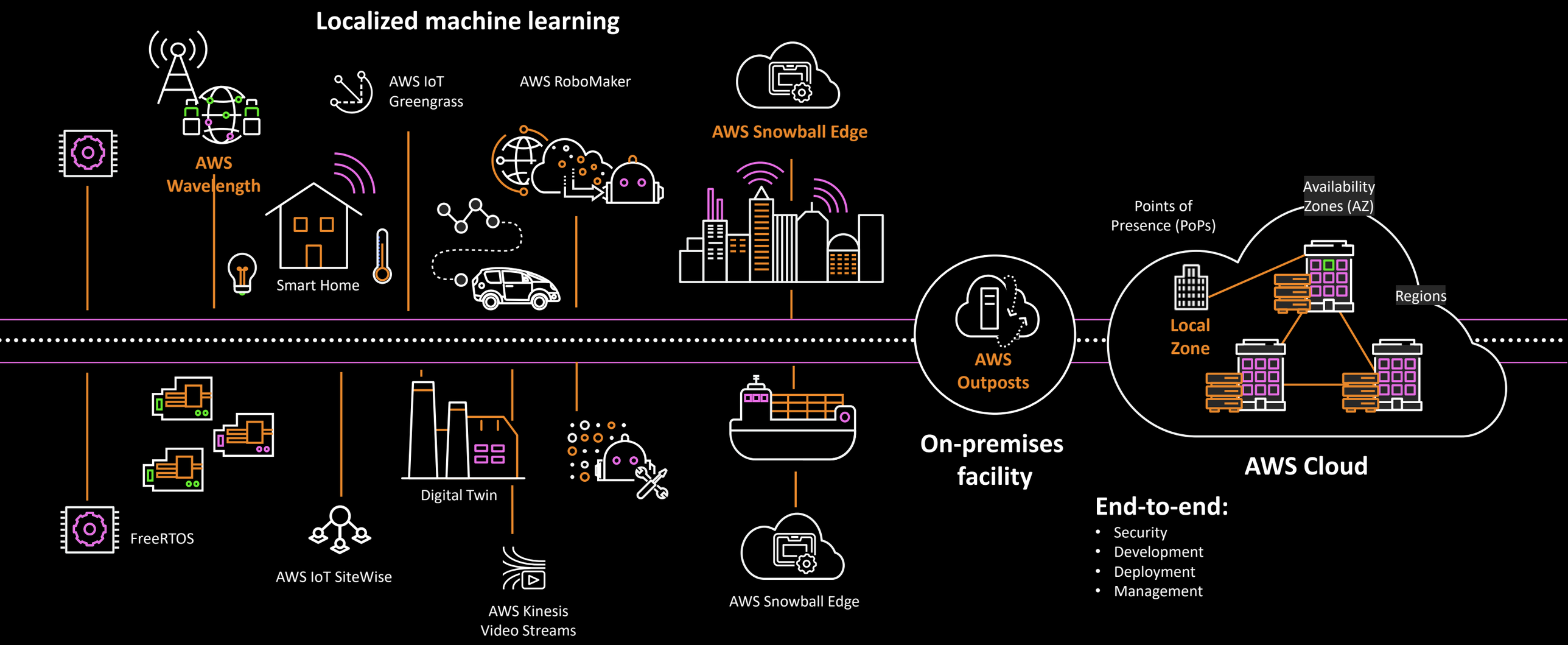
EXPERTISE

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

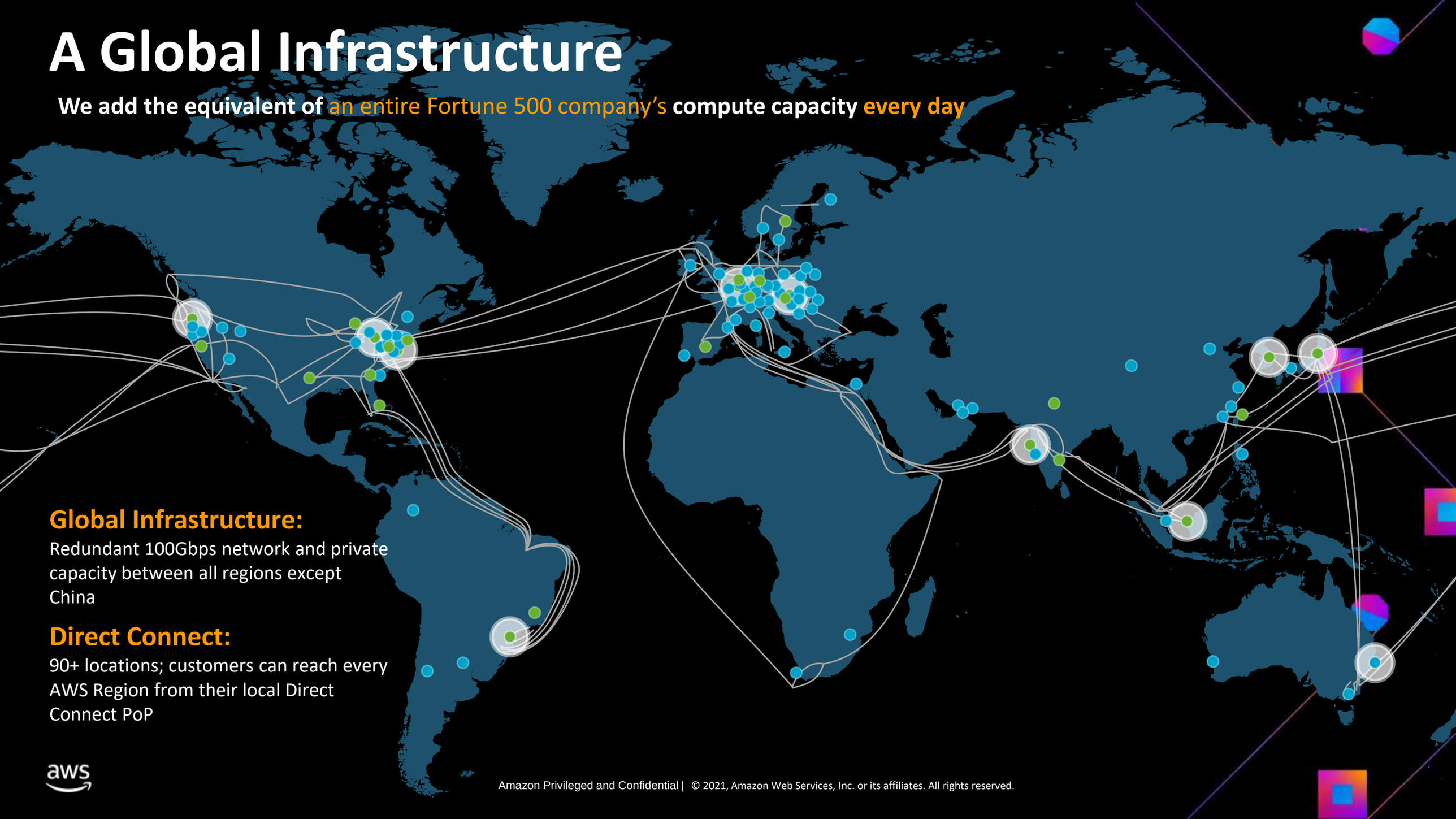


End-to-end:

- Security
- Development
- Deployment
- Management

A Global Infrastructure

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



Global Infrastructure:

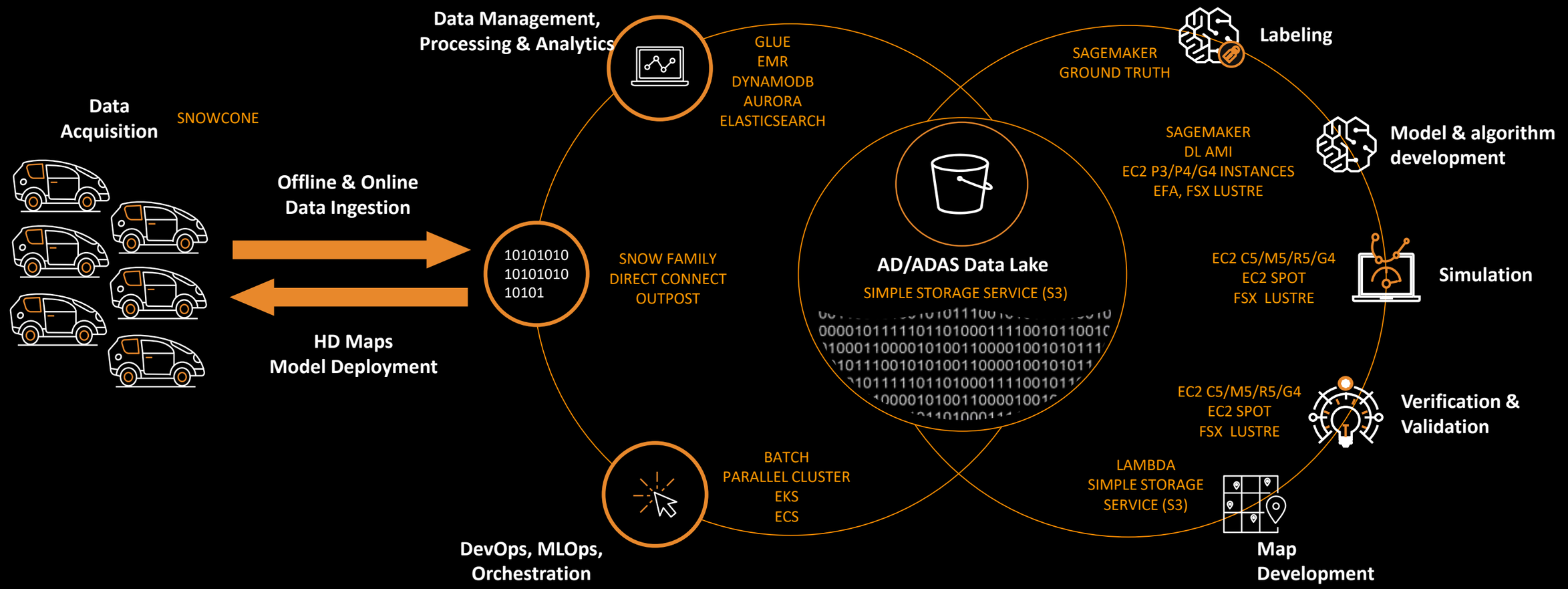
Redundant 100Gbps network and private capacity between all regions except China

Direct Connect:

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

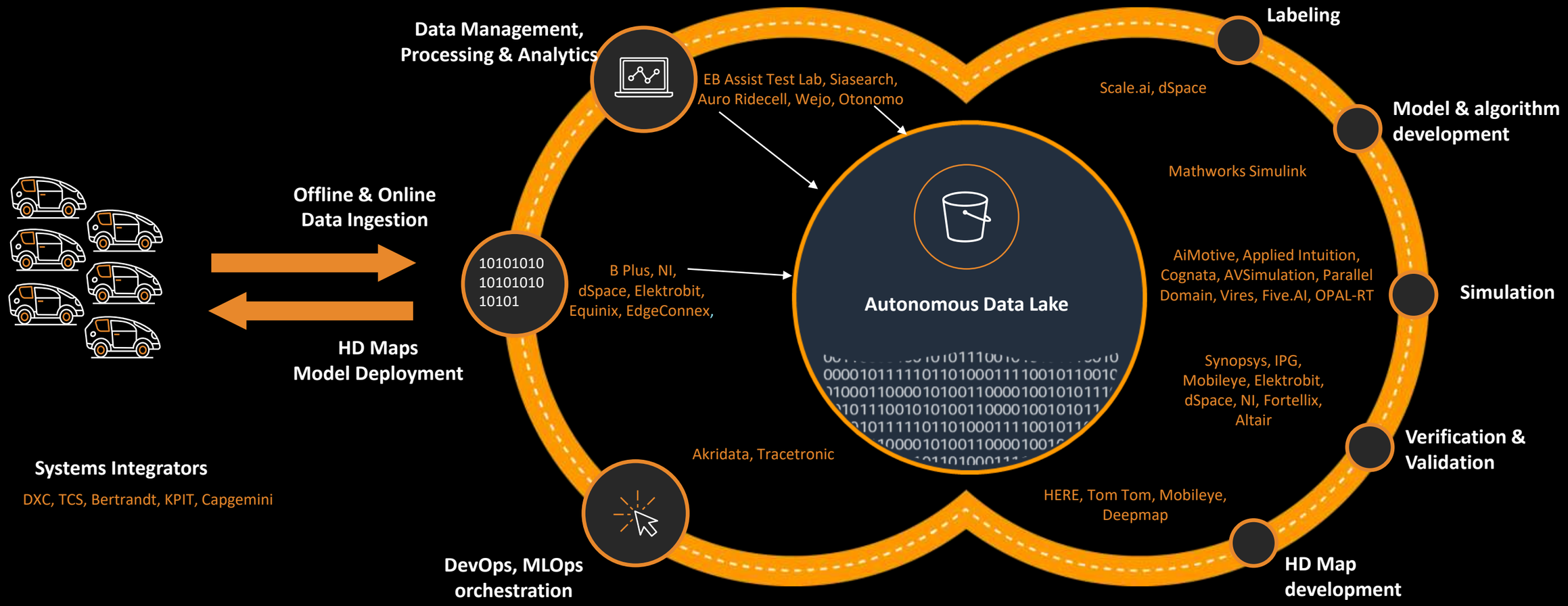
aws

AV Development Workflow



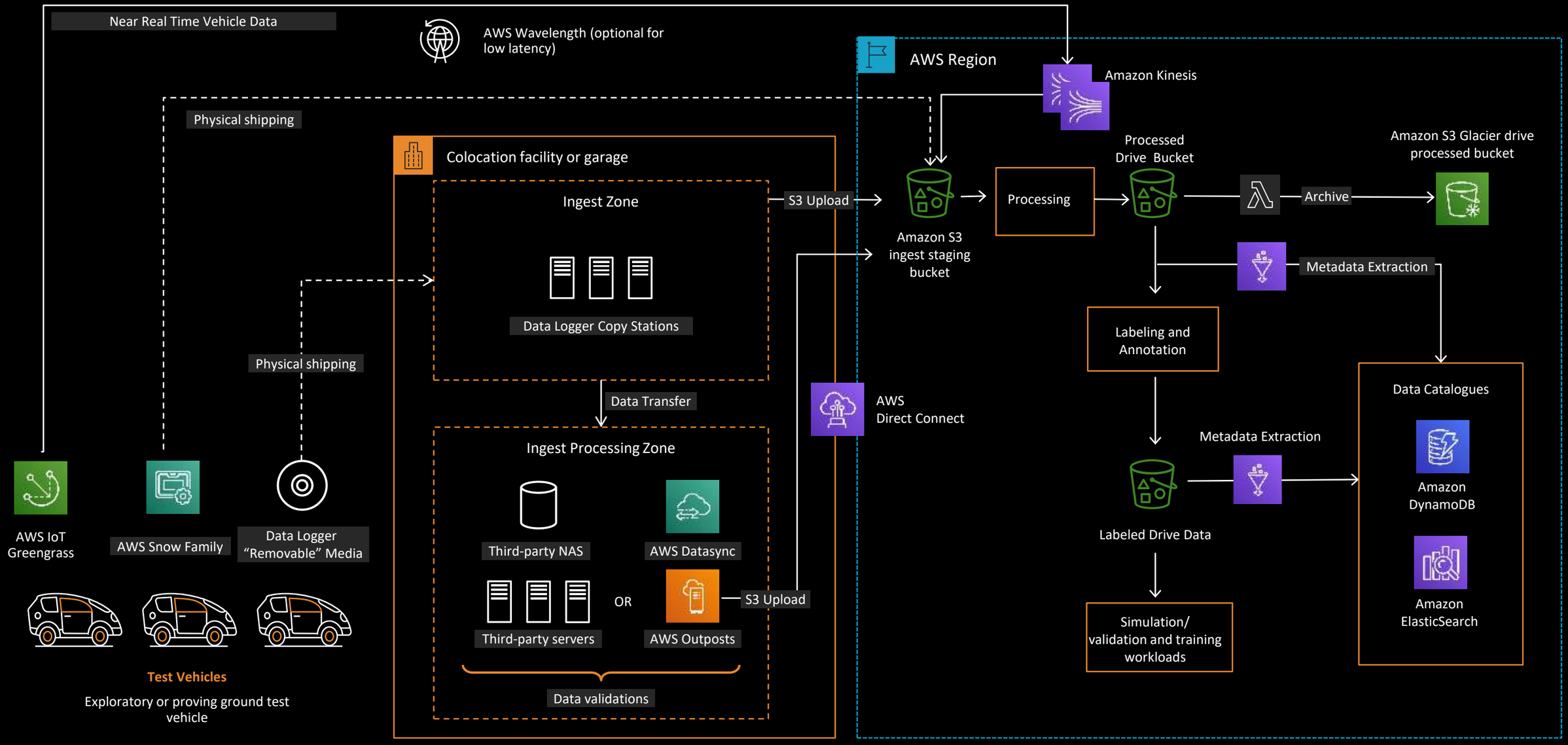
Intelligent Storage | **AI/ML Frameworks** | aws | Partners | **CI/CD Pipelines** | **Specialized Compute**

AV Partner Ecosytem



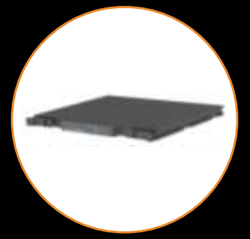
Intelligent Storage | AI/ML Frameworks | aws | Partners | CI/CD Pipelines | Specialized Compute

Autonomous vehicle ingest architecture



Autonomous vehicle Ingest

Offline transfer options



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

Data logger companies

b - plus
PIONEERING NEW MOBILITY

Elektrobit

INTREPID CONTROL SYSTEMS

NATIONAL INSTRUMENTS

dSPACE

X2E
embedded system design

VECTOR

ViGEM
VISIONS GO EMBEDDED

aws

Autonomous Vehicle Data Lake

Build data lakes quickly

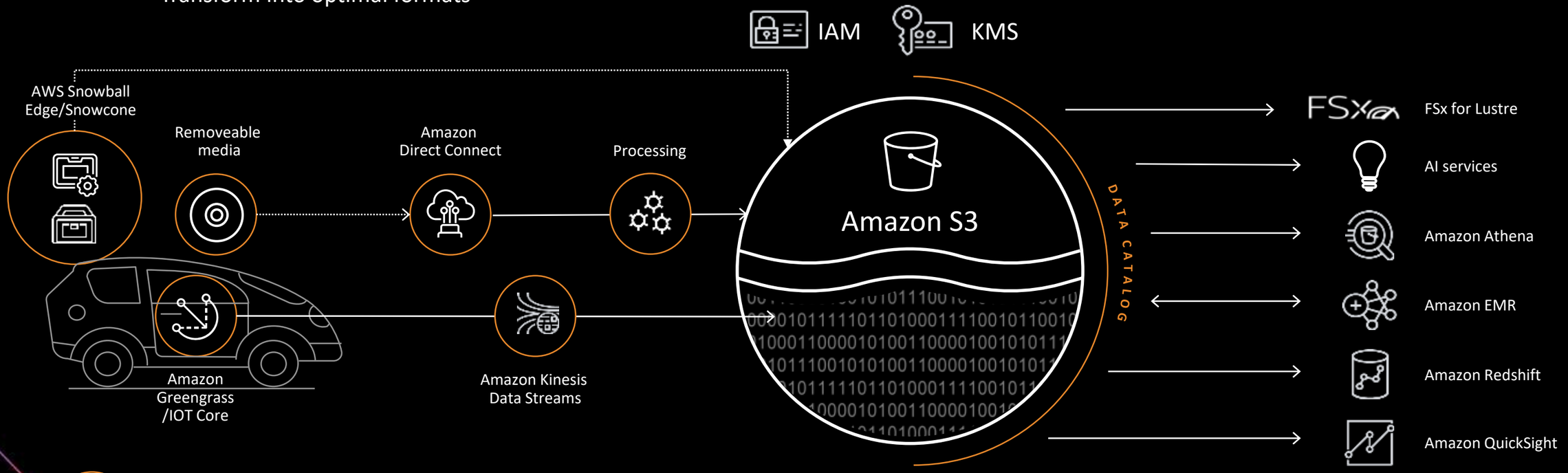
- Identify, crawl, and catalog sources
- Ingest and clean data
- Transform into optimal formats

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



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

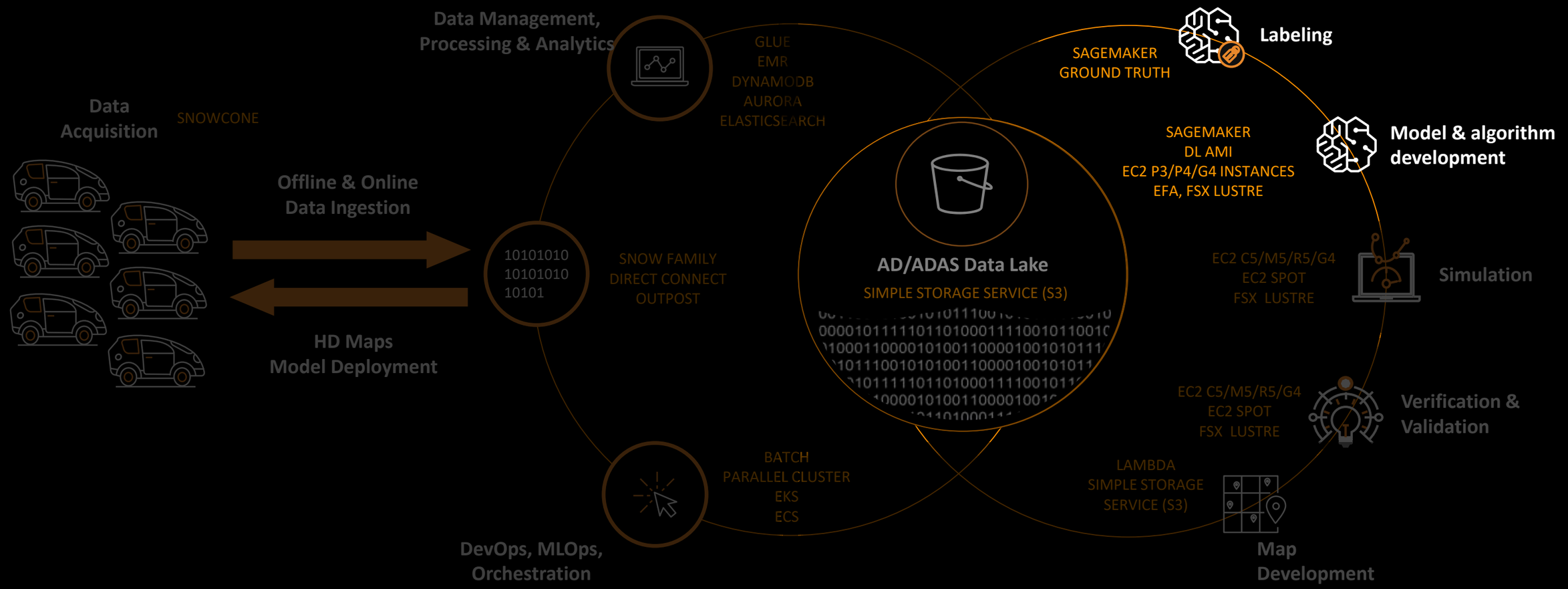
- Replicated DR data
- Infrequently accessed
- Infrequent queries
- ML model training

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



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

AV sample model stack

The usual Autonomous driving software stack is comprised of many modeling steps:

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

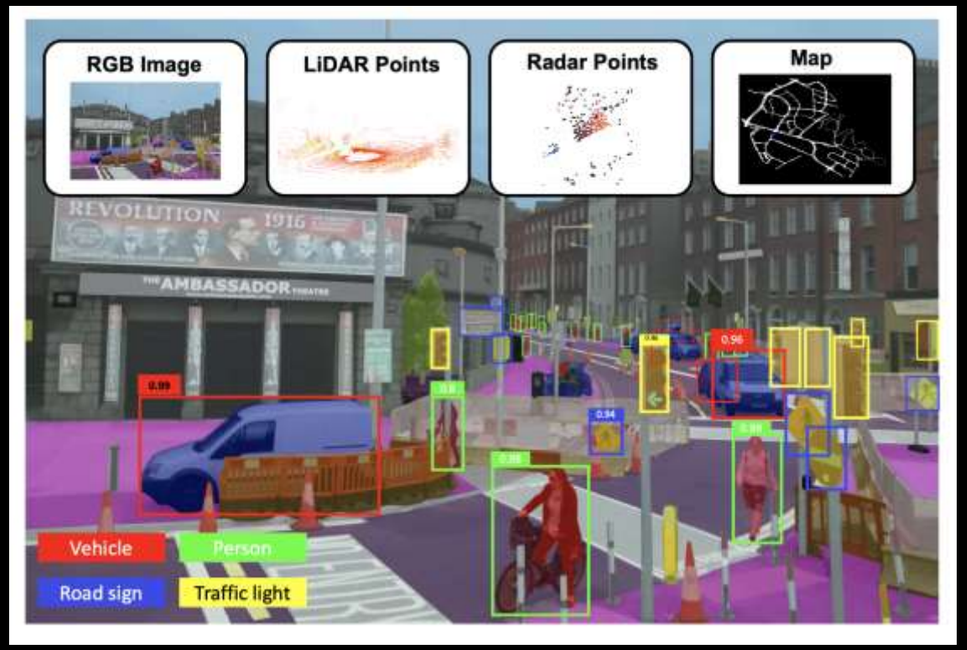
1 Sensing → 2 Perception → 3 Localization and mapping → 4 Planning → 5 Control → 6 Actuation

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



AV sample model stack

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

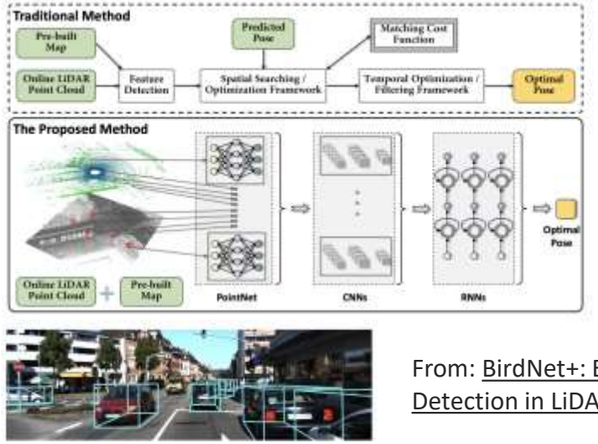
Main task is to locate environment elements around the ego

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



From: L 3-Net: Towards Learning based LiDAR Localization for Autonomous Driving

From: BirdNet+: End-to-End 3D Object Detection in LiDAR Bird's Eye View

AV sample model stack

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

Defines drive path and execution base on Localization and Perception

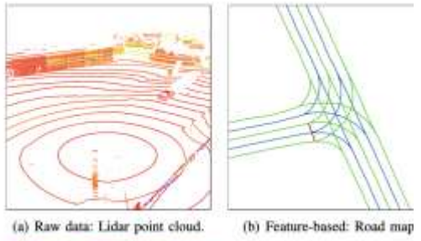
Planning can comprised layers: mission, behavior, and local planning

Control agent development based on Reinforcement or Supervised Learning

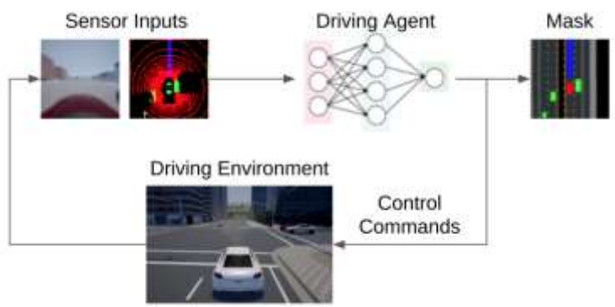
- **Main inputs:** Environment representations from mapping, objects, and tracking from perception
- **Outputs:** Path plan and drive profiles, longitudinal and lateral controls

Example heavy compute tasks:

- Deep Reinforcement learning for longitudinal and lateral control
- Domain adaptation for Simulation-to-Real deployments
- Simulated Driving Environments

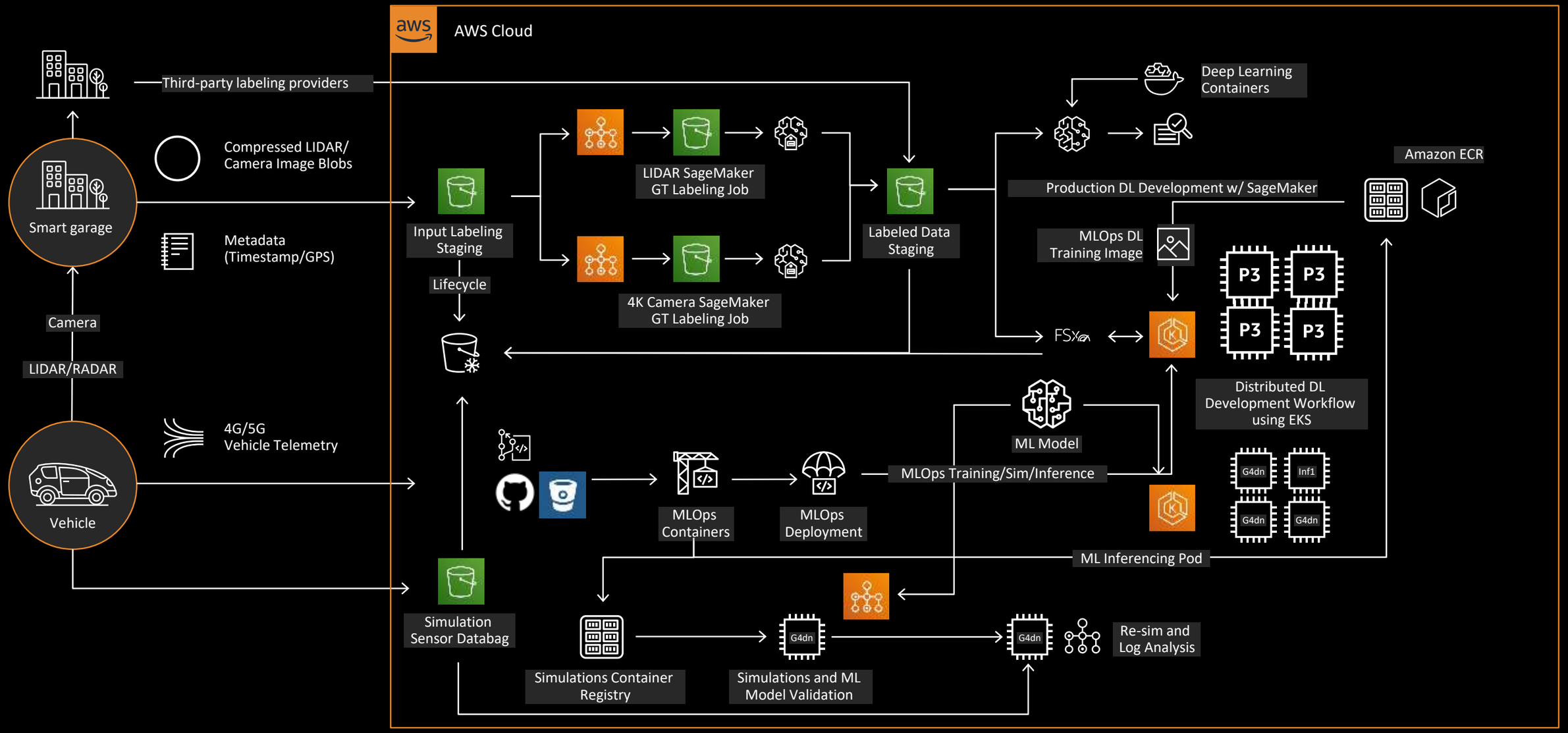


From: Design Space of Behaviour Planning for Autonomous Driving

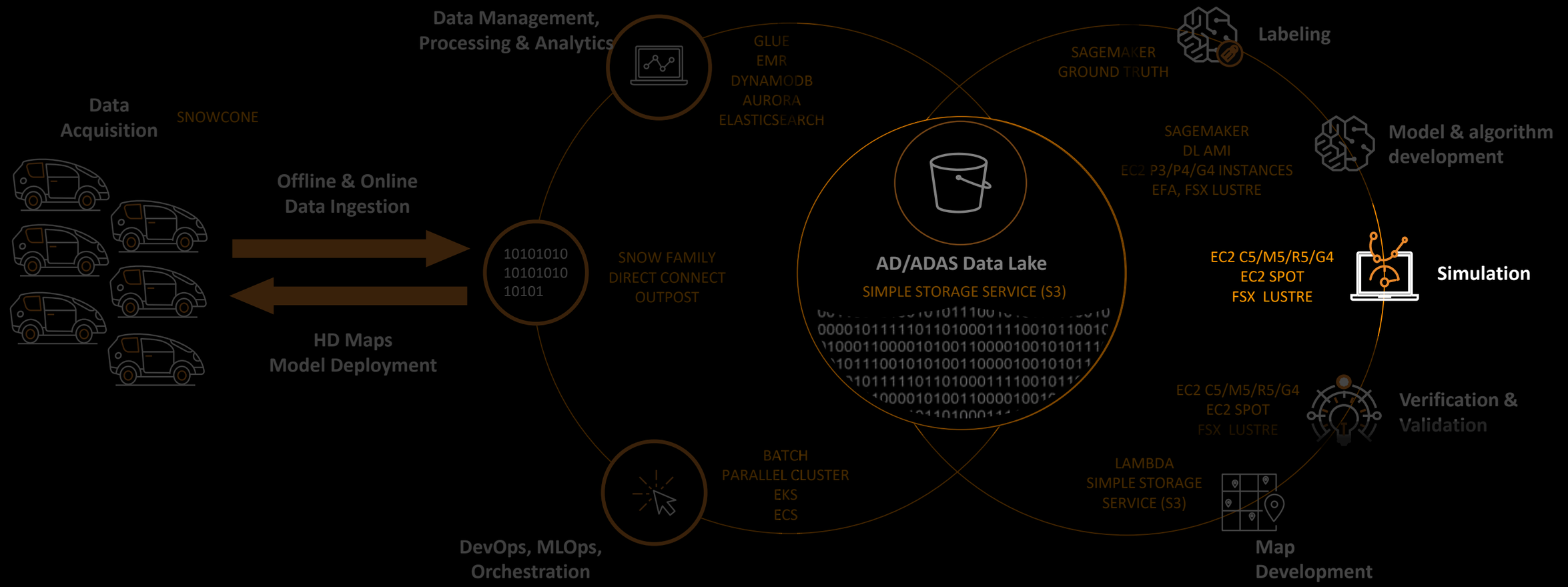


From: Interpretable End-to-end Urban Autonomous Driving with Latent Deep Reinforcement Learning

Sample Model Building Architecture

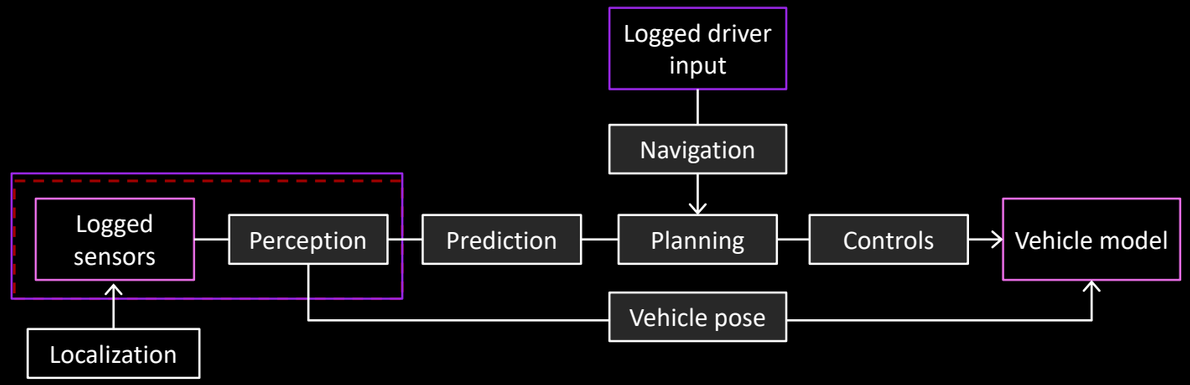


AV Development Workflow



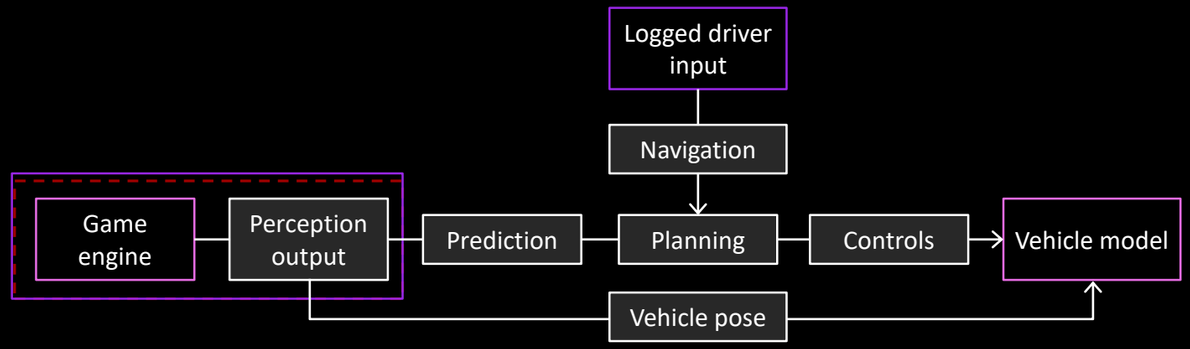
Intelligent Storage | AI/ML Frameworks | aws | Partners | CI/CD Pipelines | Specialized Compute

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.

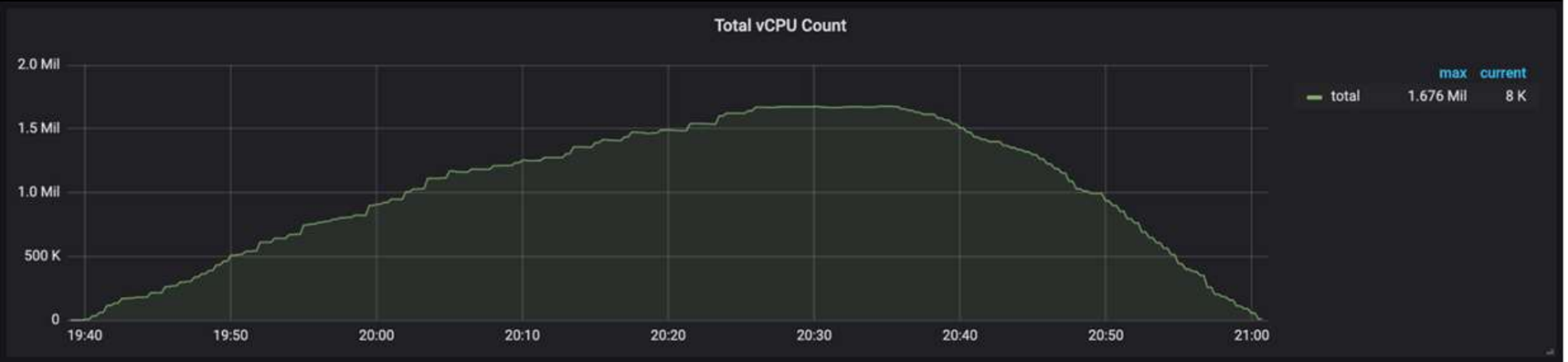
AV Simulations Typical Requirements

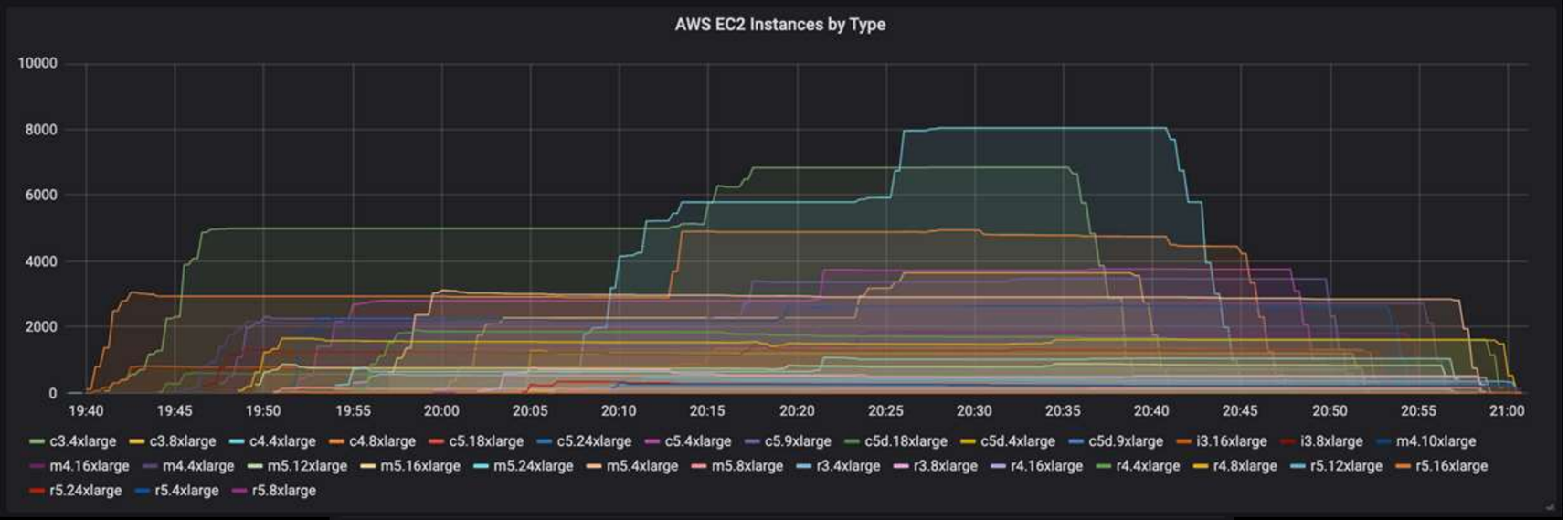
	Driving/Synthetic Simulation	Log Replay Simulation
Compute	1-4+ vCPUs (C5, M5, R5) and/or 1 GPU (P3, P4d, G4dn)	
Memory	2+GB / vCPU	
Storage	S3, Local scratch	S3, Local scratch, FSx
Runtime	1min to 1h+ per simulation	
Sims / day	100s-1M+	10k+

Large scale compute

Data intensive and GPU

Unparalleled Scale of Compute





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.

Learn more

"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

tu simple

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

Read more

> "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

lyft Level 5

Company: Lyft

Industry: Transportation & Logistics

Country: United States

Website: lyft.com

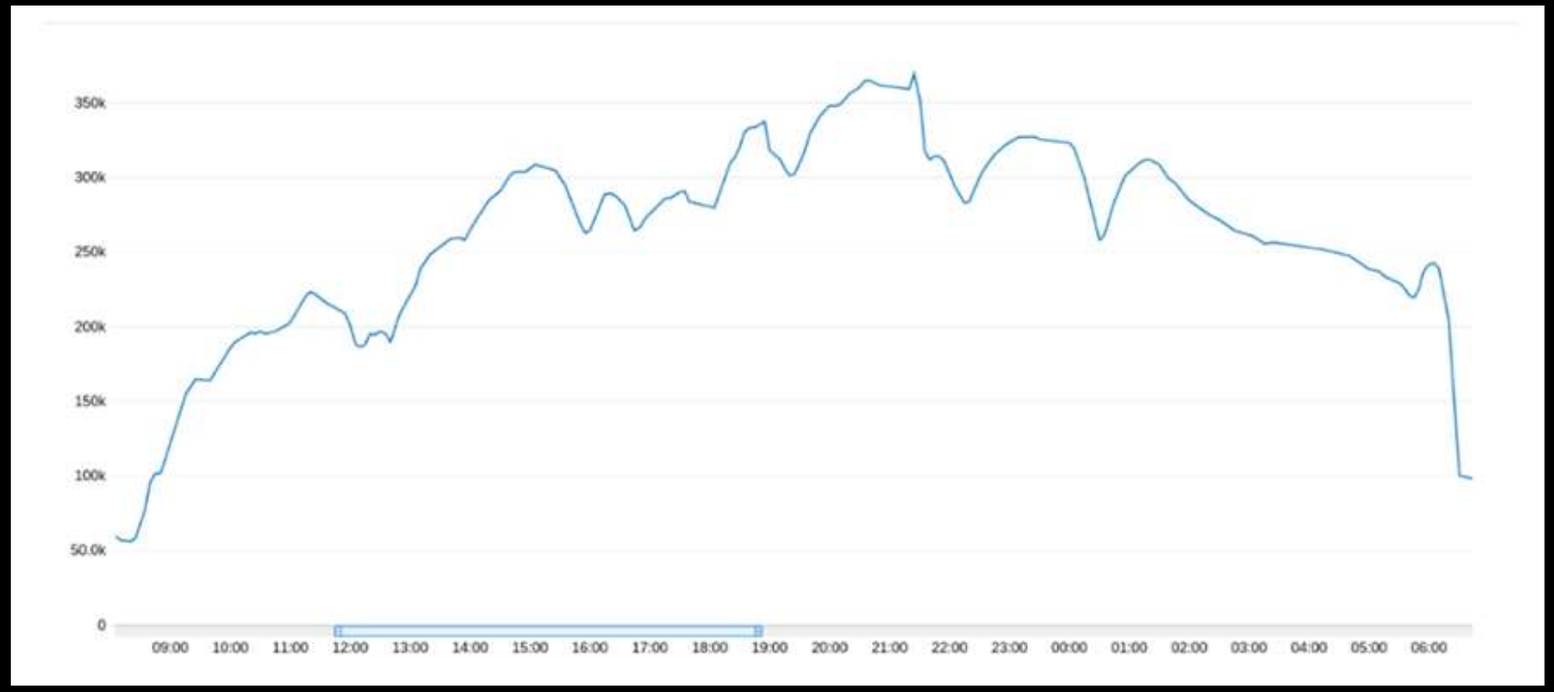
About Lyft

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.

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.

Read more



Company: Mobileye

Country: Israel/USA

Website: Mobileye.com

About Mobileye

Mobileye, an Intel company, was launched in 1999 with the belief that vision-safety technology will make our roads safer, reduce traffic congestion and save lives. With a cutting edge team of more than 1,700 employees, Mobileye has developed a range of software products that is deployed on a proprietary family of computer chips named EyeQ®.

Toyota Research Institute Accelerates Safe Automated Driving with Deep Learning at a Global Scale on AWS

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.

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.

Read more

> "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

woven planet

Company: Woven Planet (fka. Toyota Research Institute)

Country: USA/Japan

Employees: 360

Website: woven-planet.global

About TRI

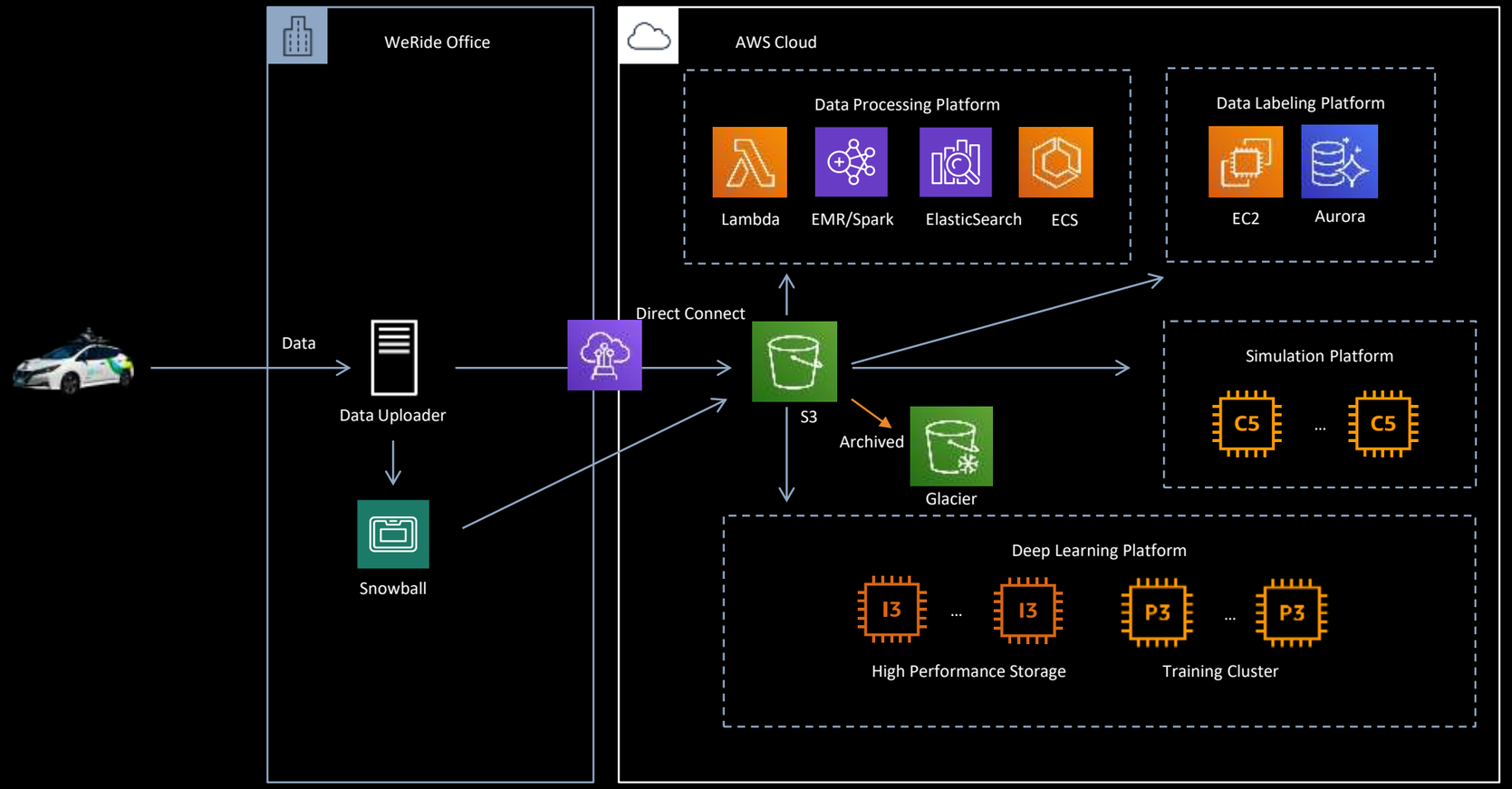
Toyota Research Institute is a wholly owned subsidiary of Toyota Motor North America under the direction of Dr. Gill Pratt. The company, established in 2015, aims to strengthen Toyota's research structure and has four initial mandates: 1) enhance the safety of automobiles, 2) increase access to cars 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%.

Read more



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.

22-23-24 JUNE | DIGITAL EVENT

Forum Teratec 2021

Unlock the future !

Merci pour votre attention.

Thank you for your attention.

PLATINUM SPONSORS

GOLD SPONSORS

SILVER SPONSORS

PARTENAIRE EUROPA VILLAGE

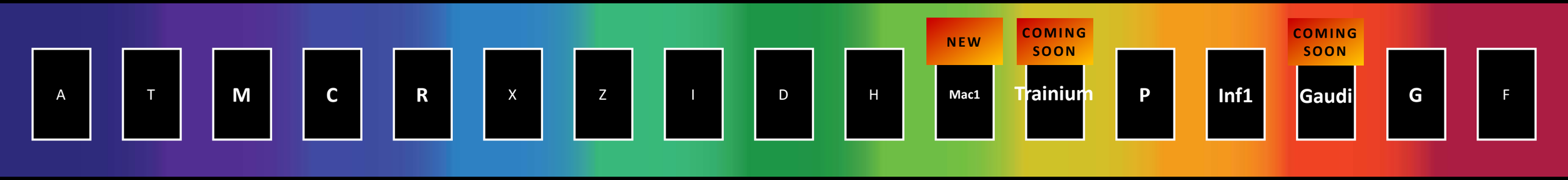
Broadest Compute Platform in the Cloud

17

INSTANCE FAMILIES

350+

INDIVIDUAL INSTANCES



Gaudi Accelerator
Cascade Lake CPU
Skylake CPU

Trainium Chip
Graviton CPU
Inferentia Chip
Nitro Hypervisor Card

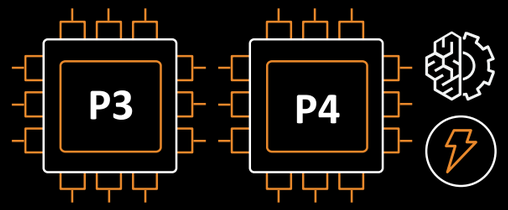
Radeon Pro GPU
EPYC CPU

Ampere A100 GPU
Tesla T4 Tensor GPU
Volta V100 GPU

Mac Mini

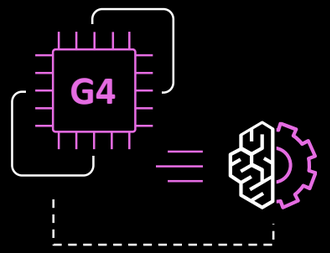
Virtex UltraScale+ FPGA

The Latest Compute Technology in the Cloud



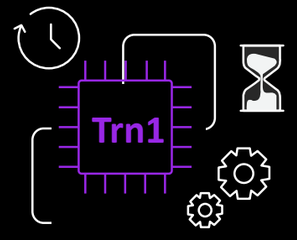
P3/P4 GPU compute instance

- **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



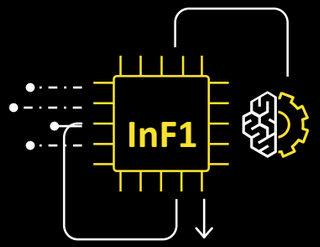
G4 GPU compute instance

- 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	C6g	R6g
General purpose workloads	Compute-intensive workloads	Memory-intensive workloads

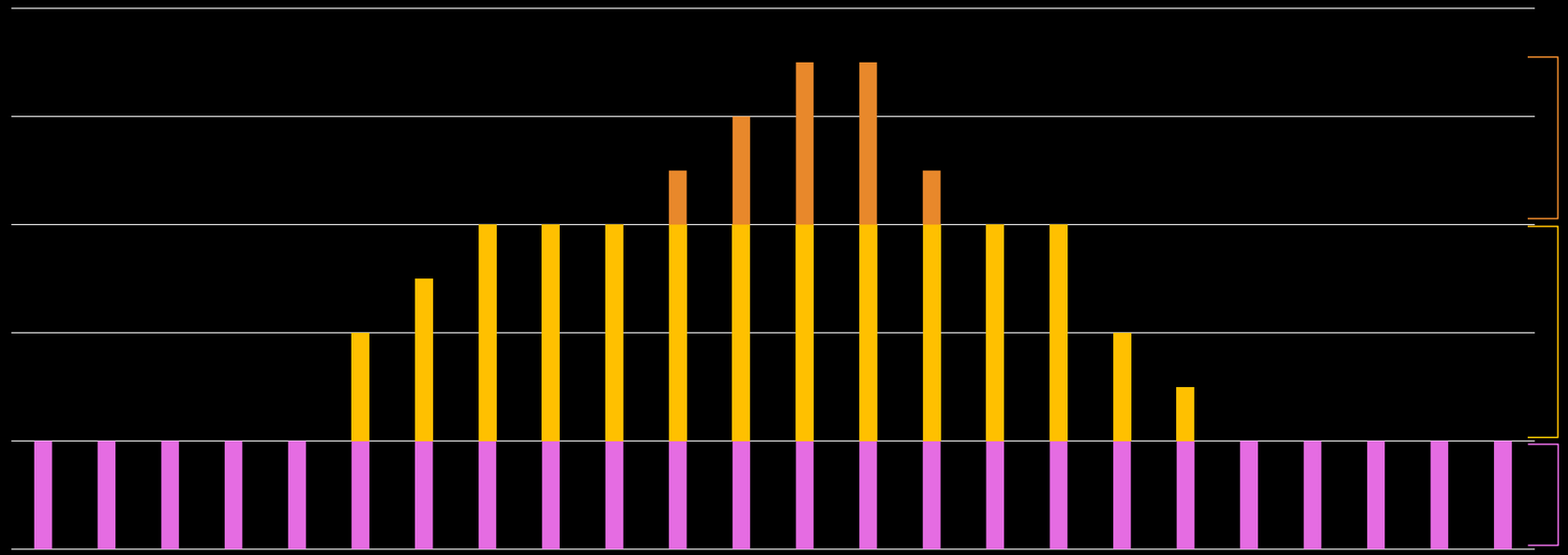
---

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
- Amazon EMR
- AWS CloudFormation
- AWS Batch