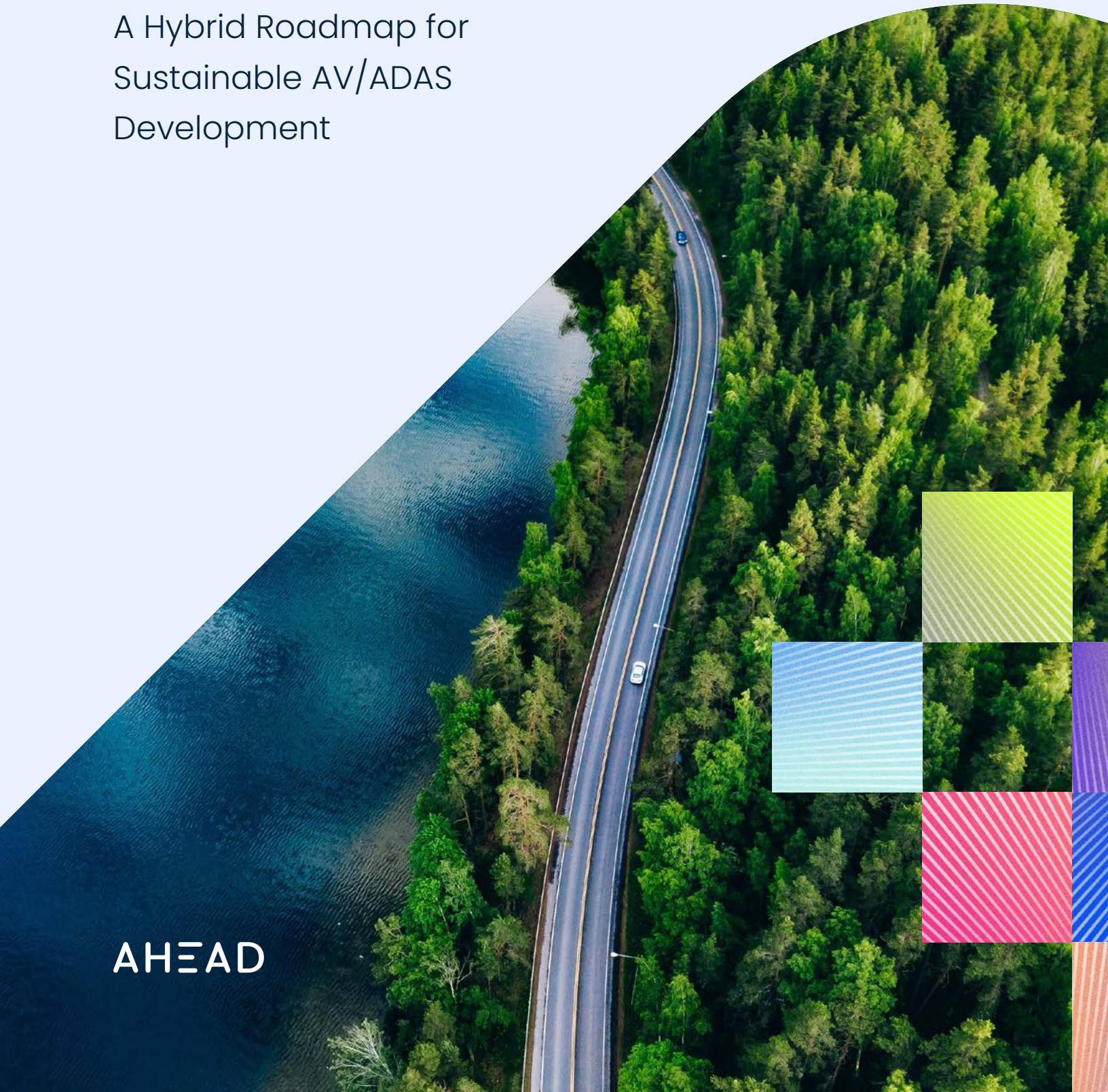
AHEAD's Physical AI Factory

A Hybrid Roadmap for Sustainable AV/ADAS Development



AHEAD

Executive Summary

The automotive industry is crossing a critical infrastructure threshold. The question is no longer whether to invest in AI-driven vehicle intelligence, but how to build a computational foundation that is economically sustainable and competitively durable.

This whitepaper provides automotive executives with a neutral, evidence-based analysis of the hybrid cloud approach to AV/ADAS development and the rising case for purpose-built, on-premises **Physical AI Factories.**

Several converging forces are reshaping infrastructure decision-making:

- GPU scarcity has made just-in-time cloud compute unreliable for safety-critical, deadline-driven development cycles.
- TCO math has shifted: sustained training workloads reach on-premise breakeven within 8-14 months; inference and fine-tuning in as few as 4 months.[1]
- Data gravity is real: test fleets generate 1.4-19 TB per vehicle per hour. Moving petabytes to the cloud for training and back down for Hardware-in-the-Loop (HIL) validation creates egress costs that structurally favor co-located storage and compute.
- Repatriation intent is high but selective: 83-86% of enterprise CIOs intend to move at least one workload on-premises, yet this does not signal a wholesale cloud exodus. [2] IDC data confirms fewer than 10% of organizations have fully repatriated anything, while public cloud spending continues to grow strongly (Gartner projects $723 billion in 2025, up 21%).

This paper analyzes the actual production workflows of industry leaders, examines the hardware trade-offs between NVIDIA, Google TPU Ironwood, and AWS Trainium/Inferentia, and precisely maps AHEAD's Foundry™, Hatch®, and Platform Engineering offerings to each stage of the AV/ADAS development loop.

1 The Infrastructure Inflection Point

From Cloud-First to Hybrid Maturity

For almost a decade, public cloud was the default home for automotive AI data. Early autonomous vehicle (AV) programs captured raw sensor streams indiscriminately and stored redundant miles by the petabyte. Cloud elasticity solved the immediate infrastructure problem while offering flexible compute options beyond GPUs.

However, as datasets and model complexity exploded, the limitations became clear. **Today, three structural shifts are driving a deliberate move toward hybrid architectures:**

1. **Data volumes are now curated rather than exhaustive. Intelligent Shadow Mode and snapshot triggering can reduce stored data by 70–90%.** [3]
2. **GPU supply remains constrained. Jensen Huang confirmed that B200 and GB200 chips are sold out through mid-2026, with a 3.6-million-unit backlog from cloud providers alone.** [4]
3. **Workloads have matured from experimental and bursty to predictable, high-utilization training loops with stable architectures.**

The result is a clear pull toward hybrid infrastructure: the cloud for collaboration, burst capacity, and global fleet management; on-premises AI Factories for sustained training, massive data gravity, and HIL validation.

The GPU Supply Crisis: A Structural Risk, Not a Cyclical Blip

The unprecedented demand for generative AI created a perfect storm. Cloud providers increased capital expenditures by more than 63% year-over-year in 2024, reaching ~$443 billion, with ~75% tied to AI infrastructure. Projections for 2026 show an additional ~36% increase. [5] The January 2025 Taiwan earthquake further highlighted this fragility exacerbating the demand problem by damaging up to 30,000 wafers at TSMC and temporarily disrupting production. For automotive OEMs operating under rigid safety-critical timelines, dependence on unpredictable cloud GPU quotas introduces unacceptable development risk. Owning a strategic portion of physical AI infrastructure has become essential for supply-chain resilience.

When Renting Becomes More Expensive Than Owning

For sustained, high-utilization workloads typical of AV model training, on-premise infrastructure now achieves breakeven against cloud rentals within 8–14 months. For inference and fine-tuning, the payback period can be as short as 4 months (Lenovo Press, "On-Premise vs. Cloud: Generative AI Total Cost of Ownership," n 1).

The following table reflects verified Q1 2026 publicly available pricing for AWS, Azure, and GCP:

FINANCIAL METRIC	AWS	AZURE	GCP	ON-PREMISE AI FACTORY
8x H100 — On-Demand	$55.04/hr[1]	$98.32/hr[2]	$88.49/hr[3]	~$12-15/hr[7]
8x H100 — 3-Yr Committed	~$30.27/hr[4]	~$59-65/hr[5]	~$40-50/hr[6]	N/A — CapEx model
Data Egress (per GB)	$0.09[8]	$0.085[8]	$0.12[8]	$0.00
Storage Standard (per TB/Mo)	$23.55	$18.84	$20.48	~$4-6

[1] AWS p5.48xlarge post-June 2025 price cut (~44% reduction). [2] Azure ND96isr H100 v5; Azure has not matched AWS cuts. [3] GCP a3-highgpu-8g; sustained-use discounts reduce this ~20% at full utilization (~$70/hr effective). [4] AWS 3-yr Savings Plans ~45% off on-demand. [5] Azure 3-yr reserved, ~35–40% off. [6] GCP 3-yr CUDs, ~45–55% off. [7] Derived TCO in Lenovo paper: DGX H100 at $350–$450K amortized 5yr + opex/power/cooling; not a listed rate. [8] Egress for first 10TB/month; rates escalate above threshold.

2 The Cloud Repatriation Signal

What the Data Actually Says

Organizations aren't abandoning the cloud entirely. What we are seeing is **selective maturation:** companies are thoughtfully moving certain predictable, high-utilization, data-heavy workloads off shared public cloud tenancy, while keeping the cloud for its strengths in elasticity, collaboration, and burst capacity. Mature autonomous vehicle model training is the textbook example of exactly the kind of workload that benefits from this balanced shift.

A number that gets quoted constantly, and often misunderstood, is the Barclays CIO Survey. In the first half of 2024, 83% of enterprises said they planned to repatriate at least one workload, rising to 86% by late 2024. Importantly, this does not mean 83% of total compute volume is leaving the cloud; it simply means most companies are bringing at least one system back (Barclays, CIO Survey, n 2). The data confirms this is a selective, not wholesale, movement: IDC shows fewer than 10% of organizations have fully repatriated anything, while public cloud spending continues to grow strongly (Gartner projects $723 billion in 2025, up 21%).

Across most industries, the main drivers for these moves are cost management and compliance. But in the AV world, the case for selective repatriation is even stronger thanks to a **few vertical-specific realities:**

Petabyte-scale data gravity: AV test vehicles generate 1.4-19 TB of sensor data per hour (Tuxera/Siemens Polarion research). A modest 10-car fleet running 8-hour shifts can easily produce 200 TB per shift, roughly 2 PB per day in the extreme (DXC Technology, "Ensuring effective autonomous vehicle data ingestion," n.3). Uploading that to the cloud for training, followed by hardware-in-the-loop testing, triggers massive egress charges of $0.09-$0.12 per GB. At petabyte scale, those costs compound fast.

Regulatory and data sovereignty requirements: GDPR and the EU AI Act put strict controls on personal vehicle data for behavioral biometrics, location history, camera footage, face and license plate details from video, and sovereign storage requirements. Running everything on-premise gives you a cleaner compliance posture that shared cloud environments can't always match without significant engineering.

GPU supply chain control: Dedicated on-premise infrastructure shields you from cloud provider quota limits, regional availability issues, geopolitical conflicts, and the kind of supply disruptions we saw in early 2025 from the Taiwan earthquake.

Model iteration speed: AV safety validation demands billions of simulated miles per software release. Every round-trip of simulation data through the cloud's egress pipes adds real delay to your development cycles.

This combination of economic, regulatory, and operational pressures makes selective repatriation not just attractive, but strategically essential for mature AV/ADAS programs.

3 THE HIDDEN COST: Cloud Egress & Hardware-in-the-Loop Validation

HIL testing remains a non-negotiable requirement in automotive development. Trained models must be validated on actual vehicle ECUs, such as Mobileye EyeQ6H boards or NVIDIA DRIVE AGX Orin, to confirm quantized behavior and satisfy ISO 26262 and SOTIF safety standards. This is done by streaming synthetic sensor data through the ECUs and verifying that the control outputs match expected vehicle behavior.

When training data resides in the cloud while HIL rigs are on-premises (a common arrangement for organizations that have not yet built dedicated AI Factories), every validation cycle requires moving large volumes of scenario data between environments. The costs and delays of this movement are frequently underestimated.

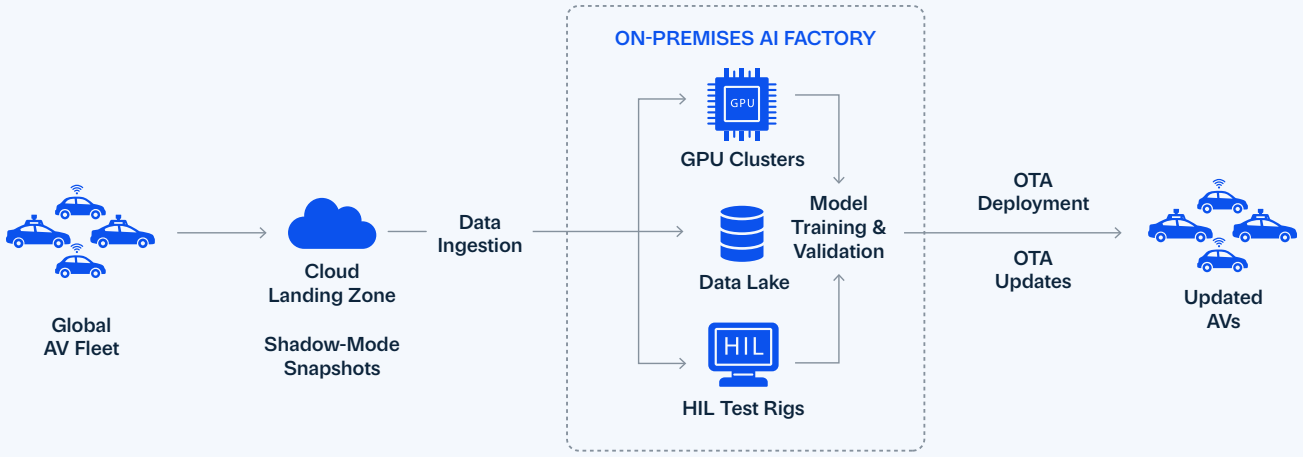
Consider a mid-tier OEM running three HIL cycles per week, with each cycle moving 50 TB of scenario data. At current AWS egress rates of $0.09 per GB, this generates approximately $4,250 per cycle, more than $660,000 annually for just one workstream. For programs running multiple concurrent validation streams, the cost multiplies quickly, and iteration speed suffers.

When HIL volumes exceed roughly 20 TB per week, the economics shift decisively in favor of co-locating training data, GPU clusters, and HIL infrastructure on-premises. As illustrated in the figure below, the optimal architecture for most mature AV/ADAS programs therefore becomes:

- Primary data lake and HIL test rigs on-premises (co-located with the model-training GPU cluster)
- Public cloud retained for burst capacity, global fleet data aggregation, collaboration, and OTA distribution

This co-located design eliminates repeated egress charges and dramatically shortens validation latency, turning what was once a hidden cost into a clear competitive advantage in development velocity.

Hybrid AI Infrastructure for AV/ADAS Development



4 COMPUTE: AV Accelerator Platform Comparison

The debate between on-premise and cloud infrastructure for autonomous vehicle workloads is already complex. It becomes even more layered when automotive infrastructure architects must also choose the right accelerator silicon for training and inference. Every platform carries distinct trade-offs in performance, ecosystem fit, cost, and on-premise availability.

A wide range of approaches are working successfully in production today. **The table below compares the four leading platforms across the dimensions that matter most for AV/ADAS teams.**

DIMENSION	NVIDIA B300 / GB300 NVL72 (BLACKWELL ULTRA)	GOOGLE TPU IRONWOOD (V7)	AWS TRAINIUM2/ TRAINIUM3	AWS DL1 HABANA GAUDI (EYEQ TARGETING)
Primary Strength	Versatility + massive ecosystem; up to 1.1 ExaFLOPS per rack; 50× more AI factory output than previous gen	Raw matrix-math efficiency and huge scale; ~30-41% lower TCO for massive training runs	Seamless SageMaker integration; 30–50% cost savings vs. equivalent GPUs	Tight integration with Mobileye EyeQ chips for a smooth training-to-vehicle pipeline
AV Training Fit	Excellent: full CUDA and PyTorch support; works with every major AV framework (NeMo, RAPIDS, DriveOS)	Strong for big transformer/ BEV models (JAX/PyTorch/ XLA required)	Strong for LLM-style models inside AWS; uses Neuron SDK	Ideal for programs targeting EyeQ SoCs; uses Intel's unified toolchain
AV Inference Fit	Best-in-class for vehicles (DRIVE AGX Orin / DRIVE Thor) and synthetic data pipelines	Great for massive cloud inference; not for vehicles	Good for high-volume cloud inference (Inferentia2); not for vehicles	Gaudi not used for vehicle inference but provides the most direct cloud training path to EyeQ6/EyeQ Ultra
On-Prem Availability	**Yes** DGX, HGX, and full SuperPOD systems (liquid cooling and 800 Gb/s networking required)	**No** GCP Cloud only	**No** AWS cloud only (Trainium4 may add hybrid options later)	**Limited** Limited server cards available, but ecosystem support lags far behind NVIDIA
Quantization Tools	Best-in-class (TensorRT-LLM, native NVFP4/INT8/ FP8)	Native BF16/FP8/INT8; no CUDA-based QAT tooling; model must be rebuilt for XLA compilation	NeuronCore supports INT8/BF16/FP8; still maturing for AV SoCs	Strongest native path to EyeQ (Intel Neural Compressor + OpenVINO)
Recommendation	Default for on-premise AI Factories and full HIL pipelines	Good for cloud-burst pre-training if you're already on GCP	Strong TCO play for AWS-native teams	Best choice for Mobileye EyeQ programs

How Leading AV Companies Are Actually Combining These Platforms

Different teams are mixing and matching accelerators to match their scale, regulatory needs, and vehicle hardware. The range of working approaches is broad, yet one clear pattern stands out: NVIDIA plays a role in nearly every stack, especially for on-premise training workloads.

Tesla trains the vast majority of its FSD models on massive NVIDIA GPU clusters (A100s, H100s, and now Blackwell-era systems). It tried its own Dojo custom chips for training, but has pivoted to focus NVIDIA on training and its custom AI5/AI6 silicon purely on vehicle inference. This gives Tesla the fastest iteration cycles while keeping vehicle hardware lean. [9]

Waymo leans heavily on Google Cloud TPUs (Ironwood-class) for large-scale pre-training and its new World Model simulation work. It still uses NVIDIA GPUs in parts of its AI Factory for certain perception tasks and validation. For the vehicle itself, Waymo relies on custom silicon chips rather than off-the-shelf inference hardware. [10]

Zoox (Amazon) runs its foundation-model training on AWS SageMaker HyperPod, mixing NVIDIA GPUs for flexibility and Trainium chips for cost savings on large multimodal models. Vehicle inference uses Zoox's purpose-built robotaxi compute, keeping the stack fully under Amazon's control. [11]

Mobileye-centric programs (Intel-owned) use Habana Gaudi accelerators in the cloud or on-prem for training, then deploy straight to EyeQ6/EyeQ Ultra chips in the vehicle. The unified Intel toolchain eliminates painful cross-vendor quantization steps. [12]

Aurora (driverless trucking) pairs NVIDIA DRIVE Thor SoCs for in-vehicle inference with NVIDIA's broader ecosystem for training and simulation. Its recent partnership with Continental and NVIDIA shows how the platform scales to high-volume manufacturing. [13]

Other players like Cruise follow similar hybrid patterns, often starting with NVIDIA DGX systems for on-prem training before adding cloud burst capacity.

The Bottom Line

Every major AV program is blending accelerators in its own way. Some stay almost entirely on NVIDIA, others mix in TPUs or Trainium for cost or scale, and a few leverage Gaudi for EyeQ-specific flows. Yet across all these approaches, NVIDIA remains the de-facto standard for on-premise AI Factory builds. Its CUDA ecosystem, full-stack continuity from training to vehicle (via DRIVE AGX/Thor), and mature quantization tools give teams the lowest risk and fastest path from data center to road.

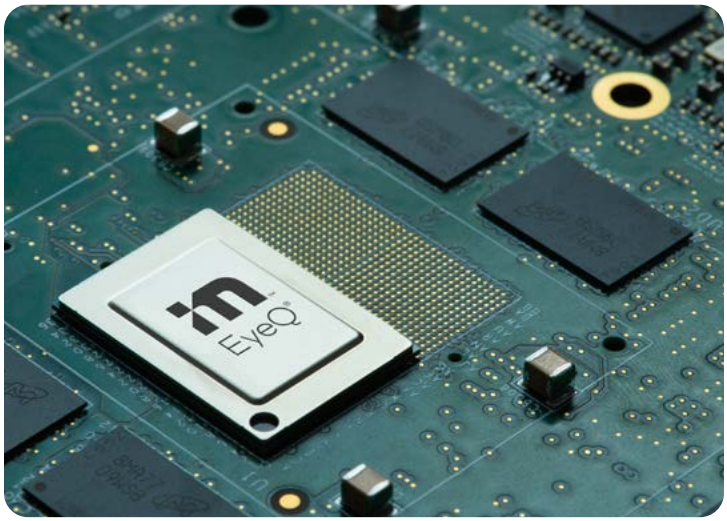
If you're planning an AV compute strategy, start with NVIDIA Blackwell Ultra for the core on-premise workloads. Then, layer in the right cloud option (TPU, Trainium, or Gaudi) only where it delivers clear TCO or ecosystem advantages. This hybrid reality is exactly why selective on-prem maturation, paired with the right silicon, has become the smartest way forward.

5 HOW INDUSTRY LEADERS ACTUALLY WORK: Production Workflow Analysis

To truly understand how mature AV/ADAS programs operate, it is essential to look beyond vendor marketing slides and examine the actual day-to-day technical architectures. The workflows below are drawn from firsthand experience, public case studies, engineering blogs, re:Invent sessions, and official announcements. They reveal a consistent pattern: successful programs have moved past the simplistic "all-cloud" or "all-on-prem" debate and built pragmatic hybrid systems that match the realities of training at scale, quantizing for vehicle silicon, and validating against strict safety standards.

Mobileye: A Proven Hybrid Architecture

Mobileye operates one of the best-documented hybrid systems in the industry. All deep-learning model training runs in the cloud on AWS, primarily using EC2 DL1 instances powered by Intel Habana Gaudi accelerators. This cluster routinely handles more than 250 production training jobs per day and has scaled to peak averages of 500,000 cores (AWS Blogs 2022–2025, n 12).



The decision to train in the cloud was deliberate: Mobileye needed essentially unlimited on-demand resources and saw development velocity increase dramatically after adopting AWS SageMaker Pipe Mode for streaming massive datasets without full data duplication.

Yet the final deployment target, the EyeQ6H chip inside millions of vehicles, is a highly constrained INT8 inference processor. To bridge this gap, Mobileye embeds quantization-aware training (QAT) directly into every cloud training job. During training on Gaudi, the model is continuously exposed to the exact INT8 rounding errors it will encounter in the vehicle. This approach "learns to tolerate" precision loss before production deployment, eliminating painful cross-vendor translation steps.

For the final, non-negotiable validation step, HIL testing on real EyeQ6H boards to meet ISO 26262 and SOTIF requirements, Mobileye keeps its massive 200-petabyte validation dataset primarily on-premises, co-located with the HIL test rigs. AWS handles active training data and overflow, but the critical validation data stays local for speed, compliance, and low-latency iteration.

This hybrid model, combining cloud infrastructure for elastic training capacity with on-premise systems for HIL validation and compliance, has become the blueprint for EyeQ-centric programs.

Tesla: Vertical Integration and the Post-Dojo Reality

Tesla initially took the opposite path, building its own massive on-premise supercomputer called Dojo to process the enormous video data stream from its fleet of more than five million vehicles. Dojo was purpose-built for long-sequence video training and featured an auto-labeling pipeline that dramatically reduced human annotation costs.

However, in 2025, Tesla made a strategic pivot. After evaluating the rapid progress of NVIDIA's GPU ecosystem, the company redirected the majority of FSD training to massive NVIDIA GPU clusters (A100s, H100s, and possibly Blackwell-era systems) while reserving its custom AI5/AI6 silicon exclusively for vehicle inference. This shift allows Tesla to leverage the full CUDA ecosystem for training while maintaining lean, purpose-built hardware in the car, delivering the fastest training-to-deployment iteration cycles in the industry (Musk on X, TechCrunch, Huang NVIDIA Blogs, n 9).

Both Mobileye and Tesla illustrate the same core truth: the winning infrastructure strategy is never purely cloud or purely on-prem. It is the mix that best aligns with three non-negotiable realities: training at hyperscale, quantizing for real vehicle silicon, and validating on production-representative hardware under strict safety standards.

6 SHADOW MODE & SNAPSHOT ARCHITECTURE: The Intelligence Flywheel

What Is Shadow Mode?

Shadow Mode is one of the most powerful tools in modern AV/ADAS development. It works by running a candidate software stack—consisting of a new perception model, path planner, or end-to-end neural network—in parallel with the production software on deployed vehicles. The candidate model receives real sensor data and makes predictions, but it has no authority over the vehicle and does not generate execution commands. This creates a continuous, real-world validation environment at fleet scale without risking safety.

The concept is not new and is used across the industry. For example, Ford implements it under the name "Continuous Learning Loop" with its BlueCruise-equipped vehicles. When a driver intervenes and takes over, the system captures the event and uses it to improve the next release. Tesla, GM, Waymo, and most other leaders use similar mechanisms. The sophistication of the triggering logic continues to increase, turning the entire production fleet into a live testbed.[14]

Snapshot Architecture: Prioritizing the High-Value Tail

Storing every mile of sensor data from an entire fleet is economically and logistically impossible. The breakthrough is Snapshot Architecture: intelligent edge triggering that captures only the highest-value moments instead of everything.

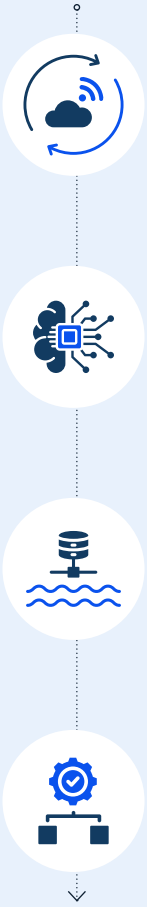
The table below shows the most common trigger categories used by leading programs:

TRIGGER CATEGORY	TRIGGER CONDITION	DATA CAPTURED
Shadow Disagreement	Candidate model diverges >X meters from production model or human driver	60–120 seconds of all sensors + ECU state
Safety Intervention	AEB, TCS, ESC activation or driver override	Pre- and post-event full sensor stream
Low-Confidence Perception	Object detection confidence below threshold or novel scene (construction, weather, etc.)	Sensor data plus confidence scores for labeling
Map Divergence	Observed road geometry diverges >N cm from HD map	Compressed camera plus lidar (≈10 kb/km)
Scenario Rarity	Scene matches a "sparse scenario" in the fleet distribution database	Selective capture based on rarity score

The effect of intelligent snapshot triggering is a dramatic reduction in the volume of data that must be stored, transmitted, and labeled to support the AV development lifecycle. Cruise has reported that fewer than 1% of raw sensor data collected from its San Francisco fleet contains actionable information, a figure that drove its investment in automated AI-based event extraction and classification.[15] Additionally, Heex Technologies' own product documentation claims deploying intelligent triggers for real-time capture at the edge delivers 70–90% data reduction compared to continuous recording.[16]

How Shadow Mode Fits the Hybrid Cloud Architecture

Shadow Mode and Snapshot architecture create a very specific data flow that directly dictates infrastructure design:



Cellular transmission to cloud: compressed snapshot packets are transmitted via 5G/LTE to a cloud-based connected vehicle platform for initial aggregation (the ideal use of cloud for global scale), which then triggers a service from the AV Ops pipeline to pick up the snapshot feeding it back into the training cycle. This is the appropriate use of cloud: global data aggregation from a distributed fleet, with no latency sensitivity.

Cloud-to-on-prem transfer: a service from the AV Ops pipeline is triggered to pick up high-priority snapshots, moving them once to the on-premise AI Factory for auto-labeling and training.

On-prem closed-loop simulation and training: all subsequent heavy compute (labeling, model training, digital-twin reconstruction) happens on-premise, co-located with the data lake and GPU cluster.

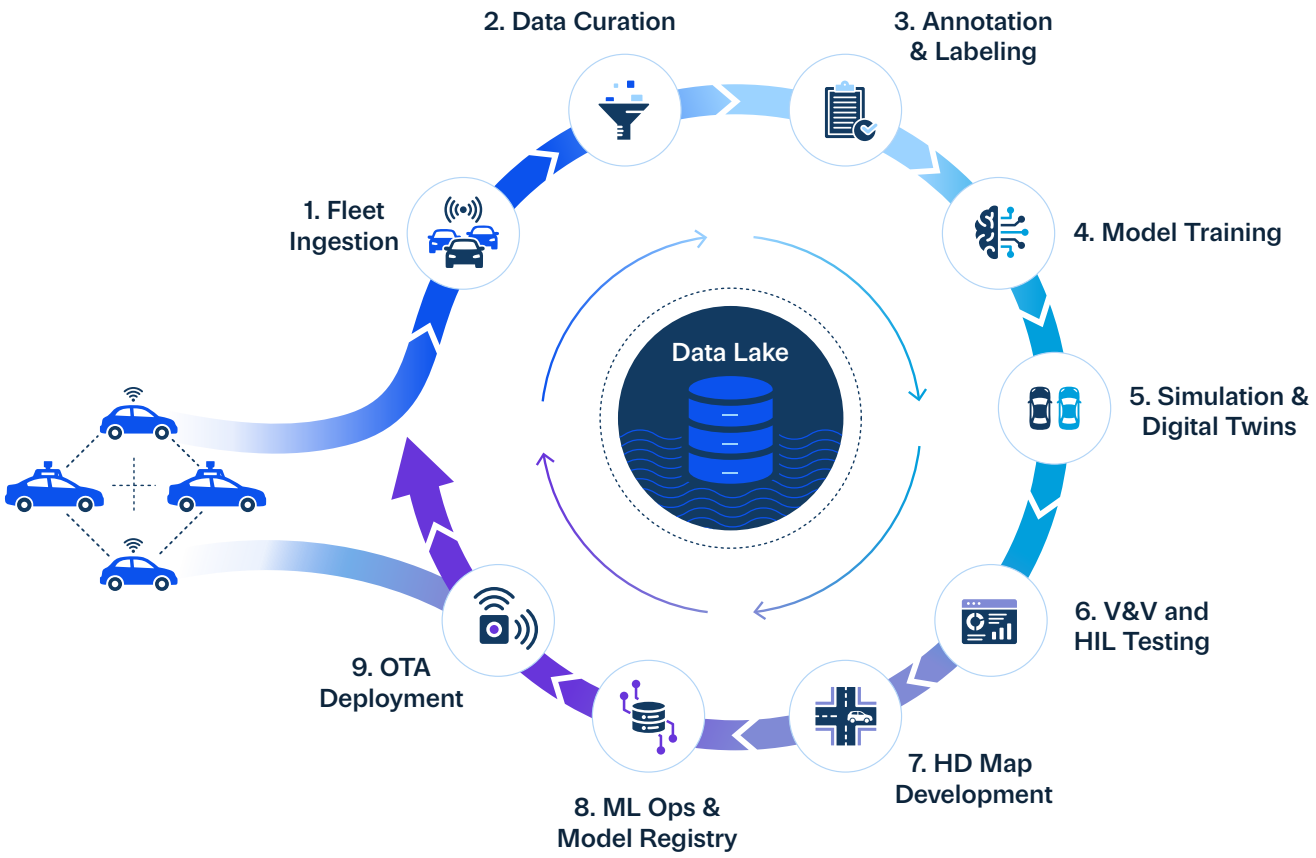
HIL validation on-prem: final safety validation runs locally on real ECUs.

This pattern explains why a pure cloud approach creates inefficiency: data lands in the cloud, gets pulled on-prem for training, results go back to the cloud for orchestration, and then must be pulled on-prem again for HIL. Each round-trip adds egress cost and latency. By using cloud infrastructure to handle fleet data ingestion and global aggregation, while keeping the heavy training and validation workloads on-premise, the hybrid model eliminates those penalties and accelerates iteration cycles.

7 The Complete AV/ADAS Development Loop

The following framework describes the full development lifecycle from vehicle sensor to OTA deployment.
It integrates Shadow Mode and Snapshot dynamics with the precise infrastructure implications of each stage, showing exactly where on-premise AI Factory capabilities deliver the greatest advantage and where cloud resources remain optimal.

The AV/AVAS Development Loop



The table below explains in detail the process at each step and maps it to AHEAD offerings.

#	STAGE	WHAT HAPPENS	INFRASTRUCTURE	AHEAD SOLUTION
1	**Shadow Disagreement**	Production and test fleets run Shadow Mode. Onboard trigger logic flags snapshots (disagreements, interventions, low-confidence perception). Data offloaded via cellular or 400 Gbps Data Transfer Terminals at service centers.	Edge: vehicle compute + 5G/LTE. Landing: cloud object store (S3/Blob) for global aggregation; on-prem high-speed ingestion terminals at depots.	Networking & Edge AI design; Data Transfer Terminal architecture
2	**Data Curation**	Ingested data is de-identified (GDPR), validated, compressed, and indexed. High-priority snapshots are transferred to on-premise data lake for processing. Rare scenario distribution is tracked to manage dataset balance.	Hybrid: cloud landing zone → on-premise primary data lake (petabyte-scale NVMe/flash storage). On-prem recommended once HIL volume justifies TCO.	AHEAD Foundry™: petabyte-scale storage design; FlashBlade + PortWorx; data pipeline engineering
3	**Annotation & Labeling**	Auto-labeling models (Teacher models on GPU cluster) generate initial ground truth for objects, lanes, and free-space. Human annotators handle low-confidence edge cases. Scenario rarity scoring updates fleet trigger thresholds.	On-premise GPU cluster (or cloud burst for overflow). Labeling infrastructure should be co-located with data lake to avoid repeated egress.	AHEAD Foundry™: GPU cluster for auto-labeling; MLOps platform for annotation workflow
4	**Model Training**	Distributed training across DGX GPU clusters. For EyeQ/SoC targets, QAT is incorporated into training to simulate deployment precision. World Foundation Model pre-training, perception model fine-tuning, and driving policy training are distinct workloads.	On-premise AI Factory (DGX H100/B200 + InfiniBand/NVLink) for sustained training. Cloud burst (AWS P5/Trainium3) for overflow or organizations not yet at CapEx threshold.	AHEAD Foundry™: DGX SuperPOD design and integration; liquid cooling; InfiniBand networking
5	**Simulation & Digital Twins**	Curated snapshots are reconstructed as interactive 3D environments using neural rendering (NeRF/3DGS). World Foundation Models generate synthetic variations. Billions of virtual miles validate safety behaviors.	Primarily on-premise (data co-location, latency requirements for closed-loop simulation). Cloud burst for overflow scenario generation.	AHEAD Platform Engineering: Omniverse/Cosmos integration; on-prem simulation cluster design

#	STAGE	WHAT HAPPENS	INFRASTRUCTURE	AHEAD SOLUTION
6	**V&V and HIL Testing**	Trained models are validated against a library of 1M+ scenarios. Hardware-in-the-Loop (HIL) testing runs models on actual ECU hardware (EyeQ6H, DRIVE AGX, custom SoCs) against synthetic sensor streams. ISO 26262 / SOTIF compliance requires this stage on production-representative hardware.	On-premise mandatory (automotive-grade HIL test rigs). Data must be available locally. Cloud egress to feed HIL creates cost and latency penalties.	AHEAD Foundry™: HIL rack integration; on-prem storage co-location; test infrastructure design
7	**HD Map Development**	Mobileye REM crowdsourcing aggregates map updates from fleet (Mobileye: 10kb/km packets). Map fusion, validation, and versioning are centralized. HD map packages are distributed to vehicle fleets.	Cloud-appropriate (global aggregation). Map fusion compute can be cloud or on-prem; map distribution via CDN.	AHEAD Platform Engineering: cloud-native map pipeline; CDN distribution architecture
8	**MLOps & Model Registry**	Validated models are version-controlled, tagged with safety confidence scores, and registered in the model repository. Performance monitoring tracks model drift across the production fleet. Automated canary deployment pipelines manage staged rollouts.	Hybrid: model registry on-prem or in cloud; monitoring infrastructure wherever telemetry aggregates.	AHEAD Platform Engineering: MLOps pipeline design; Kubeflow/MLflow integration; Hatch® for GPU asset lifecycle visibility
9	**OTA Deployment**	Models are packaged for target ECU (quantized, compressed, signed). Staged OTA rollout begins with internal test vehicles, then early adopters, then full fleet. Shadow Mode immediately activates for the new model, beginning the cycle again.	Cloud CDN for distribution. On-prem for signing infrastructure and initial staging. Fleet telemetry aggregated in cloud.	AHEAD Hatch®: lifecycle tracking of deployed model versions and GPU hardware; rollback infrastructure; OTA security architecture

8

AHEAD SOLUTIONS:

Infrastructure for the Full Development Loop

AHEAD's portfolio is purpose-built for the complexity of the AV/ADAS development lifecycle.
Three core offerings map directly to the infrastructure requirements established by this whitepaper's analysis.

AHEAD Foundry™ | hatch | AHEAD Platform Engineering

AHEAD Foundry™: Building the Physical AI Factory

AHEAD's Foundry™ is a specialized 10-megawatt facility for designing, building, and integrating custom rack-scale AI infrastructure. It is the physical manifestation of the AI Factory concept, purpose-built to handle the density, cooling, and networking requirements of modern GPU clusters.

DGX SuperPOD Design & Integration

AHEAD specializes in NVIDIA DGX H100 and B200 SuperPOD deployments, including the InfiniBand NDR networking and NVLink fabrics required for high-performance distributed training across thousands of GPUs.

Liquid Cooling Engineering

Next-generation GPU racks operate at power densities up to 250 kW per rack. AHEAD's Foundry™ designs direct-to-chip water cooling systems that dissipate heat approximately 1,000x more efficiently than air cooling, a prerequisite for NVIDIA Blackwell deployments.

Petabyte-Scale Storage

AHEAD implements high-performance unstructured storage solutions (including FlashBlade and PortWorx) designed to support 400 Gbps data ingestion terminals and feed GPU clusters without becoming an I/O bottleneck.

HIL Infrastructure Co-Location

Foundry™ designs data center layouts that co-locate HIL test rigs with the training data and GPU clusters that feed them, eliminating egress costs and minimizing validation latency.

AHEAD Hatch®: Lifecycle Intelligence for AI Hardware

AHEAD Hatch provides continuous visibility into the health, utilization, and aging of GPU assets and networking components throughout the AI Factory. For automotive programs with multi-year development cycles, this intelligence is operationally critical:

Hardware Lifecycle Planning

Hatch tracks the remaining useful life of GPU clusters and proactively identifies the optimal window for transitioning from Hopper (H100) to Blackwell (B200/B300) architecture, minimizing both obsolescence risk and premature capital cycling.

Utilization Optimization

AV training workloads have distinct utilization patterns across stages (labeling, QAT training, simulation, HIL). Hatch maps actual utilization against planned capacity to identify underutilization and guide burst-to-cloud decisions.

Model Version Tracking

As new software versions are deployed via OTA, Hatch maintains the mapping between deployed model artifacts and the hardware generations and training infrastructure that produced them, essential for root-cause analysis of safety investigations.

AHEAD Platform Engineering: The MLOps Backbone

AHEAD's Platform Engineering practice builds the software infrastructure that makes the AI Factory operational and the development loop repeatable:

MLOps Pipeline Design

AHEAD implements automated training, validation, and deployment pipelines using Kubeflow, MLflow, or NVIDIA's NeMo framework, integrated with the organization's CI/CD infrastructure and safety validation gate requirements.

Hybrid Cloud Orchestration

For organizations maintaining both on-premise AI Factory infrastructure and cloud burst capacity, AHEAD designs the networking, data replication, and workload scheduling layers that allow seamless workload movement without manual intervention.

Simulation Platform Integration

AHEAD integrates NVIDIA Omniverse and Cosmos World Foundation Model pipelines into the development loop, enabling neural simulation at scale with proper data management across the on-prem/cloud boundary.

Security & Compliance Architecture

GDPR-compliant data de-identification pipelines, model artifact signing for OTA deployment, and access control frameworks for sensitive training data are embedded into platform designs as first-class requirements.

9 STRATEGIC TRADE-OFF ANALYSIS: On-Premise AI Factory

Every mature AV/ADAS program eventually reaches the same strategic question: when does it make sense to move from cloud-first to a hybrid model anchored by a dedicated on-premise AI Factory? The answer is never binary. It depends on program scale, data volume, HIL intensity, regulatory pressure, and TCO horizon. **Below is a clear-eyed assessment of the trade-offs.**

The Case For
ON-PREMISE AI FACTORY

Predictable TCO: once initial CapEx is amortized (8-14 months for training at high utilization), the cost curve is essentially flat. Programs are protected from consumption-based "bill shock" as training scale grows.

HIL Ecosystem Integration: co-locating training data, GPU clusters, and HIL validation rigs eliminates the egress costs and latency penalties detailed in Section 3, enabling iteration cycles that cloud-only architectures cannot match.

GPU Supply Security: ownership removes exposure to cloud provider quota limits, regional availability swings, and geopolitical risk. DGX SuperPOD procurement timelines (6-10 months) are predictable and plannable.

Data Sovereignty: on-premise infrastructure provides an unambiguous GDPR and EU AI Act posture for sensitive vehicle data (camera footage, behavioral biometrics, location history) without relying on third-party contractual commitments.

Performance Optimization: Dedicated on-premise infrastructure gives organizations full control over cooling configuration and power delivery, enabling sustained peak GPU utilization. Standard shared-tenancy cloud instances cannot guarantee equivalent consistency, though dedicated bare-metal cloud offerings narrow this gap.

The Case Against
ON-PREMISE AI FACTORY

Minimum Viable Investment: Building a meaningful AI Factory typically requires $50-$100M in initial capital. This concentrates risk in a single CapEx cycle and requires board-level commitment.

Operational Burden: The OEM assumes responsibility for power, cooling, hardware maintenance, and network operations – capabilities that most automotive organizations do not have at scale in their IT organizations.

Talent Scarcity: Managing high-density GPU clusters and complex MLOps pipelines requires engineers that currently command AI-company-level compensation. Most automotive OEMs are competing with hyperscalers for this talent.

Capacity Ceiling: Unlike cloud, an AI Factory has a physical ceiling. Sudden burst demand, triggered by a critical safety investigation or accelerated program timeline, cannot be instantly accommodated without pre-planned cloud burst architecture.

Rapid Silicon Obsolescence: GPU architecture generations are currently cycling every 18-24 months (H100 → B200 → B300). On-premise CapEx decisions lock in a specific generation, whereas cloud automatically migrates to new silicon.

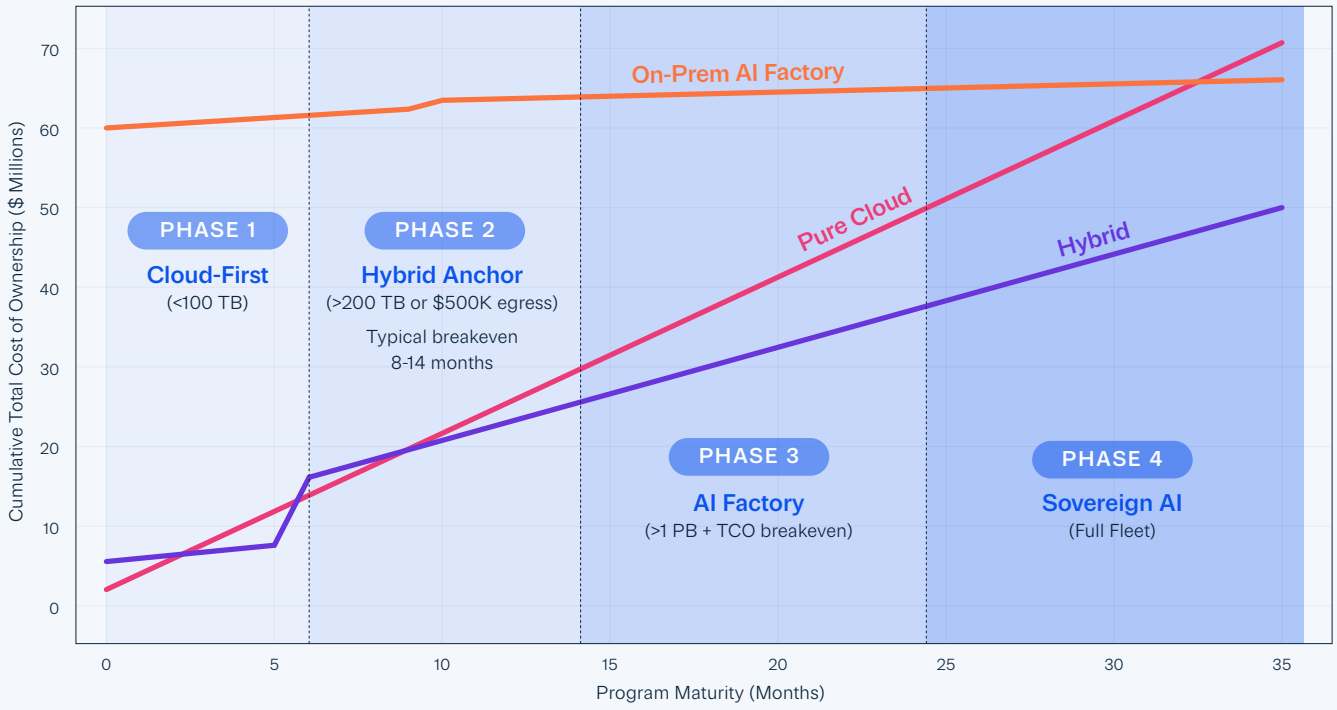
The Recommended Hybrid Maturity Model

Given these trade-offs, the optimal path is a phased, data-driven progression calibrated to program maturity and operational metrics:

PHASE	TRIGGER CONDITION	INFRASTRUCTURE MODEL	AHEAD ROLE
1. Cloud-First	Early-stage; <100 TB data; bursty training	Full cloud (AWS P5 / GCP A3); on-prem HIL hardware only	Platform Engineering: cloud MLOps architecture; HIL networking
2. Hybrid Anchor	>200 TB data; HIL egress >$500K/year; predictable training utilization	On-prem data lake plus HIL co-location; cloud burst training	Foundry™: storage design + ingestion terminals; Hatch®: utilization monitoring
3. AI Factory	>1 PB active data; TCO breakeven validated; GPU supply risk material	On-prem DGX SuperPOD + storage + HIL; cloud retained for mapping, OTA, collaboration	Foundry™: full SuperPOD build; Platform Engineering: hybrid orchestration; Hatch®: full lifecycle
4. Sovereign AI	Full fleet deployment; regulatory pressure; multi-region program	Regional AI Factory network; on-prem for all training/simulation; cloud for OTA distribution	Full AHEAD managed services across Foundry™, Hatch®, and Platform Engineering

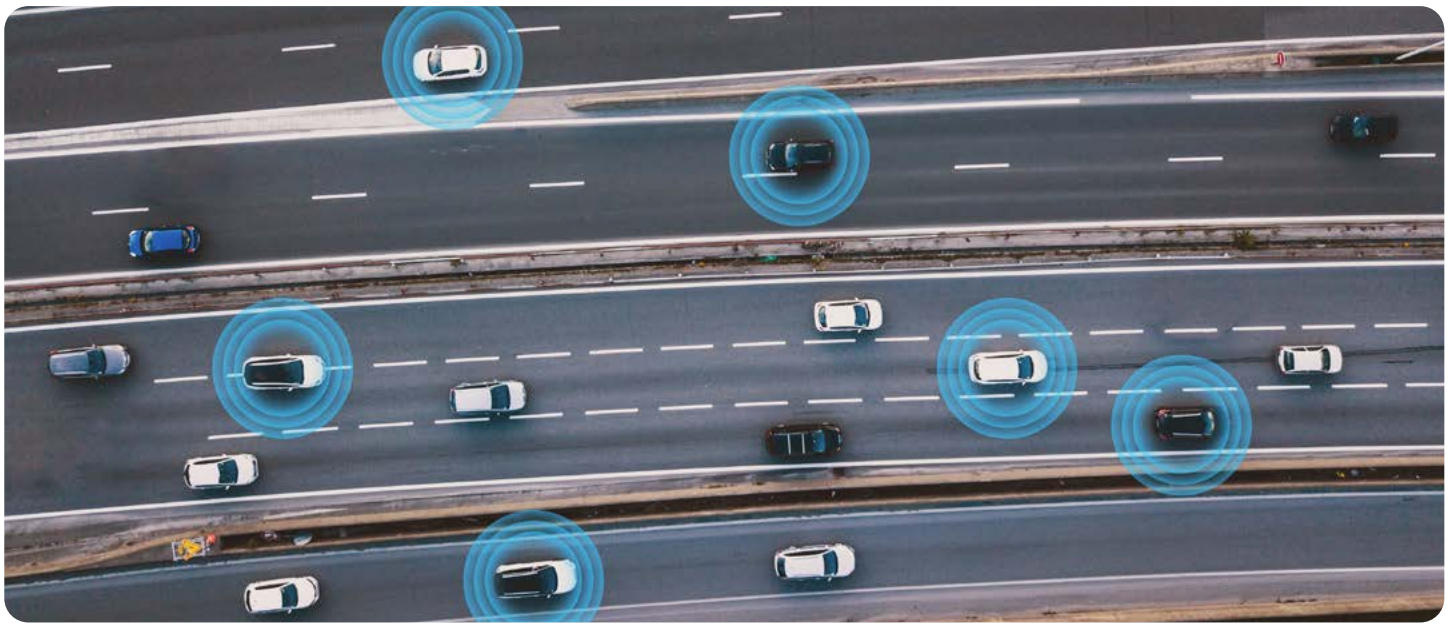
This phased model allows programs to start with low risk and scale infrastructure investment in lockstep with proven business value and safety-validation requirements.

The AV/AVAS Cost Comparison: Pure Cloud vs. Hybrid vs. On-Prem AI Factory



10 CONCLUSION: The Road to Hybrid Infrastructure Maturity

Physical AI has moved from proof-of-concept to production pipeline. Across the automotive industry, the debate over whether to invest in AI-driven vehicle intelligence has given way to a far more practical engineering and financial question: how to structure the infrastructure stack that can sustain continuous model improvement at scale, year after year, while meeting the non-negotiable demands of safety validation and regulatory compliance.



The data analyzed in this whitepaper supports a clear, nuanced conclusion.

The 83-86% repatriation signal captured by the Barclays CIO Survey is real, but it describes selective workload optimization, not a wholesale cloud exodus. The signal is strongest precisely for the workloads that define mature AV/ADAS programs: predictable high utilization, massive data gravity, and tight integration with Hardware-in-the-Loop validation.

Cloud providers are not being displaced. They remain the optimal infrastructure for global fleet data aggregation, HD map distribution, collaborative development environments, and secure OTA deployment pipelines. AWS, Azure, and Google Cloud are also innovating rapidly with cloud-native silicon (Trainium, TPU Ironwood) that delivers meaningful TCO advantages for organizations willing to invest in the associated toolchains.

What has been consistently underweighted in most infrastructure analyses is the true cost of cloud egress in the context of HIL validation. When the full cycle of training in the cloud, then repeatedly pulling scenario data back on-premise for safety-critical testing is taken into account, the breakeven timeline for on-premise investment moves significantly earlier than simple GPU-hour comparisons suggest.

NVIDIA remains the only credible choice for on-premise AI Factory deployments. The CUDA/TensorRT/DRIVE ecosystem is the only complete toolchain that spans training, quantization-aware training, simulation, and vehicle-embedded inference with full continuity. Cloud-native accelerators offer real advantages within their respective ecosystems, but do not translate to on-premise deployment at scale.

The path forward for the modern automotive OEM is therefore a hybrid maturity model: use the cloud for global fleet intelligence and elastic burst compute; build dedicated on-premise AI Factories for petabyte-scale data gravity, HIL-integrated validation, and the sustained GPU utilization that makes AI development economically predictable and competitively durable.

AHEAD is uniquely positioned to architect and deliver each layer of this model. From the physical infrastructure of Foundry™, the lifecycle intelligence of Hatch®, to the MLOps backbone of Platform Engineering, AHEAD provides the complete stack that turns infrastructure from a cost center into a true product differentiator.

In an industry where the time between a training data snapshot and a deployed software update is now a direct competitive metric, the quality of the underlying infrastructure is no longer optional. It is the foundation upon which the next generation of software-defined vehicles will be built.

[1] Lenovo Press, "On-Premise vs. Cloud: Generative AI Total Cost of Ownership," December 2025 and February 2026 editions.

[2] Barclays CIO Survey 2024–2025; IDC "Cloud Repatriation Trends" reports; Gartner "Public Cloud Services Market Forecast, Worldwide," 2025 (projecting $723 billion).

[3] DXC Technology, "Ensuring effective autonomous vehicle data ingestion," n.d. (official DXC blog).

[4] NVIDIA CEO Jensen Huang statements and Financial Content reporting, "NVIDIA's Blackwell Dynasty: B200 and GB200 Sold Out Through Mid-2026," 2025–2026.

[5] CreditSights, "Technology: Hyperscaler CapEx 2026 Estimates," November 2025; Goldman Sachs hyperscaler CapEx forecasts through 2027.

[6] Tuxera/Siemens Polarion research on modern AV sensor suites (lidar + multi-camera + radar), updated 2024–2025.

[7] SemiAnalysis, "TPUv7: Google Takes a Swing at the 900lb Gorilla," November 2025)

[8] "Optimizing cost for building AI models with Amazon EC2 and SageMaker AI," AWS Blog, March 28, 2025.

[9] Elon Musk on X (August 8–10, 2025: "All effort is focused on [AI5/AI6]... Dojo 2 was now an evolutionary dead end"; January 2026 Cortex cluster update: "~10B cumulatively just on Nvidia hardware for training"); TechCrunch, "Tesla Dojo: The rise and fall of Elon Musk's AI supercomputer," September 2, 2025 (updated with 2026 Cortex context); NVIDIA CEO Jensen Huang, CES 2026 keynote confirming Tesla's primary use of NVIDIA GPU clusters for FSD training.

[10] Waymo official blog, "The Waymo World Model: A New Frontier For Autonomous Driving Simulation," February 6, 2026.

[11] AWS re:Invent 2025 session AMZ304 – "Zoox: Building Machine Learning Infrastructure for Autonomous Vehicles" (presented December 2025).

[12] Mobileye / AWS historical case studies and engineering blog (2022–2025) on Gaudi DL1 clusters and QAT for EyeQ6H deployment.

[13] Aurora, Continental, and NVIDIA joint press release, "Aurora, Continental, and NVIDIA Partner to Deploy Driverless Trucks at Scale," January 6, 2025.

[14] "Rolling out BlueCruise 1.3: How we are using data to improve software," Omari, https://www.linkedin.com/pulse/rolling-out-bluecruise-13-how-we-using-data-improve-software-omari/.

[15] Cruise / Google Cloud official blog, "How Cruise tests its AVs on a Google Cloud platform," March 9, 2022 (authored by Mo Elshenawy, Executive Vice President of Engineering at Cruise).

[16] Heex Technologies product documentation, https://doc.heex.io/docs/getting-started/heex-overview.

AHEAD

Combining cloud-native capabilities in software and data engineering with an unparalleled track record of modernizing infrastructure, we're uniquely positioned to help accelerate the promise of digital transformation.

Visit us at [ahead.com.](ahead.com)

National Hubs

CHICAGO
444 W. Lake Street
Suite 3000
Chicago, IL 60606

NEW YORK
500 5th Avenue
Floor 17
New York, NY 10010

ATLANTA
1117 Perimeter Center
W406
Atlanta, GA 30338

SAN FRANCISCO
2000 Crow Canyon Place
Suite 250
San Ramon, CA 94583