Case Study

4th Gen Intel® Xeon® Processors
Intel® Gaudi® Accelerators

Amazon EC2 Instances Featuring Intel® Gaudi® Processors Help Reply Save Up to 40 Percent in AI Training Costs[1]

Rather than GPU-based instances, Reply chooses Amazon EC2 DL1 instances to train the open-source CARLA traffic light detection system that makes self-driving vehicles safer.

Solution Summary

- 4th Gen Intel® Xeon® processors
- Intel® Gaudi® accelerators
- Amazon EC2 DL1, and EC2 M7i
- Amazon Nitro System
- Amazon Elastic Fabric Adapter

Executive Summary

Most fatal car accidents in the United States occur at intersections. As automated vehicles grow in adoption, they must be able to use AI, onboard sensors, and cameras to identify and interpret traffic lights accurately. Reply, a leading system integrator and technology solution provider across multiple industries including automotive, seeks to address that challenge. CARLA, an open-source simulation environment, helps Reply test its autonomous driving models and use the data collected for further model training. Behind the scenes, Amazon EC2 DL1 instances supported by Habana's Intel® Gaudi® accelerators assist in the process. In addition, Amazon M7i instances using 4th Gen Intel® Xeon® processors provide Reply's customers with necessary access to API inference results. In preliminary testing, Reply found Intel Gaudi processor-based Amazon instances provided an attractive alternative to GPU-based instances. Compared to using Amazon instances like the EC2 P4d for multiple hours per day, Reply can save 40 percent in training cost using Amazon EC2 DL1 instances instead.[1]

Challenge

Making self-driving vehicles safe for occupants, other cars, and people nearby is a complex challenge. A critical piece of that puzzle centers on a vehicle's ability to correctly interpret and respond appropriately to traffic signals. The AI-based

Reply looked for options using CPUs, accelerators, and a performant cloud solution, for the AI-based CARLA environment which can simulate car sensors like GPS, LIDAR, and accelerometers.

CARLA environment can simulate car sensors like GPS, LIDAR, and accelerometers. Plus, it provides customizable driving scenarios that mimic variables like nighttime driving or varying weather conditions. GPUs often assist with enabling these capabilities in a machine learning (ML) system. However, the shortage of GPUs encouraged Reply to explore other options using CPUs and accelerators. The M7i instances on AWS also required a performant cloud platform that can scale up dynamically to accommodate high inference workloads during rush hour traffic and scale down when roads are less busy.

> "With a GPU shortage, our team sought alternate approaches to AI model training. We found that Amazon instances with Habana's Intel Gaudi accelerators offered an excellent alternative."
>
> *– Luigi De Martino, Partner at Concept Reply US*

Solution

Reply chose Amazon EC2 instances using Intel's advanced hardware and software, which provided the computational prowess CARLA needed. Reply trains its models using Amazon EC2 DL1 instances supported by Intel® Gaudi® accelerators. In addition, Reply deployed a scalable Kubernetes cluster using Amazon EC2 M7i instances with underlying 4th Gen Intel® Xeon® processors to expose the CARLA API and allow customers to access inference results. Reply also found that Intel's libraries built for CPU-based training unlock a larger pool of resources when using CPUs versus GPUs.

Results

When training models with Intel Gaudi-based Amazon EC2 instances, Reply found an exceptional substitute for commonly used GPU-based instances. Amazon S3 instances provide excellent scale for object storage services, and Amazon M7i instances offer the performance levels necessary for obtaining fast inference results. For multi-hour training processes, using Amazon EC2 DL1 instances rather than Amazon EC2 P4d instances in the automotive market, gives Reply up to a 40 percent cost savings. Reply's expertise in cloud-based solutions also paves the way for other unique offerings. For example, Reply offers a specialized vehicle damage detection system that evaluates photos to determine the extent of dents and dings and predict needed repairs.

Key Takeaways

As autonomous vehicles grow in functionality and popularity, ready-made solutions like Reply's offer excellent ROI through pre-tested, adaptable capabilities that can speed customer adoption.

While GPU-based instances often support model training, Intel Gaudi accelerators can provide a cost-effective alternative.

Outside of self-driving cars, this solution can also benefit people who are color blind or have color vision deficiency (CVD), making the roads safer for all.

For More Information

Explore Intel Xeon processors.

Learn about Habana's Intel Gaudi accelerators.

Read about Amazon EC2 instances and Amazon S3 instances.

Find out more about CARLA.

Investigate Reply's automotive offerings.

[1] Performance claim provided by Reply.

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Intel technologies may require enabled hardware, software or service activation.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

No product or component can be absolutely secure.