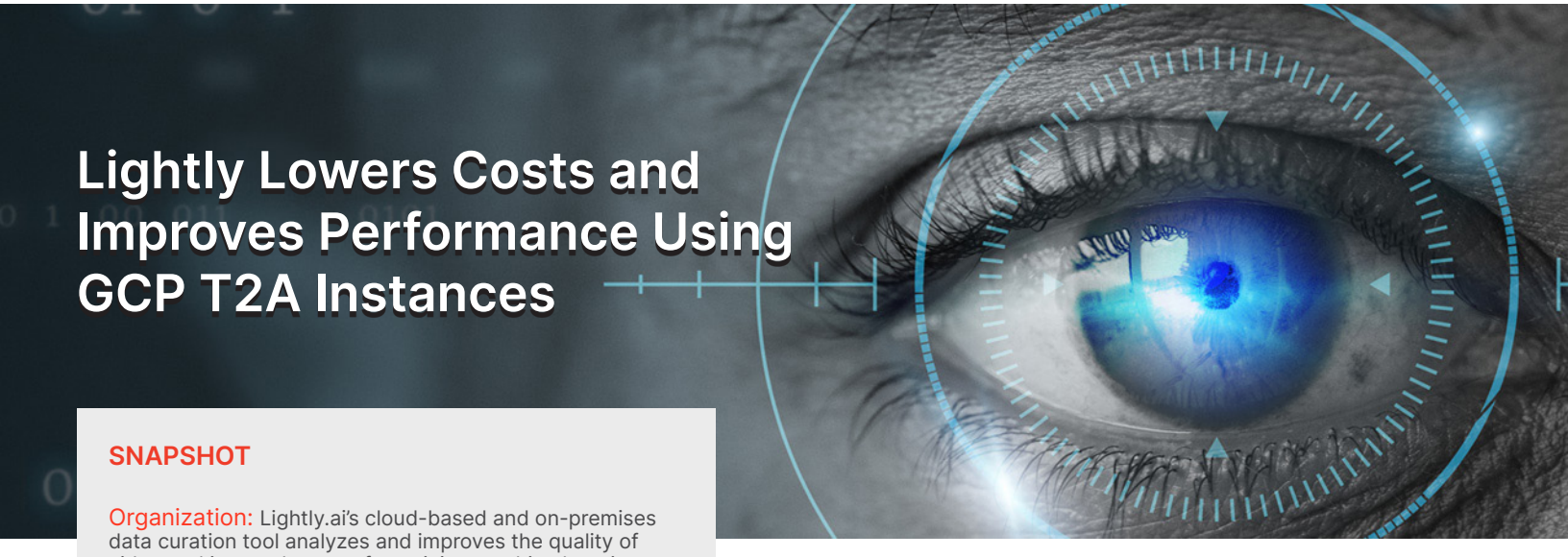
AMPERE.

LIGHTLY | Case Study



Lightly Lowers Costs and Improves Performance Using GCP T2A Instances

SNAPSHOT

Organization: Lightly.ai's cloud-based and on-premises data curation tool analyzes and improves the quality of video and image datasets for training machine learning models. They provide all necessary elements for an AI data preparation feedback loop, such as data curation (active learning), data exploration, online selection, and model monitoring. Lightly.ai facilitates efficient model training with the best dataset.

Challenge: Limitations of GPU instances (scalability, overhead, procurement) Lightly.ai's cloud-based and on-premises data curation tool typically ran on GPU instances resulting in several limitations. Firstly, GPU instances cannot be easily scaled based on the changing needs of the workload demand. A full GPU card needs to remain allocated even during low demand periods. Secondly, there is a large overhead between the CPU pre-processing (e.g., video decoding) and GPU inference processing. More complex integration efforts are needed to reduce the overhead, taking valuable developer time and resources. Thirdly, procurement of GPUs for both on-premises and cloud-based deployments is expensive and complicated.

Solution: Running Lightly's data curation tool on GCP T2A instances with Ampere AI software To overcome the limitations of AI deployments on GPU instances, Lightly collaborated with Ampere to run their curation tool on GCP T2A instances based on Ampere Cloud Native processors. This migration provided Lightly with The flexibility for fast decoding accompanied by fast model inference effectively serving all their needs with a single processing unit.

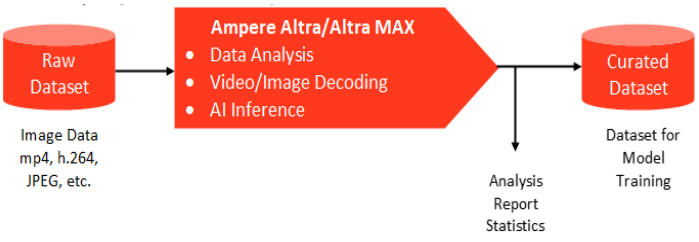
Results: Lightly.ai, which previously run on AWS only, found that GPU-free AI inference on GCP with the T2A Ampere instances not only reduced the costs, but also decreased the run-time. Ampere Altra Family of Cloud Native Processors provides sizeable advantages over the tested GPU instance [AWS g4n.2xlarge (Nvidia T4 + 8 ×86 vCPU)] reaching 7.8x better performance. Lightly.ai's customers can achieve over 3x cost reduction running on Ampere-based T2A instances on GCP. High core count of Ampere Cloud Native Processors allow for great scaling and utilizing CPU instances provides the flexibility which dedicated accelerators are lacking. Ampere - based instances can be easily provisioned and are readily available for all GCP clients. power savings, Oracle plans for most of its fleet to be converted to Ampere platforms soon.

"When running machine learning models on video data, the GPUs can't unleash their full potential as they are bottlenecked by the decoding speed. One would need to get more custom and expensive instances where the balance between decoding and processing is properly tuned. However, the flexibility of Ampere CPUs allow for fast decoding while at the same time their massive compute power allows for fast model inference. So you get the best out of both worlds in a single processing unit. We were impressed by the performance of the Ampere CPU as it easily outperformed the Nvidia T4."

– Igor Susmelj, Lightly.ai's Co-founder

INTRODUCTION

- Lightly's cloud-based and on-premises data curation tool analyzes and improves the quality of video and image datasets for training machine learning models.
- Lightly.ai's cloud-based and on-premises data curation tool typically ran on GPU instances resulting in several limitations: To overcome these challenges, Lightly collaborated with Ampere AI to run their curation tool on GCP T2A instances based on Ampere Cloud Native processors.



an easy-to-use developer toolbox. They provide all necessary elements for an AI data preparation feedback loop, such as data curation (active learning), data exploration, online selection, and model monitoring. Lightly.ai facilitates efficient model training with the best dataset. That is why companies that want to process large amounts of data in order to enable production ready and enterprise grade AI rely on Lightly.

CHALLENGES

Limitations of running GPU instances experienced by Lightly:

- GPU instances cannot be easily scaled based on the changing needs of the workload demand. A full GPU card needs to remain allocated even during low demand periods.
- There is a large overhead between the CPU pre-processing (e.g., video decoding) and GPU inference processing. More complex integration efforts are needed to reduce the overhead, taking valuable developer time and resources.
- Procurement of GPUs for both on-premises and cloud-based deployments is expensive and complicated.

"When running machine learning models on video data the GPUs can't unleash their full potential as they are bottlenecked by the decoding speed. One would need to get more custom and expensive instances where the balance between decoding and processing is properly tuned."

– Igor Susmelj, Lightly.ai's Co-founder

HOW AMPERE RESPONDED

Ampere AI engineers tested the performance of AWS g4n.2xlarge (Nvidia T4 + 8 ×86 vCPU) against that of Google Cloud T2A (Ampere Altra 48 vCPUs) processing a sample of 16 video files provided by Lightly.ai.

THE RESULTS

Instance	Run-time	$/h	Cost($)
GCP T2A 48vCPU	59.3 min	1.848	1.83
AWS g4n.2xlarge (T4 + 8 vCPU)	465.1 min	0.752	5.83

Ampere® Altra family of cloud-native processors provides sizeable advantages over the tested GPU instance. Lightly.ai's customers can achieve 7.8x better performance

CALL-TO-ACTION

Future Opportunities

Ampere provides the best-in-class solution for data preparation and data curation. With GPU-Free deployments on Ampere Cloud Native Processors customers can achieve great performance, **optimize the costs**, and invest savings into further development of their applications. With Ampere-based T2A instances now **available on GCP** this high-performance, cost-effective solution is accessible to all GCP clients.

"Lightly.ai's customers can achieve over 3x cost reduction running on Ampere T2A instances on GCP using Ampere AI software solutions for AI Inference, in addition to optimized performance. The next generation AmpereOne C3A instances on GCP will deliver on this continued value proposition."

– Igor Susmelj, Lightly.ai's Co-founder

About Ampere

Built for sustainable cloud computing, Ampere Computing's Cloud Native Processors feature a single-threaded, multiple core design that's scalable, powerful, and efficient.

Learn more

See our solutions for a variety of demanding workloads:
https://amperecomputing.com/solutions

Visit our Developer Center:
https://amperecomputing.com/developers

Disclaimer

All data and information contained in or disclosed by this document are for informational purposes only and are subject to change.

This document is not to be used, copied, or reproduced in its entirety, or presented to others without the express written permission of Ampere®.