Ampere® Boosts the Performance of Google Cloud Workloads by 30% for SADA Customers

SNAPSHOT

Organization: SADA is the market leader for Google Cloud professional services. As of 2022, Google Cloud had roughly $26B in revenue worldwide, and SADA drives a portion of that revenue as a 5X Google Cloud Partner of the Year recipient.

Challenge: As thousands of customers move mission-critical workloads to Google Cloud Platform (GCP), SADA's customers need fast, efficient, and scalable processor options to maximize performance and efficiency while keeping a lid on costs.

Solution: Ampere Cloud Native Processors deliver superior performance while consuming less power than comparable x86-based processors[1] , an ideal combination for SADA's performance focused, cost conscious, and environmentally aware customers.

Results: According to Miles Ward, Chief Technology Officer for SADA, inherent efficiency and ease of portability motivate a growing number of GCP subscribers to select Google Cloud T2A processors, powered by Ampere, for their compute-intensive workloads. In fact, SADA customer, Momento, saw a 30% improvement in performance with Ampere instances over legacy x86 VMs.

Cloud computing is not just about lowering costs, boosting performance, and shutting down on-premises data centers. Forward-looking companies look to their cloud providers to change the world. Visby leverages Google Cloud to create immersive holographic experiences. Sembly AI adopted Google Cloud to reimagine online meetings for project teams. AccuWeather relies on Google Cloud when predicting 15 billion weather events.

What do these companies have in common? All three of them—and thousands more—rely on SADA for Google Cloud implementation services. Whether it's lowering web-serving speeds on a web application or lowering time to results with analytics pipelines, these companies depend on SADA to ensure that their hardware and software systems work seamlessly. For many of these firms, Ampere® Altra® Cloud Native Processors, which sit behind Google Cloud's T2A instance family, are a vital part of the solution set.

"Our customers are eager to obtain the most efficient system design, and they often look to us for advice about the optimum processor architectures," says Ward. "We are direct users of Google tools for our internal IT functions, and we also help 3,000+ Google Cloud customers make the right decisions about their infrastructure. Ampere has excellent solutions for some of the key workloads our customers are deploying."

30% BETTER PERFORMANCE WITH FAST AND EASY TRANSITIONS

Ward and his team became enamored with high-performance Cloud Native Processors from Ampere when SADA was chosen by Google to participate in Alpha testing for the Ampere-powered T2A chip set. "We did performance and availability testing, and we got strong results from these Ampere devices almost immediately," Ward reports. "That's not always the case in alpha-testing scenarios, where you are often exposed to products that are only half-baked. The Ampere processors have been battle-tested in other environments, so Google could launch them fully formed right out of the gate. This helped us move through the analysis process very quickly."

Momento, a long-time SADA customer that runs resource-intensive caching workloads in GCP, also participated in the T2A testing. "Momento's caching workloads require extremely high performance, with request rates of millions of transactions per second," Ward says. "They saw incredible results: Ampere-based T2A VM instances outperformed Momento's x86 VMs by more than 30 percent. This allowed them to keep the price point in GCP the same as in other environments."

Momento Serverless Cache helps developers instantly supercharge their databases, with no servers to manage, configure, or maintain. Engineers from both SADA and Momento were surprised by how easy it was to transition this GCP service from older x86 systems to Ampere Cloud Native Processors. For example, Pelikan—the engine behind Momento—worked right out of the box on the Ampere processors.

"We told our engineers to carve out a couple of weeks for the migration, but it wasn't nearly as difficult as we had anticipated," Ward recalls. "Within a matter of hours, we were up and running on T2A. Momento saw the same, and was extremely happy to achieve such quick results, as well as material cost savings, thanks to the efficiency gains. Plus, they had great technical support from Ampere to optimize the configuration, such as pinning application threads and rx/tx queues to individual processor cores."

Ward expects these performance gains to improve even further as new versions of Ampere Altra processors are released, and as the Pelikan code base is tuned for the Arm architecture. "We are working with Google to add solid state disks and other capabilities to instances with these processors," Ward reveals. "There is extreme demand for more efficient building blocks and specific demand for Ampere Cloud Native Processors from hundreds of our customers. It is great to have a new competitor in the compute space pushing that ball forward."

MINIMIZING POWER CONSUMPTION AND THE CARBON FOOTPRINT

Momento's pricing policies are based on how many bytes customers send and receive from their caches. Whether those bytes are all transferred within a 3-hour window or are evenly distributed over the course of a month, the cost is the same.

Ampere Cloud Native Processors are designed to optimize performance in a small form factor with low power consumption and modest heat generation. "Electricity is a significant cost in data centers and server farms," Ward points out. "Having power-efficient servers is a huge way to reduce not only the cost of public cloud services, but also the carbon footprint of those cloud services."

These capabilities dovetail nicely with Google's sustainability initiatives: the company continually strives to improve the efficiency of its data centers, with the goal of continually reducing the amount of carbon needed to operate its global Internet business. "Google operates the most energy-efficient data centers in the world, which is only improved by the use of Ampere chips," Ward says.

REDUCING COSTS FOR SADA CUSTOMERS BY 20 to 30 PERCENT

In addition to Cloud Native Processors in its primary data centers, Google is deploying these processors in Google Distributed Cloud, a set of

fully managed hardware and software solutions that extend GCP infrastructure and services to the edge, including hybrid-cloud environments that incorporate customer data centers. "It's really useful to be able to build on Ampere in the data center," Ward explains, "because you get the same behavior when you do it remotely, and you know that you will have parity with the data center experience."

SADA uses PerfKit Benchmarker to verify the performance of key components of the GCP infrastructure. SADA reports that some benchmark tests revealed Ampere Cloud Native Processors as the fastest option, irrespective of costs. Other tests show that these processors have the best price-performance metrics, since designated workloads require less infrastructure for the same level of compute capabilities.

"We're really excited about Ampere being available on Google Cloud," Shams concludes. "We see a ton of momentum building up in Ampere Cloud Native Processors. The Momento team is excited to have the opportunity to partner with Ampere and Google Cloud on optimizations that improve our price/performance and help us deliver the best possible cache to our customers."

"We save most companies 20 to 30 percent on average with the Ampere processors," Ward states, "and we have every indication that these devices are going to get faster and faster every year."

SPURRING GCP ADOPTION WITH INDUSTRY LEADING SPECS

As a specialist in Google Cloud workload migrations throughout North America, SADA is dedicated to making GCP faster and easier to use, thereby helping thousands of customers to transform their businesses and be more productive.

"We are always trying to find the right constellation of resources—the right tools to solve problems from the available options," Ward says. "Google has a growing number of Ampere-compatible solutions, which reduces the risk to customers that want to migrate to these solutions. From what we have seen so far, customers are universally ecstatic about the price/performance benefits.

Arm-based processors such as the Ampere-based T2A VM instances are essential to remaining competitive in the cloud space."

Reflecting on the recent flurry of activity and the ramifications for the cloud computing industry, Ward states, "We spend a lot of time looking for opportunities to save customers money and help them tackle their biggest challenges. When we look closely at what software they are running, in what class of virtual machines, and under what conditions, we think that between 35 and 40 percent of our customer workloads are obvious fits for Ampere-based T2A VM instances."

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

Footnote

(1) Performance comparisons measured and published by Ampere Computing.

Details and efficiency footnotes are available here.

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®.

© 2024 Ampere® Computing LLC. All rights reserved. Ampere®, Ampere® Computing, Altra and the Ampere® logo are all trademarks of Ampere® Computing LLC or its affiliates. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies.