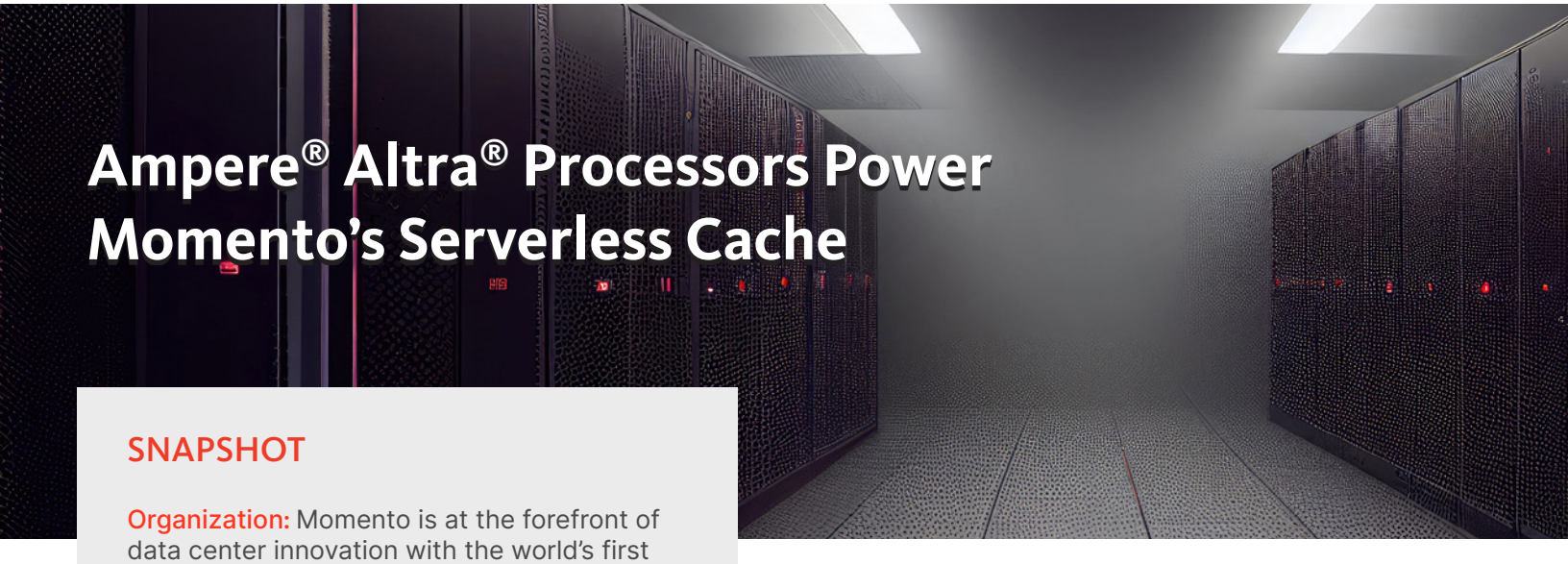
momento | Case Study



Ampere® Altra® Processors Power Momento's Serverless Cache

SNAPSHOT

Organization: Momento is at the forefront of data center innovation with the world's first serverless caching service. Built for the needs of cloud computing, Momento provides instant provisioning, auto-scaling up and down, and blazing-fast performance for database caching workloads, at any scale.

Challenge: In an attempt to ensure stable performance for erratic computing workloads, many businesses are over-provisioning infrastructure, wasting human cycles with repetitive system management tasks, yet still achieving only mediocre availability and performance for critical caching applications.

Solution: Momento's popular serverless caching service, based on Ampere Altra CPUs, helps developers instantly supercharge their databases with a turn-key cache. Developers find it easy to port their solutions to the Ampere platform. And with no servers to configure or maintain, they can spend time developing new applications rather than managing infrastructure.

Results: Momento's caching platform accelerates the developer experience and removes all distractions associated with benchmarking, configuring, scaling, and managing a database cache. Performance is demonstrably faster than other computing architectures. For example, Ampere- based T2A VM instances outperform current x86 VMs by up to 31 percent.

INTRODUCTION

As data center computing requirements continue to grow at a rapid pace, cloud providers need a new architecture to replace aging legacy x86 platforms. Demand is especially acute in the burgeoning market for managed cloud services. According to Khawaja Shams, founder and CEO at Momento, in an attempt to ensure stable performance for spikey computing workloads, many businesses are over-provisioning cloud infrastructure, wasting human cycles with repetitive system management tasks, and yet still achieving only mediocre availability and performance.

To mitigate these concerns, a growing number of service providers and cloud infrastructure providers are turning to Arm-based processors, which not only deliver superior performance, but also consume much less power than their x86 counterparts. Ampere leads the Arm processor market with its Ampere Altra and Ampere Altra Max Cloud Native Processors—the foundation of Momento's serverless caching service.

"We are particularly excited about the cost and performance of Ampere VMs," Shams says. "Processor innovation is happening faster than ever, and theuser experience of porting existing software is easier than ever. The investment required to validate new architecture just isn't as high as it used to be—and the upside is meaningfully larger."

THE WORLD'S FASTEST CACHE —AT ANY SCALE

Serverless is an application development and execution model that frees developers to focus on writing application code and developing business logic rather than managing hardware and software infrastructure. Momento's serverless cache helps developers instantly supercharge their databases with no servers to manage, configure, or maintain. As a "zero configuration" service, Momento continually optimizes each customer's infrastructure.

"We manage all of the tricky stuff on the backend," Shams explains. "If your traffic increases and your cache needs more capacity, that's on us. The only thing customers need to care about is performance, cache hit rates, and cost. We take care of the rest."

Shams describes the Momento service as the "world's fastest cache at any scale," as well as being "secure and highly available by default." That's a huge boon to the cloud computing industry, as evidenced by a recent Gartner report that revealed how many database management system (DBMS) workloads are shifting to the cloud. In 2021, revenue for managed cloud services rose to $39.2B—nearly 50 percent of all DBMS revenue[1].

As organizations steadily migrate database workloads to the cloud, Momento's serverless caching service accelerates the overall developer experience and removes distractions associated with benchmarking, configuring, and database management—creating a cost-effective, high-performance system.

"Caching is in the top five line items in the bill of almost every cloud customer we have run across," Shams notes.[2] "Customers are spending billions of dollars each year on caching infrastructure across cloud providers. Costs start to matter at the scale at which our customers operate."

EASE OF PORTABILITY TO AMPERE ON GOOGLE T2A

Momento also helps developers by empowering them to create applications quickly. It instantly scales to handle spikes and hot keys, while maintaining cache hit rates and low latencies.

No matter how many bytes are transferred, customers obtain stable and efficient transfer rates with Ampere Altra processors.

"With Ampere, the price/performance ratio often shines at high throughput," Shams said. "The ease of adoption also makes this a two-way door, enabling us to move across architectures on behalf of our customers—without them having to worry about which architecture is best for their workload at any given time."

For example, a growing number of Google Cloud customers are migrating to Compute Engine Tau T2A VMs, powered by Ampere Altra processors, which outperform current x86 VMs by up to 31 percent and lead on price-performance by up to 65 percent using on-demand pricing guidance. The Ampere-based Tau T2A VMs are designed from the ground up for predictable high performance and linear scalability using single-threaded cores, enabling demanding scale-out applications to be deployed rapidly and efficiently.

"Over the past few months, we have become intimately familiar with Google Cloud's T2A VMs," Shams says. "We were pleasantly surprised with the ease of portability to Ampere instances. The maturity of the T2A platform gives us the confidence to start using these VMs in production."[3]

Pelikan—Twitter's premiere caching framework and the engine behind Momento—worked right out of the box on the new platform. "Ease of use, performance, and cost are all very compelling reasons for us to get excited about Tau T2A," Shams adds, "and it has already been paying off for us."

COMPELLING PRICE/PERFORMANCE WITH SUPERIOR ENERGY EFFICIENCY

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.

"As far as we're concerned, you should be able to read and write your cache when you need it," Sham's stresses. "If your traffic decreases, you shouldn't have to pay the same amount of money for your low-traffic window as you did for your high-traffic window. And you most certainly shouldn't have to pay for 15 idle CPU cores on a bunch of nodes in a caching cluster just because you needed more RAM."

Strong developer ecosystems such as Momento's serverless cache will continue to help companies on their journey to cloud adoption. For example, the T2A VMs support the most popular Linux operating systems, including RHEL, CentOS Stream, Ubuntu, and Rocky Linux. Furthermore, Ampere, in partnership with Google, is working with Enterprise and Open-Source Linux OS vendors and communities to ensure these operating systems are optimized for the Ampere Cloud Native Processors.

"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."

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

Footnotes

1. Merv Adrian, "DBMS Market Transformation 2021: The Big Picture," April 16, 2022.
2. We built a serverless cache you can add to your stack before lunch," November 2, 2022.
3. "Expanding the Tau VM family with Arm-based processors," Google Cloud Blog, July 13, 2022.