momento

Cache

The first enterprise-grade serverless memory store

Whitepaper

Momento Cache is a modern memory store designed for availability, performance, and resource efficiency. It supports lightning fast key-value retrieval, as well as more complex data collections like sorted sets. Momento Cache combines a delightful, streamlined developer experience with an innovative tiered architecture to accelerate products at any scale.

Flushing stale expectations

The past decade has seen tremendous advances in cache performance as technologies like Memcache, Redis, and Elasticache push the capabilities of individual nodes and entire clusters to new heights. And yet, when we compare the state of caching technologies with that of the broader cloud industry, we see that it remains saddled with needless complexity and a rapidly-inflating total cost of ownership.

In stark contrast to the rest of the cloud landscape, no major caching technology has transitioned to the serverless paradigm, which fully abstracts away the complexity of underlying infrastructure. A properly-managed platform has zero developer-facing downtime as bugfixes, security patches, and upgrades are transparently deployed behind the scenes. This empowers developers to focus on shipping their product instead of fighting fires every day just to keep the servers running.

The inherently multi-tenant nature of a serverless platform also enables more efficient, and therefore more cost-effective, resource utilization. Frequently scaling a cluster-based service is challenging, so developers overprovision in order to avoid downtime and lost revenue. Likewise, to isolate against noisy neighbors, each team provisions and manages their own cluster, multiplying the inefficiency. A well-designed serverless platform eliminates these tensions, as workloads can safely and rapidly draw upon a vast shared pool of resources.

The constraints of self-management ultimately lead to an incomplete product that attempts to mitigate operational complexity through a smaller surface area — which is fine if one wants to build rather than buy. Still, technologies that fail to implement basic requirements like authentication and access control externalize significant burdens onto the developer. While serverless platforms may be known for opinionated, tightly-coupled integrations, it is precisely these rich feature sets that accelerate development.

Unfortunately, every major caching solution remains ensnared by this instance-based, pseudo-managed paradigm. Given advances in platform engineering and operational experience across the cloud landscape, such constraints are no longer acceptable. Imagine if S3 required you to configure instance types, to implement encryption yourself, or to suffer downtime for maintenance!

Momento Cache is the first enterprise-grade serverless memory store, designed from the ground-up to focus on what matters: ease of operation, cost-effective scalability, and unrivaled performance. Below we describe the novel techniques that make this possible for Momento Cache, as well as the batteries-included approach that defines the Momento platform.

Crafting a modern memory store

Momento Cache is all about fast, but that encompasses more than just response time! A broad range of built-in features help you get going faster, while the industry-leading architecture keeps you going fast at any scale. Here are just a few of the ways that Momento Cache modernizes the memory store for you:

Secure By Default: Momento Cache is designed to eliminate insecure handling of your users' data. We wish we didn't need to state this, but the Momento API only accepts TLS connections, with all data encrypted in memory and at rest. In addition, each request must be authenticated via an API key or short-lived tokens with an embedded permission scope.

End-To-End Control: Momento takes a principled stance that our responsibility encompasses everything from the callsite where data is passed into the Momento SDK until the moment a response is returned. While this precludes low-level integrations such as wire compatibility with RESP, we strongly believe that this is a feature. Tight control over the end-to-end experience enables us to embed optimizations that just work, such as robust connection management, automatic compression, and pre-built client configurations tailored for every environment.

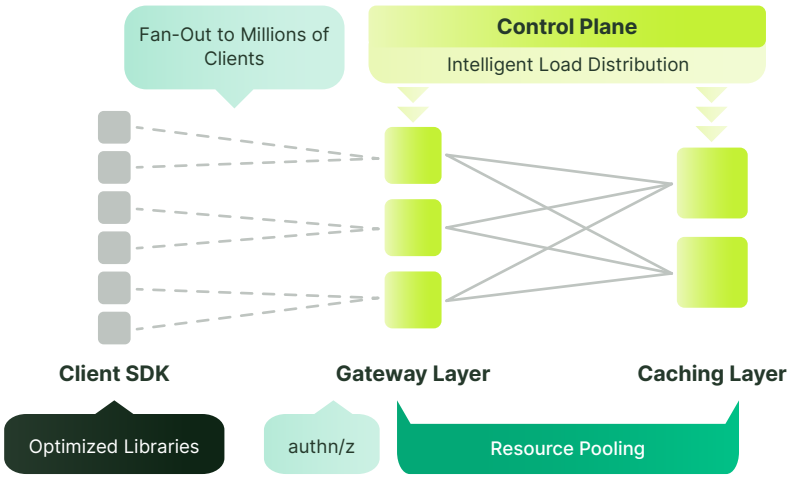
Topology-Aware Routing: Momento Cache incorporates a novel two-tiered architecture in which an independently-scalable gateway layer provides stable low-latency fan out to millions of direct connections. Internal cache topology updates are propagated to the gateway layer, which gracefully re-routes requests as cache nodes change due to scaling redeployments. This intelligent routing also enables AZ-aware connections to improve network costs and latency.

Hot Key Propagation: Momento Cache leverages a tiered caching strategy that spans both the caching and gateway layers. Hot data is propagated outwards to gateway nodes in order to quickly clear requests and prevent pressure from building up the caching layer. This can be understood as a unique form of load shedding, in which excess requests are rejected with a value rather than an error.

Together, these and other features make Momento Cache the first true enterprise-grade memory store. With Momento Cache, you can focus your attention elsewhere knowing that your caching solution has been hardened and tuned to safely handle anything you can throw at it.

A platform that's ready for prod, now

At Momento, our mission is to empower engineering teams with a better cloud platform. Like all of our services, Cache is fully-managed and built for scale. With a sreamlined development and operational experience. with fewer examples of how the Momento platform accelerates your work:



Momento Gateway: The Momento Gateway layer handles fan-out for each of our services across millions of simultaneous client connections. It integrates an expansive set of features, including authentication, access control, multiple transport protocols, and more. The Gateway layer does the heavy lifting to make your product production-ready immediately.

Client SDKs: Momento services accept connections directly from client devices, so our client SDKs incorporate robust connection management with support for gRPC, WebSockets, and HTTP, among others. In addition, these SDKs implement various optimizations and best practices, such as binary serialization and Zstandard compression.

Resource Pooling: Momento designs services to rapidly scale individual workloads by reallocating warm resources where they are most needed. This shared pool approach creates a vast, but still cost-effective, reservoir of capacity for workloads to draw upon when necessary.

Intelligent Load Distribution: Each Momento service incorporates an intelligent control plane that proactively manages the distribution of load. Deep instrumentation enables the control plane to carefully monitor and scale the system along multiple dimensions. This prevents the system from being overwhelmed while maintaining resource and cost efficiency.

The Momento platform embeds deep expertise about the experience of developing and operating global-scale products. We've built such products ourselves, and we've worked with our customers to build countless more. With the Momento platform, that expertise works for you.

momento

Get going.
Keep going.

Momento Cache has been in production for more than two years, serving billions of objects each week to tens of millions of devices for mission-critical workflows.

It offers the delightful Momento experience you know and love which has enabled enterprises to integrate and go to production in less than four weeks. Plus, our world-class customer support will be there for you the entire way!

Contact us today.