momento

Cache

Quickly improve performance, reduce costs, and handle load at any scale.

Developers get stuck using instance-based caching solutions like Redis that are painful to manage and scale—forcing them to engage in manual, error-prone tasks that run counter to the speed, elasticity, and availability benefits of caching. Momento frees developers to focus on the specific capabilities of a service without getting bogged down by the plumbing. Cloud providers have invested heavily in serverless compute, storage, databases, and almost every other part of the stack—except caching. **Momento Cache changes this.**

Configure Nothing

instead of spending days to weeks configuring, testing, and scaling clusters, go to market faster with instant-start caching. There's no architecture to build—accomplish everything with simple API calls.

Boost Performance

Accelerate your database and application performance with a highly performant cache.

Eliminate Outages

Say goodbye to cache outages with instant scalability and automatic heat management.

Key Features

- **High Scale, Performance, and Availability:** A Momento cache is ready instantly via a single API call, is optimized for tail latencies, has multi-zonal redundancy, and handles millions of requests per second (RPS) without any tuning.
- **Automatic Management:** No more tedious manual operations. Momento Cache features an intelligent proxy that handles automatic scaling, node warming, hot key mitigation, replication, and deployments—all without any maintenance windows!
- **Secure by Default:** Momento has built-in security features including end-to-end encryption, per-request authentication, VPC peering, and deep observability integrations.
- **Flexible Data Types:** Collection data types enable an abundance of use cases by providing core data structures that match up with common types in modern programming languages. Supported data types include dictionary, list, set, and sorted set.
- **Redis API Compatible:** Momento provides RESP compatibility via a sidecar proxy or drop-in SDK replacements.

Learn more about Momento Cache at gomomento.co/cache.

Or send a message to **hello@momentohq.com**.

Check out these customer stories

GAMING

Scalability and stability improvements

Highlights

- Integrated Momento Cache in **just three weeks**, making their architecture fully serverless.
- By replacing ElastiCache Redis, Momento Cache reduced the total cost of ownership for their in-game chat service by **more than 30%**.
- Momento Cache **enabled them to scale** to handle the volume of connections from Lambda functions—previously impossible.

STREAMING MEDIA

Database modernization

Highlights

- Momento Cache offered a **more elastic, modern caching solution** after they experienced several outages with their previous caching service.
- Momento Cache is **15% faster than ElastiCache and costs 52% less**.
- Achieved **multi-cloud portability** with combination of YugabyteDB and Momento Cache.

INTERNET OF THINGS

Faster object storage and ML caching

Highlights

- Relies on Momento Cache to upload hundreds of millions of images every day, **saving hundreds of thousand of dollars per year**.
- Has over 10 million cameras performing AI at the edge. **Momento Cache facilitates Wyze's video analytics workflow to enhance representative thumbnails quickly and efficiently.**

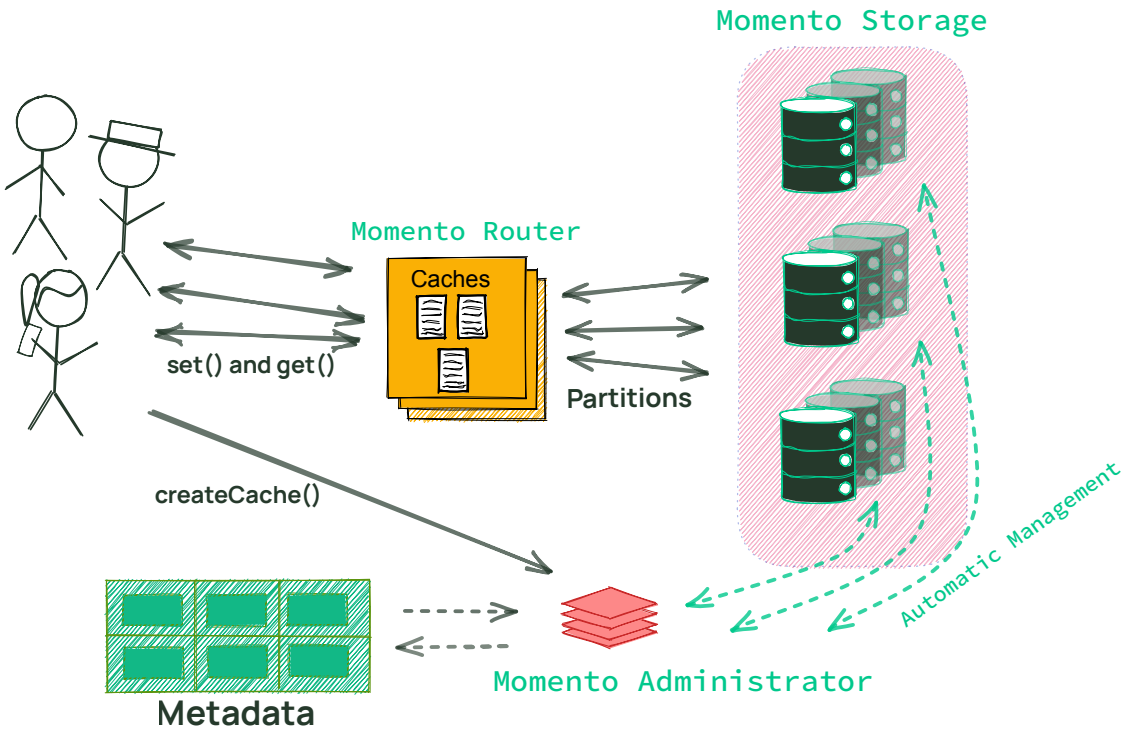
Momento is also trusted by

RAIDER.IO · gumi · GOLDSKY · VOXIE · docomo · saturn · Conoris · classmethod

gomomento.co/cache

momento

Momento architecture overview



Our team has built and operated performance-sensitive, mission-critical systems at large scale including DynamoDB, TurboTax, GitHub, and more. We have incorporated this operational expertise into the design of Momento. The simplified diagram above gives an overview of the following high-level data flows:

1. **Momento Router:** Intelligently directs data operations to caches in Momento Storage. Caches are automatically partitioned across multiple shards. This pattern is modeled after mcrouter at Meta, and twemproxy at Twitter. This distribution layer enables dynamic autoscaling, multi-zonal replication, and smooth deployments without impacting hit rates or latencies.
2. **Momento Storage:** This is where conventional storage engines reside (e.g. memcached, Redis, RocksDB). Our architecture has built-in flexibility to dynamically and non-disruptively switch between different engines based on the shape of a customer's data and load. For example, simple key-value workflows can be supported by memcached storage, whereas workflows requiring collection data types (lists, dictionaries, etc.) can be supported by Redis storage.
3. **Momento Administrator:** Acts as the "brain" of Momento's architecture. The Momento Administrator keeps track of storage health, utilization metrics, usage analytics, and the topology of the entire system. It coordinates between Momento Router and Momento Storage to dynamically manage capacity, shuffle data around to spread load, mitigate hot keys, replace components suffering infrastructure failures, and hydrate new storage nodes—all while maintaining high cache hit rates through deployments.