



Architecture Diagrams

Parallel API Composition in AWS



Parallel API Composition in AWS: Architecture Diagrams

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon's trademarks and trade dress may not be used in connection with any product or service that is not Amazon's, in any manner that is likely to cause confusion among customers, or in any manner that disparages or discredits Amazon. All other trademarks not owned by Amazon are the property of their respective owners, who may or may not be affiliated with, connected to, or sponsored by Amazon.

Table of Contents

Parallel API Composition **i**

Parallel API Composition in AWS 1

Further reading 2

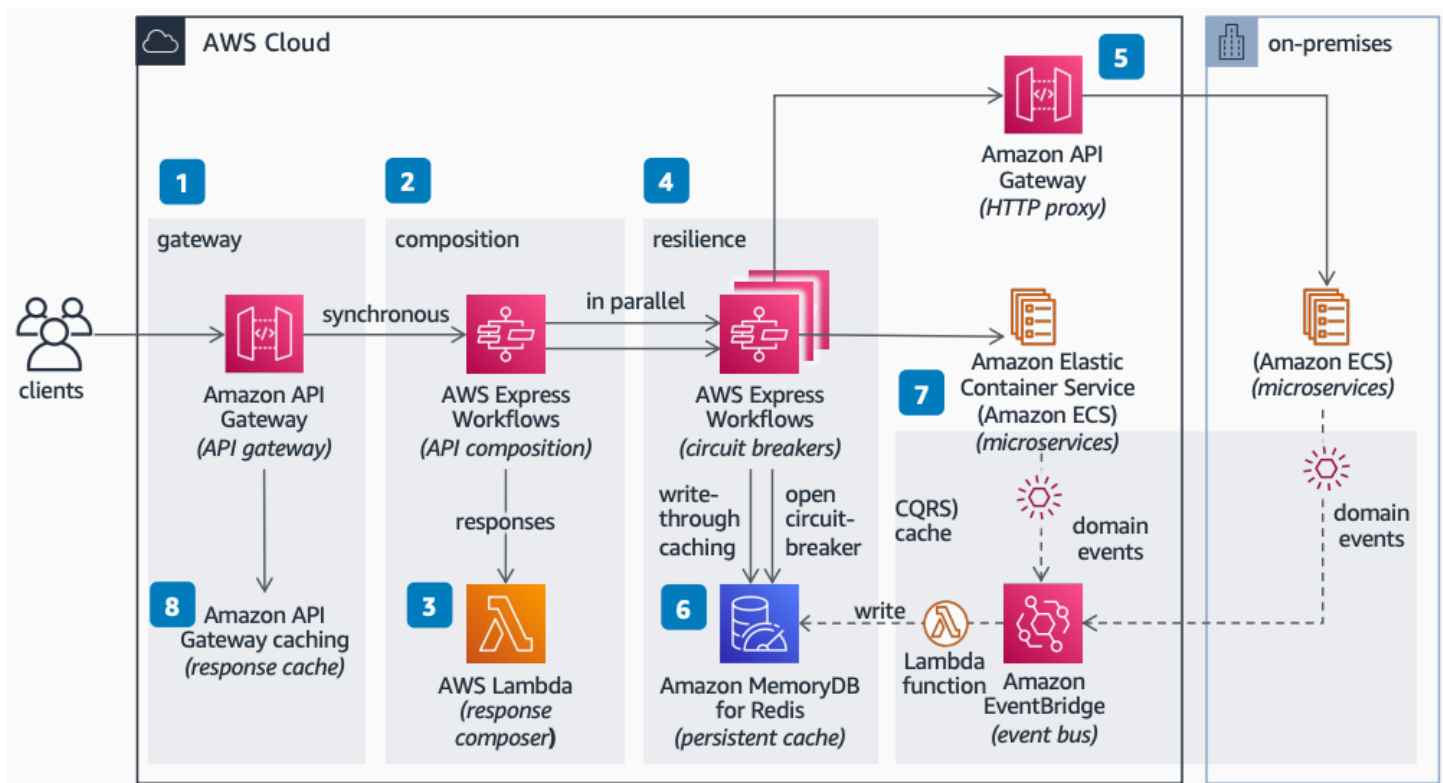
Diagram history 2

Parallel API Composition in AWS

Publication date: **May 17, 2022** ([Diagram history](#))

This architecture shows how to call multiple downstream API endpoints, in parallel or in sequence, to compose a single aggregated response. You build a generic, parameterized circuit-breaker workflow using [AWS Step Functions](#) Express Workflows and use command query responsibility segregation (CQRS) to maintain a persistent eventually-consistent cache.

Parallel API Composition in AWS



The following steps describe the architecture:

1. Create an API gateway using [Amazon API Gateway](#) to allow clients to send synchronous web requests to your microservices.
2. Use Step Functions Express Workflows to create a workflow with parallel steps that invoke multiple microservices at the same time.
3. Use an [AWS Lambda](#) function to compose the multiple responses from downstream microservices into a single aggregated response for clients.

4. Handle parallel requests with circuit breakers built with parameterized nested Step Functions Express Workflows to increase resilience.
5. Use API Gateway to proxy HTTP requests when invoking on-premises microservices.
6. Create an Amazon MemoryDB for Redis database to cache microservice responses using the [write-through caching pattern](#). When a downstream microservice becomes unavailable or its latency reaches a threshold, the circuit closes and reads all responses directly from cache.
7. Complement cached data with domain events from downstream microservices using the CQRS pattern. Create an event bus with [Amazon EventBridge](#), then handle events with a Lambda function to persist the domain aggregates in MemoryDB.
8. Enable the API Gateway response cache and adjust time-to-live (TTL) values and filters according to each request type to increase performance.

Further reading

For additional information, refer to the following resources:

- [AWS Architecture Icons](#)
- [AWS Architecture Center](#)
- [AWS Well-Architected](#)

Diagram history

To be notified about updates to this reference architecture diagram, subscribe to the RSS feed.

Change	Description	Date
Initial publication	Reference architecture diagram first published.	May 17, 2022

Note

To subscribe to RSS updates, you must have an RSS plugin enabled for the browser you are using.