



Architecture Diagrams Game Tech

Massive Scale Real-Time Messaging for Multiplayer Games



Massive Scale Real-Time Messaging for Multiplayer Games:

Architecture Diagrams Game Tech

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

Real-time messaging i

Massive Scale Real-Time Messaging for Multiplayer Games diagram 1

Further reading 2

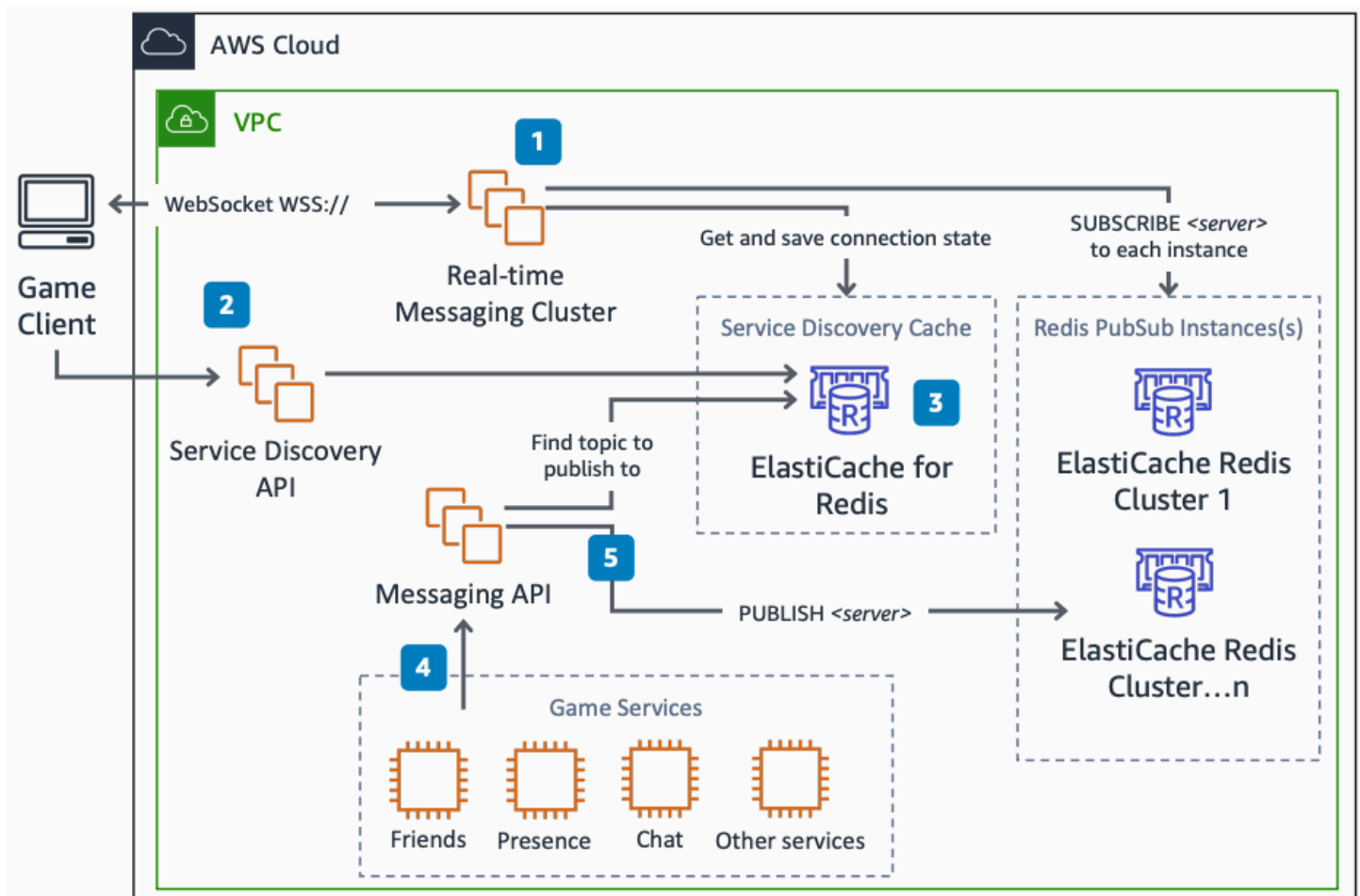
Diagram history 2

Massive Scale Real-Time Messaging for Multiplayer Games

Publication date: **November 11, 2020** ([Diagram history](#))

This architecture shows how to deliver real-time messaging for multiplayer games that can scale to support millions of concurrent users (CCU). The solution uses multiple Redis PubSub clusters with [Amazon ElastiCache for Redis](#) and WebSockets to provide horizontal scalability.

Massive Scale Real-Time Messaging for Multiplayer Games diagram



The following steps describe the architecture:

1. When a real-time messaging server starts, it checks a Service Discovery cache (using Amazon ElastiCache for Redis) for a list of available Redis PubSub instances. The server subscribes to the same topic in each PubSub instance. This approach shards traffic across multiple ElastiCache clusters while still reaching the correct server.
2. Clients send requests to a Service Discovery API to retrieve a real-time messaging server to connect to. The list of available messaging instances is stored in the Service Discovery cache.
3. The Service Discovery cache stores information about connected endpoints, including clients, real-time messaging instances, and the Redis PubSub clusters and topics that each messaging instance can receive on. If a client disconnects, this information is removed from the cache.
4. Applications use a Messaging API to deliver messages to clients. The API publishes the message to the appropriate topic that reaches the correct real-time messaging server and game client.
5. The Messaging API publishes the message to the topic in one of the Redis PubSub instances. Because each real-time messaging server listens to the same topic in every PubSub instance, the message is delivered regardless of which ElastiCache cluster receives it.

Further reading

For additional information, see 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.	November 11, 2020

 Note

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