

Explanation of the flow of data

Blocks inside the dotted perimeter are components within the API gateway. Blocks outside the dotted perimeter are resources outside the API gateway. All user requests are routed to the HTTP server thread. The HTTP server then puts the requests into the input queue of the appropriate backend service. Redis is the database used to simulate the queues.

Each backend service, once done with the processing, then relays a response through its corresponding response queue.

The API gateway contains several threads that read from each of the response queues, one thread for each. These threads get the next available message from the response queues and store them in an in-memory cache (unordered map).

The HTTP server then reads the response from the cache and returns it to the user.

Why use an unordered map to cache the response?

The cache is needed as the responses for several simultaneous requests to the API gateway may not return in order. It makes little sense for the the HTTP handler function in the HTTP server to read from the response queue directly.

How does it work exactly in code?

When the HTTP handler function receives a user request, the user request is assigned an identifier (integer) unique to the pipeline of data it is flowing in. The HTTP handler function then dispatches the request to the appropriate backend service and begins to repeatedly read from the cache until the response is found.

The backend service then sends back the same unique identifier when returning the response.

The response reader thread gets the next response from the queue, reads its unique identifier and stores the response in the map as pairs of (unique ID, byte slice of data).

Once the HTTP handler function finds that the response for the request has been found in the map using its unique ID, the HTTP handler function returns the response to the user device and deletes the response from the cache.

