Assignment 1

Author
Wanjing Hu / fng685@alumni.ku.dk

November 21, 2024

1 Question1

Provide a short description (2-4 paragraphs) of your implementation and tests, in particular focusing on:
(a) How does your implementation and tests address the all-or-nothing semantics?
(b) How did you test whether the service behaves according to the interface regardless of use of RPCs or local calls?

In regard of all-or-nothing semantics in **ratingBook**, we work on both the client side and the server side. The server calls the business level of rating books, which maintain the rating score sum and the number of the books in **Memory** inside a **java.Util.Map**, named *bookMap*. When we check the validation of the ratings and ISBN, we do inputs check first, and if any of the validation fails, the rating data will not write into the memory. For **getTopRatedBooks**, we have validation checks on parameter **K**, and we will not give a non-empty list unless there are at least K rated books in the storage. Also, the **getTopRatedBooks** returns the type of *com.acertainbookstore.business.ImmutableBook*, which guarantee that all the threads will read a same value on the getter. For **getBooksInDemand**, we transit the BookStoreBook into the immutable StockBook for thread safe before returning, too.

In regard of RPC and local call tests, we mainly apply the feature inside **setUpBeforeClass** in *BookStoreTest.java*. If the localTest property is true or not set, the storeManager and the client will all just be around the CertainBookStore, which run locally accessing on to the memory on the business level data exchanging. Otherwise, the HTTP Proxy is built, which is the stub for client and server with the functionality to communicate through HTTP and doing serializing as well as deserializing.

While testing the RPC, we should run the BookStoreHTTPServer first, and it will listen on port 8081 as we assign in the code. The the client side, we just run the tests. We should change the property in IntelliJ IDE inside the VM Options in the Run Options, and we run the test sets with *-Dlocaltest=false*. While testing the local calls, we do not need to start any server, just run the tests with the default *localtest* value undefined, or define explicitly as *-Dlocaltest=true*

2 Question2

We have stated above that the architecture achieves strong modularity. Explain this in the context of the following questions.
(a) In which sense is the architecture strongly modular? (1-2 sentences)
(b) What kind of isolation and protection does the architecture provide between the two types of clients and the bookstore service? (1-2 sentences)
(c) How is enforced modularity affected when we run clients and services locally in the same JVM, as possible through our test cases? (1-2 sentences)

Strong Modular: First the interface design separates the different functions of services, and restricts the communication to messages only. Second, the service code focuses on filling and parsing client request and service response, and they will not need to care about what happened inside the network IO, which also indicates a strong modularity.

Isolation and protections: The interface separate the client from accessing different services(methods) in the server, which is a kind of isolation, which means the server only provides a limited access through each interface, and the client would only implement one and only one kind of interface. Also, the RPC enables the clients and server to deploy on different physical devices like different pods or machines, which provides a physical isolation and a shared resource protection(such as memory and disk).

Enforced modularity: The same JVM shares memory, and there is less physical isolation and modularity in the view of the memory.

3 Question 3

(a) Is there a naming service in the architecture? If so, what is its functionality? (1-2 sentences)
(b) Describe the naming mechanism that allows clients to discover and communicate with services. (2-3 sentences)

Yes, there is probably a hard coded naming service, by assigning explicitly the IP-PORT into the serverAddress and pass to the client. It is used for the client to discover the server.

Our server register itself on localhost:8081, without identifier; and the clients query the naming service directly using the address to retrieve connection details. For the clients communicating to the server, we have URI defined by Tag, which is used for routing the different method inside server code. We also have the URL params to pass request data, and our response is serialized into the body part in HTTP.

4 Question 4

We have studied three types of RPC semantics: at-least-once, at-most-once, and exactly-once semantics. What RPC semantics is implemented in the architecture? Justify your answer. (1 paragraph)

For **At-least-once**, we guarantee the write operations are **all-or-none**, and if the data is written twice with the same values like booking rating with ISBN, the Map structure will do the filter and nothing will change. Also the synchronized method guarantees the method level concurrency safety of the bookMap[Oracle(2017)]. That makes sure the operation is idempotent. Since we do not have side effects here, **At-most-once** is not implemented. And we do not have the retry for failures, so we do not have **Exactly-once**, too.

5 Question 5

Services employing HTTP as a communication mechanism often deploy web proxy servers for scalability in the number of simultaneous client connections.
(a) Is it reasonable to use web proxy servers with the architecture of Figure 1? (yes/no)
(b) If so, explain why and describe in between which components these proxy servers should be deployed. If not, why not? (1 paragraph)
NOTE: Sometimes, web proxy servers are also used for caching. Assume for this specific question that no caching is employed, but the web proxies would only be used for scalability in the number of connections.

Yes. The Proxy level provides an isolation from the Application level, and separate the business logic with the routings with Tag. Then if the proxy is deployed separately on a different machine from the application server, the proxy machine will all be used for network connections, enabling more requests and response to deliver. And the machine for the proxy would be a high of CPU usage and network IO, which could be extended separately from the application server.

6 Question 6

Given the discussion in the question above, consider now the following questions:
(a) Is/are there any scalability bottleneck/s in this architecture with respect to the number of clients? (yes/no)
(b) If so, where is/are the bottleneck/s? If not, why can we infinitely scale the number of clients accessing this service? (1 paragraph)

Yes. The thread number in the server is a restriction for handling the client requests(referring to *CLIENT_MAX_THREADSPOOL_THREADS*, valued 250 in *BookStoreClientConstants.*), as well as each thread's max concurrency(referring to *CLIENT_MAX_CONNECTION_ADDRESS*, valued 200 in *BookStoreClientConstants*). The max concurrency describes how many requests could be inside the server at the same time. For example, if there are already 200 threads executing by the server, and then there comes the 201st request, the server will refuse that request due to the max currency constraint. When the concurrency limit is reached, the server will not handle requests any more until the request reach its timeout defined by *CLIENT_MAX_TIMEOUT_MILLISECS*. Also, since the concurrency limit is smaller than the thread pool size, there will never be a thread resource used up before the concurrency reached. The only discarding scene is the sudden max concurrency reached within the client max timeout limit, and the coming requests will be refused.

7 Question 7

Suppose the server-side of the architecture fails by a crash of the server machine where the CertainBookStore class is being run. In this context, explain the following.
(a) Would clients experience failures differently if web proxies were used in the architecture? (1 paragraph)
(b) Could caching at the web proxies be employed as a way to mask failures from clients? (1-2 sentences)
(c) How would the use of web caching affect the semantics offered by the bookstore service? (1 paragraph)

Yes, there would be difference with the proxy, where the proxy could transit the failure into another kind of failure in the response code or response body of error message with HTTP code=200. If there is no proxy, the crash will directly expose to the HTTP level, and cause some HTTP related failure like 500 series. Also there is a possibility of the proxy doing nothing to the crash, then there would be no difference.

It depends on the caching policy, that if the cache is still valid and was hit, the request will not go into the server, else the cache would be updated and still goes into server, leading to crash.

Cache would introduce inconsistencies due to outdated information, especially when the cache implementation is not thread safe. For example, cached responses might serve stale data if a client call *getBooks* after a stock update or rating change, violating the service's real-time consistency guarantees. Cache will affect coherence while concurrency refreshing occurs at the same time.

References

[Oracle(2017)] Oracle. 2017. Oracle help center. `https://docs.oracle.com/javase/tutorial/essential/concurrency/syncmeth.html`. [Online; accessed 19-December-2017].