A black and white logo

Description automatically generated

Warsaw University of Technology

Faculty of Electronics and Information Technology

Dimitrios Georgousis

Gabriel Paic

Distributed Computing and Systems

Project Concept

Introduction

Topic: 13 write – update. Implement distributed memory with read and write access, using write-update protocol.

We present a concept design for the above distributed system which reflects our current vision of implementation. The aim of the project is educational.

Assumptions

This distributed system is a developer/programmer tool. Offering an interface to use distributed memory. It uses the main memory of its Nodes and doesn’t utilize their filesystems for saving data (data is not persistent). The system is in essence a DSM simulator with a very simple cache implementation. Focus was put on showcasing the write-update and communication protocols.

It should be pointed out that from now on ‘Server’ and ‘Node’ are going to be used to refer to the same entities in our Distributed Shared Memory System.

Use Cases

|  |  |  |
| --- | --- | --- |
| name | description | return value (JSON format) |
| Use Cases | | |
| connect() | Connects to the server with a specific <server\_address> | - |
| disconnect() | Disconnects from the server that the user is already connected to | { “status”: num, “message”: str } |
| read() | Reads from a specific <memory\_address> in the system | { “status”: num, “message”: str, “data”: obj, “istatus”: str, “wtag”: num, “ltag”: num } |
| write() | Writes to a specific <memory\_address> in the system | {“status”: num, “message”: str } |
| Auxiliary Use Cases (used for testing/checking errors) | | |
| lock() | Performs an acquire\_lock() operation on the lock protecting a specific <memory\_address> | {“status”: num, “message”: str, “ret\_val”: boolean, “ltag”: str, “wtag”: str } |
| unlock() | Performs a release\_lock() operation on the lock protecting a specific <memory\_address> | {“status”: num, “message”: str, “ret\_val”: boolean, “ltag”: str, “wtag”: str } |
| dumpcache() | Returns to the user the cache contes of the Server they are connected to | {“status”: num, “message”: str, “cache”: list } |

Let’s explain the above fields:

* “status”: is 0 if the operation succeeded, non-zero otherwise
* “message”: is a short descriptive message of the operation, useful for logging events that happened in the system
* “wtag”: is the last write tag, a timestamp of the last time something was written to that address
* “ltag”: is the last lock tag, a timestamp of the last time that lock was used
* “data”: our implementation of the DSM, since it is high-level, allows any object type to be stored in the memory addresses. Effectively, we can say that each memory address can act as a whole page depending on how large the data the user stores in them is
* “istatus”: status of the data item in a specific <memory\_address>. It is either “E”: exclusively owned by its host server, or “S”: shared between the host and other Servers
* “ret\_val”: indicates if our operation managed to lock (or unlock) a given lock
* “cache”: a list of cached items on a specific Server

The system has one user type that is able to perform all operations on it. The above interfaces are provided as simple programs that take input/give output to the console. Also, we have written an abstraction which allows one to use clients in either language through it. This abstraction is provided as a Python module which can be imported and used in other code.

System - Architecture

The languages used in our implementation are Python and Java. The operating systems are Windows 11 Pro and Ubuntu 22.04.4 LTS through WSL. The Linux machine runs a Python and a Java Server. The Windows machine runs a Java Server and a Python client (or Java client, there is not much difference).

We deploy 3 Servers (Nodes) intending to showcase the functionality of the copy holder chain of memory items (data corresponding to a specific <memory\_address> in the system). If we had only 2 Servers then the copy holder chain would be trivial.

Component Design

System architecture is reflected in the component design details presented below. Also, the class diagram shown later depicts some of the interactions between these components (in the class diagram we treat them as the corresponding classes).

1. ClientLogic: A component that hides communication details and provides an interface that clients may use to perform operations on the Distributed System.
2. Client: A user interface that uses the interface provided by the `ClientLogic` component. A client may be one of our Python clients, Java clients or a Python ClientWrapper which allows users to use either Python or Java through a Python module.
3. Server (Node):
   1. MemoryItem: a single MemoryItem that represents data stored at a single memory address in our System.
   2. LockItem: a component that provides a locking abstraction to be used for managing memory accesses consistently.
   3. MemoryManager: a component that manages the operations performed on the main memory of a Node. It, also, implements the synchronization logic providing appropriate locking for its data items (MemoryItem components).
   4. Cache: a component that manages the functions of the local cache in each Node. It uses a very simple replacement algorithm: if the shared memory has size N memoryItems then memory address M is matched with position M mod N in the cache.
   5. Communication: the server handles communication with clients and serves them. It also handles situations where it needs to contact a different Server.
4. Network:
   1. Communication: message passing and data sharing between Nodes and Clients is done using TCP sockets. When a message M is to be sent. The sender first sends a message of fixed size (HEADER\_LENGTH) which informs the receiver about the actual size of M, then M is sent. Messages are in JSON format. We already showcased the types of messages returned by Servers, so now we will show the messages sent from Clients to Servers: {“type”: str, “args”: list }. “type” informs the server of what operation to perform and “args” (if present) contains the arguments to be passed to that operation.
   2. Topology: A mesh topology. Every Server can communicate freely with every other Server.

Further details of the System Architecture

Our system is influenced by the ‘Plus’ system in the DSM lecture notes. When a client makes a write request to a memory address that is not local to its Server, then that server forwards this request to the host server of the memory address. The host server then initiates the process of updating copies (write – update) having added the server that made the request in the copyholder list. We use the idea of copyholder chain from the ‘Plus’ system where each copyholder – Server is responsible for communicating the next one, thus avoiding too much network traffic on the host server. One difference is that our Servers are rather ‘stateless’ and do not retain information about pending updates. Thus, when the write – update sequence is completed instead of receiving an acknowledgement from the last Server in the chain, each Server sends an acknowledgement to its predecessor until it reaches the host server. If a server in the chain is corrupted (not able to communicate or somehow down) then the original Server is informed and will remove ALL copyholders that could not be reached from the copyholder list. If there are none remaining, then the item becomes exclusively owned again.

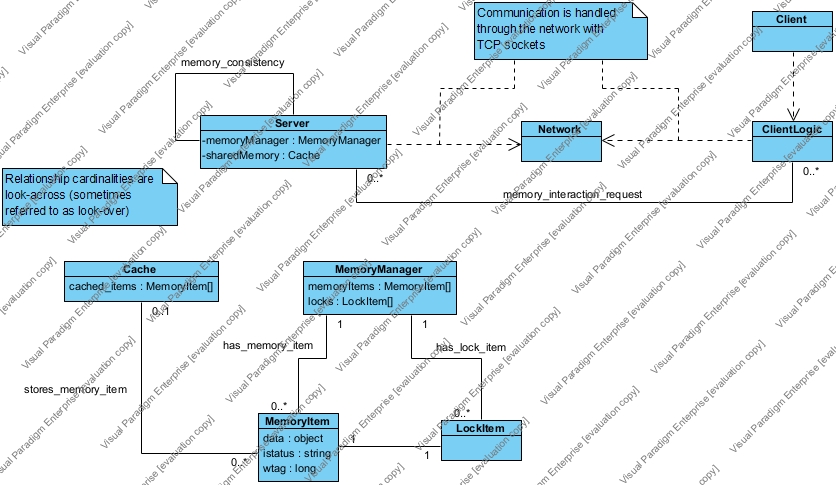
Reads to non-local memory addresses also update the local cache. Reads to non-local memory addresses that are in the cache also contact the host Server of this memory address in order to acquire the lock for it.

Locks are single – reader/single – writer for ease of implementation and this helps us enforce a consistency model on our data items.

State of Servers

The only information retained about state is in the Memory Items held by the Memory Manager and the Shared Memory. Which are the “istatus” of an item and the tags “wtag” and “ltag”. The Memory Manager also retains the copy holder chain. Apart from that, our Servers are stateless which makes implementation easier but has a greater communication overhead since our system makes a lot of connections and shared a lot of metadata each time.

Class Diagram



Relationships:

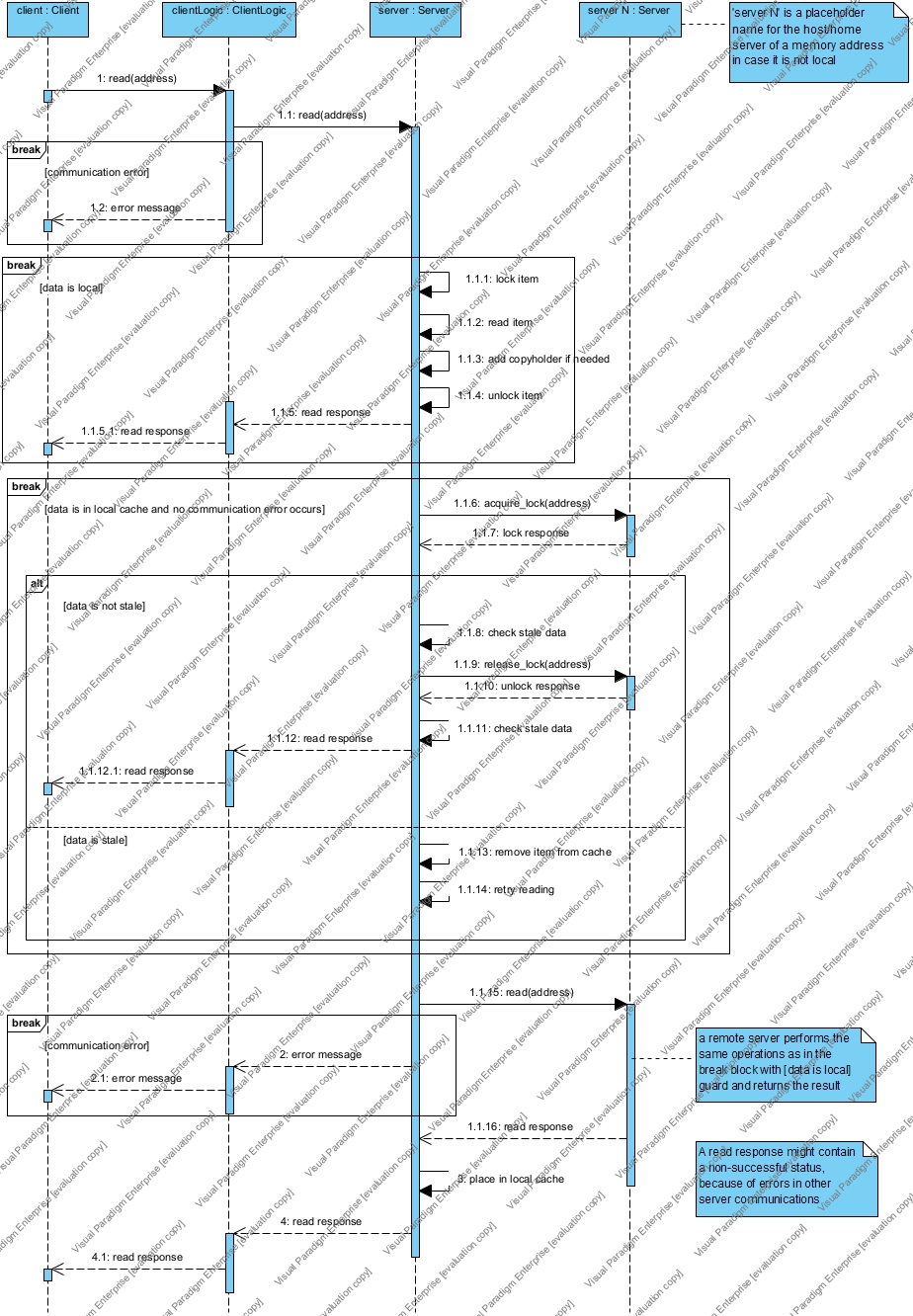
* memory\_interaction\_request: User requests some memory operation (read, write and might cause updates).
* memory\_consistency: Nodes interact with other nodes to maintain memory consistency between them.
* has\_memory\_item: A memory manager controls the behaviour of a specific memory item.
* has\_lock\_item: A memory manager has a lock item, which controls locking on a memory item.
* stores\_memory\_item: Cache memory stores locally memory items obtained from remote Servers.

Protocols used have been mentioned in the ‘Further details of the System Architecture’ section.

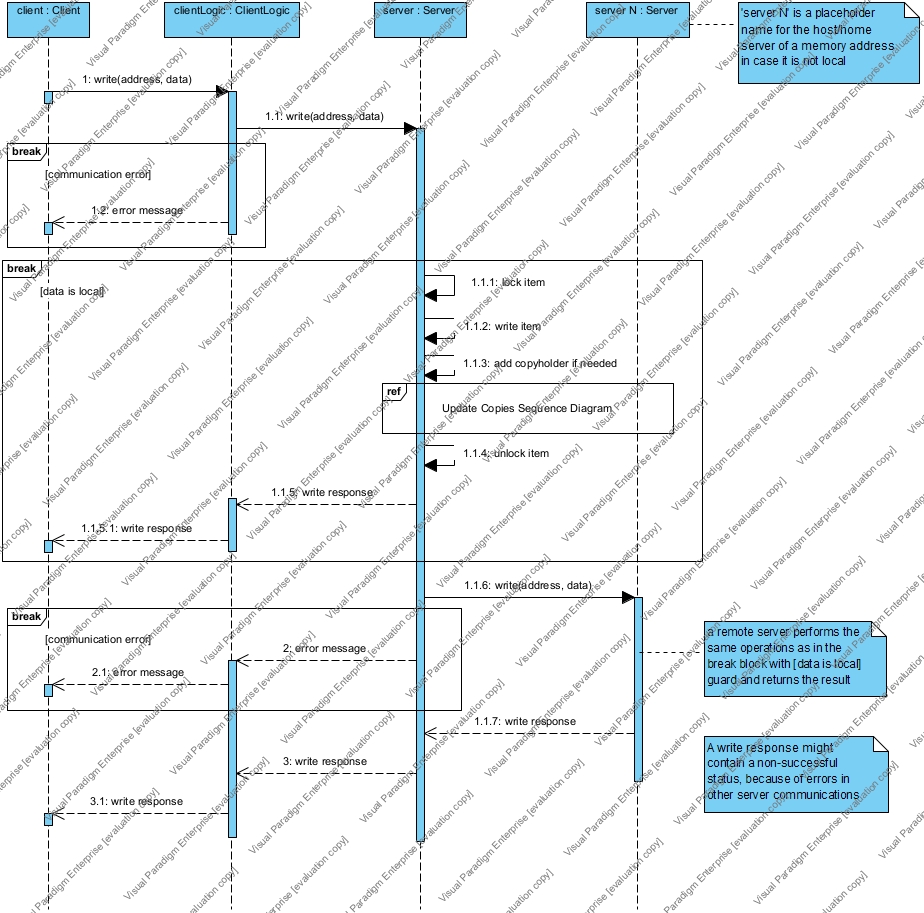
Sequence Diagrams

We show the sequence diagrams of read, write and update cache operations, since these are the main focus of the project.

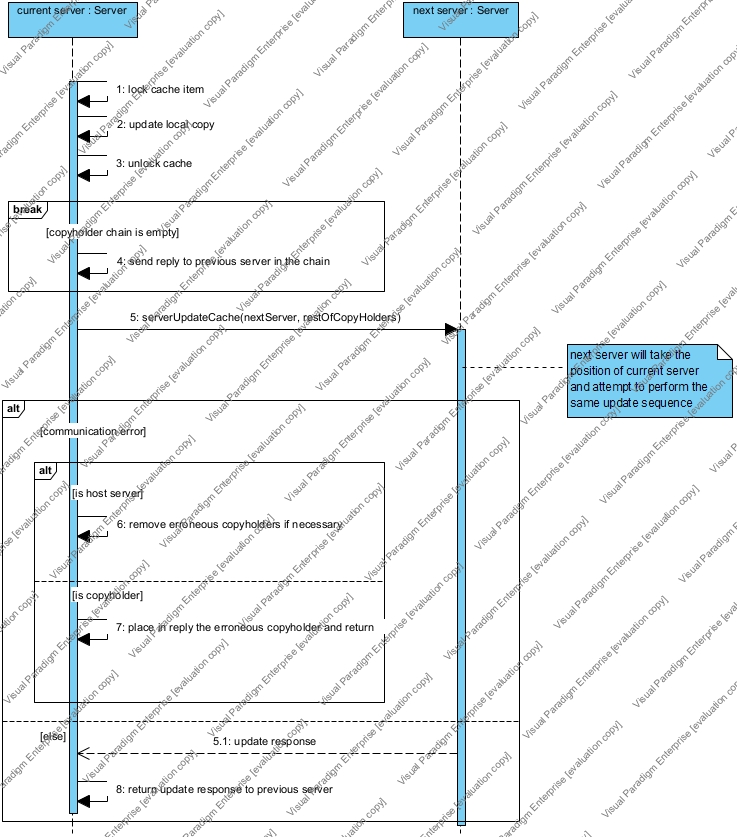
Read Sequence Diagram



Write Sequence Diagram



Update Copies Sequence Diagram



Of course, the host Server does not have a “previous server” and does not perform those actions/send those messages.

Comments

A few comments on the system. As explained read(), write(), lock() and unlock() return a “wtag” (last write tag) and an “ltag” (last lock tag). The “ltag” is incremented every time a lock is acquired and/or released. Thus the “wtag” values of a specific memory address enforce a sequence on reads and writes on a specific memory address (although it does not give us a happened-before relation between reads, since all reads between two writes have the same “wtag” value). Since, every time we read or write the lock is acquired and released, the “ltag” values of reads and writes enforce an order on all read and write operations on the same memory address. A client of this DSM may use this information to perform their own external synchronization.

One problem we have to combat is the possibility that a lock is acquired by a remote Server that becomes unavailable and thus the lock is never released. To fix this issue external locks are given with “leases” meaning that the lock will automatically be unlocked after a certain amount of time. Our Java code deploys a Timer object that schedules all these “automatic release” tasks, whereas our Python implementation uses a more software heavy approach deploying an individual “automatic release” thread after each lock is remotely acquired. The code is carefully constructed to deal with further issues that may arise, such as a remote Server attempting to re-release a lock after the timer has expired (the “ltag” is used ensure that a release request is valid) or a remote server attempting to return stale data from its cache to the Client.

Differences from Original Concept

The original Use Cases have changed. System ‘start up’ and ‘exit’ are now performed manually by executing the corresponding programs that control Servers (Nodes). Memory allocation is no longer needed; the system has a fixed amount of memory known to all clients and all clients share the same memory addresses, thus freeing memory is not needed either. Health – check is not included, but system status is implicit now, meaning if the System becomes partitioned or some Node fails in some way then a user will not learn of this situation unless they try to access data items hosted or cached (only when doing writes) by this Node. In this case they simply receive an error message. Also, new Use Cases were included such as connect, disconnect and the auxiliary ones.

There is no explicit Synchronisation Manager, Fault Manager, Mapping Manager or System Monitor now. Let us see why:

* Synchronisation Manager: the operations of this Manager are now included in the Memory Manager and the associated locking features it provides.
* Mapping Manager: we assume the IP addresses, PORTs of Nodes and Memory size are known and static. Memory addresses are given to nodes implicitly, by the way that they are created from the Server Application scripts and thus a Mapping Manager is no longer needed.
* System Monitor: Our Servers use TCP sockets for communication and each time a new request needs to be served another socket is created. Health – checking is thus performed implicitly by serve requests when they attempt to communicate with other Servers and have to handle errors that result from failure of communication.
* Fault Manager: We have already discussed about why no System Monitor exists in our current system or why the Health – Check operation is not present explicitly. The Fault Manager need not exist as its functionality is included in the Server’s error handling. We mention the CAP theorem now: when our DSM system becomes Partitioned, it preserves Consistency, but does not provide Availability meaning that Clients can perform operations on memory addresses that belong to Servers that are still accessible, but are denied such operations on unavailable Servers (even if that memory item is locally cached). In this way, the System remains partly available to the Client, but provides consistent memory accesses, while it remains partitioned and until all nodes become available again.

Test Plans

* `test.py`: This file contains some basic testing. It tests:
  + Connecting clients to server
  + Writing to local addresses
  + Reading from local addresses
  + Acquiring and releasing local locks
  + Writing to remote addresses
  + Dumping cache of Servers through clients
  + Reading from local addresses
  + Reading from remote addresses
  + Acquiring and releasing remote locks
  + Disconnecting servers

Using the output logs of Servers (in the terminal output) and the output of the test we can check if any unexpected behaviour occurs.

* + It also tests if remote reads and remote writes update the cache appropriately
  + If stale cache entries are updated appropriately
  + If corrupted copyholder are handled correctly (removal from copyholder chain)
* `test\_forgotten\_locks.py`: This file tests if locks that were acquired by a Client but never released are actually usable after the release timeout expires.
* `test\_concurrent.py`: Tests if all reads after a write return the same data. Test done both on reads that are local to the data item and ones that are remote.
* `test\_times.py`: A file that provides us some performance metrics for our System.

Performance

Metrics and results from time measurements:

Comments on performance

Notes

Project software structure and usage explained in the `README.md` file.