HW5 Report, Chordy: A Distributed Hash Table

David Fischer

October 7, 2025

1 Introduction

This assignment's goal was the implementation of *Chordy* a DHT based on a variation of *Chord*, a distributed lookup protocol designed to efficiently identify the node in a peer-to-peer network that stores a particular object through key distribution in a ring structure. *Chord* addresses many problems distributed systems face such as load balancing, decentralization, scalability, availability, and flexible naming.

2 Implementation

2.1 Ring Structure

To ease validation and development, instead of the usually intended consistent hashing function, simple numerical IDs were used for both nodes and later on keys. This still provides the random distribution of keys to nodes, which should result in even storage sizes between nodes. As opposed to the *finger table*, used in the original *Chord* paper, which allows for $O(logN)$ lookup times, a simpler routing table was implemented. Each node only keeps track of its successor and predecessor in the ring. This results in $O(N)$ lookup times. The ring structure is periodically healed through a `stabilize/3` function, by having each node ask its successor for their predecessor and adjusting its own records accordingly in situations when new nodes join. As the structure will have to wrap around at some point, the `between/3` function was implemented in a way to handle `From` being larger than `To` by checking for that condition explicitly and comparing if the specified key is either after `From` or smaller than `To`.

2.2 Storage

Storage was implemented by having each node keep a map of key-value pairs which can be added and looked up by sending corresponding messages to the nodes. These messages are routed through the ring structure until they reach the node responsible for the given key, at which point the operation is

executed and the client is notified through a message including the reference it provided using `make_ref/0`. Additionally, to support new nodes joining the ring, a handover message containing the keys split from the predecessor, based on the two nodes keys, is sent to the new node.

2.3 Failure Detection

To detect failed nodes, Erlang's failure detection mechanism was used to have each node monitor their predecessor and successor. Nodes keep a record of their successor's successor, which is provided during each stabilization round. This is done to maintain structure should their original successor fail. Should the predecessor fail, nodes set it to `nil` and wait for the predecessor's predecessor to notify them. This approach is however not ideal if two neighboring nodes fail concurrently.

2.4 Data Replication

As data is only stored in memory, node failures result in data loss. To mitigate this, nodes keep a replica of their predecessors data which is built concurrently with the main storage. The client is only notified of a successful add operation once both the main storage and the replica have been updated. Both the handover and failure handlers were updated to correctly split and merge the replica storage depending on the state of the ring. In simple terms, when a predecessor dies, the node merges the replica storage with its main storage and propagates its new state to its successor. When a new node joins, it receives the keys for both its main storage and replica storage from its predecessor.

3 Main problems and solutions

3.1 Functions and Parameters

The number of functions and parameters the mature implementations require posed difficulties with retaining a holistic view and often lead to failing tests due to return value mismatches and bugs. This was resolved through meticulous analysis and stepping through the node implementations function for function.

3.2 Replication

Implementing functional data replication posed the greatest challenge as it required considerations in every part of the system. A rudimentary early implementation used the synchronization cycles to periodically correct the replicas, however this was inefficient and would strain the network. Stepping through the implementation scenario for scenario helped identify the

necessary changes and a corresponding sketch, Figure 2, was created to aid development.

4 Evaluation

4.1 Performance

Figure 1 shows the performance of the DHT with different ring sizes and amounts of clients, each of the client machines performing 1000 operations over the first node in the ring. The heatmaps illustrate the routing tables $O(N)$ lookup times, as performance degrades more with the number of nodes in the ring than with the number of clients. The single entrypoint into the ring is also a bottleneck which could be resolved by having clients randomly hit nodes to profit from the pseudo-random key distribution.

4.2 Ring Maintenance

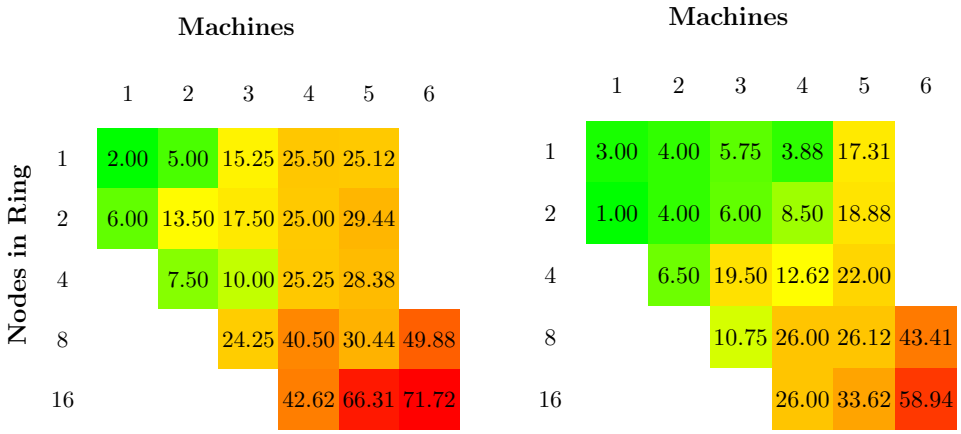
When the ring is stabilized on a short delay, the convergence is sped up and thus lookup and add operations get to the correct node faster. Key responsibility and handovers are also sped up. This however causes the system to use more resources both in terms of compute and network, specifically nodes generating messages for *request*, *notify*, and *status* each round.

5 Conclusions

The advantages of this distributed storage come both in performance through load balancing and fault tolerance through self healing and the lack of a single point of failure. Fault tolerance should however be prioritized, especially in productive applications, as data loss is often not acceptable.

Future improvements to the system include rewriting the routing table with a solution such as the *finger table* to improve lookup timings and consider network latency, although this would significantly increase complexity. A better replica and storage mechanism, ideally persisted on disk with nodes automatically restarting and recovering their state. Replicas could also be used to improve read performance, although this requires dealing with consistency issues. Finally, either an eventual consistency model could be applied to improve availability, or two-phase commits to ensure consistency, both excluding the other due to the CAP theorem.

A Performance Heatmaps



(a) Add Operations (b) Lookup Operations

Figure 1: Performance comparison in milliseconds with varying numbers of nodes and clients.

B Storage Replication Sketch

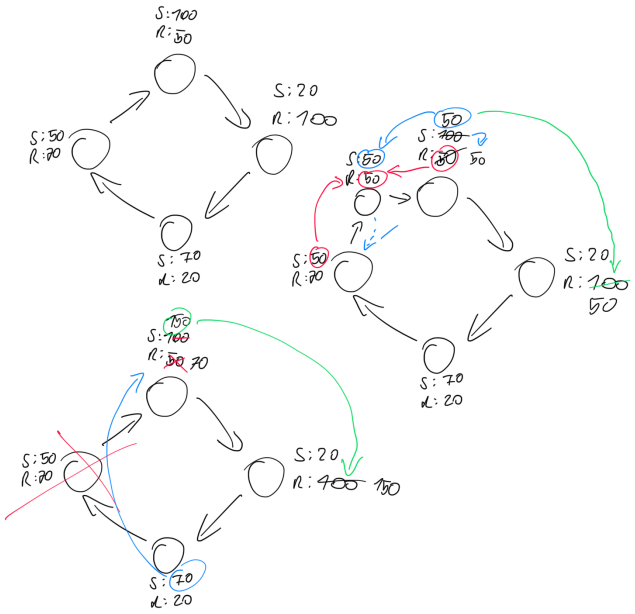


Figure 2: Sketch of the replication mechanism.

C Testing

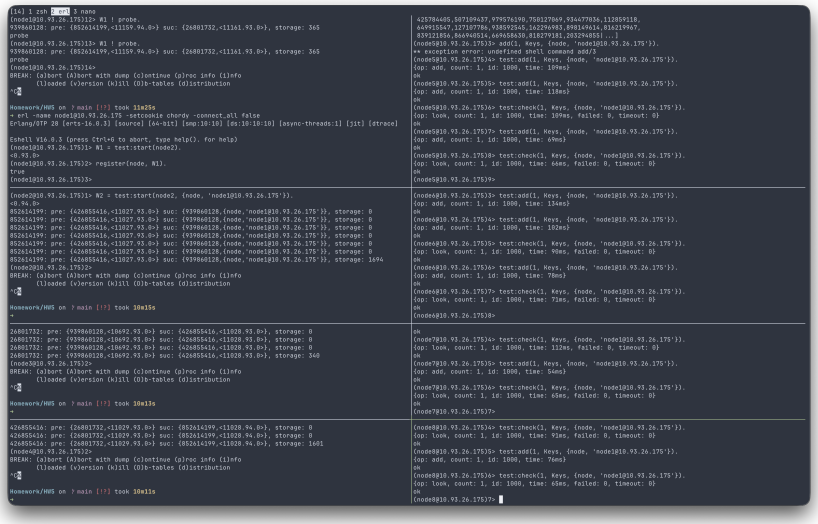


Figure 3: Single node ring test.

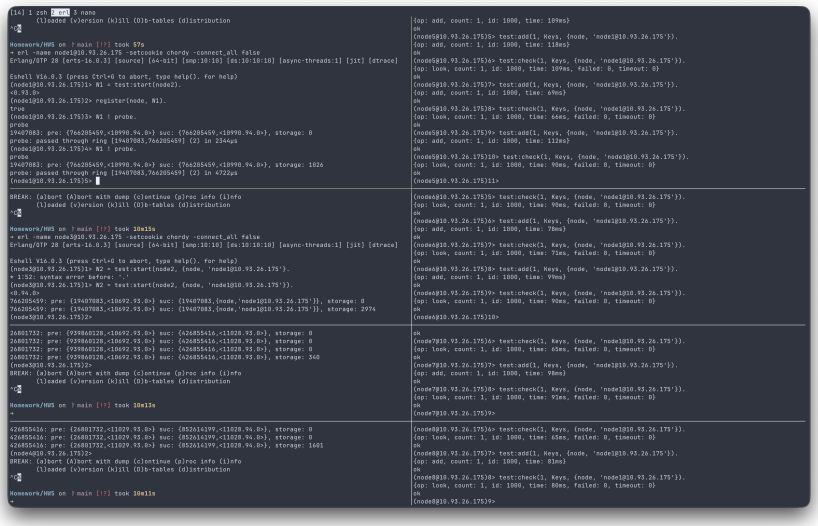


Figure 4: Two node ring test.

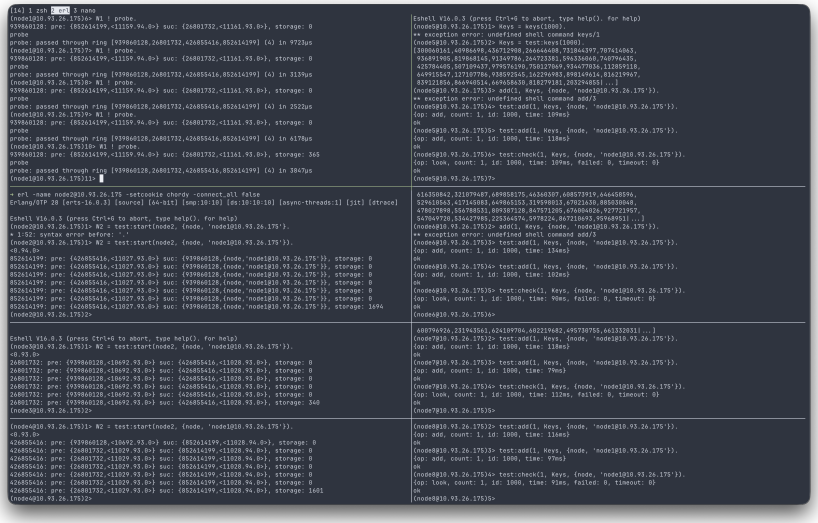


Figure 5: Four node ring test.