Chapter 1

Evaluation

1.1 Success criteria

My original success criteria for this project were:

- To have a tested library for operation-based CRDTs which represent an ordered list of characters
- To have 2 versions of a library which a text editor application might use to collaboratively edit a document with other similar applications over a network. One which will use a central server to pass data, and another where data is sent directly between clients.
- To have a library which, when used by an application, provides collaborative editing functionality with sufficiently small latency between clients to be considered 'realtime'

During the course of the project, I realised that as well as providing sufficiently small latency, it was highly desirable for my library to have reasonable memory usage and reliability as well. I will now discuss to what extend these amended criteria were met.

1.2 Core Library

1.2.1 Unit tests

To have both continuous sanity-checking and more rigorous testing of how concurrency scenarios are handled, I compiled a suite of some 50 unit tests as I developed the project. There were unit tests for some basic functions such as the operations for the helper structures (eg *opStore*), which checked for example that insertion and deletion worked as expected in given scenarios. The tests focused on concurrency involved applying some events to the CRDTs in different specified orders and checking they achieved consistent results.

1.2.2 CRDT Operation Latency

To evaluate the performance of my library, I timed how long it took *opQueue*'s consuming thread to pop and process an operation (perform, store etc.). I plotted this against the length of the document at that point. For insertion, I applied 10,000 `AddRightLocal` operations sequentially to an empty list, and timed each one (Figure 1.1). For deletion, I first inserted 10,000 vertices into the list, then applied 10,000 `DeleteLocal` operations, then reversed the results, so that the first data point is deleting from a one element list (figure 1.2). This procedure was done for the three implementations of `BaseOrderedList` I created: `LLOrderedList`, `LSEQOrderedList` and `ArrOrderedList`.

As performing a `LocalOp` vs a `RemoteOp` only differs in a few lines of code, I decided the times for the local operations would be sufficiently representative of both.

I broke the latency up into four parts:

- *Insertion* is the time it takes to apply the update to the CRDT
- *Store_op* is the time taken to put the operations in the *opStore*
- *Network_send* is the time taken to iterate over all connected peers and send the operation to them (for simplicity in the measurements there were no connected peers)
- *Recovery* is the time spent checking *heldBackOps* to see if any operations were waiting on this one to finish and if so adding them to the front of the queue (in the measurements there would be no such operations)

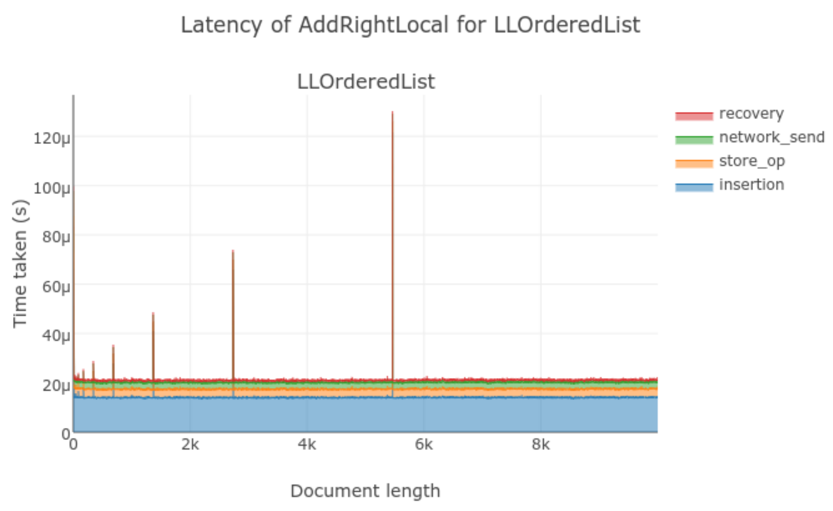
Since after performing each operation, the representation of the new list must be given to the GUI, the vertices must be scanned in order to calculate the column in which the cursor should appear, this is a linear overhead in the size of the list. This has been omitted here as it is the same for all implementations, and it dominates the total time for the first two.

For `LLOrderedList` (Figures 1.1a and 1.2a), we see that both times are largely independent of the length of the list. This is because lookup in the dictionary that maps identifiers to linked list nodes is a constant time operation, meaning so too are insertion and deletion in the linked list. The large spikes occurring for insertions are the resizing of Python's builtin dictionary. When this happens, the size is doubled, which explains why both the spacing between the spikes and their height doubles each time. This isn't seen in the deletion case, because at the point of the first measurement, 10,000 vertices have already been inserted, so the dictionary is already big enough.

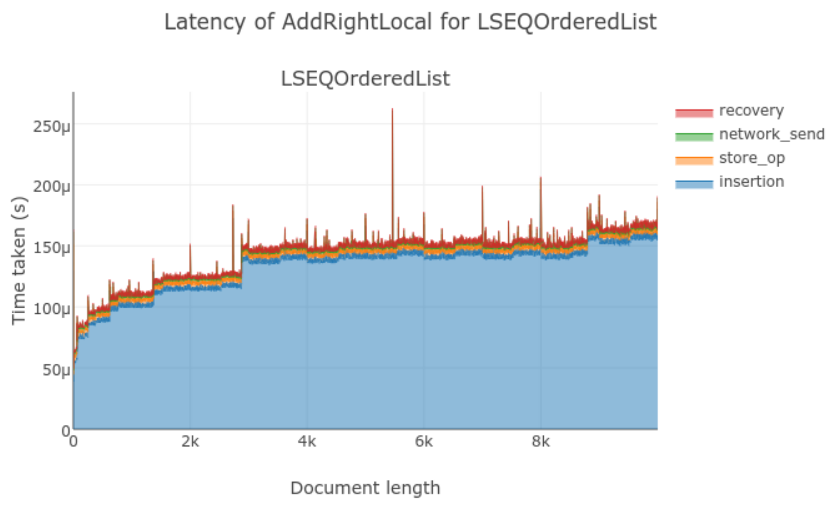
For `LSEQOrderedList` (Figures 1.1b and 1.2b), the complex data structure used to house the linked list nodes (SortedList) offers roughly logarithmic-time lookup, so this dominates both insertion and deletion cost. As such, the graphs exhibit a roughly logarithmic shape. For insertion, there are visible *chunks* of time when the cost is roughly constant, and the chunk boundaries are where some resizing is happening. The logarithmic shape is shown in more detail in Figure 1.3b, where roughly a straight line emerges with a logarithmic x-axis.

For `ArrOrderedList` (Figures 1.1c and 1.2c), since finding a node from its identifier involves a linear search on all the vertices, we see a linear relationship between the number of vertices and the time it takes to insert or delete.

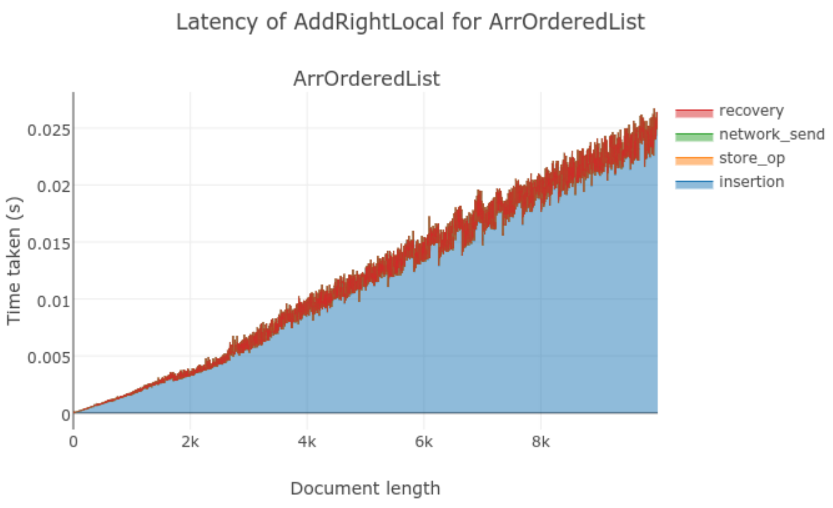
The graphs nicely show how the asymptotic complexities of performing operations for different implementations - $O(1)$, $O(\log n)$ and $O(n)$ respectively - affect real-world performance. Additionally, we can see in Figure 1.3 in both implementations that the resizing operation completely dominates for larger documents when it actually happens, but at a length of around 45,000 characters, the resizing latency is still only around 1ms. In order to get up to a resizing latency of 0.1s (on the edge of being noticeable) requires approximately 7 more resizes ($2^7 = 128$, assuming the size continues to double each time) . This happens every time the size is increased, so this time would be reached at $45,000 * 128 \approx 4$ million characters. As of December 2016, the largest wikipedia article was around 1.1MB in size, so at one byte per character this would be slightly larger than the largest Wikipedia article.



(a) `LLOrderedList`

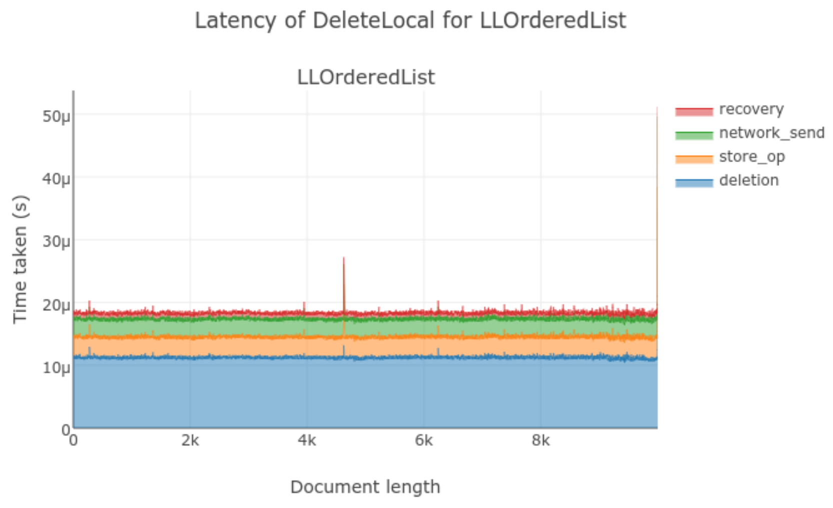


(b) `LSEQOrderedList`

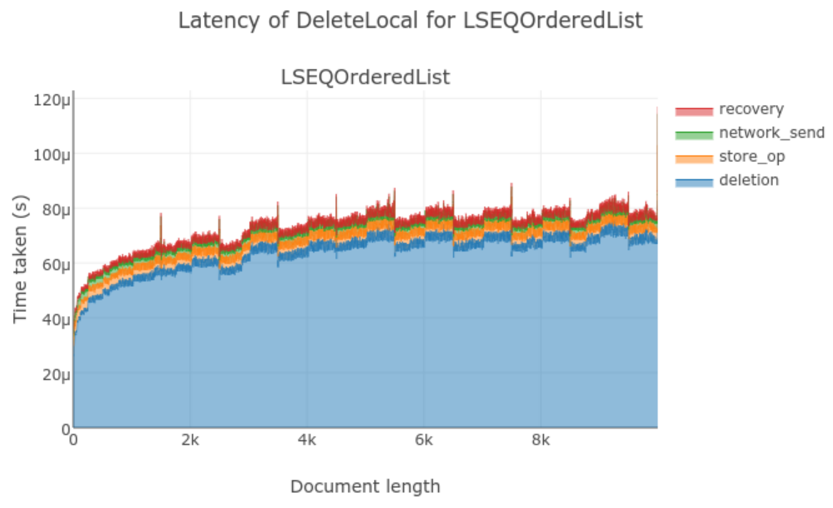


(c) `ArrOrderedList`

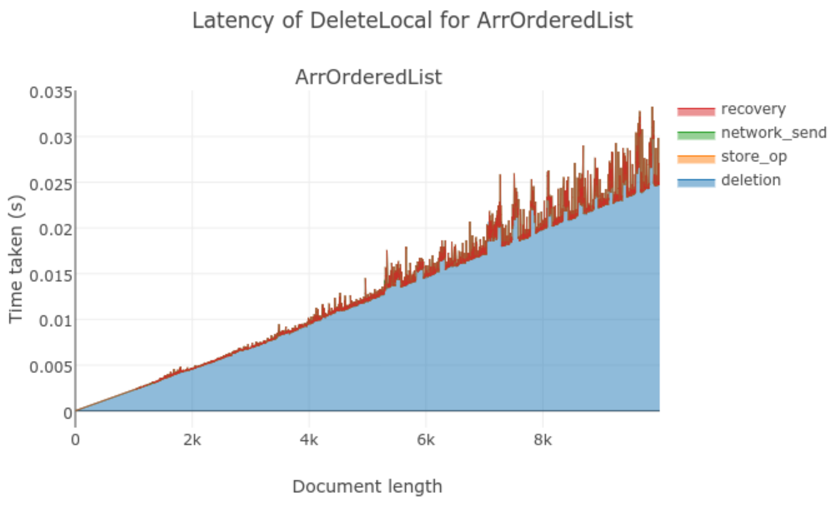
Figure 1.1: AddRightLocal latencies plotted against the length of the document. Note the difference in scales.



(a) `LLOrderedList`

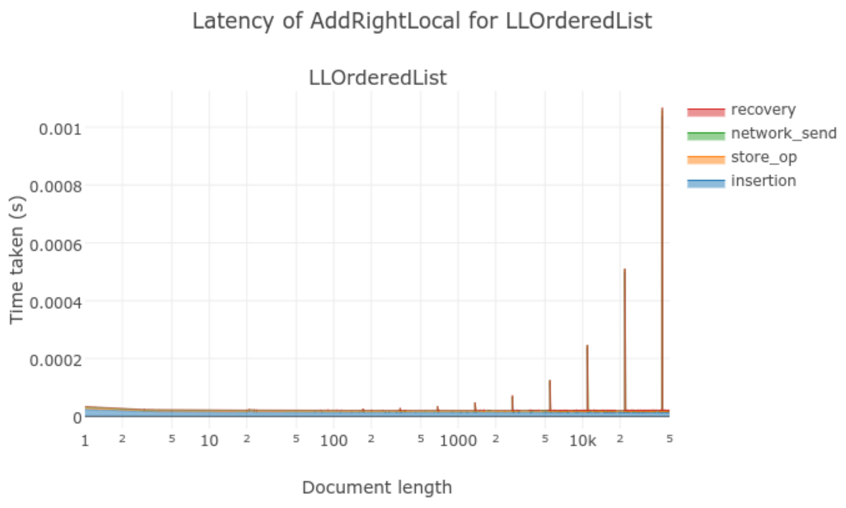


(b) `LSEQOrderedList`

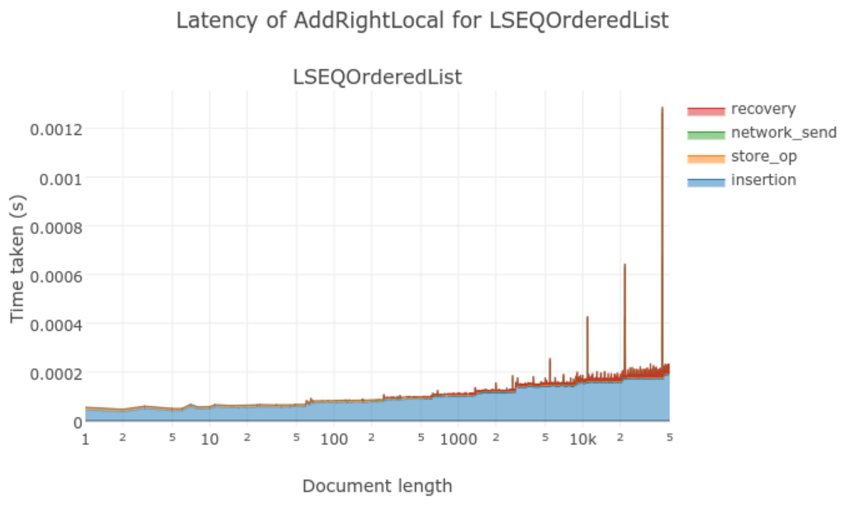


(c) `ArrOrderedList`

Figure 1.2: DeleteLocal latencies plotted against the length of the document. Again note the difference in scales.



(a) A log-scale plot of insertion time for `LLOrderedList`



(b) A log-scale plot of insertion time for `LSEQOrderedList`

Figure 1.3: Latency for 50,000 AddRightLocal operations on logarithmic scales

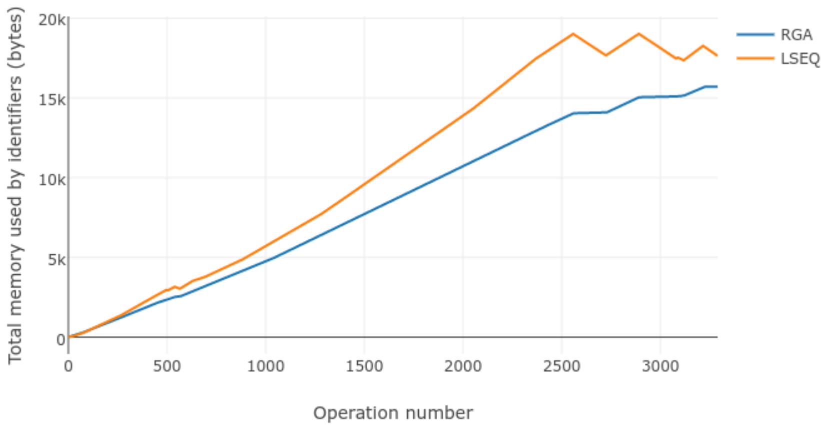
1.2.3 Memory usage

Recall that LSEQ allows vertices to be freed from memory (at the cost of more memory per vertex on average), whereas RGA allows a fixed per-vertex cost, but disallows reclaiming from memory.

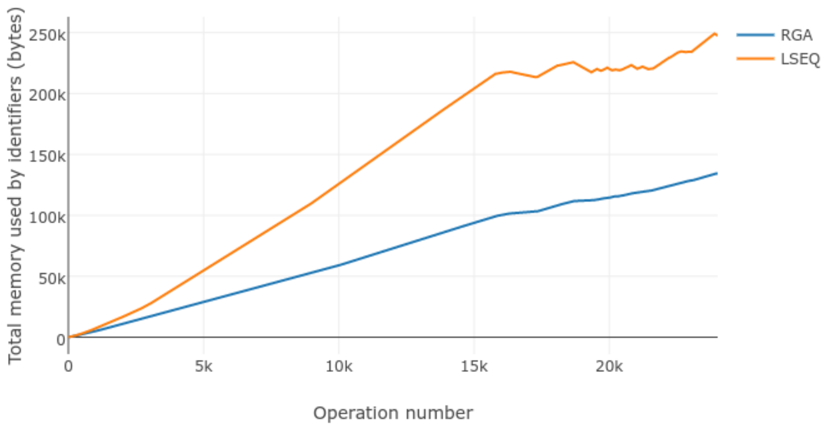
To see the real-world memory usage of the different approaches, I replicated an experiment from earlier work [**?**, **?**] simulating editing with the revision history of different Wikipedia pages. I scraped the revisions from the pages histories, then performed a character-level *diff* between consecutive versions to produce operations to transform the first version into the last that could be applied to my CRDTs.

Figure 1.4 shows the results of this. As Figures 1.4a and 1.4b show, for documents with mostly insertions, the overhead of variable size identifiers outweighs not being able to delete elements. However, for documents with more deletions performed, such as Figure 1.4c (a noticeboard reporting vandalous users which gets cleared when they are dealt with), we can see the non-decreasing memory usage by RGA is significantly higher. As such, choosing a scheme is a trade-off based on the expected proportion of deletions that will happen.

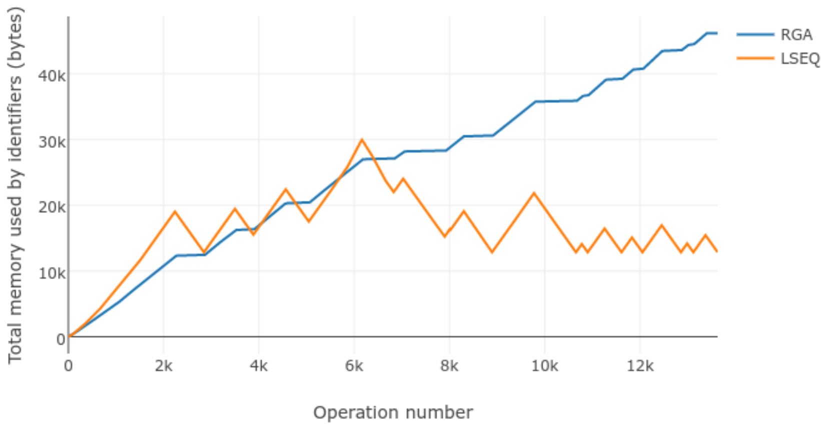
In the Oxbridge example, we see that for roughly 20,000 characters we are using around 200kB, which is 10 bytes per character for each identifier on average. Recalling the largest Wikipedia article to be $\sim$1.1MB in size and assuming all text and 1 byte per character, my scheme would use $11 + 1.1 = 12.1$ MB to store the vertices (the largest contributor), which is a reasonable figure for an application nowadays.



(a) Sehna



(b) Oxbridge



(c) Administrator intervention against vandalism

Figure 1.4: Memory usage of my implementations of RGA (`LLOrderedList`) and LSEQ when simulating editing on Wikipedia revision history.

1.2.4 Undo correctness

To show my approach to implement undo (the limited version as described in the previous chapter) is correct, I wish to prove that undoing and redoing operations will preserve consistency and all replicas will see the same state.

Firstly, due to the system of counts, an $\textsc{Add}(v)$ operation amounts to incrementing the count of v, while the $\textsc{Delete}(v)$ operation decreases it. There is then the side-effect that anything with a count of 1 is displayed in the document. So,

$$\forall others, l, op.\ p_{undo(op)} \circ p_{others} \circ p_{op}(l) \equiv_l p_{others}(l)$$

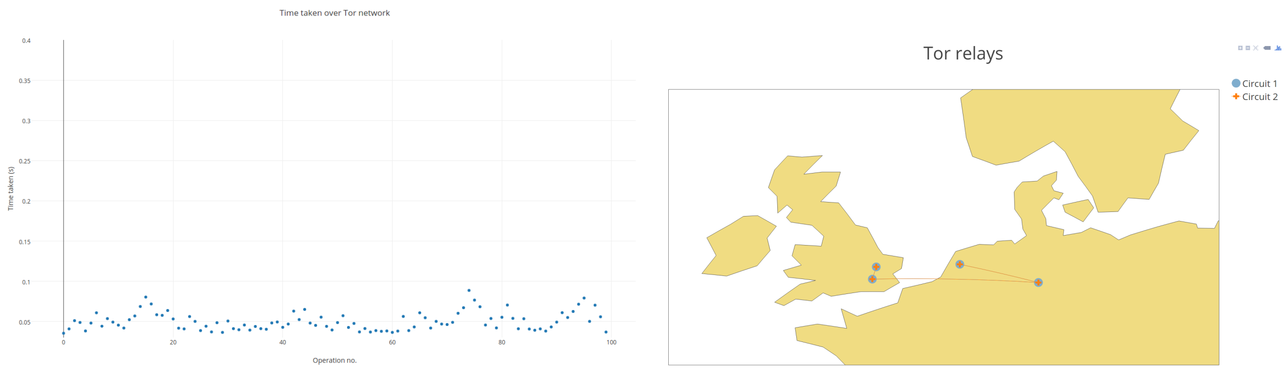
by the properties of addition (incrementing the count for one vertex at the start, then decrementing it at the end is equivalent to a no-op). Here, $p_x(l)$ returns the new state l' obtained by performing operation x on list l, and *others* is a list of operations, so that p_{others} is shorthand for $p_{others_0} \circ p_{others_1} \circ ... \circ p_{others_n}$.

There is a subtlety in the implementation that is not immediately clear, but required for the pseudocode to function correctly - the count of a vertex is always at most 1. Although $\textsc{Add}$ will increment the count of a vertex, this can happen only once if not part of an undo. So upon the first $\textsc{Add}(v)$, $\textsc{count}(v)$ becomes 1. Thereafter, an $\textsc{Add}(v)$ operation can occur at any other replica only if that replica first originates a $\textsc{Delete}(v)$ and undoes it. Since operations from the same replica are always causally ordered, any $\textsc{Add}(v)$ operations after the initial one must be preceded by a $\textsc{Delete}(v)$ at every replica. Thus, $\textsc{count}(v)$, when 1, will always be decremented before being incremented, and so can never exceed 1.

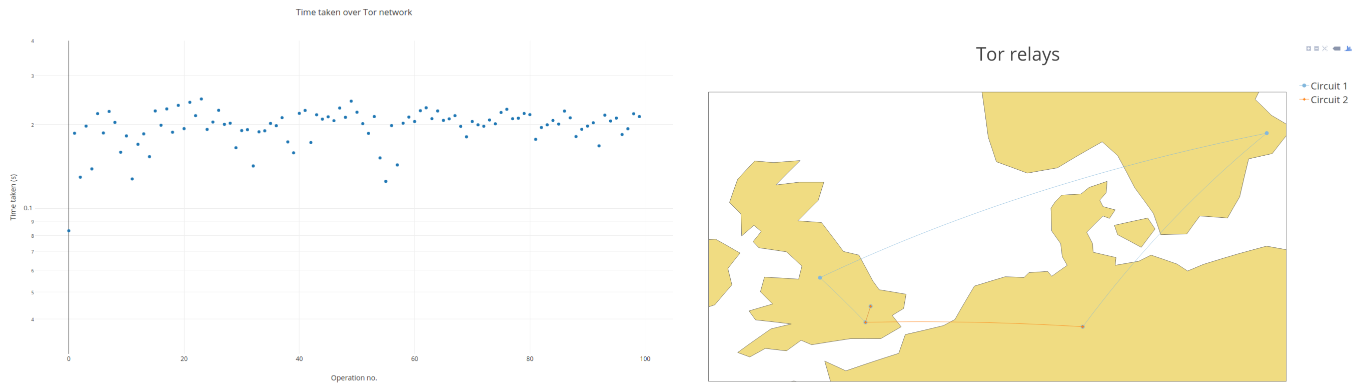
1.3 Networking

1.3.1 Network Latency (show latency profiles of different Tor circuit setups)

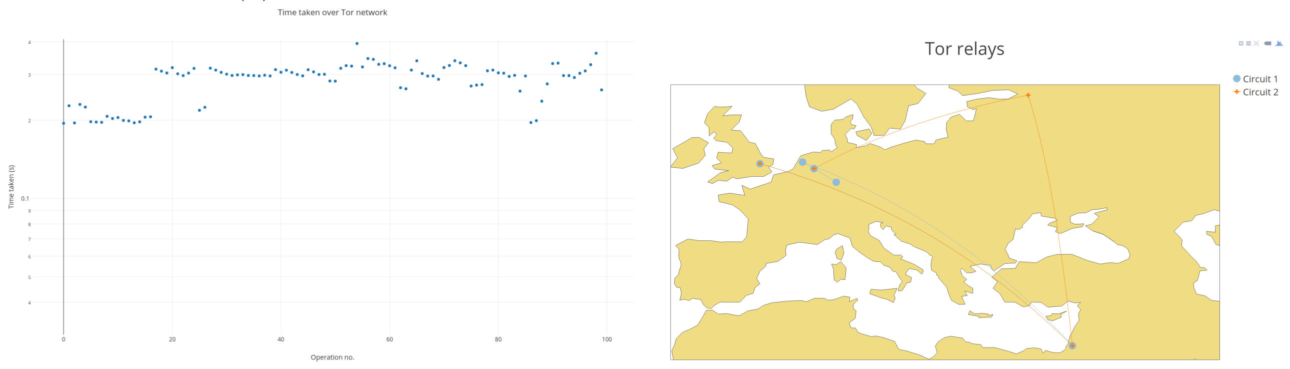
To see how quickly operations can be delivered over the Tor network. I setup a hidden service and sent 100 operations to it from the same machine. Therefore, messages would go out to the Tor network and back in, and I would record the time of departure and arrival at my machine. Three different profiles are shown in Figure 1.5 corresponding to three different choices of circuits. The maps show the locations of the relays used in the two circuits built to the rendezvous point. As the figures show, the time spent in the network is quite variable and highly dependent on the chosen locations of the relays.



(a) Relay setup 1 - mean time 0.050s, standard deviation 0.011s



(b) Relay setup 2 - mean time 0.198s, standard deviation 0.027s



(c) Relay setup 3 - mean time 0.285s, standard deviation 0.046s

Figure 1.5: Locations of Tor relays used in a Hidden Service circuit and the latency profiles for sending 100 operations through them

So, the time between typing a character and noticing another replica deleting that character involves:

- Performing the insertion locally - on occasion 0.1s in a large document, from Section 1.2.2
- Sending the operation - from the measurements above around 0.3s
- The other replica performing the operation - 0.1s

- The other replica's person noticing the change and pressing a key 0.3s for human reaction time
- Their replica performing their operation - 0.1s
- Sending their operation to you - 0.3s
- Performing their operation locally - 0.1s
- You noticing their operation - 0.3s for human reaction time

Giving a total time of around 1.6s which I deemed to be reasonable.

1.3.2 Reliability

Once a connection over Tor is established, there is nothing in the Tor specification that restricts its lifetime. Fresh circuits are by default chosen every 10 minutes, but only for new connections. Therefore, a connection opened by one replica to another would only be closed either due to one party explicitly disconnecting or due to some network event that is outside the control of the application, for example one of the relays going offline. I set up a scenario where one replica generates and sends an operation every minute to send to another running on my machine (though going through Tor), and the connection ran successfully for 23 hours before my laptop disconnected because it had to be moved, however as discussed there is no theoretical limit.

I didn't implement any retry behaviours though (beyond what TCP offers), so on such a network event, connections are lost unless the user explicitly reconnects.

1.3.3 Network Architecture

Below is a comparison of client-server and P2P architectures as implemented in my project.

Running a P2P application instead of the client-server approach has several advantages. For example, servers are inherently bottlenecks (and single points of failure), and provide privacy problems. Decentralizing means one can potentially be as anonymous as using Tor allows (providing the secondary channels used to distribute onion addresses etc. preserve anonymity) as described in this project when using Tor Hidden Services, and all data completely hidden from anyone other than the specified collaborators. Moreover, the P2P mode adds no infrastructure, so its availability only relies on the availability of the Internet and the Tor network.

However, the fully connected nature of the P2P approach means for k connected collaborators, a new peer coming online means potentially the entire causal history of the document is sent to the new peer k times (each peer may have a different set of operations so all must be queried). This is obviously wasteful as the client-server approach gives one authoritative answer.

1.3.4 Security

A man-in-the-middle (MITM) attack describes a situation where an adversary intercepts communications along a channel and relays there own messages to each endpoint. Without the use of authorization for Hidden Services, my library is vulnerable to such an attack.

To illustrate the attack, consider two peers A and B wanting to edit a document, and an adversary C who can read and write to the channel between them. If C finds out the onion address for A and B by some means, it can connect to A and B. A and B would just assume they are connected to each other, but the key point is that the authentication is unidirectional, from A to C and B to C, but not the other way round. Thus, C can intercept and modify the document any way it wants.

Authentication for Hidden Services, then, solves this by requiring C to authenticate itself to both A and B, by providing their specified cookies. The assumption is that these cookies are exchanged over a secure channel, such that C cannot get them. However, a flaw with this system is that there is no simple revocation available. Once C has a cookie, it is as authentic as any legitimate peer, and my implementation has no way to easily redistribute cookies.

One of the limitations of my project then, is that only over Tor with Client Authentication can you have an authenticated encryption scheme, as I didn't implement any verification of public keys into the non-Tor encryption protocol.