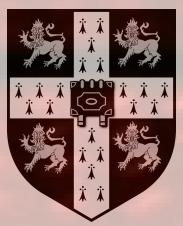
Marc Harvey-Hill

Implementing the HotStuff consensus algorithm

Computer Science Tripos Part II
Gonville and Caius College
March 2023



Declaration of Originality

I, Marc Harvey-Hill of Gonville and Caius College, being a candidate for Part II of the Computer Science Tripos, hereby declare that this dissertation and the work described in it are my own work, unaided except as may be specified below, and that the dissertation does not contain material that has already been used to any substantial extent for a comparable purpose. I am content for my dissertation to be made available to the students and staff of the University.

Signed [signature]

Date [date]

Proforma

blah blah

Contents

Chapter 1

Introduction

The power of blockchains lies in their ability to decentralise applications that were traditionally run in a centralised manner. The implications of this are far-reaching: central banks can be replaced by decentralised cryptocurrencies, traditional corporations can be replaced with decentralised autonomous organisations (DAOs), internet infrastructure like DNS servers can be decentralised, and any possible algorithm can be run on a decentralised 'world computer'. The innovation that make blockchains possible is the byzantine consensus algorithm.

Byzantine fault-tolerant consensus algorithms allows a group of parties to agree on some piece of information under adverse conditions where some messages can be lost and some parties are controlled by a malicious adversary. For example, one could create a cryptocurrency by using such an algorithm to reach consensus between many devices on a ledger of transactions like "Account Alice transfers account Bob £10"; the algorithm will ensure that transactions cannot be lost and the system cannot be sabotaged by malicious actors.

Byzantine consensus algorithms can be viewed as solutions to the byzantine generals problem [25]. In this problem, a group of generals must all agree to siege a castle at the same time, but they can only communicate via messengers that take some time to arrive and can be captured en route. Additionally, up to a third of the generals may be malicious, and try to prevent the other generals from reaching consensus on a time to attack. By following a byzantine consensus protocol the generals can reach consensus on a value like "attack at dawn". *Multi-shot* consensus algorithms allow consensus to be reached on multiple values, resulting in a continuously growing log that can never be modified or erased, only extended.

Blockchains can be either permissioned, or permissionless. Permissioned blockchains have a previously agreed set of participants in the consensus algorithm, whereas permissionless blockchains allow participants to join and leave freely. Most well-known blockchains, such as Bitcoin [26] and Ethereum [31], are of the permissionless variety. Permissioned blockchains can be deployed in a permissionless setting if they are augmented with an additional layer of security, which can be proof of work, proof of stake, or some other similar mechanism. These aim to prevent a 'Sybil attack' where a permissioned blockchain can be overrun by a large number of malicious nodes; proof of work, for example, adds a requirement for a proof of computational work in order to participate in consensus, making Sybil attacks economically and computationally infeasible. Permissioned blockchains are of interest for applications within a group or organisation, such as a company, where the participating nodes are known in advance.

HotStuff is a byzantine consensus algorithm that was notably used by Meta's Libra project [11], a cancelled permissioned blockchain-based payments system. The algorithm is relevant

because of various performance advantages over other byzantine consensus algorithms such as PBFT [14], SBFT [20], DLS [18], Tendermint [22], and Casper [12].

Building practical, well-performing implementations of consensus algorithms is non-trivial. These algorithms are usually specified in short pieces of pseudocode that may not be specified precisely and require much more code to implement in practice. Such software has a wide range of failure modes mostly due to their parallel nature, including deadlocks, resource starvation, and bugs in the implementation [15].

The main contributions of this dissertation are:

- Providing a reference implementation of HotStuff in OCaml based on a paper by Yin et. al [32].
- Synthesising information from different sources to provide a complete explanation of the HotStuff algorithm, and how it can be arrived at through modifications to simpler consensus algorithms (basic algorithm section 2.2, chained algorithm section 3.2.1, pacemaker section 3.2.2).
- Giving solutions to key practical challenges of implementation and optimisations that can be made (section 3.4), as well as showing their effectiveness (section 4.3.3)
- Giving a complete specification and proof of HotStuff (section 3.3), that adapts the pacemaker mechanism which was not sufficiently specified in the original paper. This specification synthesises the chained algorithm described in the paper (section 3.2.1), a view-change protocol based on other talks and papers (section 3.2.2), and our own changes to integrate the pacemaker into HotStuff without the need for synchronised clocks.

Chapter 2

Preparation

In this chapter we disclose my knowledge and experience prior to beginning this project (section 2.1), give a theoretical basis for understanding the HotStuff algorithm by building up from simpler consensus algorithms (section 2.2), outline the tools, libraries (section 2.3), and professional methodology (section 2.5) deployed in implementation, and highlight the requirements that the implementation should meet (section 2.4).

2.1 Starting point

I had some experience using OCaml from the IA Foundations of Computer Science course but had never used it in a project. The IB Distributed Systems course also provided some useful background knowledge, particularly as it briefly covered Raft [28], a non-byzantine consensus algorithm. I had some understanding of byzantine consensus from my own reading into Nakamoto consensus [26] and from developing a wallet application for Ethereum [31]; neither of these was directly useful to implementing HotStuff, but they gave me some wider context of the field.

2.2 HotStuff algorithm

HotStuff is a byzantine consensus algorithm; it allows a group of nodes to reach agreement on a log of values under adverse conditions, such as messages being lost, or some nodes being byzantine. In each *view* a *leader* node proposes some value by sending it the *replicas* (another word for nodes). After several messages are exchanged the log may be committed, meaning that there is consensus on the committed part of the log, and it is now immutable.

HotStuff has the following properties:

- Safety — once a log has been committed up to some point that part of the log is immutable, it can only appended to.
- Liveness — the system is guaranteed to make progress once a non-faulty leader is elected.

- Responsiveness[1] — the system is able to make progress as fast as network conditions allow once a non-faulty leader is elected; it does not have to wait for some timeout to elapse.

The system model describes the adverse conditions under which HotStuff can operate:

- Partially Synchronous — messages sent by one party will always be delivered to another within some bounded amount of time (δ) after global synchronisation time (GST) has been reached.
- Authenticated — a message source cannot be spoofed. We assume that all messages are signed, providing authentication.
- Byzantine — a maximum of f faulty nodes may be controlled by an adversary that is actively trying to prevent the nodes from reaching consensus, where $n = 3f + 1$ and n is the total number of nodes. This is the maximum number of byzantine nodes for which consensus can theoretically be reached [29][19].

Each view is composed of several *phases*. In each phase the leader broadcasts to the replicas, that respond with an acknowledgement (*ack*). The leader waits until it receives a *quorum* of $n - f$ acks[2] before proceeding to the next phase, meaning that at least $f + 1$ of the acks were from honest nodes. The basic HotStuff algorithm has five phases in each view.

The final phase of a consensus algorithm is *decide*. We assume that a *client* sends commands to the nodes, the nodes reach consensus on a log of commands, and then execute them in order; this ensures all nodes will be in the same state. The decide phase commences after the log is committed, the nodes can execute the new commands and respond to the client that the command was successfully committed.

[basic algorithm... remove stuff about view-change here] The HotStuff algorithm consists of a responsive byzantine consensus algorithm (section 2.2.3), with a view-change protocol (section 3.2.2) that allows the system to skip the view of a faulty leader. We will start by describing a simpler consensus algorithm that is not responsive, and cannot handle byzantine nodes (section 2.2.1). We then extend this algorithm to handle byzantine threats (section 2.2.2). Finally, we extend this algorithm to be responsive by removing the need for a timeout to progress (section 2.2.3).

2.2.1 Non-byzantine consensus

We will start by describing an algorithm to solve the simpler problem of reaching consensus with the stronger assumption of a crash-stop model instead of a byzantine one; this means that we assume nodes cannot be malicious but they can still crash and never come back online. Examples of similar algorithms include Raft [28] and MultiPaxos [check citations are actually multi-shot!] [23][24].

[1] To be precise HotStuff is *optimistically* responsive, we will describe this in section 2.2.3.

[2] N.B. the size of a quorum is different in the non-byzantine case.

Each view consists of three phases:

new-view — The leader learns about previously committed logs. All replicas send a *new-view* message to the leader, containing their longest previously committed log.

commit — The leader waits until it receives a quorum of greater than $\frac{n}{2}$ *new-view* messages, and then picks the longest log it received (λ) to propose. It then broadcasts a *commit* message to the replicas, proposing λ' to the nodes, where λ' is λ optionally extended with the leader's new value.

decide — Once the leader receives a quorum of acks, the log has been successfully committed. The leader can broadcast "decide" to the replicas, who can execute the new commands and respond to the client.

Crucially, this algorithm has the *safety* property. Since the leader waits to receive a quorum of greater than $\frac{n}{2}$ *new-view* messages, λ is guaranteed to be the longest log that has previously been committed. This is because the the quorum of *new-view* messages must share at least one node with the past quorum of *commit* acks for λ. The new proposal λ' will never conflict with λ, hence the algorithm is safe.

2.2.2 Byzantine consensus

In this section we extend our non-byzantine algorithm (section 2.2.1) to achieve consensus under a byzantine system model. To do this we must first introduce the *quorum certificate* (QC), a cryptographic proof that a leader has received a quorum of acks. We then consider the threats posed by byzantine nodes, give an algorithm that solves these problems, and make an argument for safety. Examples of similar algorithms include xyz

Quorum certificates

A QC is a quorum of $n - f$ acks with a matching *threshold signature*. A threshold signature combines several signatures of the same message into one [30][13]; in this case the signature from each ack is combined[3]. QCs have a key property that our byzantine algorithm will rely on:

[add a diagram of intersecting QCs!]

Property 2.2.1. *There will always be at least one honest node in the intersection of any two QCs.*

Recall that $n = 3f + 1$. The property holds since a QC contains a quorum of $n - f$ acks, at least $f + 1$ of which must be from honest nodes. To have two quorums that *do not* share an honest node would require at least $2(f + 1) = 2f + 2$ honest nodes, but our system only has $2f + 1$.

[3]Recall that our messages are signed to provide authenticated delivery.

Threats

Byzantine nodes introduce two threats that we will deal with in turn. We present ideas for solutions to these threats; it will become clear why these solutions are effective in our argument that the algorithm is safe.

1. Threat (Equivocation) — a faulty leader proposes one value to some replicas and a different value to others. For example, in the case of a cryptocurrency a malicious actor (Mallory) could carry out a double-spend attack by proposing "Mallory transfers Alice £10" to some nodes and "Mallory transfers Bob £10" to others, even if Mallory's account contains less than £20.

   Solution idea — add a new stage *prepare* which happens just before the *commit* phase, where the leader pre-proposes a log before proposing it in the *commit* phase.

2. Threat — a faulty leader proposes a log that conflicts with one that has previously been committed.

   Solution idea — when replicas receive a proposal for some log they must *lock* on it, and not accept a pre-proposal for a conflicting log.

Algorithm

Modifying the non-byzantine algorithm to include our solutions to threats 1 and 2 results in the following:

new-view — the leader learns about previously committed logs. All replicas send a *new-view* message to the leader, containing their longest previously committed log.

prepare — the leader waits δ until it receives a *new-view* message from all replicas (we will revisit this in section 2.2.3), and then picks the longest log it received (λ) to pre-propose. It then broadcasts a *prepare* message to the replicas, pre-proposing λ' to the nodes, where λ' is λ optionally extended with the leader's new value. The replicas ensure that λ' does not conflict with their *lock* before sending an ack.

commit — the leader waits until it receives a quorum of at least $n - f$ *prepare* acks. It then broadcasts a *commit* message to the replicas, proposing λ' to the nodes, and includes a QC of *prepare* acks. The replicas then *lock* on λ' and send a *commit* ack.

decide — once the leader receives a quorum of acks, λ' has been successfully committed. The leader can broadcast "decide" to the replicas, who can execute the new commands and respond to the client.

Argument for safety

We give an informal inductive argument of the safety of this algorithm based on it solving threats 1 and 2. To be safe, we must have that if some log λ is committed in view v, at no point in future will some conflicting log be committed.

Base case — in view v no log that conflicts with λ may be committed, in other words equivocation (threat 1) is not possible. For a view to commit a value, there must have been a QC of *prepare* acks, and a QC of *commit* acks. By property 2.2.1, there must be at least one honest replica in the intersection of these QCs that would not have acknowledged conflicting proposals in the same view.

Inductive step — no conflicting log can be proposed in view v' where $v' > v$ (threat 2). This holds because any log pre-proposed in view v' must receive a QC of *prepare* acks; by property 2.2.1, there must be at least one honest node in the intersection between this QC, and the QC of *commit* acks λ in view v. This honest replica is locked on λ, so would not accept a proposal that conflicts it.

From this it follows that in no view from v onwards will a log conflicting with λ be committed, so the algorithm is safe.

2.2.3 Optimistic responsiveness

The HotStuff protocol has the *optimistic responsiveness* property, which means that it can make progress as fast as network conditions allow without waiting for a timeout [cite thunderella]. It is called *optimistic* as it is only guaranteed to have this property once GST has been reached.

The algorithm described in section 2.2.2 is not responsive, as it has a timeout in the *prepare* phase. We first describe the problem that means this timeout is needed, present an algorithm that solves it, and give an informal argument of its effectiveness.

[rewrite all this!]

Problem

Consider again the *prepare* phase; the leader includes in its pre-proposal the QC from the highest view that it hears about in the *new-view* messages it receives. However, there may be some honest replica [how???] that the leader did not hear from (perhaps their message was lost) that is locked on a higher-view proposal than the one that the leader chooses to pre-propose. When this replica receives the pre-proposal, it will reject it as it is locked on a higher-view proposal, and the system will not make progress. This is why it is necessary to wait for a timeout Δ, so that the leader waits a sufficient amount of time to receive a *new-view* from all replicas.

Solution idea — blah blah

Algorithm

new-view — the leader learns about previously committed logs. All replicas send a *new-view* message to the leader, containing their longest previously committed log.

prepare — the leader waits δ until it receives a *new-view* message from all replicas (we will revisit this in section 2.2.3), and then picks the longest log it received (λ) to pre-propose. It then broadcasts a *prepare* message to the replicas, pre-proposing λ' to the nodes, where λ' is λ optionally extended with the leader's new value. The replicas ensure that λ' does not conflict with their *lock* before sending an ack.

pre-commit — blah blah

commit — the leader waits until it receives a quorum of at least $n - f$ *prepare* acks. It then broadcasts a *commit* message to the replicas, proposing λ' to the nodes, and includes a QC of *prepare* acks. The replicas then *lock* on λ' and send a *commit* ack.

decide — once the leader receives a quorum of acks, λ' has been successfully committed. The leader can broadcast "decide" to the replicas, who can execute the new commands and respond to the client.

Argument for responsiveness

This ensures that if some honest replica becomes locked on a value in the *commit* phase, then there are at least $f + 1$ honest nodes that have a 'key' for that value from the *pre-commit* phase. More specifically, they have a 'key-proof' composed of a QC over *pre-commit* acks. Replicas then send this key-proof with their *new-view* message when a new view begins. The new leader selects the key-proof with the highest view proposal and sends it along with their *prepare* message. Even if the leader does not receive a *new-view* message from the replica that is locked on the highest view proposal, they must have received the key to this proposal from some honest replica, so their *prepare* message will make progress.

2.3 Tools & libraries

In this section we outline the languages and libraries used in implementation, and justify why they were appropriate for this project.

2.3.1 OCaml

I chose OCaml [7] for this project due to its high-level nature, static type system, ability to blend functional and imperative paradigms, and good library support. OCaml's multi-paradigm nature is suitable for implementing HotStuff, as the core state machine can be elegantly expressed in a functional way, whereas interacting with the RPC library to send messages is better suited to an imperative paradigm. Additionally, the Tezos cryptocurrency [cite!!!] is written in OCaml and contains a cryptography library (section 2.3.4) that provides the functionality needed by HotStuff.

The performance bottlenecks for distributed byzantine algorithms are generally cryptography, message serialisation and network delays. This means that it is more important to choose a

language with suitable features to aid implementation, rather than picking a 'high-performance' language like C++. [mention multicore OCaml, it was not ready when this project began]

OCaml has a powerful module system that facilitates writing highly reusable code. The module system was only briefly touched upon in the tripos (in Concepts in Programming Languages from IB), so I spent time learning about these features. Modules provide an elegant interface for the core state machine to interact with the imperative parts of the program that actually send messages over the network.

There is no existing reference implementation of HotStuff in OCaml, so my project contributes to the growing OCaml ecosystem[4].

2.3.2 Lwt

Lwt [9] is a concurrent programming library for OCaml. It allows the creation of promises, which are values that will become determined in the future; these promises may spawn threads that perform computation and I/O in parallel. In order to use Lwt I had to learn about monads, which are ways of sequencing effects in functional languages [verify!!!] and are used by asynchronous promises in Lwt. Lwt is useful to this project as promises provide a way to asynchronously dispatch messages over the network and wait for their responses in different threads. Promises are cheap to create in Lwt, so one can create many lightweight threads with good performance [citation needed***].

One alternative I could have used is Jane Street's Async library [1], which has similar features but better performance; however, I chose not to due to its poor documentation. Another alternative that also has better performance than Lwt is EIO [8], but this library is new and not yet in a stable state.

2.3.3 Cap'n Proto

Cap'n Proto [2] is an RPC framework that includes a library for sending and receiving RPCs [mention message serialisation!!!], and a schema language for designing the format of RPCs that can be sent. Benchmarks for the library are presented in section 4.2.1.

[some people use xyz fancy library that is much faster but this was clearly not feasible for this project...]

2.3.4 Tezos cryptography

The Tezos cryptography library [6] provides aggregate signatures using the BLS12-381 elliptic curve construction [cite crypto paper???]. It provides functions to sign some data using a private key, aggregate several signatures into a single one, and check whether an aggregate signature is valid. The only difference from the threshold signatures needed by HotStuff is that each

[4]Since our project is implemented purely in OCaml, it could be deployed in a MirageOS unikernel [5]

individual signature in an aggregate signature can sign different data, whereas with threshold signatures each individual signature is over the same data. It is trivial to implement threshold signatures using this library by checking that the data is the same for all signatures inside the aggregate signature. Benchmarks for the library are presented in section 4.2.2.

2.4 Requirements analysis

In order to be successful the implementation should conform to the following requirements:

- Correctness — the consensus algorithm should be implemented as it is described in the paper [32]. This can be established by testing the program trace for compliance with the algorithm specification.
- Evaluation — analysis of system throughput and latency should be carried out by testing the program locally, analysing the trace, and testing in simulated network.
- Optimisation — implement features to improve transaction throughput and reduce latency over the naive implementation. This can be achieved through architectural decisions, tuning the scheduler, and ensuring cryptographic libraries are being used efficiently.

2.5 Software engineering practices

In this section we describe the professional software engineering methodology deployed during implementation, and justify that this project is ethical.

2.5.1 Development methodology

For this project we used an iterative waterfall development methodology. Objectives were chosen in accordance with the timetable set out in the proposal [reference appendix!]. Development then proceeded in cycles of implementation, testing, and analysis of the program trace and timing statements. This approach was particularly useful during the optimisation of our system (section 3.4), which required extensive analysis of the logs and rapid development of different prototypes to compare performance.

2.5.2 Testing & debugging methodology

Unit testing was carried out using 'expect tests', which compare a program trace to the correct output. A testing suite of expect tests verifies that the program behaves as specified in the HotStuff paper. This suite has 100% code coverage[5] of the consensus state machine code, the coverage report is at *_coverage/index.html*.

[5]The report says 97.89% coverage, but the only uncovered code is the testing code itself.

The Memtrace library and viewer [3] were used to profile the memory usage of the program. One can generate a flame graph of memory allocations to see which parts of the program are using the most memory.

[talk about mininet!] [4]

Due to the distributed nature of the program, normal debugging tools and profilers are not useful for debugging deadlocks and performance issues. This is because the cause of deadlocks and performance issues is often some process waiting or a backlog of work forming, but this cannot be detected by tools that just track things like CPU usage. Instead, I had to rely on manual inspection of the program trace and commands that measure the real time taken for some part of the program to run.

[integrate this with the above to make one paragraph!!!] As mentioned in section 2.5.2, the nature of the project meant that debugging had to be carried out by manual inspection of the program trace and timing sections of the program. To overcome this I carried out tests in a scientific manner, constructing a hypothesis for why the program was slow based on analysing the program trace and timing statements, then attempting to test my hypothesis while controlling other variables, and finally implementing a solution.

[CI???]

2.5.3 Source code management

I used Git for version control and regularly pushed my local changes to a private GitHub repository.

2.5.4 Ethical statement

The development of this project did not require human participants, so nobody was harmed during its implementation.

The software that has been developed contributes to an already existing blockchain ecosystem. Such software has many positive applications, but by its nature can facilitate the creation of exploitative markets. Since this software already exists, our contributions will not enable any new forms of unethical markets and products.

Chapter 3

Implementation

In this chapter we describe the architecture of our implementation (section ??), conclude our theoretical explanation of HotStuff by discussing the chained algorithm and the pacemaker (section 3.2), present a full specification for HotStuff with a proof of correctness and liveness (section 3.3), describe key optimisations implemented (section 3.4), present our load generator and experiment scripts which will be used in evaluation (section 3.5), and give an overview of the repository structure (section 3.6).

3.1 System Architecture

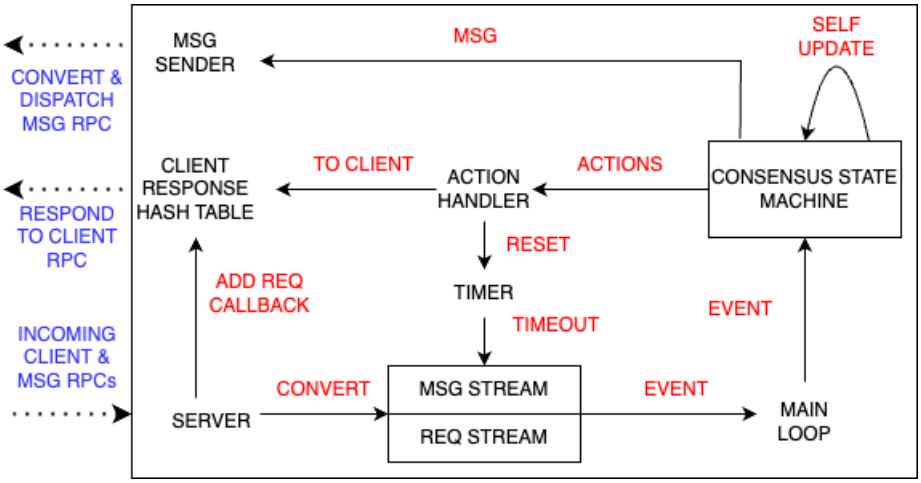


Figure 3.1: Architecture of a node.

In this section we present the system architecture that surrounds the core *consensus* module, passing it new messages and client requests, and allowing it to send messages and respond to the client. This architecture is inspired by the OCons project[1] [21], which was developed by my project supervisor.

We will follow the path of an incoming request or message as it passes through the system shown in figure 3.1.

[1]At the time I began implementation the OCons project was still under development, so I was unable to use the code in my project.

Incoming message and client RPCs — the node responds to internal messages from other nodes, and requests from a client (or a load generator, as described in section 3.5.1) containing new commands to be committed. The format of these RPCs is specified in a Cap'n Proto schema, in their custom markdown language.

Server — each node operates as a server waiting for incoming RPCs. When the server receives an RPC it must be converted from Cap'n Proto types to internal types, and added to the *message stream* or *request stream*. If the RPC was a client request, a promise for a response is added to the *client response hash table*; this will allow the system to respond to the client once the command is decided.

Message and request streams[2] - messages and requests are added to separate streams so that messages can be prioritised. Internal messages represent a backlog of work that the system has not yet completed, so we follow the general design principle of clearing this backlog before accepting new work (client requests).

Main loop — takes messages and requests from their respective streams, and delivers them to the *consensus* module. This main loop ensures that the *consensus* module is never run in parallel, which could lead to race conditions.

Consensus module — takes incoming messages and requests, outputs *actions* such as sending messages, and updates its own state. The module contains an implementation of the basic HotStuff algorithm (section 2.2), and the chained algorithm (chaining is discussed in section 3.2.1, and a full specification is given in section 3.3); both share the same signature, so can be interchanged.

Action handler — takes the actions outputted by the *consensus* module, and passes them to the appropriate handler. The three types of actions are: sending a message, responding to the client, and resetting the view timer.

Message sender — asynchronously dispatches an RPC in a new thread. To do this it must convert internal types into Cap'n Proto types, and construct an RPC that matches the schema. The message sender maintains TCP connections with all other nodes, and in the event of the connection breaking repeatedly attempts to reconnect with binary exponential back-off times.

Client request hashtable — allows client requests to be responded to. The hashtable maps each command's unique identifier to a promise, that when awoken will respond to the original client request RPC.

View timer — waits for a timeout to elapse then adds a view timeout message to the *message stream*, so that it will be delivered to the *consensus* module. The *reset* action allows the timer to be reset for a new view.

[2] A stream is thread-safe implementation of a queue in Lwt.

3.2 More HotStuff theory

In this section we conclude our theoretical explanation of HotStuff by discussing chaining, and the pacemaker. As the pacemaker was not sufficiently specified in the original paper, we will draw on other sources to give a full explanation.

3.2.1 Chaining

In this section we describe the *chained* HotStuff algorithm, which is an optimised version of the basic algorithm described in section 2.2 where different phases are pipelined. This is a standard optimisation for consensus algorithms that is described in the original paper [32].

Pipelining phases both simplifies and optimises the basic algorithm. The phases in the basic algorithm were all very similar; they involved collecting votes from replicas to form a QC that then serves in later phases. Instead of having different phases as before, we can have a single *generic* phase that collects votes, creates a *generic QC*, and sends it to the next leader; now each view is the length of a single phase and a QC can serve in multiple phases concurrently. The only exception to this is the *new-view* phase which is the same as in the basic algorithm.

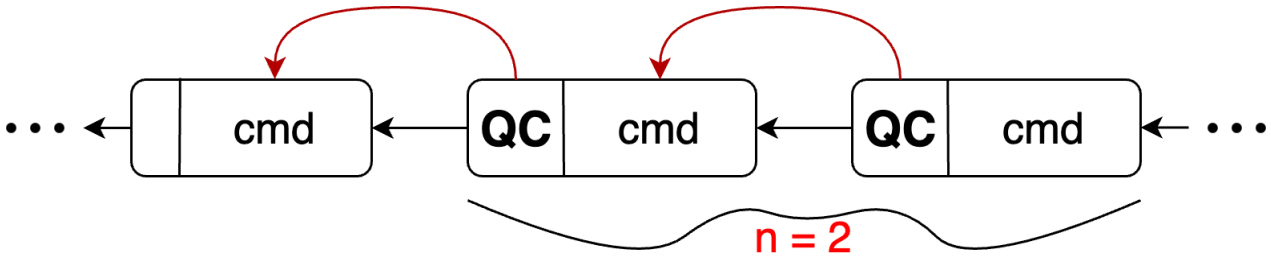


Figure 3.2: Sequence of nodes forming a 2-chain [label parent and justify links!!! fix font].

In each view a chain of nodes is extended, and different phases are carried out on nodes depending on how long a chain they form. For example, figure 3.2, shows a 2-chain, which means a proposal has already been through 2 phases[3]; this is the equivalent of being in the *commit* phase in the basic algorithm.

3.2.2 Pacemaker

In this section we discuss the pacemaker, which is responsible for the liveness of the system; ensuring that it is able to makes progress. The original paper did not sufficiently specify the pacemaker mechanism, so we have synthesised information about pacemakers from different sources (including our own experimentation) to give a full specification of HotStuff in section 3.3. Part of the pacemaker that we will explain in this section is the *view-change* protocol, which allows the view of a faulty leader to be skipped; ours is based on the pacemaker for LibraBFT [11], which is explained in a talk by Ittai Abraham [10]. We will also discuss how

[3]Not counting the new-view phase, as this can be considered to be the end of the previous view.

the pacemaker can be integrated with the HotStuff algorithm; we did this through our own modifications based on deadlocks encountered during implementation.

The notion of a pacemaker and the properties it should posses were formalised in the Cogsworth byzantine view synchronisation protocol [27]; we will later prove that our pacemaker has these properties, and that they lead to liveness (section 3.3.1). Pacemakers are related to the concept of a failure detector: a system that facilitates the detection of failed nodes [16][17]. A pacemaker extends this idea, allowing it to be used in a multi-shot setting.

View-change protocol

Once the view times out, nodes send a *complain* message to the next leader and start a new timeout for the next view. Once the next leader achieves a quorum of *complain* messages it collects them into a QC known as a *view-change proof*. This leader can then send a *view-change* message containing the *view-change proof* to all replicas, who will respond by transitioning to the next view and sending a *new-view* message to the new leader. The inclusion of the *view-change proof* prevents liveness attacks by byzantine nodes that could otherwise attack the system by constantly causing view-changes to take place and preventing non-faulty leaders from making progress.

Integration

Our approach to integrating the pacemaker is to advance a node to the next view as soon as possible once it has finished proposing or voting, or when it receives evidence that there are a quorum of nodes in a higher view. This means that the system does not require synchronised clocks, the nodes advance asynchronously as fast as the network allows.

During development we experimented with a prototype for a pacemaker that advanced views at a steady rate, but analysis of timing data indicated that this approach had inferior performance to our chosen design.

3.3 Specification

In this section we give a full specification of HotStuff based on the basic HotStuff algorithm (section 2.2), integrating chaining (section 3.2.1), and a pacemaker (section 3.2.2).

This specification implements the consensus module as described in our discussion of the system architecture (section 3.1). This module receives messages and requests from the main loop, outputs actions (sending a message, responding to a client request, or resetting the view timer), and updates its own state. We implement this as a *match* statement over the different types of incoming messages and requests, returning a new state for the consensus machine, and a list of actions.

We present the pseudocode with our additions coloured in green and modifications in pink.

We have only shown the main changes and not the other features and performance improvements we have made (such as batching), which are presented in section 3.4. The pseudocode is written to match the format of the original paper to allow for easier comparison, although as stated our actual implementation uses a match statement over the incoming messages and requests.

Algorithm 1 Modified HotStuff

function CREATELEAF(*parent, cmd, qc*)
  $b.parent \leftarrow$ branch extending with dummy nodes from *parent* to height $curView$
  $b.height \leftarrow curView + 1$
  $b.cmd \leftarrow cmd$
  $b.justify \leftarrow qc$
  return b
procedure UPDATE(b^*)
  $b" \leftarrow b^*.justify.node$
  $b' \leftarrow b".justify.node$
  $b \leftarrow b^*.justify.node$
  UPDATEQCHIGH($b^*.justify$)
  if $b'.height > b_{lock}.height$ **then**
    $b_{lock} \leftarrow b'$
  if $(b".parent = b') \wedge (b'.parent = b)$ **then**
    ONCOMMIT(b)
    $b_{exec} \leftarrow b$
procedure ONCOMMIT(b)
  if $b_{exec}.height < b.height$ **then**
    ONCOMMIT($b.parent$)
    EXECUTE($b.cmd$)
procedure ONRECEIVEPROPOSAL(MSG$_v$(GENERIC, b_{new}, qc))
  if $v =$ GETLEADER$(m.view) \wedge m.view = curView$ **then**
    $n \leftarrow b_{new}.justify.node$
    if $b_{new}.height > vheight \wedge (b_{new}$ extends $b_{lock} \vee n.height > b_{lock}.height)$ **then**
      $vheight \leftarrow b_{new}.height$
      SEND(GETLEADER(), VOTEMSG$_u$(GENERIC, b_{new}, $\perp$))
    UPDATE(b_{new})
    if not ISNEXTLEADER() **then**
      ONNEXTSYNCVIEW($curview + 1$)
procedure ONRECEIVEVOTE(VOTEMSG$_v$(GENERICACK, b, $\perp$))
  if ISLEADER$(m.view + 1) \wedge m.view \geq curView$ **then**
    if $\exists(v, \sigma') \in V_{\text{m.view}}[b]$ **then**
      return
    $V[b] \leftarrow V_{\text{m.view}}[b] \cup \{(v, m.partialSig)\}$
    if $|V_{\text{m.view}}[b]| \geq n - f$ **then**
      $qc \leftarrow QC(\{\sigma | (v', \sigma) \in V_{\text{m.view}}[b]\})$
      UPDATEQCHIGH(qc)
      ONNEXTSYNCVIEW($m.view + 1$)
function ONPROPOSE(b_{leaf}, cmd, qc_{high})
  $b_{\text{new}} \leftarrow$ CREATELEAF(b_{leaf}, cmd, qc_{high}, $b_{leaf}.height + 1$)
  BROADCAST(MSG$_v$(GENERIC, b_{new}, qc_{high}))
  return b_{new}

Algorithm 2 Modified Pacemaker

```
1: function GETLEADER
2:     return curView mod nodeCount
3: procedure UPDATEQCHIGH(qc'_high)
4:     if qc'_high.node.height > qc_high then
5:         qc'_high ← qc_high
6:         b_leaf ← qc'_high.node
7: procedure ONBEAT(cmd)
8:     if u = GETLEADER() then
9:         b_leaf ← ONPROPOSE(b_leaf, cmd, qc_high)
10: procedure ONNEXTSYNCVIEW(view)
11:     curView ← view
12:     RESETTIMER(curView)
13:     ONBEAT(cmds.take())
14:     SEND(GETLEADER(), MSG_u(NEWVIEW, ⊥, qc_high))
15: procedure ONRECEIVENEWVIEW(MSG_u(NEWVIEW, ⊥, qc'_high))
16:     UPDATEQCHIGH(qc'_high)
17: procedure ONRECIEVECLIENTREQUEST(REQ(cmd))
18:     cmds.add(cmd)
19: procedure ONTIMEOUT(view)
20:     SEND(GETNEXTLEADER(), MSG(COMPLAIN, ⊥, ⊥))
21:     RESETTIMER(view + 1)
22: procedure ONRECIEVECOMPLAIN(m = MSG(COMPLAIN, ⊥, ⊥))
23:     if ISLEADER(m.view + 1) ∧ m.view ≥ curView then
24:         if ∃(v, σ') ∈ C_m.view[b] then
25:             return
26:         C_m.view[b] ← C[b] ∪ {(v, m.partialSig)}
27:         if |C_m.view[b]| = n − f then
28:             qc ← QC({σ|(v', σ) ∈ C_m.view[b]})
29:             BROADCAST(MSG(NEXTVIEW, ⊥, qc))
30: procedure ONRECEIVEANY(m = MSG(∗, ∗, qc))
31:     if qc.view ≥ curView then
32:         ONNEXTSYNCVIEW(qc.view + 1)
```

3.3.1 Proofs

HotStuff works under a partially-synchronous system model (section 2.2), so the following proofs assume that global synchronisation time has been reached, and messages have a bounded latency of δ. [For this proof we consider the consensus machine (algorithm 1) and the pacemaker (algorithm 2), to be separate entities.] - relate this to cogsworth synchroniser!!!

Theorem 3.3.1 (View synchronisation)**.** *There exists infinite views with honest leaders that all honest replicas will be in simultaneously, and have enough time to make progress.*

Proof. From lemma 3.3.2 we have that we can always find future views v_1 and v_2 with honest leaders l_1 and l_2, and from lemma 3.3.3 we have that l_1 will eventually enter v_1. We argue that all honest replicas will simultaneously be in either v_1 or v_2. Consider the cases of how l_1 could

have entered v_1:

1. l_1 received a quorum of votes (algorithm 1, line 38) — l_1 will broadcast a QC of votes that will be received by all honest replicas within δ. These replicas will transition to v_2, and send a vote to l_2 which will also transition once it receives a quorum of votes. Hence all honest replicas will simultaneously be in v_2.
2. l_1 receives QC of COMPLAINS from itself (algorithm 1, line 32) — l_1 must have broadcast the NEXTVIEW message to all honest replicas; they will receive it within δ and all enter v_1 simultaneously.

Once all honest replicas are in a view with an honest leader, they will only transition once they have made progress, or once they timeout, which shouldn't happen if the timeout is sufficiently long. Hence they will all be in the view long enough to make progress. □

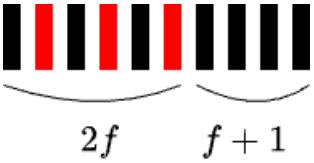


Figure 3.3: Example of round-robin leader allocation for $f = 1$. Red rectangles denote byzantine leaders. [maybe stripe red nodes so they appear in black and white???]

Lemma 3.3.2. *There exists an infinite number of consecutive assignments of two honest leaders to views. That is, we can always find future consecutive views v_1 and v_2 with honest leaders.*

Proof. We use a round-robin system to allocate leaders to views. If we attempt to alternate honest and byzantine leaders, there will always be $f + 1$ consecutive honest leaders left over (figure 3.3). Hence there will always be at least 2 consecutive honest leaders (the lemma holds trivially for $f = 0$). □

Lemma 3.3.3. *An honest replica x will eventually enter a view where it is leader.*

Proof. In the event that a byzantine node is leader and tries to prevent honest nodes from transitioning to a higher view, the honest nodes will eventually timeout and send a COMPLAIN message to the next leader (algorithm 2, line 20). This may repeat if the next leader is also byzantine. Eventually the COMPLAINs will be sent to an honest leader, that will send a NEXTVIEW message and transition all replicas into a new view (algorithm 2, line 29). Since x will always progress to a higher view, it will eventually reach a view where it is the leader [quickly justify that we cannot 'skip ahead' without a qc from a future view]. □

Theorem 3.3.4 (Synchronisation validity)**.** *The pacemaker will only advance the view if at least one honest consensus machine requests it to be advanced.*

Proof. This holds trivially for the calls to ONNEXTSYNCVIEW in algorithm 1, as the view is advanced on the request of the consensus machine.

The only other way the view can be advanced, is on the receipt of a QC (algorithm 2, line 32). For a QC to be formed a quorum of $n - f$ nodes must have complained or voted; at least one of these must have been an honest consensus machine that requested for the view to be advanced. □

3.4 Performance improvements

In this section we highlight the key performance optimisations we implemented, and describe some of the debugging and experimentation that led to them. [this is a hard problem, reference chubby again. bugs are subtle.] [15]

3.4.1 Batching

[refer to literature on batching, highlight which (practical) contributions were ours] One way to improve goodput (number of requests committed) is to *batch* requests, meaning a node may contain many commands instead of just one. This can dramatically increase goodput as now a single view can result in many commands being committed and executed instead of just one.

[reword paragraph!!!] In order to implement this change in algorithm 2 one simply has to modify line 13. Instead of taking a single element from the queue of commands waiting to be proposed, the whole queue will be 'batched' into a single proposal.

In theory this change should result in much higher goodput without any significant increase in latency. Analysis of timing data after implementing this feature revealed that latency had increased substantially. We now present some of the analysis and experimentation that was carried out to diagnose this issue.

Filtering

Further problems were uncovered by the implementation of send-to-all in the load generator (section 3.5).

Instead of reducing latency as expected the implementation of this feature actually led to dramatically increased latency, and some requests not being fulfilled at all within the time that the experiment ran. Heatmaps that show the distribution of latencies over the course of an experiment showed that latency increased exponentially over the course of the test.

Hypothesis: Send-to-all overloads the system by increasing the number of requests by a factor of n, where n is the number of nodes.

Experimentation: If our hypothesis is true, then we should see similar behaviour in a send-to-one system run at n times the throughput. This is indeed the case; running a send-to-one system with such high throughput led to exponential growth in latencies.

Further evidence: Implementing this feature resulted in Lwt errors being thrown. As described in section **??**, in order to respond to client requests, the node maintains a hash table of promises that can be awakened when the consensus state machine has successfully committed and can respond to the client. The Lwt error was caused by promises stored in the hash table being awoken multiple times, which causes an error. This provides further evidence that the same commands are being committed and responded to multiple times, which is wasted work.

Potential solution: Filter commands in the system to minimise the number of commands that are proposed by multiple nodes.

Design principle: Attempt to minimise the amount of redundant work that the system carries out by screening incoming work to check that it actually needs to be done.

[shorten this section, only give the final pseudocode and not the failed attempt] One way to implement this is to maintain a set of commands that have been committed, and before proposing or committing a node, calculating the set difference with the *commited* set to filter out redundant commands. Instead of nodes containing a list of commands, as in our naive implementation of batching, we now use a set to enable efficient computation of the difference.

Algorithm 3 Filtering implementation #1

```
1:  procedure ONCOMMIT(b)
2:      if b_exec.height < b.height then
3:          toCommit ← b.cmds \ commited
4:          commited ← commited ∪ toCommit
5:          ONCOMMIT(b.parent)
6:          EXECUTE(toCommit)
7:  procedure ONNEXTSYNCVIEW(view)
8:      curView ← view
9:      ONBEAT(cmds \ commited)
10:     cmds = {}
11:     // ...
12: procedure ONRECIEVECLIENTREQUEST(REQ(cmd))
13:     cmds ← cmds ∪ {cmd}
```

This implementation did not lead to the desired reduction in latency, the system still exhibited the exponential growth previously observed. The lack of change in observable behaviour implied that the filtering of requests was not happening successfully.

Hypothesis: A command (x) is committed four views after it is proposed. In these four views our deduplication is ineffective in screening x from being proposed again. Notably another node will not try to commit x for a second time, but the data indicates that being proposed multiple times is a more important factor affecting performance.

Experimentation: Count the elements that are filtered out from being proposed by measuring the value $|cmds| - |toPropose|$. This value was small, which supports the hypothesis that filtering was largely ineffective.

Potential solution: Screen incoming commands more aggressively by instead maintaining a set of any commands that have been *seen*. Only filter out commands when they are being proposed rather than when they are being committed, as there is no evidence that screening at

commit time gives any tangible benefit.

Algorithm 4 Filtering implementation #2

```
1: procedure ONRECEIVEPROPOSAL(MSG_v(GENERIC, b_new, qc))
2:     if v = GETLEADER(m.view) ∧ m.view = curView then
3:         seen ← seen ∪ b_new.cmds
4:         // ...
5: procedure ONNEXTSYNCVIEW(view)
6:     curView ← view
7:     ONBEAT(cmds \ seen)
8:     cmds = seen = {}
9:     // ...
10: procedure ONRECIEVECLIENTREQUEST(REQ(cmd))
11:     cmds ← cmds ∪ {cmd}
```

Note the small optimisation on line 8, the *seen* can safely be emptied as the commands have already been filtered; this reduces the amount of computation required to calculate the set difference next time. The second implementation was much more effective at screening requests and resulted in a significant reduction in latency. The difference is demonstrated in our ablation study (section 4.3.3). The effectiveness of this change led to further insight into what the bottlenecks for performance were.

Batch sizes

The effectiveness of filtering out commands to make proposals smaller implies that the size of a proposal has a significant impact on the performance of the program.

Hypothesis: Large batches of commands cause messages to become larger, which causes them to be sent slowly due to limitations in the RPC framework [serialisation!!!].

Potential mitigation: Limit the size of batches to prevent messages becoming too large.

Implementing this feature required minimal changes to algorithm 4, one simply has to take a subset of *cmds* to propose instead of the whole set. This feature gave a significant performance improvement, and largely eliminated the exponential growth in latency that had been seen previously. There is a trade-off between sending larger batches (more commands are committed, but messages take longer to send) and sending smaller batches (less commands are committed, but messages are sent faster). This trade-off is explored more in our analysis of the system with different batch sizes (section 4.3.1). Fundamentally, there are limitations in the RPC framework that give an upper bound on the performance we can hope to achieve, these limitations are evident in our benchmarking of the Cap'n Proto framework (section 4.2.1).

3.4.2 Nodes

[not strictly true, more about designing the types. make this section more about nodes, describe what they actually are!] *This section concerns the difficulties of encoding nodes in the Cap'n*

Proto schema, and designing a suitable type to represent them. As mentioned in section **??**, *Cap'n Proto requires a schema to define the format of RPCs that can be sent, and one must convert between consensus state machine internal types and Cap'n Proto types in order to send messages.*

Recursive types

The chained HotStuff algorithm has a genesis node b_0 that starts the chain of nodes. Each node contains the fields *parent*, *cmd*, *justify*, and *height*, with the *justify* field containing a quorum certificate with a 'pointer' to another node in the chain (section 3.2.1). The genesis node b_0 contains a hardcoded link to itself, so $b_0.justify.node = b_0$.

This recursion poses a problem when carrying out the conversion between Cap'n Proto types and OCaml types. It is perfectly possible to define a recursive type in OCaml, so one can represent b_0 inside the consensus state machine. However, the naive implementation of a function to convert this node into a Cap'n Proto type will not terminate, as it will infinitely recurse into the field $b_0.justify.node$. A simple solution to this problem is to add a flag to the Cap'n Proto schema *is_ b_0*; when this flag is enabled then the node is assumed to be equal to b_0. This prevents b_0 from ever having to be converted into a Cap'n Proto type or being sent over the network, it can instead be reconstructed as a recursive type in the consensus state machine of the receiver.

Node offsets

During live testing, memory usage increased rapidly, the nodes quickly exceeded the amount of memory the system had, and their processes were killed by the OS. Profiling with Memtrace (section 2.5.2) [maybe run this again, give a diagram] showed that the functions for converting between Cap'n Proto types and internal consensus types was responsible for the rapidly increasing memory usage.

This problem arose from the internal types that were designed to represent nodes. As well as the *node* type, there was a *qc* type that was used as the type of the *node.justify* field since it is a quorum certificate [use OCaml lingo, recursive sum type???]. By defining types in this way the 'pointers' in *node.justify.node* were not actually pointers, they contained a significant part of the chain. In this way the chain contained many redundant copies of parts of it, resulting in a very bloated *node* object that was very expensive to convert.

A solution to this problem is to store an offset to a node inside the *justify* field rather than the node itself - making it more like a real pointer. This offset represents how many *parent* links away the node is, and so can be used to reconstruct all of the original information. To implement this another type *node_justify* was added, that is identical to *qc*, but with the field *node* replaced with *node_offset*. One must then convert between the *node_justify* and *qc* types to reconstruct the original data and follow the *node.justify.node* link.

Node comparison

[cite merkle tree here? compare to ethereum and bitcoin...] Memtrace profiling also showed that the equality function for nodes was memory intensive. The solution to this was including a *digest* field in the node that is a hash over all of the other fields. Notably we can compute this hash over the *digest* field of the parent node rather than recursing through the whole chain. This means that we can also compare two nodes by comparing their digests without having to recurse through each chain; the digests being equal cryptographically guarantees that the whole chains are equal.

Node truncation

Our current approach sends the entire node inside internal messages, which becomes expensive as the chain grows. As highlighted in section 3.4.1, the size of messages being sent is a bottleneck in the system, making this particularly expensive.

In order to overcome this problem, one can truncate the node before sending it, cutting off the node's parent link at some chosen depth. The entire node can then be reconstructed at the receiver, the received node can be 'spliced' back together with b_{exec}, which contains the node up to the point that has been executed. However, we must ensure not to truncate the log too much, so that there is a gap between what we send and b_{exec} at the receiver, leading to commands being missed out and not executed. This is a problem in the event that some node becomes isolated from the rest; it must be able to catch up to the others once the network partition is healed.

To overcome this problem we use a TCP-style approach. We include a field containing the height of the b_{exec} node to the *PROPOSE*, *NEWVIEW*, and *COMPLAIN* messages. Each node maintains a list of the b_{exec} height of every other node. When making a proposal, the leader takes the minimum height from this list, and truncates the node up to that depth. This ensures that every node that receives the proposal has enough information to reconstruct the entire log.

There are some cases when the leader does not receive the latest b_{exec} of every other node before it makes a proposal. This means that the leader will not truncate the node as much as it could have. We optimised this by having a node send the entire list of all stored b_{exec} heights rather than just its own, allowing the heights to propagate around the system more quickly.

[give pseudocode!!!]

3.5 Implementing for evaluation

In this section we describe the infrastructure that will be used to evaluate our system in chapter 4, including the scripting developed to automate running experiments.

3.5.1 Load generator

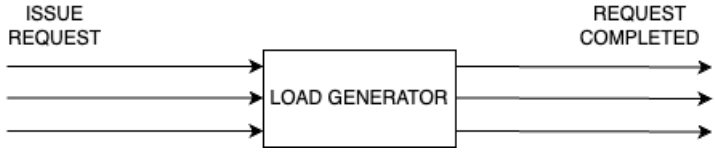


Figure 3.4: Open-loop load generator.

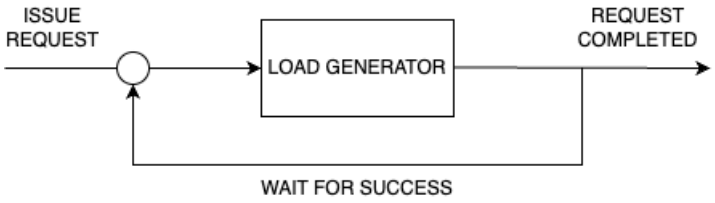


Figure 3.5: Closed-loop load generator.

The load generator is responsible for sending client requests to the nodes of the system. One is able to vary the throughput that the load generator drives the system at, and the duration that it runs for before it sends a *kill* messages to the nodes, ending the test. It is also responsible for timing and calculating statistics.

Throughput: The number of requests sent by the load generator each second.

Goodput: The number of requests that are responded to each second. This is calculated by number of responses divided by the time difference between the first response and the end of the test.

Latency: The amount of time it takes between sending a request and receiving a response. The load generator reports both mean and standard deviation of latencies.

The load generator is open-loop (figure 3.4), which means that it dispatches a request every δ seconds for the duration of the experiment, where $\delta = \frac{1}{throughput}$. This is in contrast to a closed-loop generator (figure 3.5), which must wait until it receives a response in order to send the next request. An open-loop load generator is more useful for our experiments as it allows us to overload the system and test its limits, whereas a closed-loop load generator 'waits' for the system, so cannot overload it.

SEND TO ALL [Without this feature a client would randomly chose a node to send their request to, and the command would not be proposed until the round-robin system reached that node to become the leader. Send-to-all can reduce latency by instead broadcasting a request to all nodes, meaning that the next leader can propose the command without having to wait for the round-robin system to reach them.]

The load generator uses Lwt to asynchronously dispatch requests, and stores a promise that will be fulfilled with their response. In the case of send-to-all (section **??**) the promises waiting on a response from each node are combined using *Lwt.pick*, meaning that the first node to respond will fulfil the promise and the rest will be ignored. Before beginning the experiment the load-generator sends 'dummy' requests to each node until all of them have sent a response; this ensures that all nodes are properly up and running before we start the experiment, reducing start-up effects.

3.5.2 Experiment scripts

Python scripts are used to automate the running of experiments. These scripts start the nodes and the load generator, wait for the experiment to run, kill the processes, run a script to plot graphs, then start the next experiment.

Different experiments may vary input variables such as throughput, batch size, and number of nodes. The script takes all possible permutations of the input variables, duplicates them (so that each experiment is run multiple times to get better results), and runs them in a random order. By running experiments in a random order external factors that affect performance are somewhat mitigated. For example if the experiments were not run in a random order, the system may happen to experience interference when running experiments on 8 nodes, giving distorted results for this set of experiments. If instead the experiments were run randomly, the interference would affect some random group of experiments, and it would be more apparent that these results were anomalous. [could be more concise]

3.5.3 Logging framework

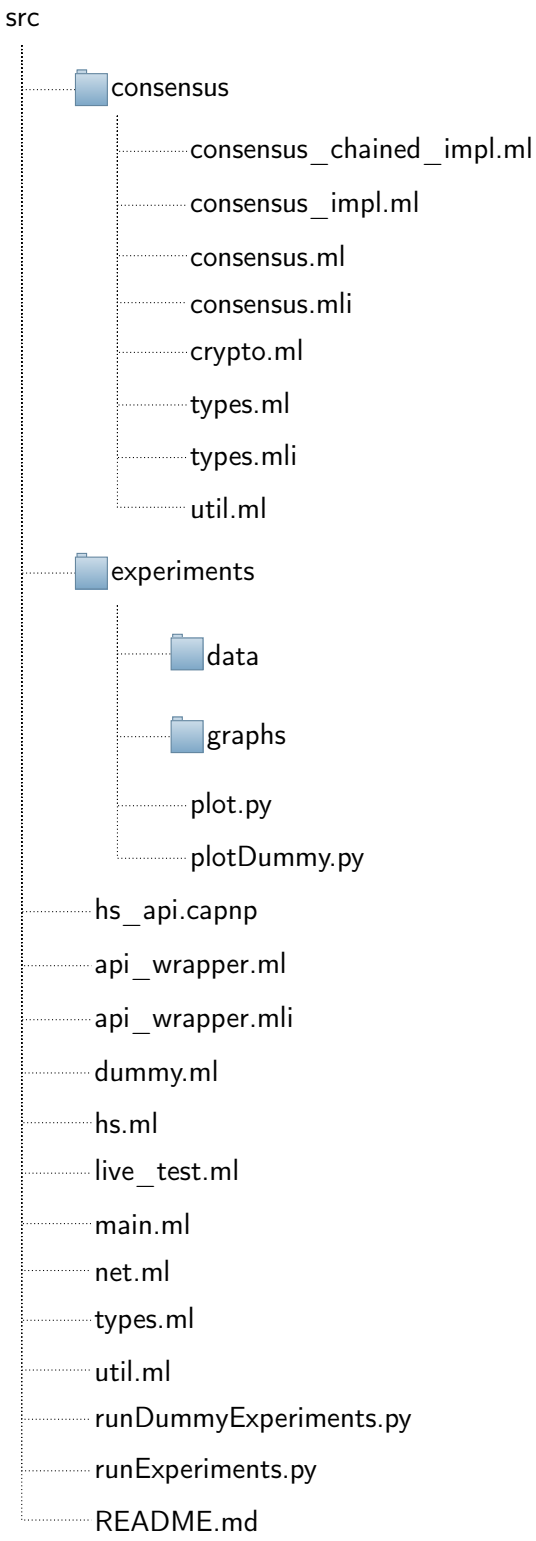
[make this section much shorter!] I added timing and logging statements to key parts of the program, allowing me to better diagnose the source of the poor performance. Running a live test with print statements has the advantage that one can quickly see when progress is not being made, or when there are pauses, as the print statements stop. However, this benefit is outweighed by the sheer volume of logs and times being printed, it becomes difficult to manage when logs are in the millions of lines long.

Moreover, debugging by print statements had a bigger problem of inconsistencies between runs making it difficult to diagnose any problem. This was because the large volume of print statements being executed had a significant effect on the performance [CPU cost!!] of the program. This is known as a *probing effect*, the behaviour of a system is altered by the act of measuring it. I initially assumed that because the print statements were not in-between the timing statements they would not affect my measurements, but they are an expensive operation that can cause delays to happen in execution where one would not expect.

This problem was overcome through the development of a simple logging framework. Key parts of the program such as the state machine advancing, actions being carried out, messages being sent, are all timed and added to lists. Critically, this list is stored and not printed out until the node is killed, so the printing cannot interfere with execution. Relevant statistics such as mean, standard deviation, and total time for each interval measured are outputted, providing crucial insight into the performance of the program.

Design principle: Minimise probing effects by carrying out the minimum possible amount of work in critical areas of the program, storing data, and moving work (such as outputting statistics) to less critical areas of the program.

3.6 Repository Overview



The *src* directory contains the main server loop and the code for interacting with Cap'n Proto to send messages. It also contains code for the load generator, and Python scripts to run experiments and benchmarks. Inside the *consensus* folder is the implementation of the consensus library based on the pseudocode presented in section 3.3. N.B. this folder also contains testing files that are omitted for conciseness. The *experiments* folder is where the data from running experiments is outputted to, and it contains scripts for plotting graphs.

In order to run an experiment, first follow the instructions in *README.md* to set up your environment. You can then run experiments by executing *python3 runExperiments.py*, and modify this script to vary the input parameters such as throughput and experiment time. These scripts will run on Linux and MacOS, but will have to be modified to work on Windows.

Chapter 4

Evaluation

In this section we highlight the methods and hardware used in evaluation (section 4.1), benchmark the performance of Cap'n Proto and the Tezos cryptography library (section 4.2), and finally evaluate the performance of our HotStuff implementation (section 4.3)

4.1 Testing methodology

Evaluation was carried out on the computer laboratory's Sofia server (2x Xeon Gold 6230R chips, 768GB RAM). Carrying out experiments on the server should help to minimise interference from other processes on the system.

Experiments were driven by an open-loop load generator (section 3.5.1), and were automated using Python scripts (section 3.5.2). In order to reduce the effect of interference, experiments were repeated 3 times, and the order of experiments was randomly permuted. In all experiments the load generator was run for 10 seconds, with a further 15 seconds after this without the load generator running to wait for any slow responses.

4.2 Library benchmarks

4.2.1 Cap'n Proto

[describe the experiment methodology, present argument, then back up with evidence from figures] We benchmarked the latency and goodput of sending messages in Cap'n Proto. We varied the size of messages sent in different tests to replicate the behaviour of the algorithm when sending 'batches' of many commands, so a message size of 600 means that the message size is approximately that of a message containing 600 commands.

The figures demonstrate that the framework has a severe drop in performance when sending large messages. For a message size of 600 the goodput goes to zero as the throughput increases, meaning that no messages are being responded to.

["Fundamentally, there are limitations in the RPC framework that give an upper bound on

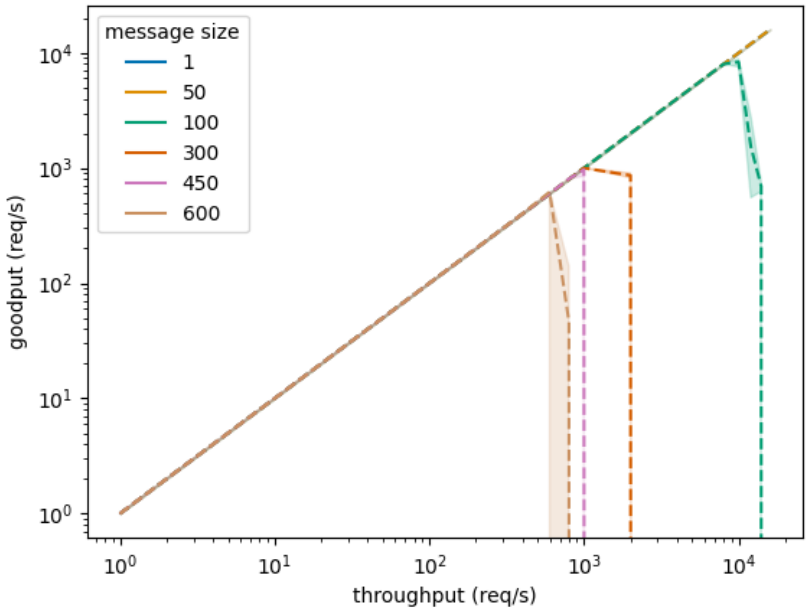


Figure 4.1: Benchmarking of Cap'n Proto server goodput for varying throughputs and message sizes.

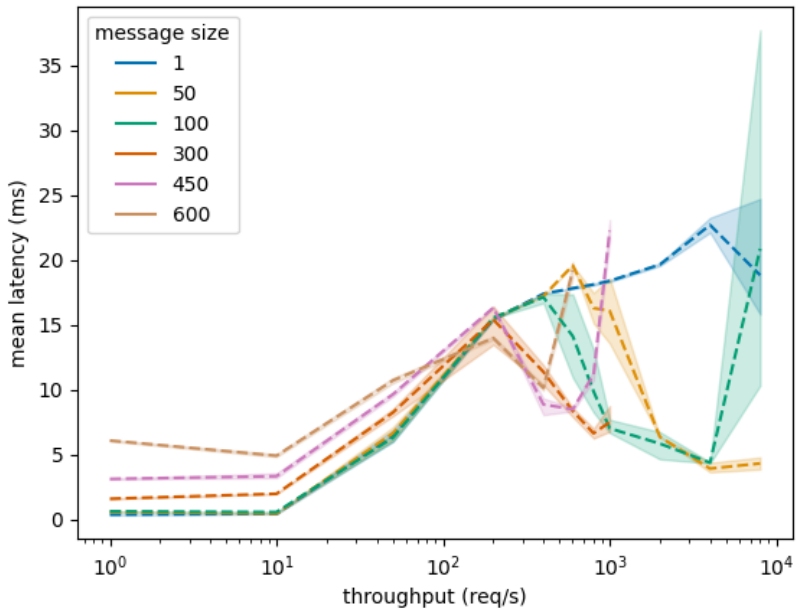


Figure 4.2: Benchmarking of Cap'n Proto server latencies for varying throughputs and message sizes. Discarded result if goodput was not within 5% of target throughput.

the performance we can hope to achieve, these limitations are evident in our benchmarking of the Cap'n Proto framework. . . "]

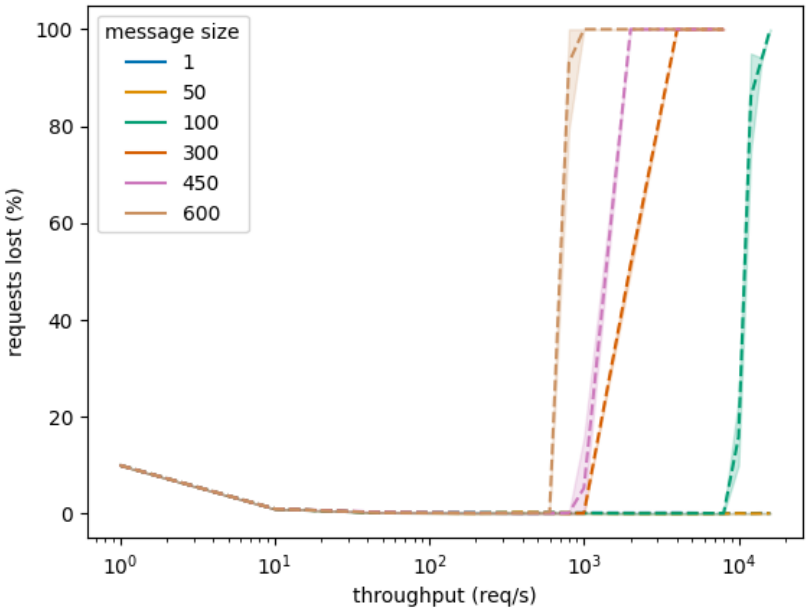


Figure 4.3: Benchmarking of Cap'n Proto server % of failed requests for varying throughputs and message sizes, run for 10s

4.2.2 Tezos Cryptography

I profiled the important functions of the library with Jane Street's Core_bench module [citation needed***]. Core_bench is a micro-benchmarking library used to estimate the cost of operations in OCaml, it runs the operation many times and uses linear regression to try to reduce the effect of high variance between runs.

Function	Time (µs)
Sign	427.87
Check	1,171.77
Aggregate (4 sigs)	302.90
Aggregate check (4 sigs)	1,179.25
Aggregate (8 sigs)	605.38
Aggregate check (8 sigs)	1,180.61

Table 4.1: Benchmarking of key functions of the Tezos Cryptography library

4.3 HotStuff implementation benchmarks

We now analyse the performance and behaviour of the system with different parameters, and under different conditions. We argue that the optimisations described in section 3.4 were effective in improving system performance, but there are fundamental limitations caused by the latency costs of Cap'n Proto serialisation (section 4.2.1) and cryptography (section 4.2.2).

In most cases, the system exhibits stable latency throughout an experiment while goodput is equal to throughput, meaning that the system is not overloaded (figure 4.4). When the

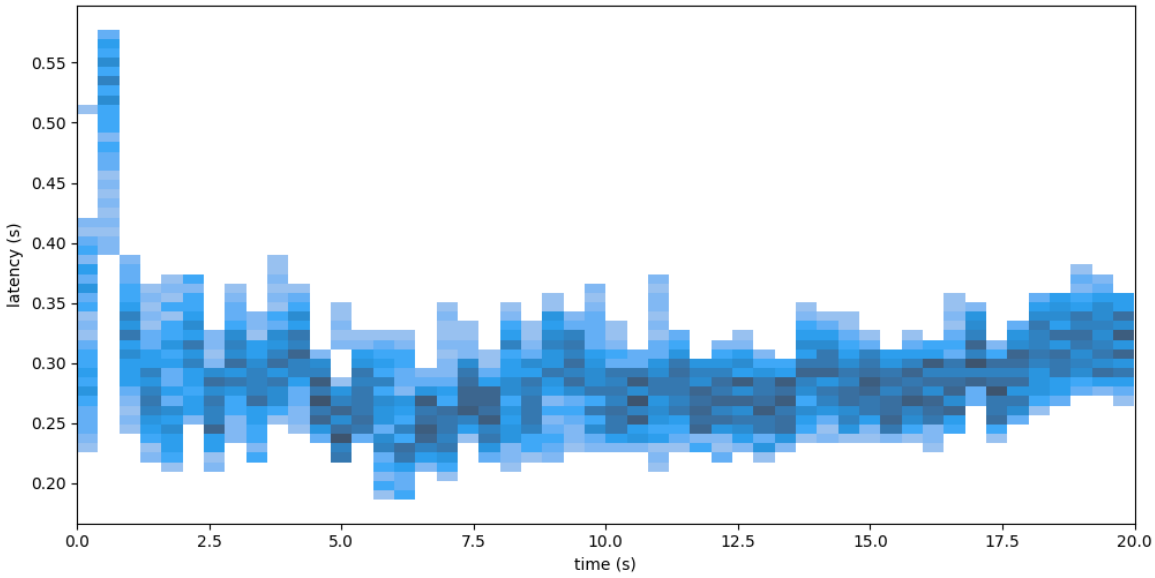


Figure 4.4: Heatmap showing relatively stable latency throughout the course of the experiment. [add parameters!!]

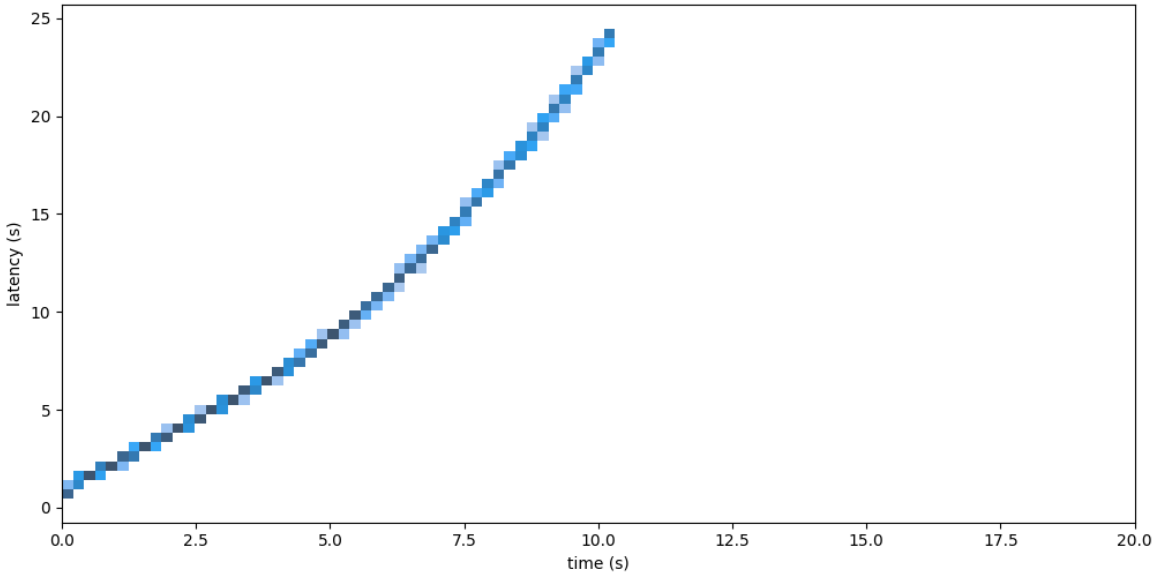


Figure 4.5: Heatmap showing linear growth in latency as the experiment progresses. [add parameters]

throughput exceeds the amount the system can keep up with, there is linear growth in latency as commands queue on the nodes (figure 4.5). Since HotStuff is a partially synchronous protocol (section 2.2), an increase in latency means that view times increase, decreasing goodput. Once the system is overloaded, the goodput levels off at around its maximum value as throughput is increased.

In our comparison of batch sizes (section 4.3.1) there is evidence that the implementation of batching (section 3.4.1) was effective, as the system is able to achieve much greater goodput with batch sizes greater than 1 (equivalent to no batching). This section also provides evidence that serialisation latency is a bottleneck, as view times begin to increase exponentially as batch sizes increase, due to messages being larger and taking longer to serialise.

Our study of node counts (section 4.3.2) gives further evidence that message serialisation is a bottleneck; higher node counts mean more internal messages being sent, causing a decline in performance due to serialisation costs. This also supports the conclusion that cryptography is a bottleneck, as more nodes means more messages must be signed and aggregated.

In our ablation study (section 4.3.3) we compare the performance of the system with different optimisations enabled, demonstrating their effectiveness in increasing goodput, and lowering latency. We also demonstrate that cryptography is a bottleneck by demonstrating the superior performance of the system with cryptography disabled.

In our wide area network (WAN) simulation study (section 4.3.4), we compare the performance of our system running locally, to a simulated mininet network (section 2.5.2) with link latency similar to what one may observe in a wide area network. [we found...]

In our view-change study (section 3.2.2) we demonstrate that the view-change protocol (section 3.2.2) is effective in ensuring the system progresses once a node has died, albeit with a significant performance penalty.

4.3.1 Batch sizes

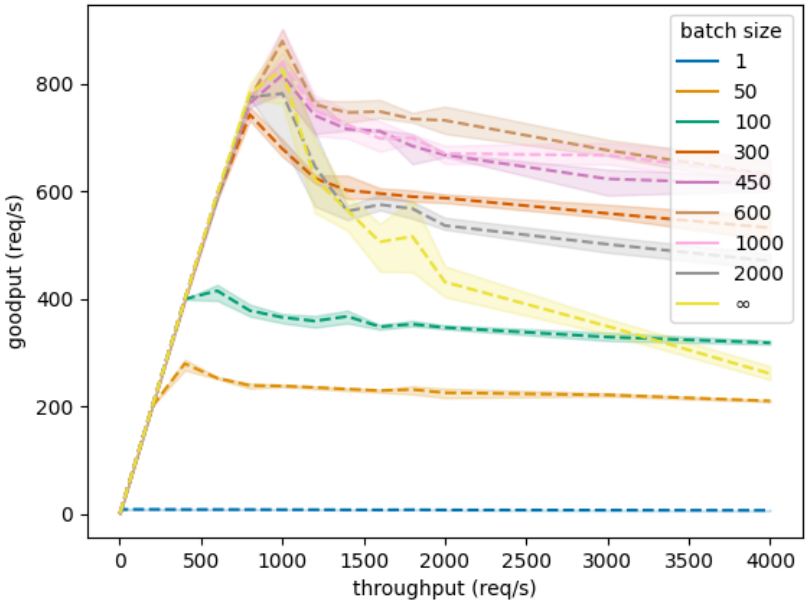


Figure 4.6: Benchmarking of goodput for varying throughputs and batch sizes.

This study compares the performance of the system for varying limits on batch sizes (section 3.4.1). All experiments were run on a network of 4 nodes.

At lower throughputs where the system is not overloaded, throughput grows linearly with goodput (figure 4.6), as the system is able to respond to all incoming requests, with roughly constant latency throughout an experiment (figure 4.4). During this period batches are not filled, so larger throughputs result in larger messages and a linear increase in latency due to increasing serialisation latency (figure 4.7). The system is able to reach a higher goodput before

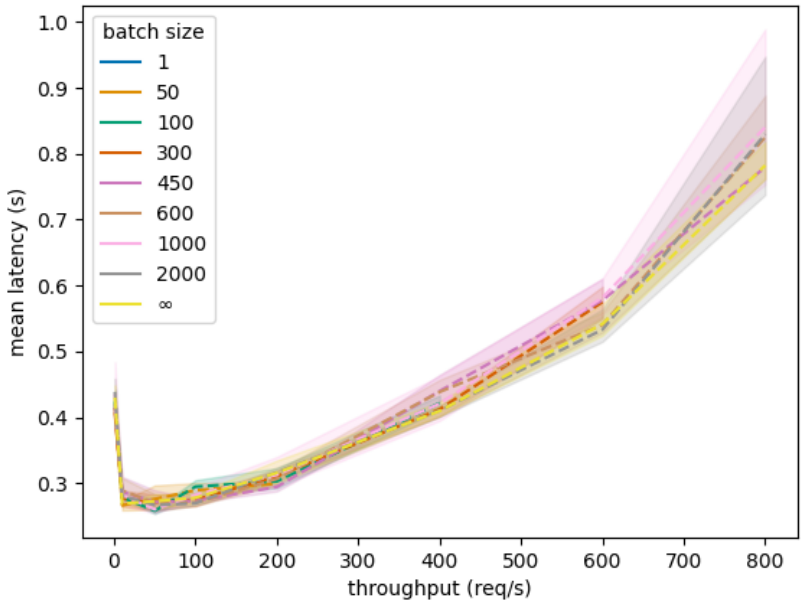


Figure 4.7: Benchmarking of mean latency while varying throughputs and batch sizes. Discarded result if goodput was not within 5% of target throughput.

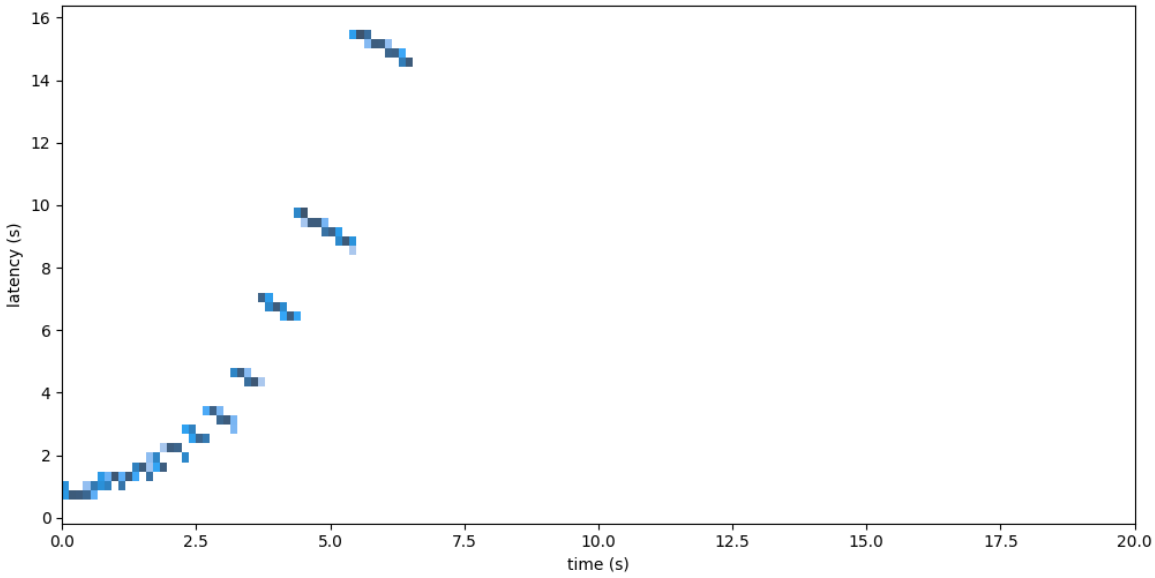


Figure 4.8: Heatmap showing exponential growth in latency as the experiment progresses. [add parameters]

being overloaded if it has a larger batch size, as each view results in more commands being committed; this supports our conclusion that batching is an effective optimisation.

Once throughput is increased enough, batches begin to be filled up and the system is overloaded. This results in the goodput flattening out (figure 4.6), as the system cannot handle the volume of requests; commands begin to queue on the nodes and latency increases linearly (figure 4.5). For higher batch size limits (especially unlimited), goodput declines once the system is overloaded rather than flattening off. This is because the benefits of larger batches are offset by messages becoming larger, causing increased serialisation latency, which increase

view times and lower goodput. For large batch sizes, view times increase exponentially, as shown by the growing vertical gaps between commands being committed in figure 4.8.

There is a clear trade-off between larger batch sizes that result in more commands being committed, and batches becoming too large and incurring exponential serialisation latency. The optimum for our system appears to be a batch size of around 600 commands, with a maximum goodput of around 900req/s. (figure 4.6).

4.3.2 Node counts

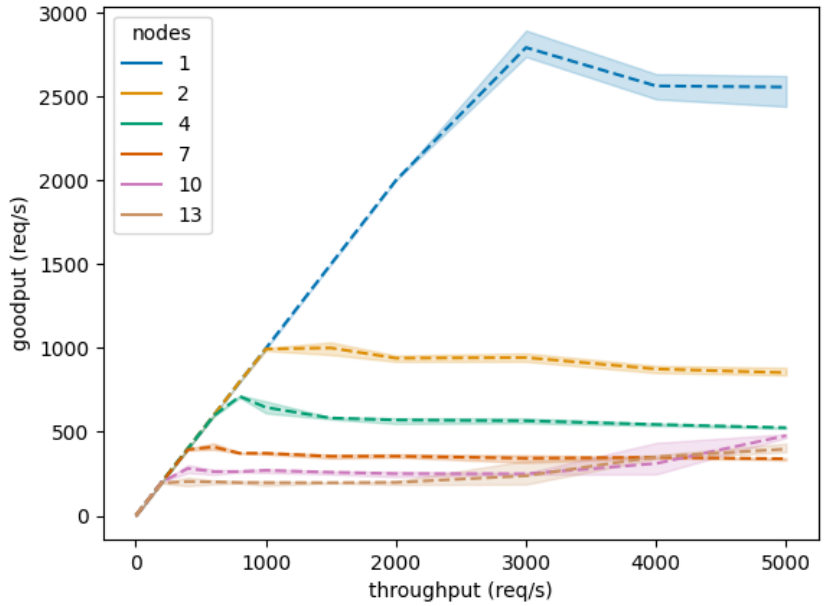


Figure 4.9: Benchmarking of goodput for varying throughputs and node counts.

This study compares the performance of the system for varying node counts. Node counts were chosen such that $n = 3f + 1$ for some f, as choosing another value would decrease performance without any benefit of increased fault-tolerance[1]. All experiments were run with a batch size of 300.

Figure 4.10 shows that as node count increases, latency increases. This is because larger node counts mean that each view requires more internal messages to be sent to progress. Sending internal messages is expensive due to the latency of serialisation and cryptography, so this results in increased overall latency. Additionally, increasing the node count increases the number of messages that must be signed, and makes aggregating signatures slower (section 4.2.2). As in our study of batch sizes (section 4.3.1), latency also increases linearly with throughput while the system is not overloaded. Notably the latency for a system of 1 node increases slowly, as there are no internal messages, just client requests and responses.

Figure 4.9 shows that the larger the node count, the lower the maximum goodput. This is again due to larger node counts resulting in more internal messages, causing more latency since

[1]A node count of 2 was also tested as it is the smallest node count where internal messages are exchanged.

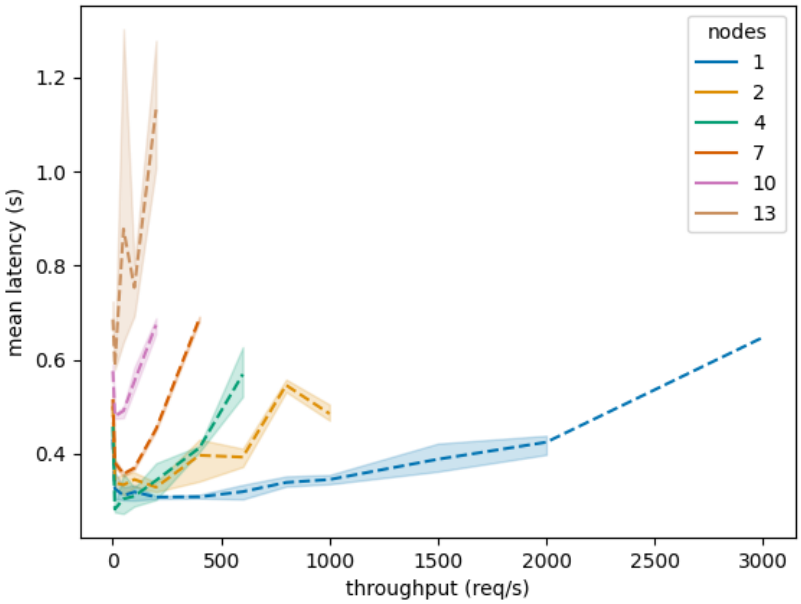


Figure 4.10: Benchmarking of mean latency while varying throughputs and node counts. Discarded result if goodput was not within 5% of target throughput.

this is a bottleneck. Increased latency causes each view to take longer, reducing the number of requests that can be responded to each second.

4.3.3 Ablation study

Version	Chaining	Truncation	Filtering	Crypto
1	✗	✗	✗	✓
2	✓	✗	✗	✓
3	✓	✗	✓	✓
4	✓	✓	✗	✓
5	✓	✓	✓	✓
6	✓	✓	✓	✗

Table 4.2: Features enabled in different versions.

This study compares the performance of different versions of the system with different optimisations enabled. The optimisations explored are chaining (section 3.2.1), node truncation (section 3.4.2), and command filtering (section 3.4.1). We also compare performance with, and without cryptography enabled. The mapping from version numbers to which features are enabled is given in table 4.2. All experiments were run with a network of 4 nodes, and unlimited batch sizes.

4.3.4 Wide area network simulation

[Give ablation graphs comparing the performance with 100ms delay, and without.]

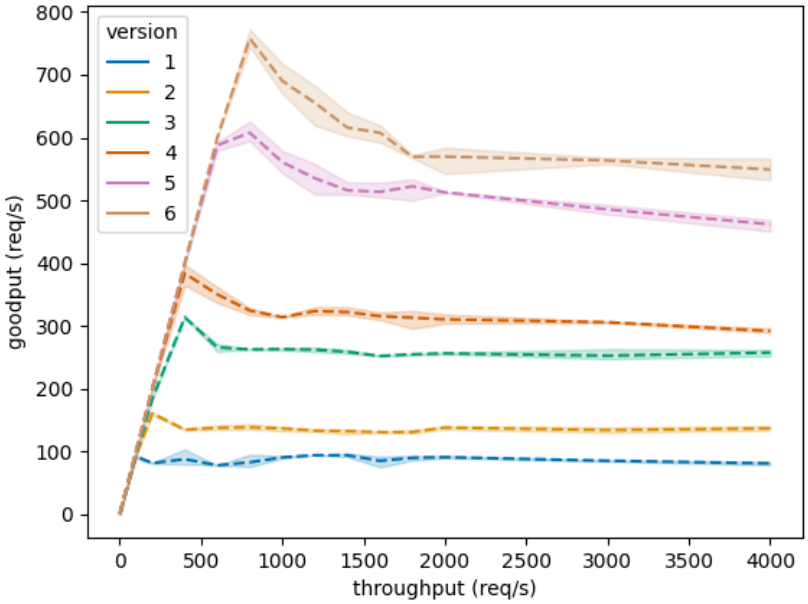


Figure 4.11: Benchmarking of goodput for varying throughputs and implementation versions, run for 10s with 4 nodes unlimited batch size.

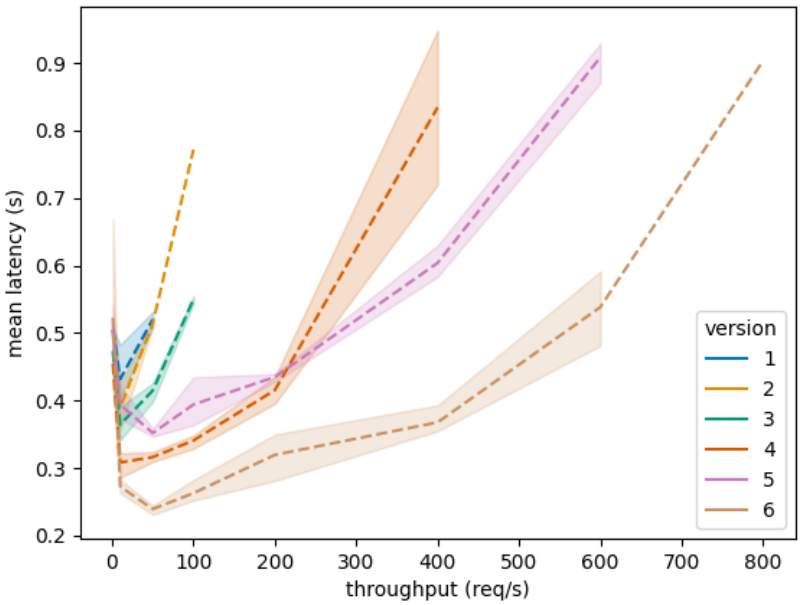


Figure 4.12: Benchmarking of mean latency while varying throughputs and implementation versions, run for 10s with 4 nodes unlimited batch size. Discarded result if goodput was not within 5% of target throughput.

4.3.5 View-changes

This study explores the behaviour of the system in the event of a node failing, where the view-change protocol (section 3.2.2) is needed to skip the faulty leader's view and make progress.

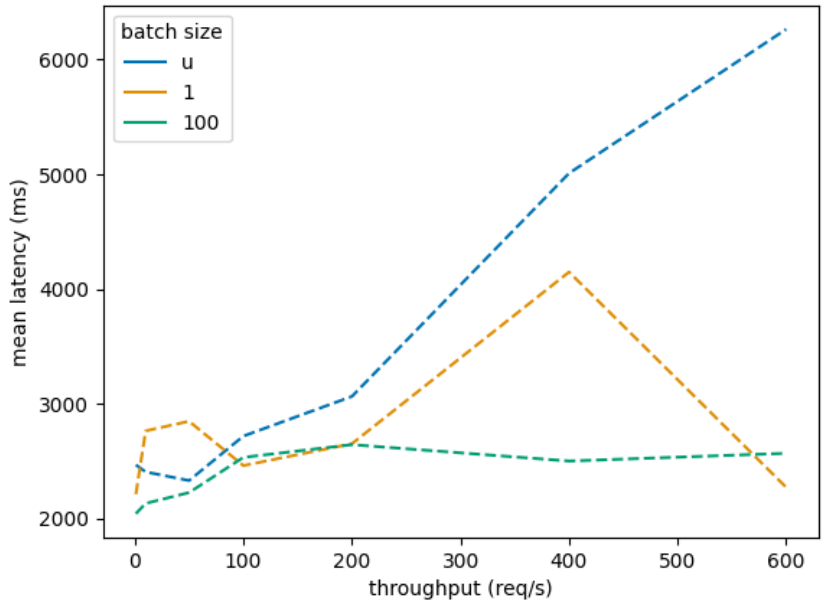


Figure 4.13: Benchmarking of mean latency while varying throughputs and batch sizes, run for 10s with 100ms network delay.

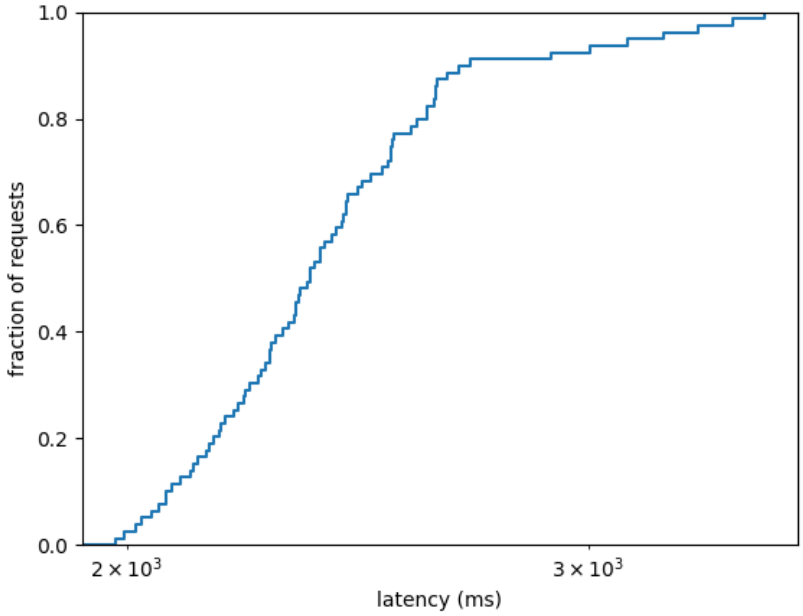


Figure 4.14: Cumulative latency plot for experiment with a throughput of 10req/s and unlimited batch size, run for 10s with 100ms network delay.

Figure 4.15 is a heatmap showing a node being killed 5s into an experiment. There is a clear jump in latency every time the killed node is the leader of the view, and the nodes must wait for the 0.5s timeout to elapse before the next view begins. The latency jump of around 1.25s is about what one would expect; 0.5s timeout and 0.75s of latency (the same as before the node was killed).

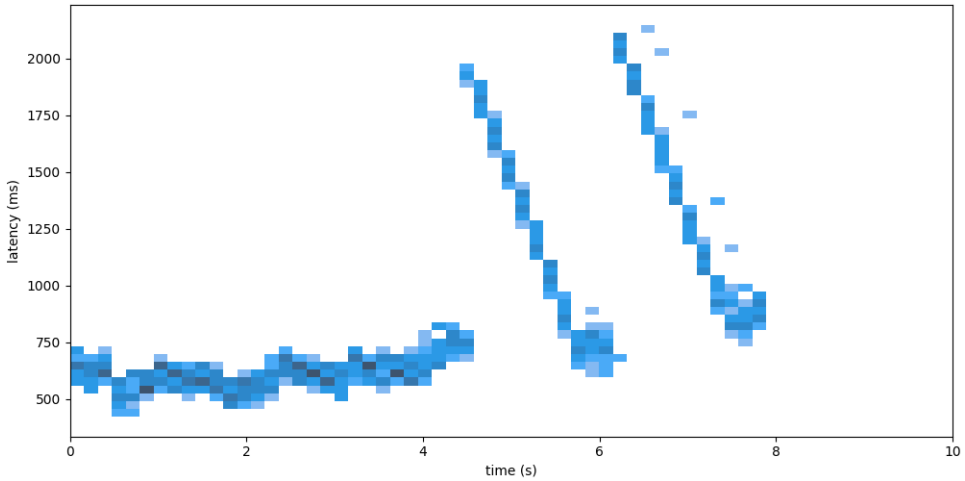


Figure 4.15: Heatmap showing distribution of latencies with a node being killed 5s in. Run for 10s with 7 nodes and a batch size of 100.

The view-change protocol is successful in allowing the system to make progress, albeit with a significant increase in latency. This is an inherent problem with the HotStuff protocol, although a failure-detector could help to minimise the number of views where the faulty node is leader.

Chapter 5

Conclusion

[consistent tense???] In this project we have given a reference implementation of the HotStuff byzantine consensus algorithm in OCaml, contributing to the wider OCaml ecosystem. The core algorithm is written in a self-contained module which could be reused by other projects with differing architectures and RPC systems. We have described some of the practical challenges of implementing HotStuff and implemented several optimisations of the basic algorithm (section 3.4). One main challenge was adapting the HotStuff pacemaker; we gave a full specification and proved that our pacemaker has desirable properties (section 3.2.2).

We have successfully met the requirements set out in section 2.4. There is significant evidence for the correctness of our implementation; our testing suite has 100% coverage of the consensus state machine code (section 2.5.2). Evaluation of the system was carried out both locally and on a simulated network, and we analysed its behaviour with different parameters, and under varying conditions (section 4). This analysis helped to identify that the system bottlenecks are message serialisation and cryptography. We also implemented several optimisations (section 3.4), and demonstrated their effectiveness in an ablation study (section 4.3.3).

Given that message serialisation and cryptography were shown to be system bottlenecks, future work could aim to overcome some of these problems to achieve better performance. One potential direction would be to implement custom message serialisation, and use a faster RPC library such as EIO [8]; both of these were out of the scope of this project. Alternative cryptography libraries could also be explored, although this may be an inherent bottleneck of any HotStuff implementation.

Future work could also explore the challenges of deploying a production-ready system based on our implementation. Although HotStuff is byzantine-fault tolerant, this project has not considered other security threats such as attacks on availability. Solving these problems would be non-trivial; some of our optimisations may be antagonistic with security considerations, for example a malicious node could repeatedly request the whole chain to be sent (instead of truncated) and bring down the system. One could also implement a proof of work or proof of stake mechanism on top of our implementation to make it resilient to Sybil attacks, allowing it to be deployed in a permissionless setting.

One lesson I learnt from this project is that graphs are an invaluable tool for analysing the performance of a distributed algorithm, and analysis of graphs should be included in the iterative passes of the waterfall development model (section 2.5.1). Much of development time was spent debugging system performance to implement the optimisations described in section 3.4. I found that as I was creating graphs and analysing performance (section 4.3), I was able

to more quickly find bugs and intuitively understand the behaviour of the system. Therefore if I were to do a similar project in future, I would further automate and parallelise the testing and graph plotting scripts to quickly gain insight during development.

In conclusion, this project has provided an implementation of a byzantine consensus algorithm, a key algorithm in the development of decentralised software. Decentralised software has far-reaching implications, and could challenge the control of large centralised authorities over critical infrastructure, platforms, and organisations.

Bibliography

[1] async.

[2] Cap'n Proto.

[3] Memtrace.

[4] Mininet.

[5] MirageOS.

[6] Tezos cryptography.

[7] Welcome to a World of OCaml.

[8] Eio – Effects-Based Parallel IO for OCaml, Apr. 2023. original-date: 2021-03-02T14:20:04Z.

[9] Lwt, Apr. 2023. original-date: 2013-07-29T20:47:51Z.

[10] Abraham, I. Byzantine fault tolerance, state machine replication and blockchains.

[11] Baudet, M., Ching, A., Chursin, A., Danezis, G., Garillot, F., Li, Z., Malkhi, D., Naor, O., Perelman, D., and Sonnino, A. State Machine Replication in the Libra Blockchain.

[12] Buterin, V., and Griffith, V. Casper the Friendly Finality Gadget, Jan. 2019. arXiv:1710.09437 [cs].

[13] Cachin, C., Kursawe, K., and Shoup, V. Random Oracles in Constantinople: Practical Asynchronous Byzantine Agreement Using Cryptography. *Journal of Cryptology 18*, 3 (July 2005), 219–246.

[14] Castro, M., and Liskov, B. Practical Byzantine Fault Tolerance.

[15] Chandra, T., Griesemer, R., and Redstone, J. Paxos made live: An engineering perspective. pp. 398–407.

[16] Chandra, T. D., Hadzilacos, V., and Toueg, S. The weakest failure detector for solving consensus. *Journal of the ACM 43*, 4 (July 1996), 685–722.

[17] Chandra, T. D., and Toueg, S. Unreliable failure detectors for reliable distributed systems. *Journal of the ACM 43*, 2 (Mar. 1996), 225–267.

[18] Dwork, C., Lynch, N., Stockmeyer, L., and Jose, S. CONSENSUS IN THE PRESENCE OF PARTIAL SYNCHRON¥.

[19] Fischer, M. J., Lynch, N. A., and Merritt, M. Easy impossibility proofs for distributed consensus problems.

[20] Golan Gueta, G., Abraham, I., Grossman, S., Malkhi, D., Pinkas, B., Reiter, M., Seredinschi, D.-A., Tamir, O., and Tomescu, A. SBFT: A Scalable and Decentralized Trust Infrastructure. In *2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)* (June 2019), pp. 568–580. ISSN: 1530-0889.

[21] Jensen, C. OCons, Mar. 2023. original-date: 2019-07-27T10:56:32Z.

[22] Kwon, J. Tendermint: Consensus without Mining.

[23] Lamport, L. The part-time parliament. *ACM Transactions on Computer Systems 16*, 2 (May 1998), 133–169.

[24] Lamport, L. Paxos Made Simple.

[25] Lamport, L., Shostak, R., and Pease, M. The Byzantine Generals Problem. *ACM Transactions on Programming Languages and Systems 4*, 3.

[26] Nakamoto, S. Bitcoin: A Peer-to-Peer Electronic Cash System.

[27] Naor, O., Baudet, M., Malkhi, D., and Spiegelman, A. Cogsworth: Byzantine View Synchronization, Feb. 2020. arXiv:1909.05204 [cs].

[28] Ongaro, D., and Ousterhout, J. In Search of an Understandable Consensus Algorithm.

[29] Pease, M., Shostak, R., and Lamport, L. Reaching Agreement in the Presence of Faults. *Journal of the ACM 27*, 2 (Apr. 1980), 228–234.

[30] Shoup, V. Practical Threshold Signatures. In *Advances in Cryptology — EUROCRYPT 2000*, G. Goos, J. Hartmanis, J. Van Leeuwen, and B. Preneel, Eds., vol. 1807. Springer Berlin Heidelberg, Berlin, Heidelberg, 2000, pp. 207–220. Series Title: Lecture Notes in Computer Science.

[31] Wood, D. G. ETHEREUM: A SECURE DECENTRALISED GENERALISED TRANSACTION LEDGER.

[32] Yin, M., Malkhi, D., Reiter, M. K., Gueta, G. G., and Abraham, I. Hotstuff: Bft consensus in the lens of blockchain, 2019.