HotStuff: BFT Consensus with Linearity and Responsiveness

Maofan Yin  
Cornell University  
VMware Research

Dahlia Malkhi  
VMware Research

Michael K. Reiter  
UNC-Chapel Hill  
VMware Research

Guy Golan Gueta  
VMware Research

Ittai Abraham  
VMware Research

ABSTRACT

We present HotStuff, a leader-based Byzantine fault-tolerant replication protocol for the partially synchronous model. Once network communication becomes synchronous, HotStuff enables a correct leader to drive the protocol to consensus at the pace of actual (vs. maximum) network delay—a property called responsiveness—and with communication complexity that is linear in the number of replicas. To our knowledge, HotStuff is the first partially synchronous BFT replication protocol exhibiting these combined properties. Its simplicity enables it to be further pipelined and simplified into a practical, concise protocol for building large-scale replication services.

CCS CONCEPTS

• Software and its engineering → Software fault tolerance; • Security and privacy → Distributed systems security.

KEYWORDS

Byzantine fault tolerance; consensus; responsiveness; scalability; blockchain

ACM Reference Format:


1 INTRODUCTION

Byzantine fault tolerance (BFT) refers to the ability of a computing system to endure arbitrary (i.e., Byzantine) failures of its components while taking actions critical to the system’s operation. In the context of state machine replication (SMR) [35, 47], the system as a whole provides a replicated service whose state is mirrored across $n$ deterministic replicas. A BFT SMR protocol is used to ensure that non-faulty replicas agree on an order of execution for client-initiated service commands, despite the efforts of $f$ Byzantine replicas. This, in turn, ensures that the $n - f$ non-faulty replicas will run commands identically and so produce the same response for each command. As is common, we are concerned here with the partially synchronous communication model [25], whereby a known bound $\Delta$ on message transmission holds after some unknown global stabilization time (GST). In this model, $n \geq 3f + 1$ is required for non-faulty replicas to agree on the same commands in the same order (e.g., [12]) and progress can be ensured deterministically only after GST [27].

When BFT SMR protocols were originally conceived, a typical target system size was $n = 4$ or $n = 7$, deployed on a local-area network. However, the renewed interest in Byzantine fault-tolerance brought about by its application to blockchains now demands solutions that can scale to much larger $n$. In contrast to permissionless blockchains such as the one that supports Bitcoin, for example, so-called permissioned blockchains involve a fixed set of replicas that collectively maintain an ordered ledger of commands or, in other words, that support SMR. Despite their permissioned nature, numbers of replicas in the hundreds or even thousands are envisioned (e.g., [30, 42]). Additionally, their deployment to wide-area networks requires setting $\Delta$ to accommodate higher variability in communication delays.

The scaling challenge. Since the introduction of PBFT [20], the first practical BFT replication solution in the partial synchrony model, numerous BFT solutions were built around its core two-phase paradigm. The practical aspect is that a stable leader can drive a consensus decision in just two rounds of message exchanges. The first phase guarantees proposal uniqueness through the formation of a quorum certificate (QC) consisting of $(n - f)$ votes. The second phase guarantees that the next leader can convince replicas to vote for a safe proposal.

The algorithm for a new leader to collect information and propose it to replicas—called a view-change—is the epicenter of replication. Unfortunately, view-change based on the two-phase paradigm is far from simple [38], is bug-prone [4], and incurs a significant communication penalty for even moderate system sizes. It requires the new leader to relay information from $(n - f)$ replicas, each reporting its own highest known QC. Even counting just authenticators (digital signatures or message authentication codes), conveying a new proposal has a communication footprint of $O(n^2)$ authenticators in PBFT, and variants that combine multiple authenticators into one via threshold digital signatures (e.g., [18, 30]) still send $O(n^2)$ authenticators. The total number of authenticators transmitted if $O(n)$ view-changes occur before a single consensus decision is reached is $O(n^3)$ in PBFT, and even with threshold signatures is $O(n^5)$. This scaling challenge plagues not only PBFT, but many other protocols developed since then, e.g., Prime [9], Zyzzyva [34], Upright [22], BFT-SMaRt [13], 700BFT [11], and SBFT [30].

HotStuff revolves around a three-phase core, allowing a new leader to simply pick the highest QC it knows of. It introduces a second phase that allows replicas to “change their mind” after voting in the phase, without requiring a leader proof at all. This alleviates the above complexity, and at the same time considerably simplifies the leader replacement protocol. Last, having (almost)
canonized all the phases, it is very easy to pipeline HotStuff, and to frequently rotate leaders.

To our knowledge, only BFT protocols in the blockchain arena like Tendermint [15, 16] and Casper [17] follow such a simple leader regime. However, these systems are built around a synchronous core, wherein proposals are made in pre-determined intervals that must accommodate the worst-case time it takes to propagate messages over a wide-area peer-to-peer gossip network. In doing so, they forego a hallmark of most practical BFT SMR solutions (including those listed above), namely optimistic responsiveness [42]. Informally, responsiveness requires that a non-faulty leader, once designated, can drive the protocol to consensus in time depending only on the actual message delays, independent of any known upper bound on message transmission delays [10]. More appropriate for our model is optimistic responsiveness, which requires responsiveness only in beneficial (and hopefully common) circumstances—here, after GST is reached. Optimistic or not, responsiveness is precluded with designs such as Tendermint/Casper. The crux of the difficulty is that there may exist an honest replica that has the highest QC, but the leader does not know about it. One can build scenarios where this prevents progress ad infinitum (see Section 4.4 for a detailed liveless scenario). Indeed, failing to incorporate necessary delays at crucial protocol steps can result in losing liveness outright, as has been reported in several existing deployments, e.g., see [2, 3, 19].

Our contributions. To our knowledge, we present the first BFT SMR protocol, called HotStuff, to achieve the following two properties:

- **Linear View Change:** After GST, any correct leader, once designated, sends only $O(n)$ authenticators to drive a consensus decision. This includes the case where a leader is replaced. Consequently, communication costs to reach consensus after GST is $O(n^2)$ authenticators in the worst case of cascading leader failures.

- **Optimistic Responsiveness:** After GST, any correct leader, once designated, needs to wait just for the first $n - f$ responses to guarantee that it can create a proposal that will make progress. This includes the case where a leader is replaced.

Another feature of HotStuff is that the costs for a new leader to drive the protocol to consensus is no greater than that for the current leader. As such, HotStuff supports frequent succession of leaders, which has been argued is useful in blockchain contexts for ensuring chain quality [28].

HotStuff achieves these properties by adding another phase to each view, a small price to latency in return for considerably simplifying the leader replacement protocol. This exchange incurs only the actual network delays, which are typically far smaller than $\Delta$ in practice. As such, we expect this added latency to be much smaller than that incurred by previous protocols that forgo responsiveness to achieve linear view-change. Furthermore, throughput is not affected due to the efficient pipeline we introduce in Section 5.

HotStuff has the additional benefit of being remarkably simple. Safety is specified via voting and commit rules over graphs of nodes. The mechanisms needed to achieve liveness are encapsulated within a Pacemaker, cleanly separated from the mechanisms needed for safety (Section 6).

2 RELATED WORK

Reaching consensus in face of Byzantine failures was formulated as the Byzantine Generals Problem by Lamport et al. [37], who also coined the term “Byzantine failures”. The first synchronous solution was given by Pease et al. [43], and later improved by Dolev and Strong [24]. The improved protocol has $O(n^3)$ communication complexity, which was shown optimal by Dolev and Reischuk [23]. A leader-based synchronous protocol that uses randomness was given by Katz and Koo [32], showing an expected constant-round solution with $(n - 1)/2$ resilience.

Meanwhile, in the asynchronous settings, Fischer et al. [27] showed that the problem is unsolvable deterministically in asynchronous setting in face of a single failure. Furthermore, an $(n - 1)/3$ resilience bound for any asynchronous solution was proven by Ben-Or [12]. Two approaches were devised to circumvent the impossibility. One relies on randomness, initially shown by Ben-Or [12], using independently random coin flips by processes until they happen to converge to consensus. Later works used cryptographic methods to share an unpredictable coin and drive complexities down to constant expected rounds, and $O(n^3)$ communication [18].

The second approach relies on partial synchrony, first shown by Dwork, Lynch, and Stockmeyer (DLS) [25]. This protocol preserves safety during asynchronous periods, and after the system becomes synchronous, DLS guarantees termination. Once synchrony is maintained, DLS incurs $O(n^4)$ total communication and $O(n)$ rounds per decision.

State machine replication relies on consensus at its core to order client requests so that correct replicas execute them in this order. The recurring need for consensus in SMR led Lamport to devise Paxos [36], a protocol that operates an efficient pipeline in which a stable leader drives decisions with linear communication and one round-trip. A similar emphasis led Castro and Liskov [20, 21] to develop an efficient leader-based Byzantine SMR protocol named PBFT, whose stable leader requires $O(n^2)$ communication and two round-trips per decision, and the leader replacement protocol incurs $O(n^3)$ communication. PBFT has been deployed in several systems, including BFT-SMaRt [13]. Kotla et al. introduced an optimistic linear path into PBFT in a protocol named Zyzzyva [34], which was utilized in several systems, e.g., Upright [22] and Byzoin [33]. The optimistic path has linear complexity, while the leader replacement protocol remains $O(n^3)$. Abraham et al. [4] later exposed a safety violation in Zyzzyva, and presented fixes [5, 30]. On the other hand, to also reduce the complexity of the protocol itself, Song et al. proposed Bosco [49], a simple one-step protocol with low latency on the optimistic path, requiring $5f + 1$ replicas. SBFT [30] introduces an $O(n^2)$ communication view-change protocol that supports a stable leader protocol with optimistically linear, one round-trip decisions. It reduces the communication complexity by harnessing two methods: a collector-based communication paradigm by Reiter [45], and signature combining via threshold cryptography on protocol votes by Cachin et al. [18].

A leader-based Byzantine SMR protocol that employs randomization was presented by Ramasamy et al. [44], and a leaderless
We consider a system consisting of a fixed set of replicas. We refer to a "broadcast," it involves the broadcaster, if correct, sending a message to all other replicas. At their core, these randomized Byzantine solutions employ randomized asynchronous Byzantine consensus, whose best known communication complexity was \( O(n^3) \) (see above), amortizing the cost via batching. However, most recently, based on the idea in this HotStuff paper, a parallel submission to PODC’19 [8] further improves the communication complexity to \( O(n^2) \).

Bitcoin’s core is a protocol known as Nakamoto Consensus [40], a synchronous protocol with only probabilistic safety guarantee and no finality (see analysis in [6, 28, 41]). It operates in a permissionless model where participants are unknown, and resilience is kept via Proof-of-Work. As described above, recent blockchain solutions hybridxize Proof-of-Work solutions with classical BFT solutions in various ways [7, 17, 26, 29, 31, 33, 42]. The need to address rotating leaders in these hybrid solutions and others provide the motivation behind HotStuff.

3 MODEL

We consider a system consisting of a fixed set of replicas indexed by \( i \in [n] \) where \( [n] = \{1, \ldots, n\} \). A set \( F \subset [n] \) of up to \( f = |F| \) replicas are Byzantine faulty, and the remaining ones are correct. We will often refer to the Byzantine replicas as being coordinated by an adversary, which learns all internal state held by these replicas (including their cryptographic keys, see below).

Network communication is point-to-point, authenticated and reliable: one correct replica receives a message from another correct replica if and only if the latter sent that message to the former. When we refer to a "broadcast," it involves the broadcaster, if correct, sending the same point-to-point messages to all replicas, including itself. We adopt the partial synchrony model of Dwork et al. [25], where there is a known bound \( \Delta \) and an unknown Global Stabilization Time (GST), such that after GST, all transmissions between two correct replicas arrive within time \( \Delta \). Our protocol will ensure safety always, and will guarantee progress within a bounded duration after GST. (Guaranteeing progress before GST is impossible [27].) In practice, our protocol will guarantee progress if the system remains stable (i.e., if messages arrive within \( \Delta \) time) for sufficiently long after GST, though assuming that it does so forever simplifies discussion.

Cryptographic primitives. HotStuff makes use of threshold signatures [14, 18, 48]. In a \((k, n)\)-threshold signature scheme, there is a single public key held by all replicas, and each of the \( n \) replicas holds a distinct private key. The \( i \)-th replica can use its private key to contribute a partial signature \( \rho_i \leftarrow tsign_i(m) \) on message \( m \). Partial signatures \( \{\rho_i\}_{i \in I} \), where \( |I| = k \) and each \( \rho_i \leftarrow tsign_i(m) \), can be used to produce a digital signature \( \sigma \leftarrow tcombine(m, \{\rho_i\}_{i \in I}) \) on \( m \). Any other replica can verify the signature using the public key and the function \( tverify \). We require that if \( \rho_i \leftarrow tsign_i(m) \) for each \( i \in I, |I| = k \), and if \( \sigma \leftarrow tcombine(m, \{\rho_i\}_{i \in I}) \), then \( tverify(m, \sigma) \) returns true. However, given oracle access to oracles \( \{tsign_i(1)|i \in [n]|, F \), an adversary who queries \( tsign_i(m) \) on strictly fewer than \( k-f \) of these oracles has negligible probability of producing a signature \( \sigma \) for the message \( m \) (i.e., such that \( tverify(m, \sigma) \) returns true). Throughout this paper, we use a threshold of \( k = 2f + 1 \). Again, we will typically leave invocations of \( tverify \) implicit in our protocol descriptions.

We also require a cryptographic hash function \( h \) (also called a message digest function), which maps an arbitrary-length input to a fixed-length output. The hash function must be collision resistant [46], which informally requires that the probability of an adversary producing inputs \( m \) and \( m' \) such that \( h(m) = h(m') \) is negligible. As such, \( h(m) \) can serve as an identifier for a unique input \( m \) in the protocol.

Complexity measure. The complexity measure we care about is authenticator complexity, which specifically is the sum, over all replicas \( i \in [n] \), of the number of authenticators received by replica \( i \) in the protocol to reach a consensus decision after GST. (Again, before GST, a consensus decision might not be reached at all in the worst case [27].) Here, an authenticator is either either a partial signature or a signature. Authenticator complexity is a useful measure of communication complexity for several reasons. First, like bit complexity and unlike message complexity, it hides unnecessary details about the transmission topology. For example, \( n \) messages carrying one authenticator count the same as one message carrying \( n \) authenticators. Second, authenticator complexity is better suited than bit complexity for capturing costs in protocols like ours that reach consensus repeatedly, where each consensus decision (or each view proposed on the way to that consensus decision) is identified by a monotonically increasing counter. That is, because such a counter increases indefinitely, the bit complexity of a protocol that sends such a counter cannot be bounded. Third, since in practice, cryptographic operations to produce or verify digital signatures and to produce or combine partial signatures are typically the most computationally intensive operations in protocols that use them, the authenticator complexity provides insight into the computational burden of the protocol, as well.

4 BASIC HOTSTUFF

HotStuff solves the State Machine Replication (SMR) problem. At the core of SMR is a protocol for deciding on a growing log of commands requests by clients. A group of state-machine replicas apply commands in sequence order consistently. A client sends a command request to all replicas, and waits for responses from \( (f + 1) \) of them. For the most part, we omit the client from the

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Correct leader</th>
<th>Authenticator complexity</th>
<th>Responsiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLS [25]</td>
<td>( O(n^3) )</td>
<td>( O(n^3) )</td>
<td>( O(n^3) )</td>
</tr>
<tr>
<td>PBFT [20]</td>
<td>( O(n^2) )</td>
<td>( O(n^2) )</td>
<td>( O(n^2) )</td>
</tr>
<tr>
<td>SSFT [36]</td>
<td>( O(n^2) )</td>
<td>( O(n^2) )</td>
<td>( O(n^2) )</td>
</tr>
<tr>
<td>Tendermint [15] / Casper [17]</td>
<td>( O(n) )</td>
<td>( O(n^3) )</td>
<td>( O(n^3) )</td>
</tr>
<tr>
<td>Tendermint / Casper</td>
<td>( O(n^2) )</td>
<td>( O(n) )</td>
<td>( O(n) )</td>
</tr>
<tr>
<td>HotStuff</td>
<td>( O(n) )</td>
<td>( O(n) )</td>
<td>( O(n) )</td>
</tr>
</tbody>
</table>

\*Signatures can be combined using threshold signatures, though this optimization is not mentioned in their original works.
discussion, and defer to the standard literature for issues regarding numbering and de-duplication of client requests.

The Basic HotStuff solution is presented in Algorithm 2. The protocol works in a succession of views numbered with monotonically increasing view numbers. Each `viewNumber` has a unique dedicated leader known to all. Each replica stores a tree of pending commands as its local data structure. Each tree node contains a proposed command (or a batch of them), metadata associated with the protocol, and a parent link. The branch led by a given node is the path from the node all the way to the tree root by visiting parent links. During the protocol, a monotonically growing branch becomes committed. To become committed, the leader of a particular view proposing the branch must collect votes from a quorum of \((n - f)\) replicas in three phases, PREPARE, PRE-COMMIT, and COMMIT.

A key ingredient in the protocol is a collection of \((n - f)\) votes over a leader proposal, referred to as a quorum certificate (or “QC” in short). The QC is associated with a particular node and a view number. The `tcombine` utility employs a threshold signature scheme to generate a representation of \((n - f)\) signed votes as a single authenticator.

Below we give an operational description of the protocol logic by phases, followed by a precise specification in Algorithm 2, and conclude the section with safety, liveness, and complexity arguments.

### 4.1 Phases

**PREPARE phase.** The protocol for a new leader starts by collecting new-view messages from \((n - f)\) replicas. The new-view message is sent by a replica as it transitions into `viewNumber` (including the first view) and carries the highest `prepare QC` that the replica received (L if none), as described below.

The leader processes these messages in order to select a branch that has the highest preceding view in which a `prepare QC` was formed. The leader selects the `prepare QC` with the highest view, denoted `high QC`, among the new-view messages. Because `high QC` is the highest among \((n - f)\) replicas, no higher view could have reached a commit decision. The branch led by `high QC.node` is therefore safe.

Collecting new-view messages to select a safe branch may be omitted by an incumbent leader, who may simply select its own highest `prepare QC` as `high QC`. We defer this optimization to Section 6 and only describe a single, unified leader protocol in this section. Note that, different from PBFT-like protocols, including this step in the leader protocol is straightforward, and it incurs the same, linear overhead as all the other phases of the protocol, regardless of the situation.

The leader uses the `createLeaf` method to extend the tail of `high QC.node` with a new proposal. The method creates a new leaf node as a child and embeds a digest of the parent in the child node. The leader then sends the new node in a `prepare message` to all other replicas. The proposal carries `high QC` for safety justification.

Upon receiving the `prepare message` for the current view from the leader, replica \(r\) uses the `safeNode` predicate to determine whether to accept it. If it is accepted, the replica sends a `prepare vote` with a partial signature (produced by `tSign`,) for the proposal to the leader.

**SAFENODE predicate.** The `SAFENODE` predicate is a core ingredient of the protocol. It examines a proposal message \(m\) carrying a QC justification \(m.justify\), and determines whether \(m.node\) is safe to accept. The safety rule to accept a proposal is the branch of \(m.node\) extends from the currently locked node `locked QC.node`. On the other hand, the liveness rule is the replica will accept \(m\) if \(m.justify\) has a higher view than the current `locked QC`. The predicate is true as long as either one of two rules holds.

**PRE-COMMIT phase.** When the leader receives \((n - f)\) `prepare votes` for the current proposal `curProposal`, it combines them into a `prepare QC`. The leader broadcasts `prepare QC` in PRE-COMMIT messages. A replica responds to the leader with PRE-COMMIT vote having a signed digest of the proposal.

**COMMIT phase.** The commit phase is similar to PRE-COMMIT phase. When the leader receives \((n - f)\) `prepare votes`, it combines them into a `precommit QC` and broadcasts it in COMMIT messages; replicas respond to it with a COMMIT vote. Importantly, a replica becomes `locked` on the `precommit QC` at this point by setting its `locked QC` to `precommit QC` (Line 25 of Algorithm 2).

This is crucial to guard the safety of the proposal in case it becomes a consensus decision.

**DECIDE phase.** When the leader receives \((n - f)\) `commit votes`, it combines them into a `commit QC`. Once the leader has assembled a `commit QC`, it sends it in a `decide message` to all other replicas. Upon receiving a `decide message`, a replica considers the proposal embodied in the `commit QC` a committed decision, and executes the commands in the committed branch. The replica increments `viewNumber` and starts the next view.

**NEXT VIEW interrupt.** In all phases, a replica waits for a message at `viewNumber` for a timeout period, determined by an auxiliary `nextView(viewNumber)` utility. If `nextView(viewNumber)` interrupts waiting, the replica also increments `viewNumber` and starts the next view.

### 4.2 Data Structures

**Messages.** A message \(m\) in the protocol has a fixed set of fields that are populated using the `Msg()` utility shown in Algorithm 1. \(m\) is automatically stamped with `curView`, the sender’s current view number. Each message has a type \(m.type \in \{\text{new-view, prepare, pre-commit, commit, decide}\}\). \(m.node\) contains a proposed node (the leaf node of a proposed branch). There is an optional field \(m.justify\). The leader always uses this field to carry the QC for the different phases. Replicas use it in `new-view` messages to carry the highest `prepare QC`. Each message sent in a replica role contains a partial signature `m.partialSig` by the sender over the tuple \((m.type, m.viewNumber, m.node)\), which is added in the `voteMsg()` utility.

**Quorum certificates.** A Quorum Certificate (QC) over a tuple \((type, viewNumber, node)\) is a data type that combines a collection of signatures for the same tuple signed by \((n - f)\) replicas. Given a QC `qc`, we use `qc.type`, `qc.viewNumber`, `qc.node` to refer to the matching fields of the original tuple.

**Tree and branches.** Each command is wrapped in a node that additionally contains a parent link which could be a hash digest of the parent node. We omit the implementation details from the
Algorithm 2. Basic HotStuff protocol (for replica).

1. for curView ← 1, 2, 3, . . . do
   \[ \text{\texttt{PRE-COMMIT phase}} \]
   2. as a leader // \( r = \text{leader(curView)} \) // we assume special new-view messages from view 0
      3. wait for \((n - f)\) new-view messages:
         \[ M \leftarrow \{ m \ | \ \text{matchingMsg}(m, \text{new-view, curView - 1}) \} \]
      4. high QC \( \leftarrow \arg \max \{ m.\text{justify}, \text{viewNumber}\} \) .justify
      5. curProposal \( \leftarrow \text{createLeaf}(\text{high QC}.\text{node}, \text{client’s command}) \)
      6. broadcast Msg(prepare, curProposal, high QC)
      7. as a replica
         8. wait for message \( m \) from leader(curView)
            \[ m : \text{matchingMsg}(m, \text{prepare, curView}) \]
      9. if \( m.\text{node} \) extends from \( m.\text{justify}.\text{node} \)
         \[ \text{SAFE}(m.\text{node}, m.\text{justify}) \text{ then} \]
            10. send voteMsg(prepare, m. node, \perp) to leader(curView)

\[ \text{\texttt{COMMIT phase}} \]

11. as a leader
   12. wait for \((n - f)\) votes:
      \[ V \leftarrow \{ v \ | \ \text{matchingMsg}(v, \text{prepare, curView}) \} \]
   13. prepare QC \( \leftarrow QC(V) \)
   14. broadcast Msg(pre-commit, \perp, prepare QC)
   15. as a replica
      16. wait for message \( m \) from leader(curView)
         \[ m : \text{matchingQC}(m.\text{justify}, \text{prepare, curView}) \]
      17. prepare QC \( \leftarrow m.\text{justify} \)
      18. send to leader(curView)
         voteMsg(pre-commit, m. justify, node, \perp)

\[ \text{\texttt{DECIDE phase}} \]

19. as a leader
   20. wait for \((n - f)\) votes:
      \[ V \leftarrow \{ v \ | \ \text{matchingMsg}(v, \text{commit, curView}) \} \]
   21. precommit QC \( \leftarrow QC(V) \)
   22. broadcast Msg(commit, \perp, precommit QC)
   23. as a replica
      24. wait for message \( m \) from leader(curView)
         \[ m : \text{matchingQC}(m.\text{justify}, \text{pre-commit, curView}) \]
      25. locked QC \( \leftarrow m.\text{justify} \)
      26. send to leader(curView)
         voteMsg(commit, m. justify, node, \perp)

\[ \text{\texttt{FINALLY}} \]

27. as a leader
   28. wait for \((n - f)\) votes:
      \[ V \leftarrow \{ v \ | \ \text{matchingMsg}(v, \text{commit, curView}) \} \]
   29. commit QC \( \leftarrow QC(V) \)
   30. broadcast Msg(decide, \perp, commit QC)
   31. as a replica
      32. wait for message \( m \) from leader(curView)
         \[ m : \text{matchingQC}(m.\text{justify}, \text{commit, curView}) \]
      33. execute new commands through \( m.\text{new-view} \), \( m.\text{node} \), respond to clients

4.3 Protocol Specification

The protocol given in Algorithm 2 is described as an iterated view-by-view loop. In each view, a replica performs phases in succession based on its role, described as a succession of “as” blocks. A replica can have more than one role. For example, a leader is also a (normal) replica. Execution of as blocks across roles can be proceeded concurrently. The execution of each as block is atomic. A new-view interrupt aborts all operations in any as block, and jumps to the “Finally” block.

Algorithm 1 Utilities (for replica r).

1. function Msg(type, node, qc)
2. \[ m.\text{type} \leftarrow \text{type} \]
3. m.\text{viewNumber} \leftarrow curView
4. m.\text{node} \leftarrow node
5. m.\text{justify} \leftarrow qc
6. return m
7. function voteMsg(type, node, qc)
8. \[ m \leftarrow \text{Msg(type, node, qc)} \]
9. m.\text{partialSig} \leftarrow \text{tsign},((m.\text{type}, m.\text{viewNumber}, m.\text{node}))
10. return m
11. procedure createLeaf(parent, cmd)
12. b.\text{parent} \leftarrow \text{parent}
13. b.\text{cmd} \leftarrow \text{cmd}
14. return b
15. function QC(V)
16. qc.\text{type} \leftarrow m.\text{type} : m \in V
17. qc.\text{viewNumber} \leftarrow m.\text{viewNumber} : m \in V
18. qc.\text{node} \leftarrow m.\text{node} : m \in V
19. qc.\text{sig} \leftarrow \text{tcombine}(\{q.\text{type}, q.\text{viewNumber}, q.\text{node})
    \quad \{\text{m.\text{partialSig} | m \in V}\})
20. return qc
21. function matchingMsg(m, t, v)
22. return \((m.\text{type} = t) \land (m.\text{viewNumber} = v)\)
23. function matchingQC(qc, t, v)
24. return \((q.\text{type} = t) \land (q.\text{viewNumber} = v)\)
25. function safeNode(node, qc)
26. return \((\text{node extends from locked QC. node}) \lor \text{safety rule}\)
27. \((q.\text{viewNumber} > \text{locked QC. viewNumber}) \lor \text{liveness rule}\)

4.4 Safety, Liveness, and Complexity

Safety. We first define a quorum certificate qc to be valid if
\[\text{tverify}(\{q.\text{type}, q.\text{viewNumber}, q.\text{node}), q.\text{sig}) \text{ is true.}\]

Lemma 1. For any valid qc1, qc2 in which qc1.\text{type} = qc2.\text{type} and
qc1.\text{node} conflicts with qc2.\text{node}, we have qc1.\text{viewNumber} \neq qc2.\text{viewNumber}.
We can now set the first lockedQC with Liveness.

Theorem 2. If w and b are conflicting nodes, then they cannot be both committed, each by a correct replica.

Proof. We prove this important theorem by contradiction. Let qc_1 denote a valid commit QC (i.e., qc_1.type = commit) such that qc_1.node = w, and qc_2 denote a valid commit QC such that qc_2.node = b. Denote v_1 = qc_1.viewNumber and v_2 = qc_2.viewNumber. By Lemma 1, v_1 ≠ v_2. W.l.o.g. assume v_1 < v_2.

We will now denote by v_3 the lowest view higher than v_1 for which there is a valid prepare QC, qc_3 (i.e., qc_3.type = prepare) where qc_3.node = w, and qc_3.node conflicts with w. Formally, we define the following predicate for any prepare QC:

E(prepare QC) := (v_1 < prepare QC.viewNumber ≤ v_2) ∧ (prepare QC.node conflicts with w).

We can now set the first switching point qc_s:

qc_s := arg min

prepare QC

prepare QC.viewNumber | prepare QC is valid ∧ E(prepare QC).

Note that, by assumption such a qc_s must exist; for example, qc_s could be the prepare QC formed in view v_2.

Of the correct replicas that sent a partial result tsign_r((qc_1.type, qc_1.viewNumber, qc_1.node)), let r be the first that contributed tsign_r((qc_3.type, qc_3.viewNumber, qc_3.node)); such an r must exist since otherwise, one of qc_1.sig and qc_3.sig could not have been created. During view v_1, replica r updates its lock QC to a precommit QC on w at Line 25 of Algorithm 2. Due to the minimality of v_2, the lock that replica r has on w is not changed before qc_s is formed. Otherwise, r must have seen some other prepare QC with lower view because Line 17 comes before Line 25, contradicting to the minimality. Now consider the invocation of safeNode in the prepare phase of view v_3 by replica r, with a message m carrying m.node = qc_s.node. By assumption, m.node conflicts with locked QC.node, and so the disjunct at Line 26 of Algorithm 1 is false. Moreover, m.justify.viewNumber > v_1 would violate the minimality of v_3, and so the disjunct in Line 27 of Algorithm 1 is also false. Thus, safeNode must return false and r cannot cast a precommit vote on the conflicting branch in view v_3, a contradiction.

Lemma 3. If a correct replica is locked such that lockedQC = precommit QC, then at least f + 1 correct replicas voted for some prepare QC matching lockedQC.

Proof. Suppose replica r is locked on precommit QC. Then, (n − f) votes were cast for the matching prepare QC in the prepare phase (Line 10 of Algorithm 2), out of which at least f + 1 were from correct replicas.

Theorem 4. After GST, there exists a bounded time period T_f such that if all correct replicas remain in view v during T_f and the leader for view v is correct, then a decision is reached.

Proof. Starting in a new view, the leader collects (n − f) new-view messages and calculates its high QC before broadcasting a prepare message. Suppose among all replicas (including the leader itself), the highest kept lock is locked QC = precommit QC*. By Lemma 3, we know there are at least f + 1 correct replicas that voted for a prepare QC* matching precommit QC*, and have already sent them to the leader in their new-view messages. Thus, the leader must learn a matching prepare QC* in at least one of these new-view messages and use it as high QC in its prepare message. By the assumption, all correct replicas are synchronized in their view and the leader is non-faulty. Therefore, all correct replicas will vote in the prepare phase, since in safeNode, the condition on Line 27 of Algorithm 1 is satisfied (even if the node in the message conflicts with a replica’s stale locked QC.node, and so Line 26 is not). Then, after the leader assembles a valid prepare QC* for this view, all replicas will vote in all the following phases, leading to a new decision. After GST, the duration T_f for these phases to complete is of bounded length.

The protocol is Optimistically Responsive because there is no explicit “wait-for-L” step, and the logical disjunction in safeNode is used to override a stale lock with the help of the three-phase paradigm.

We now provide simple constructions for leader and nextView that suffice to ensure that after GST, eventually a view will be reached in which the leader is correct and all correct replicas remain in this view for T_f time. It suffices for leader to return some deterministic mapping from view number to a replica, eventually rotating through all replicas. A possible solution for nextView is to utilize an exponential back-off mechanism that maintains a timeout interval. Then a timer is set upon entering each view. When the timer goes off without making any decision, the replica doubles the interval and calls nextView to advance the view. Since the interval is doubled at each time, the waiting intervals of all correct replicas will eventually have at least T_f overlap in common, during which the leader could drive a decision.

Liveness with two-phases. We now briefly demonstrate an infinite non-deciding scenario for a “two-phase” HotStuff. This explains the necessity for introducing a synchronous delay in Casper and Tendermint, and hence for abandoning (Optimistic) Responsiveness.

In the two-phase HotStuff variant, we omit the pre-commit phase and proceed directly to commit. A replica becomes locked when it votes on a prepare QC. Suppose, in view v, a leader proposes b. It completes the prepare phase, and some replica r_v votes.
for the prepare QC, say qc, such that qc.node = b. Hence, r0 becomes locked on qc. An asynchronous network scheduling causes the rest of the replicas to move to view v + 1 without receiving qc.

We now repeat ad infinitum the following single-view transcript. We start view v + 1 with only r0, holding the highest prepare QC (i.e. qc) in the system. The new leader l collects new-view messages from 2f + 1 replicas excluding r0. The highest prepare QC among these, qc’, has view v – 1 and b’ = qc’.node conflicts with b. l then proposes b’’ which extends b’, to which 2f honest replicas respond with a vote, but r0 rejects it because it is locked on qc, b’’ conflicts with b and qc’ is lower than qc. Eventually, 2f replicas give up and move to the next view. Just then, a faulty replica responds to l’s proposal, l then puts together a prepare QC(v + 1, b’’) and one replica, say r_{v+1}, votes for it and becomes locked on it.

Complexity. In each phase of HotStuff, only the leader broadcasts to all replicas while the replicas respond to the sender once with a partial signature to certify the vote. In the leader’s message, the QC consists of a proof of (n – f) votes collected previously, which can be encoded by a single threshold signature. In a replica’s response, the partial signature from that replica is the only authenticator. Therefore, in each phase, there are O(n) authenticators received in total. As there is a constant number of phases, the overall complexity per view is O(n).

5 CHAINED HOTSTUFF

It takes three phases for a Basic HotStuff leader to commit a proposal. These phases are not doing “useful” work except collecting votes from replicas, and they are all very similar. In Chained HotStuff, we improve the Basic Hotstuff protocol utility while at the same time considerably simplifying it. The idea is to change the view on every prepare phase, so each proposal has its own view. This reduces the number of message types and allows for pipelining of decisions. A similar approach for message type reduction was suggested in Casper [1].

More specifically, in Chained Hotstuff the votes over a prepare phase are collected in a view by the leader into a generic QC. Then the generic QC is relayed to the leader of the next view, essentially delegating responsibility for the next phase, which would have been pre-commit, to the next leader. However, the next leader does not actually carry a pre-commit phase, but instead initiates a new prepare phase and adds its own proposal. This prepare phase for view v + 1 simultaneously serves as the pre-commit phase for view v. The prepare phase for view v + 2 simultaneously serves as the pre-commit phase for view v + 1 and as the commit phase for view v. This is possible because all the phases have identical structure.

The pipeline of Basic HotStuff protocol phases embedded in a chain of Chained HotStuff proposals is depicted in Figure 1. Views v1, v2, v3 of Chained Hotstuff serve as the prepare, pre-commit, and commit Basic HotStuff phases for cmd1 proposed in v1. This command becomes committed by the end of v4. Views v2, v3, v4 serve as the three Basic HotStuff phases for cmd2 proposed in v2, and it becomes committed by the end of v5. Additional proposals generated in these phases continue the pipeline similarly, and are denoted by dashed boxes. In Figure 1, a single arrow denotes the b.parent field for a node b, and a double arrow denotes b.justify.node.

Hence, there are only two types of messages in Chained HotStuff, a new-view message and generic-phase generic message. The generic QC functions in all logically pipelined phases. We next explain the mechanisms in the pipeline to take care of locking and committing, which occur only in the commit and decide phases of Basic HotStuff.

**Dummy nodes.** The generic QC used by a leader in some view viewNumber may not directly reference the proposal of the preceding view (viewNumber – 1). The reason is that the leader of a preceding view fails to obtain a QC, either because there are conflicting proposals, or due to a benign crash. To simplify the tree structure, createLeaf extends generic QC.node with blank nodes up to the height (the number of parent links on a node’s branch) of the proposing view, so view-numbers are equated with node heights. As a result, the QC embedded in a node b may not refer to its parent, i.e., b.justify.node may not equal b.parent (the last node in Figure 2).

**One-Chain, Two-Chain, and Three-Chain.** When a node b carries a QC that refers to a direct parent, i.e., b.justify.node = b.parent, we say that it forms a One-Chain. Denote by b’ = b’.justify.node. Node b’ forms a Two-Chain, if in addition to forming a One-Chain, b’.justify.node = b’.parent. It forms a Three-Chain, if b” forms a Two-Chain.

Looking at chain b = b” justifying node, b” = b” justifying node, b” = b”.justify.node, ancestry gaps might occur at any one of the nodes. These situations are similar to a leader of Basic HotStuff failing to complete any one of three phases, and getting interrupted to the next view by nextView.

If b” forms a One-Chain, the prepare phase of b” has succeeded. Hence, when a replica votes for b”, it should remember genericQC ← b”.justify. We remark that it is safe to update genericQC even when a One-Chain is not direct, so long as it is higher than the current genericQC. In the implementation code described in Section 6, we indeed update genericQC in this case.

If b” forms a Two-Chain, then the pre-commit phase of b” has succeeded. The replica should therefore update lockedQC ← b”.justify. Again, we remark that the lock can be updated even when a Two-Chain is not direct—safety will not break—and indeed, this is given in the implementation code in Section 6.

Finally, if b” forms a Three-Chain, the commit phase of b” has succeeded, and b becomes a committed decision.

Algorithm 3 shows the pseudocode for Chained HotStuff. The proof of safety given by [50] is similar to the one for Basic HotStuff. We require the QC in a valid node refers to its ancestor. For brevity, we assume the constraint always holds and omit checking in the code.

**Algorithm 3 Chained HotStuff protocol.**

1: procedure createLeaf(parent, cmd, qc)
2: b.parent ← branch extending with blanks from parent to height curView; b.cmd ← cmd; b.justify ← qc; return b

**GENERIC phase**

3: for curView ← 1, 2, 3, . . . do
4: as a leader // r = leader(curView)
5: wait for (n – f) new-view messages:
6: M ← (m if matchingMsg(m, new-view, curView – 1))
7: // M includes the previous leader new-view message, if received
6 IMPLEMENTATION

Because of its simplicity, we can easily turn Algorithm 3 into an event-driven-style specification that is almost like the code skeleton for a prototype implementation.

As shown in Algorithm 4, the code is further simplified and generalized by extracting the liveness mechanism from the body into a module named Pacemaker. Instead of the next leader always waiting for a generic QC at the end of the generic phase before starting its reign, this logic is delegated to the Pacemaker. A stable leader can skip this step and streamline proposals across multiple heights. Additionally, we relax the direct parent constraint for maintaining the highest generic QC and locked QC, while still preserving the requirement that the QC in a valid node always refers to its ancestor. The proof of correctness is similar to Chained HotStuff and we also defer it to the appendix of [50].

Data structures. Each replica $u$ keeps track of the following main state variables:

- $V[\cdot]$ mapping from a node to its votes.
- $v_{\textnormal{height}}$ height of last voted node.
- $b_{\textnormal{lock}}$ locked node (similar to locked QC).
- $b_{\textnormal{exec}}$ last executed node.
- $q_{c_{\textnormal{high}}}$ highest known QC (similar to generic QC) kept by a Pacemaker.
- $b_{\textnormal{leaf}}$ leaf node kept by a Pacemaker.

It also keeps a constant $b_0$, the same genesis node known by all correct replicas. To bootstrap, $b_0$ contains a hard-coded QC for itself, $b_{\textnormal{lock}}, b_{\textnormal{exec}}, b_{\textnormal{leaf}}$ are all initialized to $b_0$, and $q_{c_{\textnormal{high}}}$ contains the QC for $b_0$.

Pacemaker. A Pacemaker is a mechanism that guarantees progress after GST. It achieves this through two ingredients.

The first one is "synchronization", bringing all correct replicas, and a unique leader, into a common height for a sufficiently long period. The usual synchronization mechanism in the literature [15, 20, 25] is for replicas to increase the count of $\Delta$’s they spend at larger heights, until progress is being made. A common way to deterministically elect a leader is to use a rotating leader scheme
in which all correct replicas keep a predefined leader schedule and rotate to the next one when the leader is demoted.

Second, a Pacemaker needs to provide the leader with a way to choose a proposal that will be supported by correct replicas. As shown in Algorithm 5, after a view change, in onReceiveNewView, the new leader collects new-view messages sent by replicas through onNextSyncView to discover the highest QC to satisfy the second part of the condition in onReceiveProposal for liveness (Line 18 of Algorithm 4). During the same view, however, the incumbent leader will change the new node to the end of the leaf last proposed by itself, where no new-view message is received. Based on some application-specific heuristics (to wait until the previously proposed node gets a QC, for example), the current leader invokes onBeat to propose a new node carrying the command to be executed.

It is worth noting that even if a bad Pacemaker invokes onPropose arbitrarily, or selects a parent and a QC capriciously, and against any scheduling delays, safety is always guaranteed. Therefore, safety guaranteed by Algorithm 4 alone is entirely decoupled from liveness by any potential instantiation of Algorithm 5.

Algorithm 4 Event-driven HotStuff (for replica u).

```
1: procedure createLeaf(parent, cmd, qc, height)
2:   b = new child of node with height = height
3:   b.justify = true if qa > qa
4:   procedure update(b)
5:   b = b' with height = height
6:   procedure onCommit(b)
7:   procedure onNextSyncView
8:   procedure onReceiveProposal((msg, parent, cmd, qa, height))
9:   procedure onReceiveNewView((msg, view))
10:  procedure onCommit(b)
11:  procedure onBeat(cmd)
```

Algorithm 5 Code skeleton for a Pacemaker (for replica u).

```
1: function getLeader // ...specified by the application
2: procedure updateQCHigh(qc)
3: if qa > qa then
4:   qa = qa
5:   b = new leader
6:  procedure onBeat(cmd)
7:   if u = getLeader(b) then
8:     b = onPropose(b, cmd, qc)
9:  procedure onNextSyncView
10: send msg(new-view, qa) to getLeader()
11: procedure onReceiveNewView(msg(new-view, qa))
12: procedure updateQCHigh(qc)
```

Algorithm 6 UPDATE replacement for two-phase HotStuff.

```
1: procedure update(b)
2:   b = b' with justify = true
3:   procedure updateQCHigh(b, justify)
4: if qa > qa then
5:   qa = qa
6:  procedure onCommit(b)
7:   procedure onBeat(cmd)
```

Two-phase HotStuff variant. To further demonstrate the flexibility of the HotStuff framework, Algorithm 6 shows the two-phase variant of HotStuff. Only the update procedure is affected, a Two-Chain is required for reaching a commit decision, and a One-Chain determines the lock. As discussed above (Section 4.4), this two-phase variant loses Optimistic Responsiveness, and is similar to Tendermint/Casper. The benefit is fewer phases, while liveness may be addressed by incorporating in Pacemaker a wait based on maximum network delay.

Evaluation. Due to the space limitation, we defer our evaluation results to the longer paper [50]. There, we compare our implementation to BFT-SMaRt [13], a state-of-the-art implementation based on a two-phase PBFT variant. We show that even though three-phase HotStuff has an additional phase for its responsiveness and uses digital signatures universally (where BFT-SMaRt only uses MACs for votes), it still achieves similar latency, while being able to outperform BFT-SMaRt in throughput. It also scales better than BFT-SMaRt.

ACKNOWLEDGMENTS

We are thankful to Mathieu Baudet, Avery Ching, George Danesiz, François Garillot, Zekun Li, Ben Maurer, Kartik Nayak, Dmitri Perelman, and Ling Ren, for many deep discussions of HotStuff, and to Mathieu Baudet for exposing a subtle error in a previous version posted to the ArXiv of this manuscript.

REFERENCES


