

# Be Aware of Your Leaders

Shir Cohen<sup>1,2</sup>, Rati Gelashvili<sup>1</sup>, Lefteris Kokoris Kogias<sup>1,3</sup>, Zekun Li<sup>1</sup>, Dahlia Malkhi<sup>1</sup>, Alberto Sonnino<sup>1</sup>, and Alexander Spiegelman<sup>1</sup>

<sup>1</sup> Novi Research

<sup>2</sup> Technion

<sup>3</sup> IST Austria

**Abstract.** Advances in blockchains have influenced the State-Machine-Replication (SMR) world and many state-of-the-art blockchain-SMR solutions are based on two pillars: *Chaining* and *Leader-rotation*. A predetermined round-robin mechanism used for Leader-rotation, however, has an undesirable behavior: crashed parties become designated leaders infinitely often, slowing down overall system performance. In this paper, we provide a new Leader-Aware SMR framework that, among other desirable properties, formalizes a *Leader-utilization* requirement that bounds the number of rounds whose leaders are faulty in crash-only executions.

We introduce Carousel, a novel, reputation-based Leader-rotation solution to achieve Leader-Aware SMR. The challenge in adaptive Leader-rotation is that it cannot rely on consensus to determine a leader, since consensus itself needs a leader. Carousel uses the available on-chain information to determine a leader locally and achieves Liveness despite this difficulty. A HotStuff implementation fitted with Carousel demonstrates drastic performance improvements: it increases throughput over 2x in faultless settings and provided a 20x throughput increase and 5x latency reduction in the presence of faults.

## 1 Introduction

Recently, Byzantine agreement protocols in the eventually synchronous model such as Tendermint [5], Casper FFG [6], and HotStuff [21], brought two important concepts from the world of blockchains to the traditional State Machine Replication (SMR) [12] settings, *Leader-rotation* and *Chaining*. More specifically, these algorithms operate by designating one party as *leader* of each round to propose the next block of transactions that extends a *chained* sequence of blocks. Both properties depart from the approach used by classical protocols such as PBFT [7], Multi-Paxos [13] and Raft [17] (the latter two in benign settings). In those solutions, a stable leader operates until it fails and then it is replaced by a new leader. Agreement is formed on an immutable sequence of indexed (rather than chained) transactions, organized in slots.

Leader-rotation is important in a Byzantine setting, since parties should not trust each other for load sharing, reward management, resisting censoring of submitted transactions, or ordering requests fairly [11]. The advantage of Chaining

is that it simplifies the leader handover since in the common case the chain eliminates the need for new leaders to catch up with outcomes from previous slots.

In the permissioned SMR settings [1], most existing Leader-rotation mechanisms use a round-robin approach to rotate leaders [8, 20, 21]. This guarantees that honest parties get a chance to be leaders infinitely often, which is sufficient to drive progress and satisfy *Chain-quality* [10]. Roughly speaking, the latter stipulates that the number of blocks committed to the chain by honest parties is proportional to the honest nodes’ percentage. The drawback of such a mechanism is that it does not bound the number of faulty parties which are designated as leaders during an execution. This has a negative effect on latency even in crash-only executions, as each crashed leader delays progress. Similarly to XFT [14], we seek to improve the performance in such executions, while we also maintain Chain-quality to thwart Byzantine attacks.

In this paper, we propose a leader-rotation mechanism, Carousel, that enjoys both worlds. Carousel satisfies non-zero Chain-quality, and at the same time, bounds the number of faulty leaders in crash-only executions after the global stabilization time (GST), a property we call *Leader-utilization*. The Carousel algorithm leverages Chaining to execute purely locally using information available on the chain, avoiding any extra communication. To capture all requirements, we formalize a *Leader-Aware SMR* problem model, which alongside Agreement, Liveness and Chain-quality, also requires Leader-utilization. We prove that Carousel satisfied the Leader-Aware SMR requirements.

The high-level idea to satisfy Leader-utilization is to track active parties via the records of their participation (e.g. signatures) at the committed chain prefix and elect leaders among them. However, if done naively, the adversary can exploit this mechanism to violate Liveness or Chain-quality. The challenge is that there is no consensus on a committed prefix to determine a leader, since consensus itself needs a leader. Diverging local views on committed prefixes may be effectuated, for instance, by having a Byzantine leader reveal an updated head of the chain to a subset of the honest parties. Hence, Carousel may not have agreement on the leaders of some rounds, but nevertheless guarantees Liveness and Leader-utilization after GST.

To focus on our leader-rotation mechanism, we abstract away all other SMR components by defining an SMR framework. Similarly to [19], we capture the logic and properties of forming and certifying blocks of transactions in each round in a *Leader-based round (LBR)* abstraction, and rely on a Pacemaker abstraction [4, 15, 16] for round synchronization. We prove that when instantiated into this framework, Carousel yields a Leader-Aware SMR protocol. Specifically, we show (1) for Leader-utilization: at most  $O(f^2)$  faulty leader may be elected in crash-only executions (after GST); and (2) for Chain-quality: one out of  $O(f)$  blocks is authored by an honest party in the worst-case.

We provide an implementation of Carousel in a HotStuff-based system and an evaluation that demonstrates a significant performance improvement. Specifically, we get over 2x throughput increase in faultless settings, and 20x throughput

increase and 5x latency reduction in the presence of faults. Our mechanism is adopted in the most recent version of DiemBFT [20], a deployed HotStuff-based system.

## 2 Model and Problem Definition

We consider a message-passing model with a set of  $n$  parties  $\Pi = \{p_1, \dots, p_n\}$ , out of which  $f < \frac{n}{3}$  are subject to failures. A party is *crashed* if it halts prematurely at some point during an execution. If it deviates from the protocol it is *Byzantine*. An *honest* party never crashes or becomes Byzantine. We say that an execution is *crash-only* if there are no Byzantine failures therein.

For the theoretical analysis we assume an eventually synchronous communication model [9] in which there is a global stabilization time (GST) after which the network becomes synchronous. That is, before GST the network is completely asynchronous, while after GST messages arrive within a known bounded time, denoted as  $\delta$ .

As we later describe, we abstract away much of the SMR implementation details by defining and using primitives. Therefore, our Leader-rotation solution is model agnostic and the adversarial model depends on the implementation choices for those primitives.

### 2.1 Leader-Aware SMR

In this section we introduce some notation and then define the Leader-Aware SMR problem. Roughly speaking, Leader-Aware SMR captures the desired properties of the Leader-rotation mechanism in SMR protocols that are leader-based.

An SMR protocol consists of a set of parties aiming to maintain a growing chain of *blocks*. Parties participate in a sequence of rounds, attempting to form a block per round. In Leader-Aware SMR, each round is driven by a leader. We capture these rounds via the Leader-based round (LBR) abstraction defined later.

A block consists of transactions and the following meta-data:

- A (cryptographic) link to a *parent* block. Thus, each block implicitly defines a chain to the genesis block.
- A round number in which the block was formed.
- The author id of the party that created the block.
- A certificate that (cryptographically) proves that  $2f + 1$  parties endorsed the block in the given round and with the given author. We assume that it is possible to obtain the set of  $2f + 1$  endorsing parties<sup>4</sup>.

Note that having a round number and the author id as a part of the block is not strictly necessary, but they facilitate formalization of properties and analysis.

<sup>4</sup> This can be achieved by multi-signature schemes which are practically as efficient as threshold signatures [3].

For example, an *honest block* is defined as a block authored by an honest party and a *Byzantine block* is a block authored by a Byzantine party.

We assume a predicate  $\text{certified}(B, r) \in \{\text{true}, \text{false}\}$  that locally checks whether the block has a valid certificate, i.e. it has  $2f+1$  endorsements for round  $r$ . If  $\text{certified}(B, r) = \text{true}$  we say that  $B$  is a *certified* block of round  $r$ . When clear from context, we say that  $B$  is *certified* without explicitly mentioning the round number.

An SMR protocol does not terminate, but rather continues to form blocks. Each block  $B$  determines its *implied* chain starting from  $B$  to the genesis block via the parent links. We use notation  $B \rightarrow B'$ , saying  $B'$  *extends*  $B$ , if block  $B$  is on  $B'$ 's implied chain. Honest parties can *commit* blocks in some rounds (but usually not all). A committed block indirectly commits its implied chain. An SMR protocol must satisfy the following:

**Definition 1 (Leader-Aware SMR).**

- **Liveness:** *An unbounded number of blocks are committed by honest parties.*
- **Agreement:** *If an honest party  $p_i$  has committed a block  $B$ , then for any block  $B'$  committed by any honest party  $p_j$  either  $B \rightarrow B'$  or  $B' \rightarrow B$ .*
- **Chain-quality:** *For any block  $B$  committed by an honest party  $p_i$ , the proportion of Byzantine blocks on  $B$ 's implied chain is bounded.*
- **Leader-Utilization:** *In crash-only executions, after GST, the number of rounds  $r$  for which no honest party commits a block formed in  $r$  is bounded.*

The first two properties are common to SMR protocols. While most SMR algorithms satisfy the above mentioned Liveness condition, a stronger Liveness property can be defined, requiring that each honest party commits an unbounded number of blocks. This property can be easily be achieved by an orthogonal forwarding mechanism, where each honest leader that creates a block explicitly sends it to all other parties. A notion of Chain-quality that bounds the adversarial control over chain contents was first suggested by Garay et al. [10]. We introduce the Leader-utilization property to capture the quality of the Leader-rotation mechanism in crash-only executions.

### 3 Leader-Aware SMR: The Framework

In order to isolate the Leader-rotation problem in Leader-Aware SMR protocols, we abstract away the remaining logic into two components. First, similar to [18, 19] we capture the logic to form and commit blocks by the *Leader-based round (LBR)* abstraction (Section 3.1). We follow [4, 16] and capture round synchronization by the Pacemaker abstraction (Section 3.2). These two abstractions can be instantiated with known implementations from existing SMR protocols.

In Section 3.3 we define the core API for Leader-rotation and combine it with the above components to construct an SMR protocol. In Section 4 we present a Leader-rotation algorithm that can be easily computed based on locally available information and makes the construction a Leader-Aware SMR.

### 3.1 Leader-based round (LBR)

The LBR abstraction exposes to each party  $p_i$  an API to invoke  $LBR(r, \ell)$ , where  $r \in \mathbb{N}$  is a round number and  $\ell$  is the leader of round  $r$  according to party  $p_i$ . Intuitively, a leader-based round captures an attempt by parties to certify and commit a block formed by the leader<sup>5</sup> - which naturally requires sufficiently many parties to agree on the identity of the leader. We assume that non-Byzantine parties can only endorse a block  $B$  with round number  $r$  and author  $\ell$  by calling  $LBR(r, \ell)$ .

Every LBR invocation returns within  $\Delta_l > c\delta$  time, where  $c$  depends on the specific LBR implementation (i.e., each round requires a causal chain of  $c$  messages to complete). That is,  $\Delta_l$  captures the inherent timeouts required for eventually synchronous protocols. We say that round  $r$  has  $k \leq n$  *LBR-synchronized*( $\ell$ ) invocations if  $k$  honest parties invoke  $LBR(r, \ell)$  after GST and within  $\Delta_l - c\delta$  time of each other with the same party  $\ell$ <sup>6</sup>.

The return value of an LBR invocation in round  $r$  is always a block with a round number  $r' \leq r$ . The intention is for LBR invocations to return gradually growing committed chains. Occasionally, there is no progress, in which case the invocations are allowed to return a committed block whose round  $r'$  is smaller than  $r$ . Formally, the output from LBR satisfies the following properties:

**Definition 2 (LBR).**

- **Endorsement:** For any block  $B$  and round  $r$ , if  $\text{certified}(B, r) = \text{true}$ , then the set of endorsing parties of  $B$  contains  $2f + 1$  parties.
- **Agreement:** If  $B$  and  $B'$  are certified blocks that are each returned to an honest party from an LBR invocation, then either  $B \rightarrow B'$  or  $B' \rightarrow B$ .
- **Progress:** If there are  $k \geq 2f + 1$  LBR-synchronized( $\ell$ ) invocations at round  $r$  and  $\ell$  is honest, then they all return a certified  $B$  with round number  $r$  authored by  $\ell$ .
- **Blocking:** If a non-Byzantine party  $\ell$  never invokes  $LBR(r, \ell)$ , then no  $LBR(r, \ell)$  invocation may return a certified block formed in round  $r$ .
- **Reputation:** If a non-Byzantine party  $p$  never invokes LBR for round  $r$ , then any certified block  $B$  with round number  $r$  does not contain  $p$  among its endorsers.

The LBR definition intends to capture just the key properties required for round abstraction in SMR protocols but leaves room for various interesting behavior. For example, if the progress preconditions are not met at round  $r$ , then some honest parties may return a block  $B$  for round  $r$  while others do not. Moreover, in this case the adversary can *hide* certified blocks from honest parties and reveal them at any point via the LBR return values.

<sup>5</sup> Existing SMR protocols may have separate rounds (and even leaders) for forming and committing blocks, but this distinction is not relevant for the purposes of the paper and LBR abstraction is defined accordingly.

<sup>6</sup> LBR-synchronized requires that the corresponding execution intervals have a shared intersection lasting  $\geq c\delta$  time.

### 3.2 The Pacemaker

The Pacemaker [4, 15, 16] component is a commonly used abstraction, which ensures that, after GST, parties are synchronized and participate in the same round long enough to satisfy the LBR progress. We assume the following:

**Definition 3 (Pacemaker).** *The Pacemaker eventually produces  $\text{new\_round}(r)$  notifications at honest parties for each round  $r$ . Suppose for some round  $r$  all  $\text{new\_round}(r)$  notifications at non-Byzantine parties occur after GST, the first of which occurs at time  $T_f$ , and the last of which occurs at time  $T_l$ . Then no non-Byzantine party receives a  $\text{new\_round}(r + 1)$  notification before  $T_l + \Delta_p$  and  $T_l - T_f \leq \delta$ . The Pacemaker can be instantiated with any parameter  $\Delta_p > 0$ .*

To combine the LBR and Pacemaker components in to an SMR protocol in Section 3.3 we fix  $\Delta_p = \Delta_l$ . Note that by using the above definition, the resulting protocol is not responsive since parties wait  $\Delta_p$  before advancing rounds. This can easily be fixed by using a more general Pacemaker definitions from [4, 15, 16]. However, we chose the simplified version above for readability purposes since the Pacemaker is orthogonal to the thesis of our paper.

### 3.3 Leader-rotation - the missing component

In Algorithm 1 we show how to combine the LBR and Pacemaker abstractions into a leader-based SMR protocol. The missing component is the Leader-rotation mechanism, which exposes an  $\text{choose\_leader}(r, B)$  API. It takes a round number  $r \in \mathbb{N}$  and a block  $B$  and returns a party  $p \in \Pi$ . The  $\text{choose\_leader}$  procedure is locally computed by each honest party at the beginning of every round.

The Agreement property of Algorithm 1 follows immediately from the Agreement property of LBR, regardless of  $\text{choose\_leader}$  implementation. In Appendix A we prove that Algorithm 1 satisfies liveness as long as all honest parties follow the same  $\text{choose\_leader}$  procedure and that this procedure returns the same honest party at all of them infinitely often. In the next section we instantiate Algorithm 1 with Carousel: a specific  $\text{choose\_leader}$  implementation to obtain a Leader-Aware SMR protocol. That is, we prove that Algorithm 1 with Carousel satisfies liveness, Chain-quality, and Leader-utilization.

---

#### Algorithm 1 Constructing SMR: code for party $p_i$

---

```

1:  $\text{commit\_head} \leftarrow \text{genesis}$ 
2: upon  $\text{new\_round}(r)$  do
3:    $\text{leader} \leftarrow \text{choose\_leader}(r, \text{commit\_head})$ 
4:    $B \leftarrow \text{LBR}(r, \text{leader})$ 
5:   if  $\text{commit\_head} \rightarrow B$  then
6:      $\text{commit } B \triangleright$  all blocks in  $B$ 's implied chain that were not yet committed.
7:      $\text{commit\_head} \leftarrow B$ 

```

---

## 4 Carousel: A Novel Leader-Rotation Algorithm

In this section, we present Carousel—our Leader-rotation mechanism. The pseudo-code is given in [Algorithm 2](#), which combined with [Algorithm 1](#) allows to obtain the first Leader-Aware SMR protocol.

We use reputation to avoid crashed leaders in crash-only executions. Specifically, at the beginning of round  $r$ , an honest party checks if it has committed a block  $B$  with round number  $r - 1$ . In this case, the set of endorsers of  $B$  are guaranteed to not have crashed by round  $r$ . For Chain-quality purposes, the  $f$  latest authors of committed blocks are excluded from the set of endorsers, and a leader is chosen deterministically from the remaining set.

If an honest party has not committed a block with round number  $r - 1$ , it uses a round-robin fallback scheme to elect the round  $r$  leader. Notice that different parties may or may not have committed a block with round number  $r - 1$  before round  $r$ . In fact, the adversary has multiple ways to cause such divergence, e.g. Byzantine behavior, crashes or message delays. As a result parties can disagree on the leader identity, and potentially compromise liveness. We prove, however, that Carousel satisfies liveness, as well as leader utilization and Chain-quality. Specifically, we show that (1) the number of rounds  $r$  for which no honest party commits a block formed in  $r$  is bounded by  $O(f^2)$ ; and at least one honest block is committed  $5f + 2$  rounds. The argument is non-trivial, since for example, we need to show that the adversary cannot selectively alternate the fallback and reputation schemes to control the Chain-quality.

---

**Algorithm 2** Leader-rotation: code for party  $p_i$

---

```

8: procedure choose_leader( $r, commit\_head$ )
9:    $last\_authors \leftarrow \emptyset$ 
10:  if  $commit\_head.round\_number \neq r - 1$  then
11:    return ( $r \bmod n$ ) ▷ round-robin fallback
12:   $active \leftarrow commit\_head.endorsers$ 
13:   $block \leftarrow commit\_head$ 
14:  while  $last\_authors < f \wedge block \neq genesis$  do
15:     $last\_authors \leftarrow last\_authors \cup \{block.author\}$ 
16:     $block \leftarrow block.parent$ 
17:   $leader\_candidates \leftarrow active \setminus last\_authors$ 
18:  return  $leader\_candidates.pick\_one()$  ▷ deterministically pick from the set

```

---

### 4.1 Correctness

**Leader-Utilization.** In this section, we are concerned with the protocol efficiency against crash failures. We consider time after GST, and at most  $f$  parties that may crash during the execution but follow the protocol until they crash (i.e., non-Byzantine). We say that a party  $p$  crashes in round  $r$  if  $r + 1$  is the

minimal number for which  $p$  does not invoke  $LBR$  in [line 4](#). Accordingly, we say that a party is *alive* at all rounds before it crashes. In addition, we say that a round  $r$  occurs after GST if all `new_round` ( $r$ ) notifications at honest parties occur after GST.

We start by introducing an auxiliary lemma which extends the LBR Progress property for crash-only executions. Since in a crash-only case faulty parties follow the protocol before they crash, honest parties cannot distinguish between an honest leader and an alive leader that has not crashed yet. Hence, the LBR Progress property hold even if the leader crashes later in the execution. Formal proof of the following technical lemma, using indistinguishability arguments, appears in [Appendix A](#).

**Lemma 1.** *In a crash-only execution, let  $r$  be a round with  $k \geq 2f + 1$   $LBR\text{-synchronized}(\ell)$  invocations, such that  $\ell$  is alive at round  $r$ , then these  $k$  invocations return a certified  $B$  with round number  $r$  authored by  $\ell$ .*

Furthermore, if no party crashes in a given round and the preconditions of the adapted LBR Progress conditions are met a block is committed in that round and another alive leader is chosen.

**Lemma 2.** *If the preconditions of [Lemma 1](#) hold and no party crashes in round  $r$ , then  $k \geq 2f + 1$  honest parties commit a block for round  $r$  and return the same leader  $\ell'$  at [line 3](#) of round  $r + 1$  and  $\ell'$  is alive at round  $r$ .*

*Proof.* By [Lemma 1](#),  $k$  honest parties return from  $LBR(r, \ell)$  with a certified block  $B$  with round number  $r$  authored by  $\ell$ . Then, since `commit_head`  $\rightarrow B$ , they all commit  $B$  at [line 6](#) of round  $r + 1$ . By the LBR Reputation property, the set of  $B$ 's endorsers does not include parties that crashed in rounds  $< r$ . Since no party crashes in round  $r$ ,  $B$ 's endorsers are all alive in round  $r$ . Since these  $2f + 1$  parties each committed block  $B$  with round number  $r$ , in `choose_leader` in [Algorithm 1](#), they all use the reputation scheme ([line 18](#)) to choose the round  $r + 1$  leader, that we showed is alive at round  $r$ .

Next, we utilize the latter to prove that in a round with no crashes, it is impossible for a minority of honest parties to return with a certified block from an LBR instance. Namely, either no honest party returns a block, or at least  $2f + 1$  of them do.

**Lemma 3.** *In a crash-only execution, let  $r$  be a round after GST in which no party crashes. If one honest party returns from LBR with a certified block  $B$  with round number  $r$ , then  $2f + 1$  honest parties return with  $B$ .*

*Proof.* Assume an honest party returns a certified block  $B$  with round number  $r$  after invoking  $LBR(r, \ell)$ . By the LBR Blocking property,  $\ell$  itself must have invoked  $LBR(r, \ell)$  and by assumption it was *alive* at round  $r$ . By the LBR Endorsement property, the set of endorsing parties of  $B$  contains  $2f + 1$  parties. Since we consider a crash-only execution, it follows by assumption that  $2f + 1$  party called  $LBR(r, \ell)$ . Due to the use of Pacemaker, these calls are  $LBR\text{-synchronized}(\ell)$  invocations. Finally, by [Lemma 1](#) all these calls return a certified  $B$  with round number  $r$  authored by  $\ell$ .



We prove that in a window of  $f + 2$  rounds without crashes, there must be a round with the sufficient conditions for a block to be committed for that round.

**Lemma 4.** *In a crash-only execution, let  $R$  be a round after GST such that no party crashes between rounds  $R$  and  $R + f + 2$  (including). There exists a round  $R \leq r \leq R + f + 2$  for which there are  $2f + 1$  LBR-synchronized( $\ell$ ) invocations with a leader  $\ell$  that is alive at round  $r$ .*

*Proof.* First, let us consider the LBR invocations for round  $R$ . By Lemma 3, if one honest party returns with a block  $B$  with round number  $R$ , then  $2f + 1$  honest parties return with  $B$ , commit it and update `commit_head` accordingly (line 7). In this case, there are  $2f + 1$  `choose_leader( $R + 1, B$ )` invocations, which all return at line 18. Otherwise, no party return a block with round number  $R$ , and thus they all return at line 11. By the code and since a block implies a unique chain, in both cases  $2f + 1$  honest parties return the same leader  $\ell$  in `choose_leader( $R + 1, B$ )` (either by reputation or round-robin). By the Pacemaker guarantees and since  $R + 1$  occurs after GST, there are at least  $2f + 1$  LBR-synchronized( $\ell$ ) invocations. If  $\ell$  is alive at round  $R + 1$ , we are done. Otherwise,  $\ell$  must have been crashed before round  $R$  by the alive definition and lemma assumptions. Thus, by the LBR Blocking property no honest party commits a block for round  $R$  and they all choose the same leader for the following round at line 11. The lemma follows by applying the above argument for  $R + f + 2 - R + 1 = f + 1$  rounds.

Finally, we bound by  $O(f^2)$  the total number of rounds in a crash-only execution for which no honest party commits a block:

**Lemma 5.** *Consider a crash-only execution. After GST, the number of rounds  $r$  for which no honest party commits a block formed in  $r$  is bounded by  $O(f^2)$ .*

*Proof.* Consider a crash-only execution and let  $R_1, R_2, \dots, R_k$  the rounds after GST in which parties crash ( $k \leq f$ ). For ease of presentation we call a round for which no honest party commits a block formed in  $r$  a *skipped* round. We prove that the number of skipped rounds between  $R_i$  and  $R_{i+1}$  for  $1 \leq i < k$  is bounded. If  $R_{i+1} - R_i < f + 4$ , then there are at most  $f + 4$  rounds and hence at most  $f + 4$  skipped rounds. Otherwise, we show that at most  $f + 2$  rounds are skipped between rounds  $R_i$  and  $R_{i+1}$ .

First, by Lemma 4, there exists a round  $R_i < R_i + 1 \leq r \leq R_i + 1 + f + 2 < R_{i+1}$  for which there are  $2f + 1$  LBR-synchronized( $\ell$ ) invocations with a leader  $\ell$  that is *alive* at round  $r$ . By Lemma 2, since no party crashes in round  $r$ ,  $2f + 1$  honest parties return the same leader  $\ell'$  at line 3 of round  $r + 1$  and  $\ell'$  is alive at round  $r$ . Since no party crashes at round  $r + 1$  as well (because  $R_{i+1} - R_i \geq f + 4$ ),  $\ell'$  is alive at round  $r + 1$ . By the Pacemaker guarantees and since we consider rounds after GST, we conclude that there are at least  $2f + 1$  LBR-synchronized( $\ell'$ ) invocations for round  $r + 1$ . By Lemma 2 applied again for round  $r + 1$ ,  $2f + 1$  honest parties commit a block for round  $r + 1$ . Thus, round  $r + 1$  is not *skipped*. We repeat the same arguments until round  $R_{i+1}$ , and

conclude that in each of these rounds a block is committed. Hence, the rounds that can possibly be skipped between  $R_i$  and  $R_{i+1}$  are  $R_i \leq r' < r$ . Thus there are  $O(f)$  skipped round between  $R_i$  and  $R_{i+1}$ . For  $R_k$  we use similar arguments but since no party crashes after  $R_k$ , we apply [Lemma 2](#) indefinitely. We similarly conclude that there are  $O(f)$  skipped rounds after  $R_k$ . All in all, since  $k \leq f$ , we get  $O(f^2)$  skipped rounds.

We immediately conclude the following:

**Corollary 1.** *Algorithm 1 with Algorithm 2 satisfies Leader-utilization.*

**Chain-Quality.** For the purposes of the Chain-quality proof, we say that a block is committed when some honest party commits it. We say that a block  $B$  with round number  $r$  is *immediately committed* if an honest party commits  $B$  in round  $r$ . When we refer to a leader elected in of [Algorithm 2](#) from the round-robin mechanism we mean [line 11](#), and when we refer to a leader elected from the reputation mechanism, we mean [line 18](#).

We begin by showing that each round assigned with an honest round-robin leader implies a committed block in that round or the one that precedes it (not necessarily an honest block).

**Lemma 6.** *Let  $r$  be a round after GST such that  $p_i = (r \bmod n)$  is honest. Then, either Byzantine block with round number  $r - 1$  or an honest block with round number  $r - 1$  or  $r$  is immediately committed.*

*Proof.* If a block is immediately committed with round number  $r - 1$  then we are done. Otherwise, no honest party commits a block with round number  $r - 1$  in round  $r - 1$ , and they all elect the round  $r$  leader  $\ell$  using the round-robin mechanism. By the assumption,  $\ell$  is honest.

By the Pacemaker, all honest invocations of  $LBR(r, \ell)$  in [line 4](#) are LBR-synchronized( $\ell$ ). Since there are at least  $2f + 1$  honest parties, by the LBR Progress property, all honest invocations return the same certified block  $B$  with round number  $r$  authored by  $\ell$ . Then, the honest parties commit  $B$  at [line 6](#).

If there are two consecutive rounds assigned with honest round-robin leaders and in addition the last  $f$  committed blocks are Byzantine, then an honest block follows, as proven in the following lemma.

**Lemma 7.** *Let  $r'$  be a round after GST such that  $p_i = (r' \bmod n)$  and  $p_j = (r' + 1 \bmod n)$  are honest. Suppose  $f$  blocks with round numbers in  $[r, r')$  with different Byzantine authors are committed. For a block  $B$  with round number  $r'$  or  $r' + 1$  that is immediately committed, there is an honest block with round number  $[r, r' + 1]$  on  $B$ 's implied chain.*

*Proof.* By the LBR endorsement assumption and property, the author of block  $B$  should be either a reputation-based, or a round-robin leader of round  $r'$  or  $r' + 1$ . If it is a round-robin leader, then by the lemma assumption, the leader

is honest and since  $B$  is the head of its implied chain, the proof is complete. Thus, in the following we assume that  $B$ 's author is a reputation-based leader. By the SMR Agreement property and the lemma assumption,  $B$ 's implied chain contains  $f$  blocks with different Byzantine authors and rounds numbers in  $[r, r')$ . By the code of the reputation-based mechanism, either all  $f$  byzantine authors are excluded from the *leader\_candidates* which implies that  $B$  has an honest author, or that there is an honest block with round number in  $[r, r')$  on  $B$ 's implied chain.

Lastly, the following lemma proves that in any window of  $5f + 2$  rounds an honest block is committed.

**Lemma 8.** *Let  $r$  be a round after GST. At least one honest block is committed with a round number in  $[r, r + 5f + 2]$ .*

*Proof.* Suppose for contradiction that no honest block with round number in  $[r, r + 5f + 2]$  is committed. There are at least  $f$  rounds  $r'$  in  $[r, r + 3f + 1)$ , such that rounds  $r' - 1$  and  $r'$  are allocated an honest leader by the round-robin mechanism. By Lemma 6, a block with round number  $r' - 1$  or  $r'$  is immediately committed. Due to Lemma 6 and the contradiction assumption, for any such round  $r'$ , a Byzantine block with round number  $r' - 1$  is immediately committed. Since  $r' - 1$  has an honest round-robin leader, the block must be committed from the reputation mechanism.

It follows that  $f$  Byzantine blocks with round numbers in  $[r, r + 3f + 1)$  are immediately committed from the reputation mechanism, and consequently, they all must have different authors. Note that there exists  $r' \in [r + 3f + 1, r + 5f + 2)$  (in a window of  $2f + 1$  rounds), such that the round-robin mechanism allocates honest leaders to rounds  $r'$  and  $r' + 1$ . By Lemma 6, a block  $B$  with round number  $r'$  or  $r' + 1$  is immediately committed. Lemma 7 concludes the proof.

We conclude the following:

**Corollary 2.** *Algorithm 1 with Algorithm 2 satisfies Chain-quality and Liveness.*

Taken jointly, Corollary 1, Corollary 2, and the Agreement property proved in Section 3.3 yield the following theorem:

**Theorem 1.** *Algorithm 1 with Algorithm 2 implements Leader-Aware SMR.*

## 5 Implementation

We implement Carousel on top of a high-performance open-source implementation of HotStuff<sup>7</sup> [21]. We selected this implementation because it implements

<sup>7</sup> <https://github.com/asonnino/hotstuff>

a Pacemaker [21], contrarily to the implementation used in the original HotStuff paper<sup>8</sup>. Additionally, it provides well-documented benchmarking scripts to measure performance in various conditions, and it is close to a production system (it provides real networking, cryptography, and persistent storage). It is implemented in Rust, uses Tokio<sup>9</sup> for asynchronous networking, ed25519-dalek<sup>10</sup> for elliptic curve based signatures, and data-structures are persisted using RocksDB<sup>11</sup>. It uses TCP to achieve reliable point-to-point channels, necessary to correctly implement the distributed system abstractions. By default, this HotStuff implementation uses traditional round-robin to elect leaders; we modify its `LeaderElector` module to use Carousel instead. Implementing our mechanism requires to add less than 200 LOC, and does not require any extra protocol message or cryptographic tool. We are open-sourcing Carousel<sup>12</sup> along with any measurements data to enable reproducible results<sup>13</sup>.

## 6 Evaluation

We evaluate the throughput and latency of HotStuff equipped Carousel through experiments on Amazon Web Services (AWS). We then show how it improves over the baseline round-robin leader-rotation mechanism. We particularly aim to demonstrate that Carousel (i) introduces no noticeable performance overhead when the protocol runs in ideal conditions (that is, all parties are honest) and with small committees, and (ii) drastically improves both latency and throughput in the presence of crash-faults. Note that evaluating BFT protocols in the presence of Byzantine faults is still an open research question [2].

We deploy a testbed on AWS, using `m5.8xlarge` instances across 5 different AWS regions: N. Virginia (us-east-1), N. California (us-west-1), Sydney (ap-southeast-2), Stockholm (eu-north-1), and Tokyo (ap-northeast-1). Parties are distributed across those regions as equally as possible. Each machine provides 10Gbps of bandwidth, 32 virtual CPUs (16 physical core) on a 2.5GHz, Intel Xeon Platinum 8175, 128GB memory, and run Linux Ubuntu server 20.04.

In the following sections, each measurement in the graphs is the average of 5 independent runs, and the error bars represent one standard deviation. Our baseline experiment parameters are: 10 honest parties, a block size of 500KB, a transaction size of 512B, and one benchmark client per party submitting transactions at a fixed rate for a duration of 5 minutes. We then crash and vary the number of parties through our experiments to illustrate their impact on performance. The leader timeout value is set to 5 seconds for committees of 10 and 20, and increased to 10 seconds for committees of 50. When referring to *latency*, we mean the time elapsed from when the client submits the transaction to when

<sup>8</sup> <https://github.com/hot-stuff/libhotstuff>

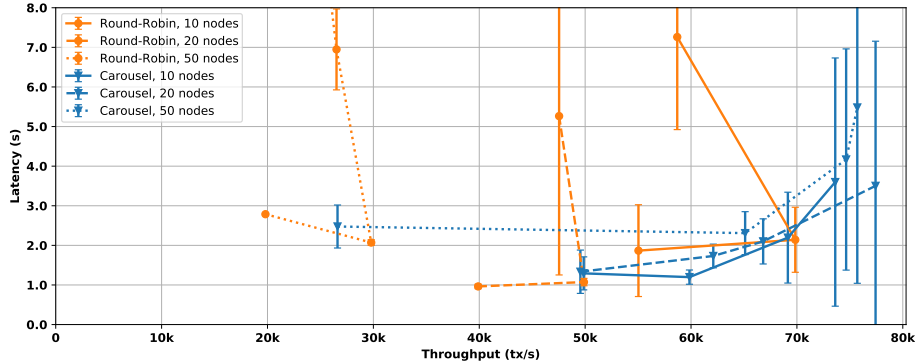
<sup>9</sup> <https://tokio.rs>

<sup>10</sup> <https://github.com/dalek-cryptography/ed25519-dalek>

<sup>11</sup> <https://rocksdb.org>

<sup>12</sup> <https://github.com/asonnino/hotstuff/tree/leader-reputation>

<sup>13</sup> <https://github.com/asonnino/hotstuff/tree/leader-reputation/data>



**Fig. 1.** Comparative throughput-latency performance of HotStuff equipped with Carousel and with the baseline round-robin. WAN measurements with 10, 20, 50 parties. No party faults, 500KB maximum block size and 512B transaction size.

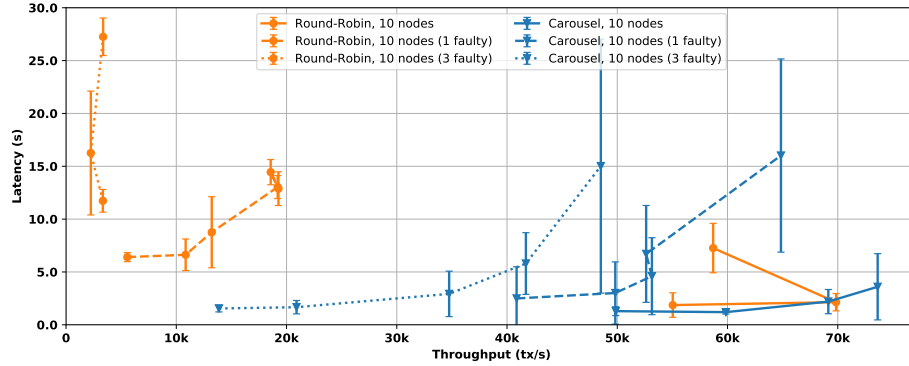
the transaction is committed by one party. We measure it by tracking sample transactions throughout the system.

### 6.1 Benchmark in Ideal Conditions

Figure 1 depicts the performance of HotStuff with both Carousel and the baseline round-robin running with 10, 20 and 50 honest parties. For small committees (10 parties), the performance of the baseline round-robin HotStuff is similar to HotStuff equipped with Carousel. We observe a peak throughput around 70,000 tx/s with a latency of around 2 seconds. This illustrates that the extra code required to implement Carousel has negligible overhead and does not degrade performance when the committee is small. When increasing the committee (to 20 and 50 parties), HotStuff with Carousel greatly outperforms the baseline: the bigger the committee, the bigger the performance improvement. With 50 nodes, the throughput of our mechanism based HotStuff increases by over 2x with respect to the baseline, and remains comparable to the 10-parties testbed. After a few initial timeouts, Carousel has the benefit to focus on electing performant leaders. Leaders on more remote geo-locations that are typically slower are elected less often, the protocol is thus driven by the most performant parties. Latency is similar for both implementations and around 2-3 seconds.

### 6.2 Performance under Faults

Figure 2 depicts the performance of HotStuff with both Carousel and the baseline round-robin when a set of 10 parties suffers 1 or 3 crash-faults (the maximum that can be tolerated). The baseline round-robin HotStuff suffers a massive degradation in throughput as well as a dramatic increase in latency. For three faults, the throughput of the baseline HotStuff drops over 30x and its latency increases 5x compared to no faults. In contrast, HotStuff equipped with Carousel maintains a good level of throughput: our mechanism does not elect crashed leaders, the protocol continues to operate electing leaders from the remaining



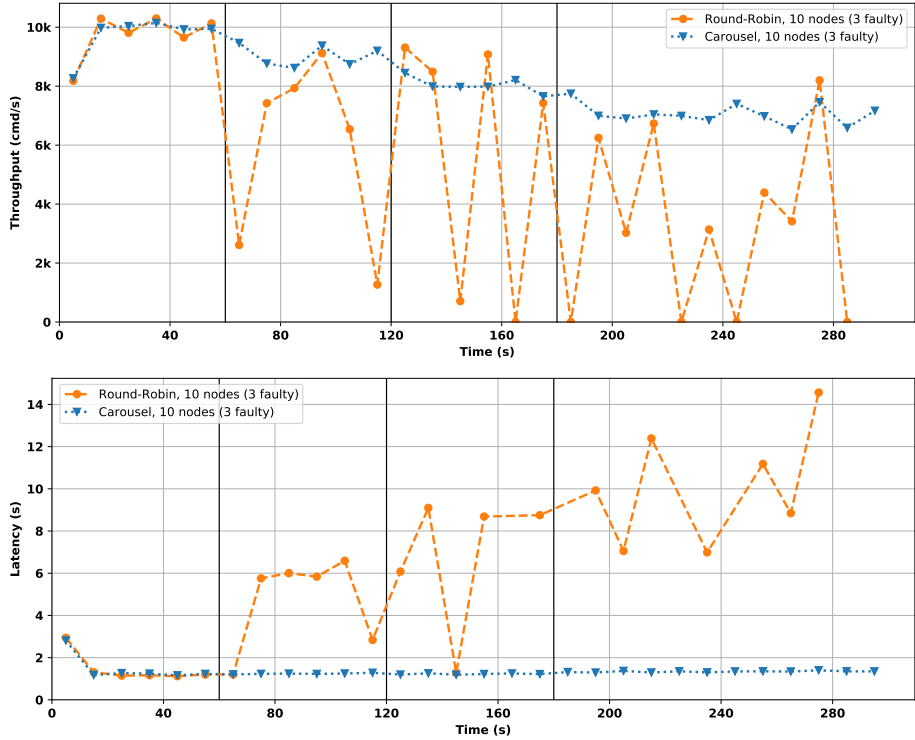
**Fig. 2.** Comparative throughput-latency performance of HotStuff equipped with Carousel and with the baseline round-robin. WAN measurements with 10 parties. Zero, one and three party faults, 500KB maximum block size and 512B transaction size.

active parties and is not overly affected by the faulty ones. The reduction in throughput is in great part due to losing the capacity of faulty parties. When operating with 3 faults, Carousel provides a 20x throughput increase and about 5x latency reduction with respect to the baseline round-robin.

Figure 3 depicts the evolution of the performance of HotStuff with both Carousel and the baseline round-robin when gradually crashing nodes through time. For roughly the first minute, all parties are honest; we then crash 1 party (roughly) every minute until a maximum of 3 parties are crashed. The input transaction rate is fixed to 10,000 tx/s throughout the experiment. Each data point is the average over intervals of 10 seconds. For roughly the first minute (when all parties are honest), both systems perform ideally, timely committing all input transactions. Then, as expected, the baseline round-robin HotStuff suffers from temporary throughput losses when a crashed leader is elected. Similarly, its latency increases with the number of faulty parties, and presents periods where no transactions are committed at all. In contrast, HotStuff equipped with Carousel delivers a stable throughput by quickly detecting and eliminating crashed leaders. Its latency is barely affected by the faulty parties. This graph clearly illustrates how Carousel allows HotStuff to deliver a seamless client experience even in the presence of faults.

## 7 Conclusions

Leader-rotations mechanisms in chaining-based SMR protocols were previously overlooked. Existing approaches degraded performance by keep electing faulty leaders in crash-only executions. We captured the practical requirement of leader-rotation mechanism via a Leader-utilization property, use it define the Leader-Aware SMR problem, and described an algorithm that implements it. That is, we presented a locally executed algorithm to rotate leaders that achieves both: Leader-utilization in crash-only executions and Chain-quality in Byzantine ones. We evaluated our mechanism in a Hotstuff-based open source system



**Fig. 3.** Comparative performance of HotStuff equipped with Carousel and with the baseline round-robin when gradually crashing nodes through time. The input transactions rate is fixed to 10,000 tx/s; 1 party (up to a maximum of 3) crashes roughly every minute. WAN measurements with 10 parties, 500KB maximum block size and 512B transaction size.

and demonstrated drastic performance improvements in both throughput and latency compared to the round-robin baseline.

## References

1. Elli Androulaki, Artem Barger, Vita Bortnikov, Christian Cachin, Konstantinos Christidis, Angelo De Caro, David Enyeart, Christopher Ferris, Gennady Laventman, Yacov Manevich, et al. Hyperledger fabric: a distributed operating system for permissioned blockchains. In *Proceedings of the thirteenth EuroSys conference*, pages 1–15, 2018.
2. Shehar Bano, Alberto Sonnino, Andrey Chursin, Dmitri Perelman, and Dahlia Malkhi. Twins: White-glove approach for bft testing. *arXiv preprint arXiv:2004.10617*, 2020.
3. Dan Boneh, Manu Drijvers, and Gregory Neven. The modified BLS multi-signature construction. <https://crypto.stanford.edu/~dabo/pubs/papers/BLSmultisig.html>, 2018.
4. Manuel Bravo, Gregory Chockler, and Alexey Gotsman. Making byzantine consensus live. In *34th International Symposium on Distributed Computing (DISC 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.
5. Ethan Buchman. *Tendermint: Byzantine fault tolerance in the age of blockchains*. PhD thesis, 2016.
6. Vitalik Buterin and Virgil Griffith. Casper the friendly finality gadget.
7. Miguel Castro, Barbara Liskov, et al. Practical byzantine fault tolerance. In *OSDI*, volume 99, pages 173–186, 1999.
8. Benjamin Y Chan and Elaine Shi. Streamlet: Textbook streamlined blockchains. In *Proceedings of the 2nd ACM Conference on Advances in Financial Technologies*, pages 1–11, 2020.
9. Cynthia Dwork, Nancy Lynch, and Larry Stockmeyer. Consensus in the presence of partial synchrony. *Journal of the ACM (JACM)*, 35(2):288–323, 1988.
10. Juan Garay, Aggelos Kiayias, and Nikos Leonardos. The bitcoin backbone protocol: Analysis and applications. In *Annual international conference on the theory and applications of cryptographic techniques*, pages 281–310. Springer, 2015.
11. Mahimna Kelkar, Fan Zhang, Steven Goldfeder, and Ari Juels. Order-fairness for byzantine consensus. In *Annual International Cryptology Conference*, pages 451–480. Springer, 2020.
12. Leslie Lamport. Time, clocks, and the ordering of events in a distributed system. In *Communications of the ACM*, volume 21, page 558–565. 1978.
13. Leslie Lamport et al. Paxos made simple. *ACM Sigact News*, 32(4):18–25, 2001.
14. Shengyun Liu, Paolo Viotti, Christian Cachin, Vivien Quéma, and Marko Vukolić. {XFT}: Practical fault tolerance beyond crashes. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 485–500, 2016.
15. Oded Naor, Mathieu Baudet, Dahlia Malkhi, and Alexander Spiegelman. Cogsworth: Byzantine view synchronization. *arXiv preprint arXiv:1909.05204*, 2019.
16. Oded Naor and Idit Keidar. Expected linear round synchronization: The missing link for linear byzantine smr. In *34th International Symposium on Distributed Computing (DISC 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.
17. Diego Ongaro and John Ousterhout. In search of an understandable consensus algorithm. In *2014 USENIX Annual Technical Conference (USENIX ATC 14)*, pages 305–319, 2014.
18. Alexander Spiegelman. In search for a linear byzantine agreement. *arXiv preprint arXiv:2002.06993*, 2020.



19. Alexander Spiegelman, Arik Rinberg, and Dahlia Malkhi. Ace: Abstract consensus encapsulation for liveness boosting of state machine replication. In *24th International Conference on Principles of Distributed Systems (OPODIS 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.
20. The Diem Team. Diembft v4: State machine replication in the diem blockchain. <https://developers.diem.com/docs/technical-papers/state-machine-replication-paper.html>.
21. Maofan Yin, Dahlia Malkhi, Michael K Reiter, Guy Golan Gueta, and Ittai Abraham. Hotstuff: Bft consensus with linearity and responsiveness. In *Proceedings of the 2019 ACM Symposium on Principles of Distributed Computing*, pages 347–356, 2019.

## Appendix A Correctness

**Lemma 9.** *If `choose_leader` returns the same honest party at all honest parties for infinitely many rounds, then each honest party commits an unbounded number of blocks.*

*Proof.* If `choose_leader` returns the same honest party at all honest parties for infinitely many rounds, then there are infinitely many rounds after GST for which it does so. Let  $r$  be such a round. By the Pacemaker guarantees, all honest parties make  $LBR\text{-synchronized}(\ell)$  invocations with the same honest leader  $\ell$  returned from the `choose_leader` procedure. By the LBR Progress property, they all return a certified block  $B$  and commit it at [line 6](#).

**Lemma 1.** *In a crash-only execution, let  $r$  be a round with  $k \geq 2f + 1$   $LBR\text{-synchronized}(\ell)$  invocations, such that  $\ell$  is alive at round  $r$ , then these  $k$  invocations return a certified  $B$  with round number  $r$  authored by  $\ell$ .*

*Proof.* Let  $\pi_1$  be a crash-only execution, such that round  $r$  has  $k \geq 2f + 1$   $LBR\text{-synchronized}(\ell)$  invocations with a leader  $\ell$  that is alive at round  $r$ . If  $\ell$  is honest, then the LBR Progress property concludes the proof.

Otherwise,  $\ell$  is faulty and by definition it crashes in round  $> r$ . Let  $\pi_2$  be a crash-only execution that is identical to  $\pi_1$  until  $\ell$  crashes, and the rest of  $\pi_2$  is an arbitrary execution where the honest parties in  $\pi_1$  remain honest but  $\ell$  never crashes and is also honest. Thus, in  $\pi_2$  the preconditions of the LBR Progress property hold and all  $k$   $LBR\text{-synchronized}(\ell)$  invocations return a certified  $B$  with round number  $r$  authored by  $\ell$ .

An  $LBR(r, \ell)$  invocation by any party  $p$  completes within  $\Delta_l$  time, and starts immediately after Pacemaker's `new_round(r)` notification at  $p$  (because `choose_leader` is computed locally and takes 0 time). By Pacemaker's guarantees, no party receives `new_round(r+1)` notification until  $\Delta_p = \Delta_l$  time after the last `new_round(r+1)` notification at some party, hence all  $LBR(r, \ell)$  invocations must complete before any party receives a `new_round(r+1)` notification.

$\pi_1$  and  $\pi_2$  are identical until  $\ell$  crashes, which must happen after  $\ell$  receives its `new_round(r+1)` notification from the Pacemaker. This is because  $\ell$  is alive in round  $r$  and follows the protocol, invoking  $LBR$  in round  $r+1$  after receiving the `new_round(r+1)` notification. As a result,  $\pi_1$  and  $\pi_2$  are indistinguishable to all  $LBR(r, \ell)$  invocations, and the  $k$   $LBR\text{-synchronized}(\ell)$  invocations in  $\pi_1$  return certified block  $B$  with round number  $r$  authored by  $\ell$  as in  $\pi_2$ , as desired.