Let It TEE: Asynchronous Byzantine Atomic Broadcast with $n \geq 2f + 1$

Marc Leinweber
Institute of Information Security and Dependability (KASTEL), Karlsruhe Institute of Technology (KIT), Germany

Hannes Hartenstein
Institute of Information Security and Dependability (KASTEL), Karlsruhe Institute of Technology (KIT), Germany

Abstract
Asynchronous Byzantine Atomic Broadcast (ABAB) promises simplicity in implementation as well as increased performance and robustness in comparison to partially synchronous approaches. We adapt the recently proposed DAG-Rider approach to achieve ABAB with $n \geq 2f + 1$ processes, of which $f$ are faulty, with only a constant increase in message size. We leverage a small Trusted Execution Environment (TEE) that provides a unique sequential identifier generator (USIG) to implement Reliable Broadcast with $n > f$ processes and show that the quorum-critical proofs still hold when adapting the quorum size to $\left\lfloor \frac{n}{2} \right\rfloor + 1$. This first USIG-based ABAB preserves the simplicity of DAG-Rider and serves as starting point for further research on TEE-based ABAB.

1 Introduction
Atomic Broadcast primitives play a crucial role for Byzantine-fault tolerant (BFT) State Machine Replication (SMR). A prominent example for BFT SMR in the partially synchronous model is PBFT [6]. By use of small Trusted Execution Environments (TEE) that generate and sign unique sequential identifiers on each process, called USIGs, Veronese et al. [13] showed that PBFT’s communication complexity can be reduced and the fault tolerance can be increased to $n \geq 2f + 1$ while still tolerating Byzantine faults. The authenticity/integrity of TEEs can be verified remotely and, thus, TEEs are assumed to only fail by crashing. However, as shown by Miller et al. [12], Asynchronous Byzantine Atomic Broadcast (ABAB) outperforms approaches based on the partially synchronous model particularly under faults and tends to show a simpler design. While it is known that any asynchronous crash fault-tolerant algorithm can be compiled to withstand Byzantine faults using TEEs [3, 7], the proposed compilers show either a polynomial or an exponential overhead in runtime. We are interested in a simple and straightforward design of a USIG-enhanced ABAB that does not add further message rounds and only adds a constant number of bits to each message (essentially a counter value and a signature). To this end, we adapt DAG-Rider [10] to provide ABAB with $n \geq 2f + 1$ processes. We give a quick recap on DAG-Rider and explain the adaption TEE-Rider. Besides using TEE-based Reliable Broadcast and changing the required quorums from $2f + 1$ to $\left\lfloor \frac{n}{2} \right\rfloor + 1$, we leave DAG-Rider unchanged. We show that the quorum-based arguments of DAG-Rider still hold for TEE-Rider.
2 TEE-Rider: Transforming DAG-Rider to $n \geq 2f + 1$

We make use of the following definition of Atomic Broadcast for a set of processes $P, n := |P|$. The processes communicate over authenticated point-to-point links with eventual delivery.

**Definition 1** (Atomic Broadcast). Each process $p_i \in P$ receives client transactions $t$ via events `clientRequest(t)`. Correct processes deliver tuples $(t, r, p_j)$, where $t$ is a client transaction, $r \in \mathbb{N}_0$ a round number, and $p_i \in P$ the process that initially received $t$, satisfying the following properties:

- **Agreement**: If a correct process $p_i \in P$ delivers $(t, r, p_j)$, then every other correct process $p_k \in P, k \neq i$ eventually delivers $(t, r, p_j)$ with probability 1.
- **Integrity**: For each round $r \in \mathbb{N}_0$ and process $p_j \in P$, a correct process $p_i \in P$ delivers $(t, r, p_j)$ at most once.
- **Validity**: If a correct process $p_i \in P$ receives an event `clientRequest(t)`, then every correct process $p_k \in P$ eventually delivers $(t, r, p_i)$ with probability 1.
- **Total Order**: Let $m_1$ and $m_2$ be any two valid tuples that are delivered by any two correct processes $p_i, p_j \in P$. If $p_i$ delivers $m_1$ before $m_2$, then $p_j$ delivers $m_1$ before $m_2$.

### 2.1 Changes in Assumptions, Building Blocks, and Setup

In addition to the assumptions of DAG-Rider, we assume that each process is equipped with a USIG [13] that may only fail by crashing. It implements a signature service that binds a unique counter value to each signature it produces. The USIG is used for Reliable Broadcast with a fault tolerance of $n > f$ as, e.g., implemented in [8, Algorithm 1]: it is a ‘single echo’ algorithm with USIG-signed messages and attached counter value. Correct processes relay a message once and reject messages with invalid USIG signatures or with counter values already received which prevents equivocating messages for the same counter. An instance of the Reliable Broadcast abstraction has two functions: broadcast($r, m$) to reliably broadcast exactly one arbitrary message $m$ for round $r$ to all processes in $P$, and delivered() which returns all messages that were received by the instance since the last call to delivered(). We expect the Reliable Broadcast abstraction to fullfil the following properties:

**Definition 2** (Reliable Broadcast). A sender $p_s \in P, n := |P|$ can call broadcast($m$). Correct processes deliver $(c, m)$ where $c \in \mathbb{N}_0$ and $m$ an arbitrary message satisfying the following properties:

- **RB-Agreement**: If a correct process $p_i \in P$ delivers $(c, m)$, then every other correct process $p_k \in P, k \neq i$ eventually delivers the same $(c, m)$.
- **RB-Integrity**: For each $c \in \mathbb{N}_0$, a correct process $p_k \in P$ delivers $(c, m)$ at most once.
- **RB-Validity**: If a correct sender calls broadcast($m$), then every correct process $p_i \in P$ eventually delivers $(c, m)$.

Additionally, we assume an asynchronous common coin, e.g. as defined by Cachin et al. [5], that produces a uniformly distributed common random number $p$ out of $\{p \mid p \in \mathbb{N}_0: p < n\}$ for all correct processes and a name $i \in \mathbb{N}_0$ as soon as $f + 1$ processes invoked toss($i$); repetitive calls with same the $i$ yield the same $p$. We further assume a trusted setup of the common coin and the USIGs using a public key infrastructure (to set up the common coin’s threshold signature scheme, dealerless variants [4] and those with an asynchronous setup [1] exist).
2.2 The Algorithm

DAG-Rider uses \( n \) Reliable Broadcast instances to disseminate process messages and to construct a directed acyclic graph (DAG) that captures the communication history of all processes. In a second step, each process derives consensus on the order of transactions using a Common Coin based on the graph structure. The adapted DAG-Rider algorithm executed by a correct process \( p_i \in P \) is shown in Algorithm 1. Utility functions are listed in Algorithm 2. The core of the approach is the construction and interpretation of a (local) DAG that captures received transactions and the observed communication sequence between processes. The DAG is structured in rounds and a round contains at maximum one vertex per process, i.e., \( n \) vertices. Rounds are addressed in an array style and the local view of a process \( p_i \) on the DAG is indicated by an index \( i \). The very first round \( \text{DAG}_i[0] \) is initialized with \( n \) hard-coded “genesis” vertices. A vertex in round \( r \) has two types of edges: strong edges point to vertices of round \( r - 1 \) and weak edges point to vertices of any round \( r' \leq r - 2 \).

As soon as \( p_i \) received \( \lfloor \frac{n}{2} \rfloor + 1 \) valid vertices for a round \( r \), i.e., \( \lfloor \frac{n}{2} \rfloor + 1 \) vertices referencing \( \lfloor \frac{n}{2} \rfloor + 1 \) vertices of round \( r - 1 \) as strong edges (\( v.\text{strong} \), l. 11), for which it also knows its predecessors (l. 15), \( p_i \) will complete round \( r \) and transition to round \( r + 1 \). Now, as soon as \( p_i \) receives a client transaction, it will become the payload of a vertex \( v \) which is created and broadcast by \( p_i \) for round \( r + 1 \) (ll. 44 and 21-27). The vertex \( v \) connects to all vertices \( p_i \) received for round \( r \) (l. 23). If \( p_i \) received vertices \( u \) for older rounds that are not reachable from the newly created vertex using the transitive closure of strong and weak edges (a ‘path’), \( u \) will become a weak edge of \( v \) (\( v.\text{weak} \), l. 26). The new vertex is broadcast using Reliable Broadcast instance \( i \) to all processes (l. 27). Every fourth round a so-called wave, consisting of four rounds, is completed (l. 19) and the DAG structure is used to derive a total order on the transactions (ll. 28-42). Each wave \( w \) has exactly one wave leader \( v \) which is chosen calling \( \text{coin.toss}(w) \) from the vertices of \( w \)'s first round \( (w, 1) \). The random number is used to select the process whose vertex is to be used as wave leader. If \( v \) was not (yet) received or
there are no $\lfloor \frac{w}{2} \rfloor + 1$ vertices in the $w$’s fourth round($w, 4$) that have $v$ in their transitive closure of strong edges (a ‘strong path’), i.e. the direct commit rule is not fulfilled, the wave cannot be committed (l. 30). If wave $w$ can be committed, process $p_i$ checks first if there are wave leaders of waves $w'$ between the last wave that was committed (variable decidedWave) and the current wave $w$ that were received in the meantime and are connected to the leader of the wave $w' + 1$ (retrospective commit, ll. 33–36). The wave leaders are used as the root for a deterministic graph traversal to determine the total order of transactions (ll. 38–42). An example for a resulting graph with $n = 5$ processes, i.e. $f \leq 2$, is shown in Figure 1.

3 Correctness Argument

Lemmas 1 and 2 of the original DAG-Rider publication [10] are crucial for Total Order and Agreement and rely on quorum intersection arguments. The following Lemmas 3 and 4 show wave leaders of waves $w$. For a round $r \geq$ round($w + 1, 1$) will have a strong path to $v$.

Proof. Since $p_i$ commits $v$ in round($w, 4$), the direct commit rule is fulfilled (l. 30): $\exists U \subseteq \text{DAG}_i[\text{round}(w, 4)]; |U| \geq |\frac{w}{2}| + 1 \land \forall u \in U: \text{strongPath}(u, v)$. A valid vertex must reference at least $|\frac{n}{2}| + 1$ distinct vertices of the previous round with a strong edge (l. 11). Thus, a process $p_j \in P$ broadcasting a valid vertex $v_j$ for round($w + 1, 1$) selected at least $|\frac{n}{2}| + 1$ vertices of round($w, 4$) as strong edges for $v_j$. Any two subsets of size $|\frac{n}{2}| + 1$ of a superset of size $n$ intersect at least in one element. Thus, every valid vertex of a process broadcast for round($w + 1, 1$) must have at least one edge to a vertex of $U$, and, via $U$ to $v$. As every valid vertex of round($w + 1, 1$) has a strong path to $v$, and every valid vertex of round($w + 1, 2$) connects to at least $|\frac{n}{2}| + 1$ vertices of round($w + 1, 1$), by induction, any valid vertex $v'$ of any process $p_j \in P$ broadcast for a round $r \geq$ round($w + 1, 1$) has a strong path to $v$.

Lemma 4. When a correct process $p_i \in P$ completes round($w, 4$) of wave $w$, then $\exists V_1 \subseteq \text{DAG}_i[\text{round}(w, 1)], V_3 \subseteq \text{DAG}_i[\text{round}(w, 4)]; |V_1| \geq |\frac{n}{2}| + 1 \land |V_4| \geq |\frac{n}{2}| + 1 \land (\forall v_1 \in V_1, \forall v_4 \in V_4: \text{strongPath}(v_4, v_1))$.

Proof. By use of Reliable Broadcast and validity checks in ll. 11 and 15, faulty processes are limited to omission faults. Thus, the get-core argument of Attiya and Welch [2, Sec. 14.3.1] still holds [2, Sec. 14.3.3]: Let $A \in \{0, 1\}^{n \times n}$ be a matrix that contains a row for each possible vertex of round($w, 3$) and a column for each possible vertex of round($w, 2$). Let $A[j, k] = 1$ if the vertex of process $p_j$ of round($w, 3$) has a strong edge to the vertex of process $p_k$ of round($w, 2$) or $p_j$ sends no vertex (or an invalid one) but $p_k$ sends a valid vertex for round($w, 2$). As there are at least $\lfloor \frac{n}{2} \rfloor + 1 \leq n - f$ correct processes, each row of $A$ contains at least $\lfloor \frac{n}{2} \rfloor + 1$ ones and $A$ contains at least $n(\floor{\frac{n}{2}} + 1)$ ones. Since there are $n$ columns, there must be a column $l$ with at least $\floor{\frac{n}{2}} + 1$ ones. This implies there is a vertex $v_l$ by process $p_l$ in round($w, 2$) s.t. $\forall V_3 \subseteq \text{DAG}_i[\text{round}(w, 3)]; |V_3| \geq \floor{\frac{n}{2}} + 1 \land \forall v_3 \in V_3: \text{strongPath}(v_3, v_l)$. As at most $f$ vertices in $V_3$ belong to faulty processes that may commit send omission faults for round($w, 3$) and $\floor{\frac{n}{2}} + 1 \geq f + 1$, by quorum section at least one vertex of $V_3$ is received by any correct process $p_j \in P$ before it sends its vertex for round($w, 4$). Thus, every valid vertex in $\text{DAG}_i[\text{round}(w, 4)]$ has at least one strong edge to a vertex of $V_3$. Since $v_l$ must be valid and thus has a strong edge to each vertex of a set $V_1 \subseteq \text{DAG}_i[\text{round}(w, 1)], |V_1| \geq \floor{\frac{n}{2}} + 1,$
Algorithm 1 TEE-Rider pseudocode for process $p_i \in P, n := |P|, n \geq 2f + 1$

1: state $DAG$: array of sets of vertices, $DAG[0]$ initialized with “genesis” vertices
2: state $r$: $\mathbb{N}$, initialized with 1
3: state decidedWave: $\mathbb{N}_0$, initialized with 0
4: state transactionsToPropose: queue of client transactions $t$, initialized empty
5: state buffer: set of vertices, initialized empty
6: state rb: array of $n$ Reliable Broadcast instances with delivered() and broadcast($r, m$)
7: state coin: common coin instance with toss($w$)
8: while True do
9:   for $k \leftarrow 0$ up to $n - 1$ do
10:      for $m = (r', v) \in rb[k]$.delivered() do
11:         if $|v$.strong$| < \lceil \frac{n}{2} \rceil + 1$ then continue
12:         v.source $\leftarrow p_k$; v.round $\leftarrow r'$; v.delivered $\leftarrow$ False
13:         buffer.add(v)
14:      for $v \in$ buffer do
15:         if $v$.round $\geq r \lor \exists u \in v$.strong $\lor v$.weak: $u \notin \cup_{i \geq 0}DAG[r']$ then continue
16:         $DAG[v$.round$]$.add(v)
17:         buffer.remove(v)
18:      if $|DAG[r]| < \lceil \frac{n}{2} \rceil + 1$ then continue
19:      if $r \mod 4 = 0$ then waveReady($\frac{r}{2}$)
20:   $r \leftarrow r + 1$
21:   wait until $\neg$transactionsToPropose.isEmpty()
22: $v \leftarrow$ new vertex
23: $v$.block $\leftarrow$ transactionsToPropose.dequeue(); $v$.strong $\leftarrow$ $DAG[r - 1]$
24: for $r' \leftarrow r - 2$ down to 1 do
25:      for $u \in DAG[r']$ do
26:         if $\neg$path($v, u$) then $v$.weak.add(u)
27:      $rb[i]$.broadcast($r, v$)
28: function waveReady($w$)
29:   $v \leftarrow$ getWaveLeader($w$) $\triangleright$ if round($w, 1$) vertex of chosen process is not in $DAG$
30:   if $v = \perp \lor \exists u \in DAG[\text{round}(w, 4)]$: strongPath($u, v$) $| < \lceil \frac{n}{2} \rceil + 1$ then return
31: leadersStack $\leftarrow$ new stack; leadersStack.push($v$)
32: for $w' \leftarrow w - 1$ down to decidedWave + 1 do
33:    $u \leftarrow$ getWaveLeader($w'$)
34:    if $u \neq \perp \land$ strongPath($v, u$) then
35:       leadersStack.push($u$); $v \leftarrow u$
36: decidedWave $\leftarrow w$
37: while $\neg$leadersStack.isEmpty() do
38:   $v \leftarrow$ leadersStack.pop()
39: verticesToDeliver $\leftarrow \{ u \mid u \in \cup_{r' \geq 0}DAG[r'] \land$ path($v, u$) $\land \neg u$.delivered $\}$
40: for $u \in$ verticesToDeliver in deterministic order do
41:   u.delivered $\leftarrow$ True; deliver ($u$.block, $u$.round, $u$.source)
42: upon clientRequest ($t$)
43: transactionsToPropose.enqueue($t$)
Algorithm 2 Utility functions pseudocode

1: function path(v, u) : boolean
2: return exists a sequence of vertices \( (v_1, v_2, ..., v_k) \in \cup_{r' \geq 0} \text{DAG}[r'] \) such that
3: \( v_1 = v \land v_k = u \land \forall i \in [2, k] : v_i \in v_{i-1}.\text{strong} \cup v_{i-1}.\text{weak} \)

4: function strongPath(v, u) : boolean
5: return exists a sequence of vertices \( (v_1, v_2, ..., v_k) \in \cup_{r' \geq 0} \text{DAG}[r'] \) such that
6: \( v_1 = v \land v_k = u \land \forall i \in [2, k] : v_i \in v_{i-1}.\text{strong} \)

7: function getWaveLeader(w) : vertex or ⊥
8: \( p_j \leftarrow \text{coin.toss}(w) \)
9: if \( \exists v \in \text{DAG}[\text{round}(w, 1)] : v.\text{source} = p_j \) then return v
10: return ⊥
11: function round(w, i) : N
12: return \( 4(w - 1) + i \)

any valid vertex of rounds \( r \geq \text{round}(w, 4) \) has a strong path to every vertex, including \( V_1 \), reached by \( v_l \) via strong paths. Please note that the construction of the set \( V_1 \) is valid for all correct processes that complete the wave and, thus, represents the ‘common core’. ▶

4 Discussion and Conclusion

DAG-Rider shows the power of causal order broadcast to implement consensus. The adaption for TEEs preserves the simplicity of DAG-Rider while increasing its fault tolerance and reducing the communication effort (i.e., from ‘double echo’ to ‘single echo’ Reliable Broadcast). The ease of adaption of DAG-Rider for TEEs make it a perfect textbook example for TEE-based ABAB. Follow-up work to DAG-Rider, Tusk [9], addresses a major deployability issue, namely garbage collection, shortens the wave length, and replaces the underlying Reliable Broadcast with a communication scheme that leverages the graph structure to achieve linear communication complexity in the happy case. Additionally, to the best of our knowledge, there exists no TEE-based, dealerless, and asynchronous common coin primitive. In summary, investigating a TEE-based dealerless setup as well as transforming the follow-ups of DAG-Rider for empirical studies to investigate the assumed superiority of asynchronous TEE-based approaches, e.g., in comparison to MinBFT [13], is a promising line of research.

References


