Ordered Consensus with Equal Opportunity

Yunhao Zhang,[†] Haobin Ni,* Soumya Basu,[‡] Shir Cohen,[†] Maofan Yin*, Lorenzo Alvisi,[†] Robbert van Renesse,[†] Qi Chen,[§] and Lidong Zhou[§]

[†]Cornell University

*University of Washington

*University of California

 \S Microsoft Research

[‡]Vigil Markets

*University of California, Santa Barbara

Abstract

The specification of state machine replication (SMR) has no requirement on the final total order of commands. In blockchains based on SMR, however, order matters, since different orders could provide their clients with different financial rewards. Ordered consensus augments the specification of SMR to include specific guarantees on such order, with a focus on limiting the influence of Byzantine nodes. Real-world ordering manipulations, however, can and do happen even without Byzantine replicas, typically because of factors, such as faster networks or closer proximity to the blockchain infrastructure, that give some clients an unfair advantage. To address this challenge, this paper proceeds to extend ordered consensus by requiring it to also support equal opportunity, a concrete notion of fairness, widely adopted in social sciences. Informally, equal opportunity requires that two candidates who, according to a set of criteria deemed to be relevant, are equally qualified for a position (in our case, a specific slot in the SMR total order), should have an equal chance of landing it. We show how randomness can be leveraged to keep bias in check, and, to this end, introduce the secret random oracle (SRO), a system component that generates randomness in a fault-tolerant manner. We describe two SRO designs based, respectively, on trusted hardware and threshold verifiable random functions, and instantiate them in Bercow, a new ordered consensus protocol that, by approximating equal opportunity up to within a configurable factor, can effectively mitigate the well-known ordering attacks in SMR-based blockchains.

1 Introduction

This paper extends *ordered consensus* [80] by motivating, expressing, and enforcing *equal opportunity*, a concrete notion of fairness that applies to how a state machine replication (SMR) [71] protocol *orders* client requests.

SMR is the most general technique for building fault-tolerant services. At its core, SMR requires a group of replicas to agree on the same, totally ordered sequence of client requests (*i.e.*, a ledger). Therefore, it is unsurprising that SMR has become a standard paradigm for permissioned blockchains. Applying SMR to this new context, however, poses fresh challenges.

As long as all requests from correct clients eventually appear in the ledger, their specific order is immaterial when

SMR is used for fault-tolerance: all that matters is for all correct replicas to process client requests in the same order. In blockchains, however, the specific order matters, as it can determine the financial rewards associated with the transactions recorded in the ledger.

Ordered consensus aims to make order a first-class citizen of the SMR specification. Specifically, each replica is required to associate with each command an *ordering indicator* (e.g., a timestamp), which the replica can use to express how it would like to order commands with respect to one another. Leveraging ordering indicators, it is possible to prove [80] that, while Byzantine influence over the ledger's order cannot be completely eliminated, it can be curtailed. In particular, it is possible to guarantee the ordering of commands stored in the ledger satisfies *ordering linearizability*: if the lowest timestamp that any correct replica assigns to command c_2 is larger than the highest timestamp that any correct node assigns to command c_1 , then c_1 will precede c_2 in the ledger—independent of the actions of Byzantine replicas.

The starting point of this paper is the observation that, while limiting the influence of Byzantine nodes is a necessary first step towards providing fairness, unfairness can and does arise in practice even when all replicas are correct.

Consider, for example, the practice known in financial markets as *front-running*, where a party, aware of the existence of a large buy order for some stock, places beforehand its own buy order for the same stock. This party is then able to buy low and later sell high, once the stock's value has been driven up by the ensuing large buy order. It has been widely reported how a faster network can enable front-running not just in traditional financial markets [56], but also in decentralized ones [37, 74]. No Byzantine replica is necessary for these attacks to succeed in a blockchain based on SMR: when using timestamps as ordering indicators, the difference in network latency (either due to physical proximity or access to faster network facilities) between clients and replicas may provide some clients with a systemic advantage over others.

This paper's first contribution is to offer a framework to reason about and address this kind of systemic bias. Taking inspiration from social sciences, which have a long history of reasoning about bias and unfairness, we observe that, whenever ranking is involved, the position of an entry in the ranking depends on the entry's specific characteristics (or

1

features). Some of the features are relevant to the stated purpose of the ranking, while others may be irrelevant. For example, US employers and lending agencies are legally forbidden to consider certain irrelevant features (protected classes) in making decisions [1, 2]. Intuitively, a fair ranking is one that relies only on the entries' relevant features and ignores all the other features. Entries with indistinguishable relevant features should then have an equal chance of being ranked ahead of each other.

Building on the expressiveness offered by ordered consensus, this paper instantiates this general notion of fairness in the concrete context of SMR-based blockchains. Since we aim for the ledger to reflect an order respecting the time clients issued requests, we deem the *time of issue* as the relevant feature in determining the ledger's order. Other features, such as geographic location, are considered irrelevant.

Unfortunately, existing protocols for ordered consensus neither distinguish between relevant and irrelevant features, nor reason about such distinctions. As a result, protocols like Pompē [80], Aequitas [50], and Themis [49] are all vulnerable to parties that leverage irrelevant features, such as faster network facilities, to engage in front-running and in its close relative, *sandwich attacks* (see Section 2.1).

Although preventing Byzantine replicas from conducting such attacks is provably impossible [80], we show that distinguishing between relevant and irrelevant features when assigning ordering preference can mitigate the problem.

Concretely, our second contribution is to specify ϵ -Ordering equality, a new ordering property based on this distinction, and to propose a protocol that enforces it. Intuitively, ϵ -Ordering equality requires the likelihood of all permutations of client requests with indistinguishable relevant features to differ by at most ϵ . To enforce it, we use a Secret Random Oracle (SRO), an abstraction that offers a fault-tolerant and unbiased source of randomness, to add some noise to the final order computed from the replicas' ordering indicators. We provide two SRO designs—one design uses the Trusted Execution Environment (TEE), while the other design relies purely on cryptography, using threshold verifiable random functions [30, 40].

Given that the profitability of front-running and sandwich attacks depends on the ledger recording a specific permutation of transactions, adding randomness, by altering the probability of adopting *that* specific permutation, can reduce the effectiveness of these attacks significantly.

While adding randomness mitigates the attacks, it can also compromise the role of a client request's *relevant* feature in determining its position in the ledger. Our third contribution is to quantify this tension as a trade-off between ϵ -Ordering equality and another ordering property, Δ -Ordering linearizability. Like ordering linearizability [80], Δ -Ordering linearizability is an ordering guarantee robust to Byzantine tampering. However, while ordering linearizability depends on timestamps reflecting when client commands are received,

and thus conflates both relevant and irrelevant features, Δ -Ordering linearizability applies to timestamps that reflect the real time when clients' commands are issued: it ensures that the ledger will respect the invocation order of two client command issued at least Δ time apart.

Ideally, we would like both the ϵ and the Δ in the respective fairness guarantees to be as small as possible. Unfortunately, however, there is a trade-off between them: adding more random noise can decrease ϵ , but at the price of potentially expanding the Δ interval necessary to ensure that, independent of Byzantine ploys, two commands Δ or more apart will be correctly ordered in the ledger.

Finally, to explore the practical implications of this trade-off, we design, implement, and evaluate Bercow, a new ordered consensus protocol designed to operate in the *partially synchronous* model [38] introduced to sidestep the impossibility of safe, live, and fault-tolerant consensus in asynchronous systems [39]. Bercow is always safe, and, during periods where progress is possible (formally, after some unknown Global Stabilization Time; practically, during long-enough synchronous intervals), it is also live, and enforces both ϵ -Ordering equality and Δ -Ordering linearizability. Specifically, Bercow modifies Pompē by adding SRO-generated random noise to the fault-tolerant timestamp that Pompē uses to order commands.

Our evaluation demonstrates that Bercow can be effective in mitigating front-running and sandwich attacks while incurring moderate performance overhead. For example, when adding a random noise sampled from $[0, \Delta_{net} * 5]$, where Δ_{net} is a bound on the message delay experienced during synchronous intervals, Bercow can keep ϵ under 10% - a threshold considered acceptable in other contexts where equal opprtunity is to be enforced [2, 3] — while matching Pompē's throughput; the added random noise increases median consensus latency by about 14% in a setup of 49 nodes.

2 Equal opportunity

Consider a system in which clients invoke commands. The system aims to produce, as its output, a total order that reflects the real time at which commands are invoked—earlier commands should precede later ones.

2.1 Motivating equal opportunity

Informally, if two commands have the same invocation time, equal opportunity says that the two possible orders should be equally likely to appear in the system output. Similarly, if three invocations all have the same invocation time, the six possible orders should be equally likely. To show how equal opportunity is often violated in the real world, we analyze publicly available traces of Ethereum. While Ethereum is a permissionless blockchain, none of the issues we identify depend on Ethereum being permissionless.

System output	Victim's profit	Attacker's profit
i_2, i_1, i_3	-\$500	\$800
i_3, i_1, i_2	\$700	-\$400
other order	\$300	\$0

Figure 1. An example of sandwich attacks where the victim invokes i_1 and the attacker invokes i_2 and i_3 . The semantics of the three invocations are explained in Section 5.2.

Case #1: Two invocations. Violating equal opportunity for two invocations may not only indicate bias but provide opportunities for front-running [37, 74]. Empirical studies show that both phenomena have been significant factors in the allocation of \$89M over 32 months in the Ethereum blockchain [66].

For example, an invocation from Europe is likely to be ordered before a simultaneous one from Australia because more Ethereum nodes are located in Europe. If the system orders the invocation from Europe earlier in its output more than half the time, we say the system is biased toward Europe. While the geographical location is typically an irrelevant feature in the context of equal opportunity, such bias has been observed and reported in other blockchains as well [60].

Geographical bias can lead to undesirable consequences on blockchain liquidations. In real life, liquidations occur when an individual goes bankrupt. For example, if someone cannot pay their debts but owns a house, a court can sell the house to repay the debts. If the market price of the house is \$1.2M, the court may sell it for only \$1M. Therefore, many parties would compete for pocketing a \$200K profit. Similar liquidations happen on blockchains, where they provide a common way of making a profit in the stable coin [6] and lending[11] applications. The buyer whose command is ordered first on the blockchain is typically the one that realizes the profit.

Consider two clients from Europe and Australia. Suppose they invoke the liquidation command simultaneously, and the system is biased toward Europe, ordering the European command first with a 75% chance. In that case, the expected value of the European client will be $0.2M \cdot 0.75 = 150K$, and that will be $0.2M \cdot 0.25 = 50K$ for the Australian client. In other words, geographical bias could cause very different profits to clients who should be treated equally.

Similarly, a client intent on becoming the beneficiary of a liquidation's profits could leverage faster network connections to violate equal opportunity and front-run other clients.

Case #2: Three invocations. In this case, violating equal opportunity among three simultaneous commands could enable sandwich attacks [81]. Empirical studies show that victims of sandwich attacks have lost more than \$174M over 32 months in the Ethereum blockchain [66].

Figure 1 shows an example of sandwich attacks observed in the decentralized exchange applications. Right after the

victim invokes command i_1 , the attacker invokes commands i_2 and i_3 . The attacker only profits if the order the system outputs is i_2 , i_1 , i_3 . Therefore, the key to sandwich attacks is making i_2 , i_1 , i_3 a much more likely output than equal opportunity would allow. A common strategy to influence the odds is to privately relay i_2 and i_3 to colluding nodes, which will then exclusively propose blocks containing the sequence i_2 , i_1 , i_3 [66].

Of course, the attacker is always free to decide which specific commands to invoke as their trading strategy, and different trading strategies may still lead to different expected profits – but, crucially, the system should not allow attackers to tamper with the odds of its different possible outputs: all six permutations of the three commands should be roughly equally likely.

2.2 Impartiality and Consistency

The notion of equal opportunity derives from the combination of two well-known principles in economics [78] – *impartiality* and *consistency*. When applied to our settings, impartiality informally requires that the order of commands should not be influenced by irrelevant features, such as clients' geolocation. Consistency instead requires that the invocation of a new command should not cause the relative order of existing commands to change.

We now introduce these notions more formally.

Client invocation. An *invocation* is a pair $\langle c, \vec{f_r} \rangle$ where c is a command and $\vec{f_r}$ is a vector of relevant features, *i.e.*, of the only features that should be considered in determining how a client's commands should be ordered. Typically, features to be ignored include a client's identifier, geolocation, wealth, and network facilities. In blockchains, relevant features typically include invocation time and transaction fee. An *invocation profile*, denoted as \mathcal{I} , is a set of invocations.

Node preference. Nodes observe invocations and express preferences. The *preference* of a node is a set of $\langle i, o \rangle$ pairs, where i is an invocation and o is an ordering indicator (*i.e.*, a piece of metadata such as a score or a timestamp). The preference of a node represents how a node would like to order the invocations. A *preference profile*, denoted as \mathcal{P} , is a vector of preferences from all correct nodes.

World and chance relation. A *world* is a pair of $\langle I, \mathcal{P} \rangle$, representing the scenario where clients invoke commands I and correct nodes express preferences \mathcal{P} . For all worlds, \mathcal{P} is well-formed under I, meaning all the invocations in \mathcal{P} should also appear in I.

It is uncertain which world will actually happen. When one admits that nothing is certain, one must also add that some things are more nearly certain than others [45, 68]. We thus introduce *chance relations*: for any two worlds, w_1 and w_2 , $w_1 >_c w_2$ denotes that w_1 has a higher chance than w_2 , and $w_1 \sim_c w_2$ denotes that the two worlds have an equal chance of happening in the system.

3

$$I = \{\langle c_1, \langle 5pm \rangle \rangle_{i_1}, \langle c_2, \langle 5pm \rangle \rangle_{i_2}, \langle c_3, \langle 5:01pm \rangle \rangle_{i_3} \}$$

$$\mathcal{P}_1 = \langle \{\langle i_1, 1 \rangle, \langle i_2, 2 \rangle, \langle i_3, 3 \rangle \} \rangle$$

$$\mathcal{P}_2 = \langle \{\langle i_2, 1 \rangle, \langle i_1, 2 \rangle, \langle i_3, 3 \rangle \} \rangle$$

Figure 2. An example of invocation and preference profiles for a system with one node (n = 1, f = 0). The system uses invocation time as the only relevant feature and sequence numbers as ordering indicators. $I' = \{i_1, i_2\}$ for impartiality because they have the same relevant feature 5pm. If there is more than one correct node, \mathcal{P}_2 could permutate I' for one or more entries in vector \mathcal{P}_1 .

Impartiality. A system is *impartial* if and only if, for all world $\langle I, \mathcal{P}_1 \rangle$, for all $I' \subseteq I$, for all \mathcal{P}_2 permutating I' in \mathcal{P}_1 , if all invocations in I' have the same \vec{f}_r , then $\langle I, \mathcal{P}_1 \rangle \sim_c \langle I, \mathcal{P}_2 \rangle$.

Impartiality is the first pillar of equal opportunity, and Figure 2 shows an example. Since 5pm is the relevant feature of both i_1 and i_2 , and \mathcal{P}_2 swaps i_1 and i_2 from \mathcal{P}_1 , impartiality says that the two worlds, $\langle I, \mathcal{P}_1 \rangle$ and $\langle I, \mathcal{P}_2 \rangle$ should be equally likely to happen in the system. In other words, the order of i_1 and i_2 should be based on invocation time and independent of irrelevant features.

Consistency is the second pillar of equal opportunity. Figure 3 shows an example. $\langle I, \mathcal{P}_1 \rangle$ and $\langle I, \mathcal{P}_2 \rangle$ are two worlds where c_3 is invoked but never received by the correct nodes so that i_3 is missing from the preferences. Consistency says that $\langle I, \mathcal{P}_1 \rangle \succ_c \langle I, \mathcal{P}_2 \rangle \iff \langle I, \mathcal{P}_3 \rangle \succ_c \langle I, \mathcal{P}_4 \rangle$. In other words, the order of i_1 and i_2 should be based on their *own* features and be independent of the features of i_3 , even the invocation time of i_3 . Let $\mathcal{P}_1 =_I \mathcal{P}_2$ denote \mathcal{P}_1 equals to \mathcal{P}_2 over invocations in I and $I = \{I_1, I_2\}$ denote a partition of I. We now define consistency formally.

Consistency. A system is *consistent* if and only if, for all worlds $\langle I, \mathcal{P}_1 \rangle$, $\langle I, \mathcal{P}_2 \rangle$, $\langle I, \mathcal{P}_3 \rangle$, $\langle I, \mathcal{P}_4 \rangle$, for all partition $I = \{I_1, I_2\}$, $\mathcal{P}_1 = I_1$ $\mathcal{P}_3 \wedge \mathcal{P}_2 = I_1$ $\mathcal{P}_4 \wedge \mathcal{P}_1 = I_2$ $\mathcal{P}_2 \wedge \mathcal{P}_3 = I_2$ \mathcal{P}_4 implies $\langle I, \mathcal{P}_1 \rangle \succ_c \langle I, \mathcal{P}_2 \rangle \iff \langle I, \mathcal{P}_3 \rangle \succ_c \langle I, \mathcal{P}_4 \rangle$.

A straightforward approach to achieving both impartiality and consistency is to establish a *point system*. In a point system, the system designer decides a formula mapping each invocation to a *score*, and this score only depends on the relevant features of the invocation. The output of a point system orders all the invocations by their scores and breaks ties by uniformly sampling a permutation. Impartiality is guaranteed because the score assigned to each invocation only depends on its relevant features, so the same relevant features lead to the same scores. Consistency is guaranteed because the score of each invocation depends on its own features and is thus independent of the other invocations. Indeed, it has been proved that, when it comes to ranking, a point system is the only mechanism that can satisfy both impartiality and consistency [78].

```
I = \{\langle c_1, 1pm \rangle_{i_1}, \langle c_2, 1:01pm \rangle_{i_2}, \langle c_3, 1:05pm \rangle_{i_3} \}
\mathcal{P}_1 = \langle \{\langle i_1, 1 \rangle, \langle i_2, 2 \rangle \} \rangle \quad \mathcal{P}_3 = \langle \{\langle i_1, 1 \rangle, \langle i_2, 2 \rangle, \langle i_3, 3 \rangle \} \rangle
\mathcal{P}_2 = \langle \{\langle i_1, 2 \rangle, \langle i_2, 1 \rangle \} \rangle \quad \mathcal{P}_4 = \langle \{\langle i_1, 2 \rangle, \langle i_2, 1 \rangle, \langle i_3, 3 \rangle \} \rangle
```

Figure 3. An example of consistency where $I_1 = \{i_1, i_2\}$ and $I_2 = \{i_3\}$. This example corresponds to real-world scandals: some stock exchanges introduced special commands like c_3 to illegally help certain trading firms profit by manipulating the order of c_1 and c_2 [56].

2.3 Properties for distributed systems

In distributed systems, invocation time is typically the only relevant feature: commands are ordered according to when they are invoked. Equal opportunity can be achieved with a point system directly using the invocation time as the score. Unfortunately, in practice, the invocation time of commands cannot be measured accurately. The invocation time can only be approximately measured by the time nodes observe the invocation. Therefore, such a measurement reflects the invocation time and also irrelevant features such as geolocation. To accommodate such measurement inaccuracy in distributed systems, we relax a point system with two parameters, ϵ and Δ , and define two properties, ϵ -Ordering equality and Δ -Ordering linearizability.

 ϵ -Ordering equality. For all invocation profile I and subset $I' \subseteq I$, for all two total orders of I' denoted as \succ_1 and \succ_2 , if all invocations in I' have the same invocation time as their relevant feature, then $|Pr[\succ_1] - Pr[\succ_2]| \le \epsilon(|I'|)$.

In this definition, ϵ is a function with the cardinality of I' as its parameter, and $Pr[\succ]$ denotes the probability of \succ appearing in the system output under the condition that I is invoked. Different systems could enforce this property with different ϵ functions. Intuitively, by making ϵ approach 0, clients with the same relevant features would have similar chances, reflecting the tie-breaking part of a point system. We now define another property that reflects how a point system deals with different scores.

 Δ -**Ordering linearizability.** For all invocation profile I and two invocations $i_1, i_2 \in I$, if i_1 is invoked more than Δ time before i_2 , then i_1 will appear before i_2 in the output.

Ideally, Δ can be 0 so that commands invoked earlier will always appear earlier in the output. Consider a naive setup in which all the nodes are correct, and they measure the invocation time accurately. The ideal could then be achieved, as stated by the following theorem.

Theorem 2.1. If all nodes are correct and accurately measure the invocation time of all invocations, ordering equality and ordering linearizability with $\epsilon = \Delta = 0$ can be achieved using a point system.

Proof. By properties of a point system.

```
Reveal(Int k, Set<Signature> s) \rightarrow Int | Error Generate(Int k) \rightarrow Proof Verify(Int k, Proof p, Int r) \rightarrow Bool
```

Figure 4. The interface of a secret random oracle (SRO).

3 Secret Random Oracle

The point system mechanism suggests that randomness can be crucial to ordering equality. We thus specify a system component, *secret random oracle* (SRO), that generates random numbers and keeps the random numbers secret until other components have finished producing their outputs. By keeping those numbers secret, the outputs of other system components cannot depend on them.

Design overview. Consider two clients who both invoke a command at time t. Say the first client uses faster network facilities for front-running, and the two commands are received by all correct nodes before $t+200\,ms$ and $t+400\,ms$, respectively. If two timestamps are chosen from $[t, t+200\,ms]$ and $[t, t+400\,ms]$ for ordering, the system could sample two random numbers independently and add to the timestamps, so the probability of the two commands being ordered one way or the other is close. This $random\ noise$ affects the ϵ of ordering equality and the Δ of ordering linearizability. As the intensity of the random noise approaches infinity, ϵ will approach 0 because ordering is then dominated by randomness, but Δ will unfortunately approach infinity. We will study such a trade-off between ϵ and Δ quantitatively.

SRO interface and guarantees. Consider a set of n nodes, at most f of which can behave arbitrarily. Every node holds a private key and knows the public keys of all other nodes. They provide a secret random oracle service with an interface shown in Figure 4.

A random function (not shown) maps an integer k to a pseudorandom number. Reveal is invoked to reveal the random number after all nodes in a quorum demonstrate, by providing signatures of k, that they wish to reveal the random number. Generate takes integer k and returns a cryptographic proof. Given a proof, Verify verifies whether the random number returned by Reveal is correct. An SRO provides the following guarantees:

Uniqueness. Reveal maps every integer k to a random number. Multiple queries of Reveal with the same k and any valid set of signatures return the same random number. A set of signatures is valid if it contains valid signatures of k from at least n - f different nodes. Reveal returns an error when given an invalid set of signatures.

Secrecy. For all integers k, if r is the unique integer that Reveal would return and the adversary does not have valid signatures of k from n-f different nodes, then it is computationally infeasible for the adversary to distinguish r from a uniform random sample with non-negligible probability.

Randomness. For all integers k, the unique integer r that Reveal would return is a cryptographically secure random number in that r is a non-error uniform random sample from the codomain of Reveal.

Validity. For all integers k, if r is the unique integer returned by Reveal and p is the proof returned by Generate, then $Verify(k,p,r) \rightarrow True$ and it is computationally infeasible to find integer $r' \neq r$ making $Verify(k,p,r') \rightarrow True$.

We show two SRO designs, one using trusted hardware and another using cryptography. We then integrate an SRO with Pompē—a state-of-the-art ordered consensus protocol—and prove ordering equality and ordering linearizability. We further demonstrate a quantitative trade-off between the two ordering properties, which helps system designers decide how much random noise to add.

3.1 An SRO design using trusted hardware

Trusted Execution Environments (TEEs) provide a hardware enclave that protects the integrity and confidentiality of user software. Using TEEs to enforce secrecy in blockchains has been actively advocated [8, 21] and practically adopted by the Ethereum testnet [13]. Here, we show how to implement an SRO based on TEEs.

Initialization. Consider that every node has a TEE running the SRO software. During initialization, each TEE generates a random number with special CPU instructions (*e.g.*, using RDRAND in x86) and runs a distributed consensus protocol to agree on one such number. This number is kept confidential within the TEEs and will be used as the seed of a pseudorandom function denoted as RAND.

Reveal, Generate and Verify. Given an integer and a set of signatures, a node invokes Reveal by forwarding the arguments to its local TEE. The TEE returns RAND(seed, k) or an error depending on whether the second parameter contains enough valid signatures of k. Similarly, Generate forwards k to the TEE, which returns HASH(RAND(seed, k)) where HASH is a one-way function. Lastly, Verify returns whether parameter p equals HASH(r).

Correctness. The same seed and the deterministic pseudorandom function ensures uniqueness and randomness. Secrecy is ensured by the integrity and confidentiality of the TEEs, which keep the seed and random numbers secret and only reveal the random numbers after seeing enough valid signatures. The security properties of one-way functions ensure validity.

We tacitly assume that the initialization (*i.e.*, consensus on the random seed) eventually finishes. This assumption ensures liveness: all invocations of the SRO functions on a correct node eventually return a result.

3.2 An SRO design using threshold VRF

A Threshold Verifiable Random Function (or threshold VRF) is a cryptographic construction used by several Byzantine

```
Threshold VRF node-side function

Produce(Int k) → Share

Threshold VRF client-side functions

Combine(Set<Share> s) → Int | Error

Valid(Int k, Int node_id, Share s) → Bool

The modified node-side function

Produce(Int k, Set<Signature> s) → Share | Error
```

Figure 5. The interface of (modified) threshold VRF.

agreement protocols [30, 40]. TVRFs have been used to select a random set of nodes as a committee and to ensure safety and liveness under a fully asynchronous network model. We show how to use a threshold VRF to construct an SRO.

Let TVRF denote a function from an integer to a pseudorandom number, similar to RAND in the first design. Figure 5 shows the interface of threshold VRF for evaluating TVRF(k). Each node invokes Produce(k), which produces a *share* using its private key. After collecting a sufficient number of shares, a client invokes Combine, which returns TVRF(k). However, shares from Byzantine nodes could be invalid. To this end, Valid checks, using the corresponding public key, whether a share from a node is valid or not. Threshold VRFs provide two important properties, informally [30]:

Robustness. For all integers k, it is computationally infeasible for an adversary to produce enough valid shares such that the integer output of Combine is not TVRF(k).

Unpredictability. Without enough valid shares for TVRF(k), an adversary cannot distinguish TVRF(k) from a uniform random sample with non-negligible probability.

We make a slight modification to threshold VRF. For the Produce interface, we add a parameter, a set of signatures of k to be verified. If verification fails, correct nodes must return an error instead of a share. We can now design an SRO as follows: Reveal forwards the two parameters to all nodes, collects enough valid shares, and invokes Combine. Generate returns the set of public keys. Verify takes all the public keys and a set of shares as input and returns whether all the shares are valid (using the Valid interface).

Correctness. Robustness implies uniqueness and validity: for all integers k, combining enough valid shares can only produce TVRF(k) since an adversary cannot create valid shares leading to a combined value different from TVRF(k). The properties of threshold VRF have been proven under the random oracle model, which implies randomness—informally, hash values are cryptographically secure pseudorandom numbers. Unpredictability implies secrecy because, without valid signatures of k out of n-f nodes, an adversary cannot gain enough valid shares and—without enough valid shares—it has no information about TVRF(k). Liveness is ensured, assuming all network messages are eventually delivered.

3.3 Bercow: Integrating an SRO with Pompē

We now introduce Bercow, which integrates an SRO with Pompē [80], a state-of-the-art ordered consensus protocol. The goal is to enforce the properties in Section 2.3.

The Pompē protocol. Pompē employs any standard leader-based BFT SMR protocol (*e.g.*, [32]) that offers a primitive to agree on a value for each slot in a sequence of consensus decisions. Pompē transforms such a protocol into a new one that enforces ordering properties.

Specifically, Pompē associates the slots with consecutive time intervals. For example, one slot may be associated with time interval $[t,\ t+500\,ms)$, and the next slot could be associated with interval $[t+500\,ms)$, and the next slot could be associated with interval $[t+500\,ms,\ t+1000\,ms)$. For simplicity, Pompē assumes that such a mapping from slots to time intervals is common knowledge. Within each consensus slot, the value to agree on is a set of $\langle c,\ ats \rangle$ pairs where c is a command and ats is called an $assigned\ timestamp$. Pompē provides two guarantees for this assigned timestamp: (1) ats falls in the time interval associated with the slot; (2) ats is bounded by the lowest and highest timestamps provided by correct nodes. The commands are then ordered by their assigned timestamps.

Pompē requires 3f+1 nodes and achieves these guarantees by collecting 2f+1 timestamps from different nodes for each command c. The *median* of the 2f+1 timestamps is chosen as the assigned timestamp ats for command c. Since there are at most f faulty nodes, the median of any 2f+1 timestamps is upper-bounded and lower-bounded by timestamps provided by correct nodes.

Integrating Pompē with an SRO. Bercow adds a random number to the assigned timestamp of each command. Concretely, after consensus is reached for slot k in Pompē, a correct node obtains a set of signatures, which proves that consensus has been reached. With k and this set of signatures, a correct node in Bercow invokes the Reveal SRO interface and obtains a random number used to seed in a random number generator. Importantly, no node can determine what seed the correct nodes will use until it has received sufficient signatures, and therefore, the seed is independent of the consensus decision.

The random number generator assigns a pseudorandom number r to each command, each uniformly sampled from $[0, \Delta_{noise})$. Section 3.4 describes how to select Δ_{noise} . Instead of directly ordering commands by their assigned timestamps, commands are ordered by ats + r.

We call a command c stable (or finalized) if the output ledger produced contains c. In Pompē, commands are ordered by the assigned timestamps, and commands in a slot become stable when this slot reaches consensus. After adding random noise, it takes longer for a command to become stable in Bercow. Suppose the latest slot that reaches consensus is associated with time interval [ts, ts'), a command c becomes

stable in Bercow if ats + r < ts', meaning that command c and all commands before c can be produced to the ledger.

Safety and liveness. Pompē guarantees the same safety and liveness properties as the classic SMR protocols [32, 54]. Bercow provides the same guarantees and differs only by how commands are ordered. We will now focus on proving the new ordering properties.

3.4 Quantifying a trade-off between ε-Ordering equality and Δ-Ordering linearizability

The remaining question is how to decide Δ_{noise} . We give a quantitative answer based on the upper bound Δ_{net} on message delivery latency, node processing time, and clock drift of correct nodes, defined by the partial synchrony model [38].

Partial synchrony model. One variant of the partial synchrony model introduces the *Global Stabilization Time* (GST). Specifically, there is an unknown time GST such that, after this time, there is a known bound Δ_{net} on network latency and processing time. The safety and liveness of Pompē are proven under this model. We now analyze the ordering properties of Bercow in the same model. More precisely:

Assumption 3.1. After the global stabilization time (GST), if a command is invoked at time T, correct nodes will provide timestamps in the range $[T, T + \Delta_{net}]$ for this command.

Note that a simple clock synchronization protocol has been given as part of the Pompē protocol. We now prove Δ -Ordering linearizability and ϵ -Ordering equality with $\Delta = \Delta_{net} + \Delta_{noise}$ and $\epsilon = 1 - (1 - \Delta_{net}/\Delta_{noise})^2$ in the two invocation case. This result implies a trade-off: as Δ_{noise} approaches infinity, ϵ will approach 0 while Δ will approach infinity.

Lemma 3.1. The assigned timestamp of a command is bounded by timestamps provided by correct nodes.

Proof. See [80].

Theorem 3.1. (Δ -Ordering linearizability) After the global stabilization time (GST), for all invocations i_1 and i_2 , if i_1 is invoked more than $\Delta_{net} + \Delta_{noise}$ earlier than i_2 , then i_1 is guaranteed to appear before i_2 in the output.

Proof. Suppose i_1 and i_2 are invoked at time T_1 and T_2 . By Assumption 3.1 and Lemma 3.1, the assigned timestamp of i_1 is in the range $[T_1, T_1 + \Delta_{net}]$. After adding the random noise, the resulting timestamp is in the range $[T_1, T_1 + \Delta_{net} + \Delta_{noise}]$. Similarly, the resulting timestamp for i_2 is in the range $[T_2, T_2 + \Delta_{net} + \Delta_{noise}]$. Therefore, if $T_2 > T_1 + \Delta_{net} + \Delta_{noise}$, i_2 will have a higher timestamp and appear after i_1 in the output.

Theorem 3.2. ($\epsilon(2)$ -Ordering equality) After the global stabilization time (GST), for all invocations i_1 and i_2 invoked at the same time, $|Pr[i_1 < i_2] - Pr[i_2 < i_1]| \le 1 - (1 - \Delta_{net}/\Delta_{noise})^2$.

Proof(sketch). Suppose i_1 and i_2 are both invoked at time T. By Assumption 3.1 and Lemma 3.1, the assigned timestamps

are in range $[T, T + \Delta_{net}]$. By assigning T to i_2 and $T + \Delta_{net}$ to i_1 , probability $Pr[i_1 < i_2]$ will be minimized. Therefore,

to
$$l_1$$
, probability $Pr[l_1 < l_2]$ will be infinitized. Therefore,
$$Pr[i_1 < i_2] \ge \frac{1}{\Delta_{noise}^2} \int_{T+\Delta_{net}}^{T+\Delta_{noise}} \int_{T}^{T+\Delta_{noise}} (t_1' < t_2') dt_2' dt_1'$$

$$\ge \frac{1}{2} (1 - \Delta_{net}/\Delta_{noise})^2$$

The ϵ parameter of ordering equality can be derived.

$$|Pr[i_2 < i_1] - Pr[i_1 < i_2]| = |1 - 2 * Pr[i_1 < i_2]|$$

 $\leq 1 - (1 - \Delta_{net}/\Delta_{noise})^2$

We provide the full proof in appendix A, which also proves the general theorem below for n > 2. As suggested in Section 2.1, real-world concerns about ordering could also arise due to violating ordering equality with three invocations (*i.e.*, enforcing $\epsilon(3)$ -Ordering equality is necessary to limit sandwich attacks).

Theorem 3.3. ($\epsilon(n)$ -Ordering equality) After the global stabilization time (GST), for all invocations $i_1..i_n$ invoked at the same time, for all two total orders of $i_1..i_n$ denoted as $>_1$ and $>_2$, $|Pr[>_1] - Pr[>_2]| \le 1/n!((1+\alpha)^n - (1-\alpha)^n - n\alpha^n)$ where α denotes the ratio $\Delta_{net}/\Delta_{noise}$.

Choosing the Δ_{noise} parameter. With Pompē and many other protocols, system designers tune their systems by choosing Δ_{net} . Suppose they now want to mitigate systemic bias, particularly front-running and sandwich attacks. In this case, they would assume n=3 and tune Δ_{noise} based on a target ϵ in Bercow. In certain legal contexts, $\epsilon \approx 0.1$ is used for equal opportunity: the so-called four-fifth rule [2, 3] says that if two candidates from different ethnic groups are equally qualified for a job, the difference between their chances of getting an offer cannot be more significant than 45% vs. 55%. This rule has also influenced machine learning fairness (e.g., demographic parity) [27].

3.5 Generalizing to multiple relevant features

The two ordering properties use invocation time as the only relevant feature. However, in the real world, there could be more relevant features such as *transaction fee*. Specifically, a client must pay specific fees to the nodes for executing its commands, and many blockchains have fixed transaction fees. In contrast, other blockchains allow clients to pay higher fees in exchange for a more favorable position in the ledger. While system designers make subjective decisions on what features are relevant, our results can be generalized to scenarios with multiple relevant features.

Specifically, system designers decide a formula that takes the measurement of each relevant feature and outputs a *score* (*e.g.*, a formula with invocation time and transaction fee as input). Similar to time, the score is a number, and the score given by a correct node may be inaccurate. We can thus regard the Δ_{net} parameter as the maximum difference

7

Clients from different geolocation (or using different network facilities) would be treated equally by Bercow.	
The expected profit of sandwich attacks could decrease significantly in Bercow.	
Baselines could all be significantly biased and there is a trade-off in Bercow between the two ordering properties.	
For performance, Bercow maintains the same throughput as Pompē and incurs moderate latency overhead.	§5.4, 5.5

Figure 6. Summary of evaluation results.

between an accurate score and a score provided by a correct node when reasoning about the ordering properties.

4 Implementation

We implement two SRO variants based on Section 3.1 and Section 3.2. For the first variant, we choose Intel's software guard extensions (SGX) [18] as the TEE. Random numbers are generated by the function SHA256 (seed + k) where seed is the random seed decided during SRO initialization and k is the parameter of Reveal. The SHA256 function is provided by the official SGX library for Linux [17]. This library does not provide a SHA512 function, which would make the SRO more secure in the blockchain context. Note that the Int type in Figure 4 and Figure 5 does not have to be the typical 4-byte integer. In blockchains, 32-byte and 64-byte integers are both commonly used.

For the second variant, we start with an implementation of threshold VRF in C++ [12, 40] and modify the Produce interface by adding signature verification. Random numbers here are the SHA512 hash of signatures combining a threshold of shares. The default configuration of this implementation uses the mcl cryptographic library [19] and the BN256 curve. Unlike many BFT systems, cryptographic libraries are not the performance bottleneck of Bercow. As we will explain later in the evaluation, the overhead is dominated by the noise and trade-off described in Section 3.4. Therefore, we have chosen the cryptographic libraries based on the convenience of implementation. In the two SRO variants, signature verification in the Reveal interface is implemented with the secp256k1 library from Bitcoin [20], which is also used by the Pompē implementation.

Ease of integration. We integrate the two SRO variants with a Pompē implementation based on HotStuff [10]. Recall that Pompē employs any leader-based BFT SMR protocol and transforms it into a new protocol that enforces ordering properties. This Pompē implementation employs HotStuff because HotStuff is the foundation of Diem [9], an influential blockchain project. Our modifications involve modest system effort and do not modify the complex components that achieve consensus. Instead, we only modify the component producing the totally ordered output, and we wrap this component into a new one that applies the random noise and waits for the new stability condition, as described in Section 3.3. Our experience suggests that it could be easy to integrate other consensus systems with an SRO for the purpose of equal opportunity.

5 Experimental evaluation

We ask two main questions in our evaluation: (1) How do the new ordering properties mitigate front-running, geographical bias, and sandwich attacks? Why are baselines vulnerable?; and (2) What is the end-to-end performance of Bercow? Figure 6 shows a summary of our findings.

We choose three baselines: HotStuff [77], Pompē [80] and Themis [49] representing different existing fairness concepts. HotStuff adopts a *rotating leadership* fairness concept, and we configure HotStuff by making all nodes serve as the leader for the same amount of time. Themis guarantees a fairness property called γ -batch-order-fairness, and we choose $\gamma=1$ which informally means that if all correct nodes receive i_1 before i_2 , then i_1 should be ordered before i_2 in the system output. Themis only provides the simulation code instead of an actual system implementation. Pompē adopts the concept of *removing oligarchy* and, as shown in Section 3.3, the output order in Pompē is not unilaterally decided by a leader. We implement two variants of SRO as described in Section 4 and integrate them with Pompē as Bercow.

Configuration and metrics. We run Bercow and the three baselines on 52 machines in CloudLab [14] (m400, 8-core ARMv8, 64GB memory, Ubuntu Linux 20.04 LTS). We run the SROs on a separate set of machines with the Intel Xeon Silver 4410Y and Intel Xeon D-1548 CPUs because our two SRO implementations rely on cryptographic acceleration, and the TEE-based SRO uses Intel SGX. Network latency across different geolocations is emulated using the Linux traffic control (tc) utility.

Statistics show that the top countries running Ethereum nodes are: US (40%), Germany (12%), Singapore (5%), UK (4%), Netherlands (4%), France (3%), Japan (3%), Canada (3%), Australia (3%) and Finland (3%) [15]. To answer the evaluation question about fairness, we run 80 nodes simulating these statistics and then evaluate the impact of bias or attacks. For the US, we simulate 15 nodes in Washington, 15 in San Francisco, and 10 in Austin, representing the eastern, western, and central US. For other countries, we simulate all the nodes in one of their major cities. Latency information is from WonderNetwork [16]. The maximum latency in such a configuration is 296ms from Canberra in Australia to Oulu in Finland. We thus assume $\Delta_{net} = 300 \, ms$ which is similar to the choice of 400ms in a real-world blockchain protocol [23]. We use this configuration of Ethereum to reflect the uneven distribution of blockchain nodes despite all systems we evaluate being permissioned systems.

Our metric for fairness for two commands is the difference between the probabilities of the two possible relative orders of the two commands in the ledger. When two commands are invoked at the same time in Washington and London, the probability of the Washington invocation appearing first in the ledger is 0.76 (denoted as Pr[W < L] = 0.76) in our measurement of HotStuff. Therefore, Pr[W < L] - Pr[L < W] = 0.76 - 0.24 = 0.52, which is much higher than a reasonable target ϵ (e.g., 0.1).

To answer the evaluation question about performance, we measure the latency and throughput of Pompē and Bercow by scaling the systems from 4 to 49 nodes. In these experiments, each node runs on a separate machine, and each machine is configured with a 150 ms outbound network latency with the tc utility. We fix a setup with 800 concurrent clients, and each client invokes commands in a closed loop (i.e., it waits for the consensus result of its currently outstanding command before invoking the next one). We aim to measure the overhead of Bercow over Pompē in a geo-distributed setup. As for batching, we use 1500 ms as the time interval associated with each consensus slot, and each assigned (i.e., median) timestamp is associated with only one command.

5.1 Bias and front-running

Figure 7 shows that geographical bias could be significant in the baselines. The output order produced by Pompē and Themis is deterministic: simultaneous invocations from the four cities are always ordered as Washington < London < Munich < Tokyo. The reason is that in our configuration, simulating the node distribution of Ethereum, Washington has a lower network latency than most of the 80 nodes in the other cities. Therefore, in Pompē, the median timestamp of any quorum for the Washington invocation will be lower. The fairness property of Themis directly implies that, in this configuration, Washington should be ordered first.

HotStuff rotates leadership, and the invocation from Tokyo would have a chance to appear first in the ledger when a Tokyo node serves as the leader. However, this chance is still low compared to Munich invocations. Specifically, we find that Pr[M < T] = 0.925, leading to the 0.85 in the table. This difference increases to 0.93 when compared to Washington. While Tokyo and Washington are geographically distant, a similar bias can happen between nearby cities. London and Munich are both in Europe, but Pr[L < M] is 0.725 leading to the 0.45 in the table.

Figure 8 shows how geographical bias could be effectively reduced in Bercow. By adding a random noise sampled from $[0, \Delta net]$ (*i.e.*, [0, 300ms]) to the median timestamp of Pompē, all probability differences could be controlled under 0.4. If $\epsilon = 0.1$ is a target for the system, system designers could then choose $\Delta_{noise} = 5 * \Delta_{net}$ (*i.e.*, 1500*ms*), and the worst case bias across the four cities could be controlled under 0.087 as shown in the third line. In real-world deployments, Δ_{noise}

Baselines	HotStuff	Pompē/Themis
Pr[W < L] - Pr[L < W]	0.52	1
Pr[W < T] - Pr[T < W]	0.93	1
Pr[L < M] - Pr[M < L]	0.45	1
Pr[M < T] - Pr[T < M]	0.85	1

Figure 7. Geographical bias measured in HotStuff, Pompē and Themis. W, L, M and T stand for Washington, London, Munich and Tokyo. For two simultaneous invocations from two cities, Pr[City1 < City2] stands for the probability of the invocation from City1 being the first in the ledger.

Bercow	$\Delta_{noise} = \Delta_{net}$	$\Delta_{noise} = 5 * \Delta_{net}$
Pr[W < L] - Pr[L < W]	0.036	0.007
Pr[W < M] - Pr[M < W]	0.158	0.033
Pr[W < T] - Pr[T < W]	0.399	0.087
Pr[L < M] - Pr[M < L]	0.119	0.025
Pr[L < T] - Pr[T < L]	0.367	0.082
Pr[M < T] - Pr[T < M]	0.269	0.056

Figure 8. Geographical bias measured in Bercow with $\Delta_{noise} = \Delta_{net}$ (*i.e.*, 300ms) and $\Delta_{noise} = 5 * \Delta_{net}$ (*i.e.*, 1500ms).

could be empirically chosen by considering the most distant clients for fairness.

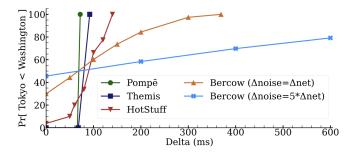
Front-running typically occurs when one client has lower network latencies to most nodes than another, as Washington does in this configuration. By adding random noise, Bercow could give all the clients involved in the liquidation events (explained in Section 2.1) more balanced chances of obtaining the \$42.6M profit over 32 months.

5.2 Sandwich attacks

An exchange maintains a pool of some token A (*e.g.*, USD) and some token B (*e.g.*, CNY). For example, people traveling from the US to China may put some USD into the pool and take some CNY away. Similar pools exist on blockchains, and the trading volume of Uniswap, a decentralized exchange, has exceeded one trillion dollars [7].

Pools in these exchanges follow a constraint: *amount of tokenA* * *amount of tokenB* = *constant*, which is called the automated market makers (AMMs) approach [22]. Suppose the constant is 1800; the number of tokens A and B in the pool could be, for instance, $\langle 45, 40 \rangle$, $\langle 60, 30 \rangle$, or $\langle 75, 24 \rangle$. Say $\langle 75, 24 \rangle$ is the current status, and Alice needs 15 token A. Alice can put 6 token B into the pool and take 15 token A away so that the pool state moves to $\langle 60, 30 \rangle$.

The sandwich attack works as follows. After seeing Alice's transaction, an attacker, Bob, first buys 15 token A, moving the pool status to $\langle 60, 30 \rangle$. Alice now needs to pay 10 (instead of 6) token B in order to exchange for 15 token A and move the pool status to $\langle 45, 40 \rangle$. Bob can thus exchange the 15 token A back to 10 token B, making a 4 token B profit. The three steps reflect the three invocations in Figure 1. If the



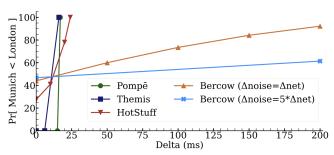


Figure 9. Trade-off between ϵ -Ordering equality and Δ -Ordering linearizability in Bercow. The left figure shows the results for two clients in Washington and Tokyo. The right figure shows the results for two clients in London and Munich.

market prices of the tokens are \$100 and \$200, respectively, we will get the dollar values in Figure 1.

The success of sandwich attacks depends on whether the attacker's first invocation (i_1 in Figure 1) can front-run the victim's invocation since the attacker can always close the attack by delaying its last invocation. Suppose the network conditions of the attacker and victim are similar to those of London and Munich. Figure 7 shows that, in systems like Pompē or Themis, the attacker can succeed with a high probability. In Bercow, the attacker would have a lower expected profit because of ϵ -Ordering equality.

5.3 Trade-off between ϵ -Ordering equality and Δ -Ordering linearizability

Figure 9 shows the relation between ordering and invocation time for two invocations from Washington and Tokyo (or London and Munich). Specifically, the x-axis represents how long the Tokyo client invokes its commands *before* the Washington client. The y-axis represents the probability of the Tokyo invocation being ordered first. Intuitively, when the Tokyo client invokes its command early enough, its probability of being ordered first will be 100%.

For Bercow, with $\Delta_{noise} = \Delta_{net}$, the two invocations are treated less equally when the x-axis is 0 compared to $\Delta_{noise} = 5 * \Delta_{net}$, but it leads to a lower Δ (*i.e.*, 369ms) for ordering linearizability. In other words, for Bercow with $\Delta_{noise} = \Delta_{net}$, the y-axis will reach 100% when the x-axis is 369ms. With $\Delta_{noise} = 5 * \Delta_{net}$, the probability of the Tokyo invocation ordered first is only 69.8% even if it is invoked 400ms earlier. This shows the trade-off between removing the influence of irrelevant features and preserving the signal strength of the relevant features when adding the random noise in Bercow.

The results for Pompē and Themis look similar. For Pompē, the Tokyo invocation is guaranteed to be ordered first if it is invoked more than 72ms earlier, while this threshold is 92ms for Themis. The reason is that Pompē and Themis order commands based on when nodes receive the invocations. When the difference in invocation time is above this threshold, most of the nodes in our configuration will receive the

	s =0 (base)	s =200
TEE	3us	base+20.2ms
TVRF (67/100)	95.2+4.3ms	base+19.9ms
TVRF (133/200)	185.7+9.7ms	base+19.9ms

Figure 10. Latency of the Reveal interface in different SRO implementations. |s| denotes the number of signatures to be verified in the second parameter of Reveal. The numbers in the parentheses are the threshold and total number of nodes for TVRF. The base case of TVRF consists of two latency results for generating and combining shares.

Tokyo invocation earlier. The result for HotStuff starts from 3.8% when the x-axis is 0 and grows to 100% when the x-axis is 140*ms*. This result follows the intuition that the Tokyo invocation will have better chances of reaching the leader node earlier if it is invoked earlier.

The right part of Figure 9 shows the results of invocations from Munich and London. Compared to the left part, we make two observations. First, for HotStuff and Bercow, the chances are closer to 50% when the x-axis is 0. Second, for Pompē and Themis, the threshold becomes 17ms, which is lower than 72ms or 92ms. Both observations are because these two cities are geographically closer.

Given $\epsilon = 0.1$ as a target, our results in Section 3.4 show that $\Delta_{noise} = 20 * \Delta_{net}$ is necessary to provide a guarantee of this target. This is because our results assume the worst case: two simultaneous invocations would obtain timestamps that differ by Δ_{net} . In practice, the worst-case difference could be much lower than Δ_{net} so that system designers can choose this parameter according to their actual deployment setup.

5.4 Latency of secret random oracles

Figure 10 shows the latency of the two SRO implementations. We show two cases of |s|=0 and |s|=200, where |s| denotes the number of signatures to be verified in the second parameter of Reveal. When |s|=0, Reveal does not verify signatures, and the latency is solely for generating random

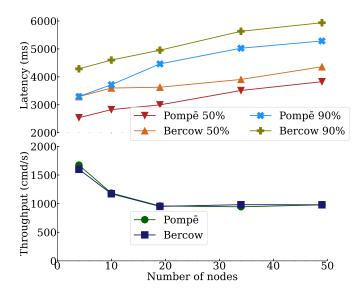


Figure 11. End-to-end performance of Pompē and Bercow with $\Delta_{noise} = 5 * \Delta_{net}$ (*i.e.*, 1500 ms).

numbers. For the TEE variant, the latency consists of entering an SGX enclave and computing a SHA256 function, which takes only 3us. For the threshold VRF variant, the latency consists of three parts: (1) generating shares, (2) collecting the shares over the network, (3) combining a threshold of shares. The results in Figure 10 show (1) and (3). Under a setup of 100 nodes and 67 as the threshold, the latency of producing a share is 95.2ms, and the latency of combining 67 shares is 4.3ms. When moving to a setup of 200 nodes with 133 as the threshold, the two latencies roughly double. This is because threshold VRF algorithms have a high constant factor in their computational complexity, which dominates latency in these setups. Lastly, when |s| = 200, the latency of verifying the 200 signatures is around 20ms, both within and outside an SGX enclave.

These results show a trade-off between performance and decentralization. The SRO based on SGX has a much lower latency, making it more practical, but it requires trusting a centralized party, Intel. In the following experiments, we choose performance in this trade-off and use the TEE variant of SRO, since we have observed TEE-based projects running on the Ethereum Sepolia testnet [13].

5.5 End-to-end performance of Bercow

Figure 11 shows the latency and throughput of Bercow and Pompē when scaling from 4 to 49 nodes. The configurations of these experiments are described at the beginning of this section. The x-axis represents the number of nodes. The y-axis of the upper part represents end-to-end latency in milliseconds. The y-axis of the lower part represents system throughput in command per second.

Bercow achieves the same throughput as Pompē, meaning that integrating with an SRO does not impact throughput.

This is expected because obtaining random numbers from an SRO is much cheaper than consensus, so the consensus part of Bercow is the throughput bottleneck.

As for latency, the median latency for 49 nodes increases from 3827 ms in Pompē to 4361 ms in Bercow, increasing by 534 ms. The 90 percentile tail latency increases from 5285 ms to 5939 ms, increasing by 654 ms. The relative increases are only 14% and 12.4%, respectively.

In real-world blockchains, the typical end-to-end latency is on the magnitude of minutes (*e.g.*, Bitcoin) or seconds (*e.g.*, Ethereum) which are higher than the latency results in Figure 11. However, the random noise being added could be similar to our experiments (*e.g.*, $\Delta_{noise} = 1500 \, ms$), which would lead to a potentially lower relative latency increase.

6 Related work

Traditional BFT SMR systems. There is a long line of work on traditional BFT SMR systems [28, 33, 35, 36, 43, 44, 46, 51, 52, 57–61, 65, 72, 73, 76], starting with the seminal work of PBFT [32]. These works focus on enforcing safety and liveness, removing or constraining Byzantine influence, improving performance or theoretical complexity, etc. Our work focuses on how to choose the output order in SMR, which is not considered by the traditional specification. Some works [28, 34, 47, 55] use trusted hardware to increase the ratio of Byzantine nodes that the system can tolerate. Other works [30, 40, 42] use randomness to elect a committee or achieve safety and liveness in a fully asynchronous model. Unlike these works, our work uses trusted hardware to provide a fault-tolerant source of randomness and applies randomness for equal opportunity.

Rotating leadership. Some works adopt a leader and rotate the leadership frequently, but they focus on reducing theoretical complexity or preventing faulty leaders from degrading performance. Aardvark [36] employs periodic leader changes to ensure a certain degree of performance in the presence of faulty leaders. Aardvark sets an expectation on the minimal throughput that a leader must ensure and triggers a leader change if the current leader fails to meet such an expectation. Different from Aardvark, HotStuff [77] employs leader rotation and optimizes the communication complexity. Specifically, HotStuff's communication complexity is linear in the number of nodes, which makes it more suitable for blockchains. Adopted by Diem [9], rotating leadership in HotStuff aims to provide some sense of fairness in a permissioned blockchain. Our work instead specifies and enforces a concrete notion of fairness, namely equal opportunity, and our evaluation results show that rotating leadership could cause significant bias in a real-world deployment.

Removing oligarchy. Leaderless protocols argue against having a leader node who can unilaterally decide the output order. EPaxos [62], or Egalitarian Paxos, is an SMR protocol that attempts to make the system egalitarian. While the

concept of egalitarianism is closely related to equal opportunity, EPaxos does not specify nor enforce egalitarian ideals except being a leaderless protocol. Byzantine oligarchy [80] is a first attempt to specify the goal of leaderless protocols in the context of ordering and Pompē is the first leaderless protocol that provably removes a Byzantine oligarchy. To achieve this, Pompē requires a client to collect timestamps from a quorum of nodes, and the median is used to order a command. However, our evaluation shows that using such a median timestamp could make the system even more biased and less equal than prior work such as HotStuff. In contrast to EPaxos and Pompē, Bercow enforces a notion of fairness that is unbiased by definition. These definitions and the design of Bercow were first published in a thesis [79].

Decentralized first-come-first-serve. Some recent works define and enforce fairness concepts related to first-come-first-serve [31, 48–50, 63, 67]. Specifically, these protocols enforce variants of the receive-order-fairness property [50], which essentially says that if a majority of the nodes receive an invocation first, it should be ordered first in the output. We argue that this property can be unfair because, without distinguishing relevant features from irrelevant ones, it can amplify systemic bias in real-world blockchains, as shown in the evaluation.

While these works enforce specific properties related to first-come-first-serve, the framework of ordered consensus makes it possible to prove that it is impossible, in general, to prevent Byzantine replicas from manipulating the order (e.g., from conducting front-running) [80]. This result is inspired by the Arrow's impossibility theorem [26] and the Gibbard-Satterthwaite impossibility theorem [41, 69] from the field of social choice theory. In the past two decades, computer scientists became interested in social choice theory, leading to the creation of the field of computational social choice [29].

Game theory. The BAR model [24, 25] explores how to connect Byzantine fault tolerance to game theory. The core of the connection is adopting the Nash theorem [64], which states that every normal-form game must have an equilibrium. The Nash theorem connects the rationality model [70, 75] with the Brouwer fixed-point theorem (*i.e.*, a fixed point is interpreted as an equilibrium). As practitioners, we find the concept of equal opportunity and its violations more prevalent in real-life scenarios than Nash equilibrium. We do reuse the key concept of the expected value from the rationality model when explaining violations of equal opportunity in Section 2.1. We are also inspired by one of the axioms at the core of the rationality model [70] when defining consistency in Section 2.2. However, unlike BAR, we do not reuse the concept of Nash equilibrium in this work.

Equal opportunity in real-life scenarios. The principles of impartiality and consistency have been scrutinized in the context of how society allocates resources. In his book [78], Young studies the two principles in a variety of contexts, from

employment to kidney exchange, and discusses how they are embodied in key pieces of legislation [1, 2, 4, 5]. The book includes a proof that a point system is the only mechanism that satisfies both impartiality and consistency. Our work adopts this book's approach: we use the invocation time as the score for ordering and use a secret random oracle to break ties.

Financial regulation laws require financial exchanges to be impartial to all traders. A recent exposé [56], however, has concluded that high-frequency traders have routinely engaged in market-exploiting behaviors, such as aggressive latency optimizations and front-running by exploiting their location or the availability of fast network connections—which our framework would stigmatize as irrelevant features.

Kidneys for transplant used to be exchanged through a free market, and wealthy patients had a better chance of getting kidneys. This raised fairness concerns and led to legislation that transferred the operation to the government in order to overcome such bias towards wealthy patients [4]. While this law enforced *impartiality* between the rich and poor, *consistency* became a concern. The first algorithm proposed was inconsistent, and the order of two patients getting kidneys could be switched due to a third patient joining the system. As a result, a new algorithm was proposed as an amendment later, enforcing the consistency principle [5]. Details of the relevant features and the point system in this case have been discussed in Chapter 2 of this book [78].

In the Olympic games, a game decides an order of athletes and needs to give equal opportunities. The Olympic rules specify which relevant features should be measured in each game through equipment or human judges. A key irrelevant feature is nationality, and judgments should not be biased toward any nation. Besides, due to cognitive bias, a judge may make correlated decisions when judging a sequence of observations. For example, in diving, athletes dive alternatively, and an athlete's performance may impact the scores given to the next athlete [53]. Judges typically need professional training in order to combat such implicit bias and make consistent judgments.

7 Conclusion

This paper introduces equal opportunity, a concrete notion of ordering fairness in SMR based on the distinction between relevant and irrelevant features. Existing protocols – including ones that attempt to provide some fairness – can be significantly biased and vulnerable to ordering attacks, even without Byzantine replicas.

We design, implement, and evaluate Bercow, a new ordered consensus protocol that guarantees two ordering properties for equal opportunity, ϵ -Ordering equality and Δ -Ordering linearizability. Bercow effectively mitigates the well-known ordering attacks while eliminating such attacks in the presence of Byzantine influence is impossible.

References

- [1] Civil rights act.
 - https://www.congress.gov/bill/88th-congress/house-bill/7152, 1964.
- [2] Equal employment opportunity act. https: //www.eeoc.gov/history/equal-employment-opportunity-act-1972, 1972.
- [3] Uniform guidelines on employee selection procedures. https://www.ecfr.gov/current/title-29/subtitle-B/chapter-XIV/part-1607, 1978
- [4] National organ transplant act. https://www.congress.gov/bill/98th-congress/senate-bill/2048, 1984.
- [5] National organ transplant program extension act. https://www.congress.gov/bill/101st-congress/house-bill/5146, 1990.
- [6] MakerDAO. https://makerdao.com, 2017.
- [7] Uniswap. https://uniswap.org/, 2018.
- [8] Blockchains in trusted execution environments (TEEs).
 https://medium.com/@nadeem.bhati/blockchains-in-trusted-execution-environments-tees-9343b6c3f9e8, 2019.
- [9] Diem. https://www.diem.com/en-us/, 2019.
- [10] A Pompē implementation based on HotStuff. https://github.com/Pompe-org/Pompe-HS, 2020.
- [11] Aave liquidity protocol. https://aave.com/, 2020.
- [12] A threshold VRF implementation. https://github.com/fetchai/research-dvrf, 2020.
- https://github.com/fetchai/research-dvrf, 2020.
 [13] Block building inside SGX.
- https://writings.flashbots.net/block-building-inside-sgx, 2023.
- [14] CloudLab. https://cloudlab.us/, 2023.
- [15] Ethereum mainnet statistics. https://ethernodes.org/countries?synced=1, 2023.
- [16] Global ping statistics: Ping times between WonderNetwork servers. https://wondernetwork.com/pings, 2023.
- [17] Intel software guard extensions for Linux OS. https://github.com/intel/linux-sgx, 2023.
- [18] Intel software guard extensions (SGX). https://www.intel.com/content/www/us/en/architecture-and-technology/software-guard-extensions.html, 2023.
- [19] mcl: A portable and fast pairing-based cryptography library. https://github.com/herumi/mcl, 2023.
- [20] The secp256k1 library from Bitcoin. https://github.com/bitcoin-core/secp256k1, 2023.
- [21] The sting framework (SF). https://initc3org.medium.com/the-sting-framework-sf-ef00702c88c7, 2023.
- [22] What are automated market makers (AMMs)? https://chain.link/education-hub/what-is-an-automated-market-maker-amm, 2023.
- [23] Alpenglow: A New Consensus for Solana. https://www.anza.xyz/blog/alpenglow-a-new-consensus-for-solana, 2025.
- [24] Ittai Abraham, Lorenzo Alvisi, and Joseph Y Halpern. Distributed computing meets game theory: Combining insights from two fields. ACM Sigact News, 42(2):69–76, 2011.
- [25] Amitanand S Aiyer, Lorenzo Alvisi, Allen Clement, Mike Dahlin, Jean-Philippe Martin, and Carl Porth. BAR fault tolerance for cooperative services. In Proceedings of the ACM Symposium on Operating Systems Principles (SOSP), pages 45–58, 2005.
- [26] Kenneth J Arrow. Social choice and individual values, volume 12. Yale University Press, 1951.
- [27] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and Machine Learning: Limitations and Opportunities. MIT Press, 2023.
- [28] Johannes Behl, Tobias Distler, and Rudiger Kapitza. Hybrids on steroids: SGX-based high performance BFT. In Proceedings of the ACM European Conference on Computer Systems (EuroSys), pages 222–237, 2017.

- [29] Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D Procaccia. Handbook of computational social choice. Cambridge University Press, 2016.
- [30] Christian Cachin, Klaus Kursawe, and Victor Shoup. Random oracles in constantinople: Practical asynchronous Byzantine agreement using cryptography. In Proceedings of the ACM Symposium on Principles of Distributed Computing (PODC), page 123–132, 2000.
- [31] Christian Cachin, Jovana Mićić, Nathalie Steinhauer, and Luca Zanolini. Quick order fairness. arXiv:2112.06615, 2022.
- [32] Miguel Castro and Barbara Liskov. Practical Byzantine fault tolerance and proactive recovery. ACM Transactions on Computer Systems (TOCS), 20(4):398–461, 2002.
- [33] Miguel Castro, Rodrigo Rodrigues, and Barbara Liskov. BASE: Using abstraction to improve fault tolerance. ACM Transactions on Computer Systems (TOCS), 21(3):236–269, 2003.
- [34] Byung-Gon Chun, Petros Maniatis, Scott Shenker, and John Kubiatowicz. Attested append-only memory: Making adversaries stick to their word. In *Proceedings of the ACM Symposium on Operating Systems Principles (SOSP)*, pages 189–204, 2007.
- [35] Allen Clement, Manos Kapritsos, Sangmin Lee, Yang Wang, Lorenzo Alvisi, Mike Dahlin, and Taylor Riche. UpRight cluster services. In Proceedings of the ACM Symposium on Operating Systems Principles (SOSP), pages 277–290, 2009.
- [36] Allen Clement, Edmund Wong, Lorenzo Alvisi, Mike Dahlin, and Mirco Marchetti. Making Byzantine fault tolerant systems tolerate Byzantine faults. In Proceedings of the USENIX Symposium on Networked Systems Design and Implementation (NSDI), pages 153–168, 2009
- [37] Philip Daian, Steven Goldfeder, Tyler Kell, Yunqi Li, Xueyuan Zhao, Iddo Bentov, Lorenz Breidenbach, and Ari Juels. Flash boys 2.0: Frontrunning, transaction reordering, and consensus instability in decentralized exchanges. In Proceedings of the IEEE Symposium on Security and Privacy (S&P), pages 910–927, 2020.
- [38] Cynthia Dwork, Nancy Lynch, and Larry Stockmeyer. Consensus in the presence of partial synchrony. Journal of the ACM (JACM), 35(2), 1988.
- [39] Michael J. Fischer, Nancy A. Lynch, and Michael S. Paterson. Impossibility of distributed consensus with one faulty process. In Proceedings of the Symposium on Principles of Database Systems, pages 1–7, 1983.
- [40] David Galindo, Jia Liu, Mihair Ordean, and Jin-Mann Wong. Fully distributed verifiable random functions and their application to decentralised random beacons. In *Proceedings of the IEEE European* Symposium on Security and Privacy (EuroS&P), pages 88–102, 2021.
- [41] Allan Gibbard. Manipulation of voting schemes: a general result. Econometrica: Journal of the Econometric Society, pages 587–601, 1973.
- [42] Yossi Gilad, Rotem Hemo, Silvio M Micali, Georgios Vlachos, and Nickolai Zeldovich. Algorand: Scaling Byzantine agreements for cryptocurrencies. In Proceedings of the ACM Symposium on Operating Systems Principles (SOSP), pages 51–68, 2017.
- [43] Guy Golan Gueta, Ittai Abraham, Shelly Grossman, Dahlia Malkhi, Benny Pinkas, Michael K. Reiter, Dragos-Adrian Seredinschi, Orr Tamir, and Alin Tomescu. SBFT: A scalable decentralized trust infrastructure for blockchains. arxiv:1804/01626v1, 2018.
- [44] Andreas Haeberlen, Petr Kouznetsov, and Peter Druschel. PeerReview: Practical accountability for distributed systems. In Proceedings of the ACM Symposium on Operating Systems Principles (SOSP), pages 175–188, 2007.
- [45] Joseph Y Halpern. Reasoning about Uncertainty. MIT Press, 2017.
- [46] James Hendricks, Shafeeq Sinnamohideen, Gregory R Ganger, and Michael K Reiter. Zzyzx: Scalable fault tolerance through Byzantine locking. In Proceedings of the Internal Conference on Dependable Systems and Networks (DSN), pages 363–372, 2010.

- [47] Rüdiger Kapitza, Johannes Behl, Christian Cachin, Tobias Distler, Simon Kuhnle, Seyed Vahid Mohammadi, Wolfgang Schröder-Preikschat, and Klaus Stengel. CheapBFT: Resource-efficient Byzantine fault tolerance. In Proceedings of the ACM European Conference on Computer Systems (EuroSys), pages 295–308, 2012.
- [48] Mahimna Kelkar, Soubhik Deb, and Sreeram Kannan. Order-fair consensus in the permissionless setting. In Proceedings of the ACM on ASIA Public-Key Cryptography Workshop, pages 3–14, 2022.
- [49] Mahimna Kelkar, Soubhik Deb, Sishan Long, Ari Juels, and Sreeram Kannan. Themis: Fast, strong order-fairness in Byzantine consensus. In Proceedings of the ACM Conference on Computer and Communications Security (CCS), page 475–489, 2023.
- [50] Mahimna Kelkar, Fan Zhang, Steven Goldfeder, and Ari Juels. Order-fairness for Byzantine consensus. In Proceedings of the International Cryptology Conference (CRYPTO), pages 451–480, 2020.
- [51] Ramakrishna Kotla, Lorenzo Alvisi, Mike Dahlin, Allen Clement, and Edmund Wong. Zyzzyva: Speculative Byzantine fault tolerance. In Proceedings of the ACM Symposium on Operating Systems Principles (SOSP), pages 45–58, 2007.
- [52] Ramakrishna Kotla and Mike Dahlin. High throughput Byzantine fault tolerance. In Proceedings of the Internal Conference on Dependable Systems and Networks (DSN), pages 575–584, 2004.
- [53] Robin SS Kramer. Sequential effects in Olympic synchronized diving scores. Royal Society Open Science, 4(1):160812, 2017.
- [54] Leslie Lamport. The part-time parliament. ACM Transactions on Computer Systems (TOCS), 16(2):133–169, 1998.
- [55] Dave Levin, John R. Douceur, Jacob R. Lorch, and Thomas Moscibroda. TrInc: Small trusted hardware for large distributed systems. In Proceedings of the USENIX Symposium on Networked Systems Design and Implementation (NSDI), pages 1–14, 2009.
- [56] Michael Lewis. Flash boys: A wall street revolt. W. W. Norton & Company, 2014.
- [57] Jinyuan Li and David Maziéres. Beyond one-third faulty replicas in Byzantine fault tolerant systems. In Proceedings of the USENIX Symposium on Networked Systems Design and Implementation (NSDI), pages 131–144, 2007.
- [58] J. Liu, W. Li, G. O. Karame, and N. Asokan. Scalable Byzantine consensus via hardware-assisted secret sharing. *IEEE Transactions on Computers*, 68(1), 2019.
- [59] Shengyun Liu, Paolo Viotti, Christian Cachin, Vivien Quéma, and Marko Vukolic. XFT: Practical fault tolerance beyond crashes. In Proceedings of the USENIX Symposium on Operating Systems Design and Implementation (OSDI), pages 485–500, 2016.
- [60] Shengyun Liu, Wenbo Xu, Chen Shan, Xiaofeng Yan, Tianjing Xu, Bo Wang, Lei Fan, Fuxi Deng, Ying Yan, and Hui Zhang. Flexible advancement in asynchronous BFT consensus. In Proceedings of the ACM Symposium on Operating Systems Principles (SOSP), page 264–280, 2023.
- [61] Jean-Philippe Martin and Lorenzo Alvisi. Fast Byzantine consensus. IEEE Transactions on Dependable and Secure Computing, 3(3):202–215, 2006
- [62] Iulian Moraru, David G. Andersen, and Michael Kaminsky. There is more consensus in egalitarian parliaments. In Proceedings of the ACM Symposium on Operating Systems Principles (SOSP), pages 358–372, 2013.
- [63] Ke Mu, Bo Yin, Alia Asheralieva, and Xuetao Wei. Separation is good: A faster order-fairness Byzantine consensus. In Proceedings of the Network and Distributed System Security Symposium (NDSS), 2024.
- [64] John Nash. Non-cooperative games. *Annals of Mathematics*, pages 286–295, 1951.
- [65] Daniel Porto, João Leitão, Cheng Li, Allen Clement, Aniket Kate, Flavio Junqueira, and Rodrigo Rodrigues. Visigoth fault tolerance. In Proceedings of the ACM European Conference on Computer Systems (EuroSys), pages 1–14, 2015.

- [66] Kaihua Qin, Liyi Zhou, and Arthur Gervais. Quantifying blockchain extractable value: How dark is the forest? In Proceedings of the IEEE Symposium on Security and Privacy (S&P), pages 198–214, 2022.
- [67] Geoffrey Ramseyer and Ashish Goel. Fair ordering via streaming social choice theory. arXiv:2304.02730, 2023.
- [68] Bertrand Russell. Am I an Atheist or an Agnostic? 1950.
- [69] Mark Allen Satterthwaite. Strategy-proofness and arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10(2):187–217, 1975
- [70] Leonard J Savage. The Foundations of Statistics. Courier Corporation, 1972.
- [71] Fred B. Schneider. Implementing fault-tolerant services using the state machine approach: A tutorial. ACM Computing Surveys, 22(4):299–319, 1990.
- [72] Joao Sousa, Alysson Bessani, and Marko Vukolic. A Byzantine fault-tolerant ordering service for the hyperledger fabric blockchain platform. In Proceedings of the Internal Conference on Dependable Systems and Networks (DSN), pages 51–58, 2018.
- [73] Florian Suri-Payer, Matthew Burke, Zheng Wang, Yunhao Zhang, Lorenzo Alvisi, and Natacha Crooks. Basil: Breaking up BFT with ACID (transactions). In Proceedings of the ACM Symposium on Operating Systems Principles (SOSP), pages 1–17, 2021.
- [74] Christof Ferreira Torres, Ramiro Camino, and Radu State. Frontrunner jones and the raiders of the dark forest: An empirical study of frontrunning on the Ethereum blockchain. In *Proceedings of the USENIX Security Symposium*, pages 1343–1359, 2021.
- [75] John von Neumann and Oskar Morgenstern. Theory of Games and Economic Behavior. Princeton University Press, 1947.
- [76] Jian Yin, Jean-Philippe Martin, Arun Venkataramani, Lorenzo Alvisi, and Mike Dahlin. Separating agreement from execution for Byzantine fault tolerant services. In *Proceedings of the ACM Symposium on Operating Systems Principles (SOSP)*, pages 253–267, 2003.
- [77] Maofan Yin, Dahlia Malkhi, Michael K Reiter, Guy Golan Gueta, and Ittai Abraham. HotStuff: BFT consensus with linearity and responsiveness. In Proceedings of the ACM Symposium on Principles of Distributed Computing (PODC), pages 347–356, 2019.
- [78] H Peyton Young. Equity: In Theory and Practice. Princeton University Press. 1995.
- [79] Yunhao Zhang. Ordered Consensus with Equal Opportunity. PhD thesis, May 2024.
- [80] Yunhao Zhang, Srinath Setty, Qi Chen, Lidong Zhou, and Lorenzo Alvisi. Byzantine ordered consensus without Byzantine oligarchy. In Proceedings of the USENIX Symposium on Operating Systems Design and Implementation (OSDI), pages 633–649, 2020.
- [81] Liyi Zhou, Kaihua Qin, Christof Ferreira Torres, Duc V Le, and Arthur Gervais. High-frequency trading on decentralized on-chain exchanges. In Proceedings of the IEEE Symposium on Security and Privacy (S&P), pages 428–445, 2021.

Appendix A Full proofs and remarks

A.1 Protocol overview

The protocols we consider here are based on Pompē. In Pompē, every command c_i is associated with a timestamp t_i , and all the commands are ordered based on their timestamps. Pompē guarantees that every t_i is bounded on both sides by a timestamp from a correct node.

To reduce systemic bias, our new protocol independently samples a random noise η_i from the probabilistic distribution Φ for each command and orders all the commands based on the modified timestamp $t_i' = t_i + \eta_i$, where t_i is the timestamp produced by Pompē.

For clarity, we call the timestamp given by Pompē the *assigned timestamp* (of a command) and the timestamp with added noise *modified timestamp* (of a command).

For simplicity, this write-up analyzes the case where Φ is a uniform distribution on $[0, \Delta_{noise}]$ with Δ_{noise} being a protocol parameter we can choose.

A.2 Proofs

We first assume a time bound Δ_{net} , which includes the network latency and additional slack for clock drifts across nodes.

Assumption A.1. The assigned timestamp given to any command c sent by a client at time T is in the interval $[T, T+\Delta_{net}]$.

Pompē can guarantee this assumption in a synchronous network with appropriate parameter $\Delta_{\text{net}}.$ Pompē does not guarantee which value in this interval will be chosen as the assigned timestamp, so we assume the adversary can pick any value from the interval. We also assume the adversary can know the result modified timestamp once it has picked the assigned timestamp and use that information to make further choices.

A.2.1 \triangle -Ordering linearizability.

Definition A.1 (Δ -Ordering Linearizability). For any two commands, c_1 and c_2 , sent at least Δ apart, i.e. $|T_1 - T_2| > \Delta$, then it is guaranteed that the earlier command will be ordered before the later command.

Theorem A.1. Our protocol guarantees ($\Delta_{net} + \Delta_{noise}$)-Ordering Linearizability.

Proof. By our assumption, the minimum assigned timestamp for a command sent at time T is T, and the maximum assigned timestamp is $T + \Delta_{\text{net}}$.

By the definition of our protocol, the minimum modified timestamp is T, and the maximum modified timestamp is $T + \Delta_{\text{net}} + \Delta_{\text{noise}}$.

Thus, it is impossible for two commands sent more than $\Delta_{net} + \Delta_{noise}$ apart to be ordered in the reverse order.

A.2.2 ϵ -Ordering equality.

Definition A.2 (ϵ -Ordering Equality). There exists a function $\epsilon(n)$ such that, for any set of n commands $C = \{c_1, c_2, ..., c_n\}$ all sent at the same time T and any two total order, \prec_1 and \prec_2 , of commands in C, the difference in probabilities of the protocol outputting \prec_1 and \prec_2 is bounded by $|Pr[\prec_1] - Pr[\prec_2]| \le \frac{\epsilon(n)}{n!}$.

Because the adversary can assign an assigned timestamp to any command sent at time T up to $T + \Delta_{net}$, we must have $\Delta_{noise} > \Delta_{net}$ to prevent the adversary from dictating the ordering between any two commands.

We first look at the case where n = 2.

We use the symbol $c_1 < c_2$ to denote the event that command c_1 is ordered before command c_2 in the output order.

Theorem A.2. For any two commands, c_1 and c_2 sent at the same time T, we have $|Pr[c_1 < c_2] - Pr[c_2 < c_1]| \le 1 - (1 - \frac{\Delta_{\text{net}}}{\Delta_{\text{net}}})^2$.

Proof. Let t_i be the assigned timestamp and t'_i be the modified timestamp for command c_i . By the definition of our protocol, t'_i is uniformly sampled from the interval $[t_i, t_i + \Delta_{\text{noise}}]$. Thus, we have:

$$Pr[c_1 < c_2] = \frac{1}{\Delta_{\text{noise}}^2} \int_{t_1}^{t_1 + \Delta_{\text{noise}}} \int_{t_2}^{t_2 + \Delta_{\text{noise}}} (t_1' < t_2') dt_2' dt_1'$$

By the assumption, $t_1, t_2 \in [T, T + \Delta_{\mathsf{net}}]$. The optimal strategy for the adversary who controls assigned timestamps is assigning $T + \Delta_{\mathsf{net}}$ to c_1 and T to c_2 to minimize the probability of $c_1 < c_2$. Thus, we have:

$$Pr[c_1 < c_2] \ge \frac{1}{\Delta_{\text{noise}}^2} \int_{T+\Delta_{\text{not}}}^{T+\Delta_{\text{noise}}} \int_{T}^{T+\Delta_{\text{noise}}} (t_1' < t_2') dt_2' dt_1'$$

Eliminate *T* and, because of $\Delta_{\text{net}} < \Delta_{\text{noise}}$, we have:

$$Pr[c_1 < c_2] \ge \frac{1}{\Delta_{\text{noise}}^2} \int_{\Delta_{\text{noise}}}^{\Delta_{\text{noise}}} \int_{x}^{\Delta_{\text{noise}}} 1 dy dx = \frac{1}{2} (1 - \frac{\Delta_{\text{net}}}{\Delta_{\text{noise}}})^2$$

Because this is a tight lower bound following this strategy, the difference in probabilities of those two output orders is given by:

$$|Pr[c_1 < c_2] - Pr[c_2 < c_1]| = |1 - 2Pr[c_1 < c_2]| \le 1 - (1 - \frac{\Delta_{\text{net}}}{\Delta_{\text{noise}}})^2$$

We also prove the general case.

Theorem A.3. Our protocol satisfies $\epsilon(n) = (1 + \alpha)^n - (1 - \alpha)^n - n\alpha^n$ -ordering equality.

Proof. It suffices to prove a lower bound and an upper bound on any Pr[<]. For convenience, let $\alpha = \frac{\Delta_{\text{net}}}{\Delta_{\text{noise}}}$. We also use the following lemma:

Lemma A.1. For any n independent random variables uniformly sampled from the same interval [a, b], all orderings of those n variables have the same probability of $\frac{1}{n!}$.

Lemma A.2 (Tight lower bound). $Pr[\prec] \geq \frac{1}{n!} (1-\alpha)^n$.

Proof. Consider the time interval $[T + \Delta_{\text{net}}, T + \Delta_{\text{noise}}]$. Any command sent at time T has a probability of $1 - \alpha$ being assigned a modified timestamp uniformly sampled from the interval because the assigned timestamp, t, is picked from in $[T, T + \Delta_{\text{net}}]$ and the additional noise, η , is sampled uniformly from $[0, \Delta_{\text{noise}}]$. Thus, the probability of all n modified timestamps being contained in this interval is $(1 - \alpha)^n$. Using the lemma above, all n! possible permutations of these n commands will have equal probabilities, which gives the lower bound of $\frac{1}{n!}(1 - \alpha)^n$.

This lower bound is tight by assigning the first command in < an assigned timestamp of $T + \Delta_{\text{net}}$ and the last command in < an assigned timestamp of T. In this case, all commands' modified timestamps must be included in this interval for < to be the output order.

Lemma A.3 (Tight upper bound). $Pr[\prec] \leq \frac{1}{n!}((1+\alpha)^n - n\alpha^n)$

Proof. We prove this tight upper bound by induction on n. The base case, n = 1, holds trivially.

For the inductive case, consider how the adversary can pick assigned timestamps to maximize the probability of <.

For the first command in $\langle c_1 \rangle$, because the adversary would like its modified timestamp to be as small as possible, the optimal strategy is assigning an assigned timestamp of exactly T and thus a modified timestamp uniformly sampled from $[T, T + \Delta_{\text{noise}}]$.

There are two possible outcomes for t_1' : $t_1' \le T + \Delta_{\text{net}}$ and $t_1' > T + \Delta_{\text{net}}$.

In the first case, the adversary can assign an assigned timestamp to every other command of at least t_1' to ensure c_1 is the first command in the permutation. But the adversary must still make the other commands follow \prec . The optimal strategy is to achieve the tight upper bound of n-1 commands with a different $\alpha' = \alpha - \frac{t_1' - T}{\Delta_{\text{noise}}}$, which gives the following integral over the probability density function:

$$\int_0^\alpha \frac{1}{(n-1)!} ((1+(\alpha-x))^{n-1} - (n-1)(\alpha-x)^{n-1}) dx$$

$$= \int_0^\alpha \frac{1}{(n-1)!} ((1+y)^{n-1} - (n-1)y^{n-1}) dy$$

$$= \frac{1}{n!} ((1+\alpha)^n - 1 - (n-1)\alpha^n)$$

In the second case, where $t_1' > T + \Delta_{\text{net}}$, the optimal strategy is to delay the latter commands as much as possible. Thus, every other command gets an assigned timestamp of exactly $T + \Delta_{\text{net}}$, and their modified timestamps are all sampled uniformly from $[T + \Delta_{\text{net}}, T + \Delta_{\text{net}} + \Delta_{\text{noise}}]$. To have < being the output order, all other commands must be in the interval $[t_1', T + \Delta_{\text{net}} + \Delta_{\text{noise}}]$, which has a probability of $(1 + \alpha - \frac{t_1'}{\Delta_{\text{noise}}})^{n-1}$. Furthermore, those commands need to satisfy <. Because all those timestamps are sampled uniformly

from the same interval, by the lemma above, we have the following integral:

$$\int_{\alpha}^{1} \frac{1}{(n-1)!} (1+\alpha-x)^{n-1} dx$$

$$= \int_{\alpha}^{1} \frac{1}{(n-1)!} y^{n-1} dy$$

$$= \frac{1}{n!} (1-\alpha^{n})$$

Adding the probabilities of the two cases together, we have:

$$Pr[<] \le \frac{1}{n!}((1+\alpha)^n - 1 - (n-1)\alpha^n) + \frac{1}{n!}(1-\alpha^n) = \frac{1}{n!}((1+\alpha)^n - n\alpha^n)$$

This bound is tight following our construction.

Thus, given the tight lower and upper bounds, the difference $\frac{\epsilon(n)}{n!}$ is bounded by $\frac{1}{n!}((1+\alpha)^n-(1-\alpha)^n-n\alpha^n)$, $\epsilon(n)\sim 2n\alpha$.

A.3 Remarks

- 1. For any n, the ordering equality bound converges to 0 when $\Delta_{\text{noise}} \rightarrow \infty$. This matches our intuition that infinitely large random noise could achieve equality trivially.
- 2. The above results show the trade-off between ordering linearizability and ordering equality: A larger Δ_{noise} gives a better ordering-equality bound ($\sim \frac{2}{(n-1)!} \cdot \frac{\Delta_{\text{neit}}}{\Delta_{\text{noise}}}$) but a worse ordering-linearizability bound ($\Delta_{\text{net}} + \Delta_{\text{noise}}$) and, thus, worse latency.
- 3. Given a fixed Δ_{noise} , the bound on the difference between probabilities of different output orders, $\epsilon(n)$, weakens as the number of commands n grows. This is because the adversary has more control over the output order for more commands, as it can pick the assigned timestamps for all commands.