Indicus: Unchaining Byzantine Databases

Your N. Here
Your Institution

Second Name
Second Institution

Abstract

Your abstract text goes here. Just a few facts. Whet our appetites. Not more than 200 words, if possible, and preferably closer to 150.

1 Introduction

There is an increasing tension between the desire to share data online, and the security concerns it entails. This paper asks the question: how can we enable *mutually distrustful parties* to consistently and reliably share data, while minimizing centralization?

The ability to share data online offers exciting opportunities. In banking, systems like SWIFT enable financial institutions to quickly and accurately receive information such as money transfer instructions; and in manufacturing, online data sharing can improve accountability and auditing amongst the globally distributed supply chain.

Increased data sharing, however, raises questions of how to *decentralize trust*. Banking institutions must currently place their trust in the centralized SWIFT's network to issue payment orders. Sometimes there is even no identifiable source of trust. Consider the supply chain for the latest iPhone: it spans three continents, and hundreds of different contractors []; neither Apple nor these contractors trust each other, yet all must be willing to agree and share information about the construction of the same product.

Recognizing this challenge by both the research and industry communities, much effort has focused on enabling shared computation between mutually distrustful parties, in the context of byzantine fault tolerance (BFT), and blockchains. Systems proposed in the literature of BFT[][] provide the abstraction of a totally ordered log; the log is agreed upon by the n participants in the system, of which at most f can misbehave. Each participant executes operations that may touch one to multiple objects in the log. In the blockchain world, Bitcoin and Ethureum have become popular distributed computing platforms providing the same log abstraction and aiming for decentralizing trust. Furthermore, Microsoft Azure has launched projects[] that leverage these blockchain platforms and extend the digital transformation beyond the companies' four walls in a supply chain.

This paper argues that there exists a fundamental mismatch between the implementation of a totally ordered log and the reality of much large-scale distributed processing. Many large-scale distributed systems consist primarily of unordered and unrelated operations. For example, a product supply chain consists of many concurrent steps that do not require ordering. Imposing an ordering on non-conflicting operations is not only often unnecessary, but costly: participants in the shared computation must vote to order operations, store the full state of the system, and replay the full log for auditing.

While there exists work on mitigating this scalability bottleneck through sharding [], the latent total order requirement introduces unnecessary coordination overhead, as coordination is performed twice, at the level of individual shards, and across shards. Callinicos [] and Omniledger [], for instance, runs a full BFT protocol for every operation. This is especially problematic when workloads are geo-replicated [], or when, as in BFT, the replication factor is high. Further, these systems support transactions under the assumption that their read and write operations are known a priori, which limits the set of applications that they can support.

As another research trend of mitigating the scalability bottleneck, EPaxos[], TAPIR[] and CURP[] only consider the ordering between potentially conflicting operations, instead of commutative operations. However, these systems assume the crash-failure model and are non-trivial to be extended to the Byzantine model, so that they cannot directly solve the problem of data sharing among mutually distrustful parties.

Existing research, in essence, is either attempting to build concurrency control and sharding functionalities over BFT replication, or integrating these functionalities into a crash-failure replication protocol. In this paper, we will show how to build these desiring functionality inside a BFT replication protocol. Specifically, our goal is to *provide the illusion of a centralized shared log, rather than the non-scalable reality of*

a totally ordered log.

Looking beyond treating transactions as black box requests and leveraging existing commutativity is not a new idea. However, while copious efforts exist to design and build more decentralized systems in order to exploit transaction knowledge for the Crash Failure model, few, if any attempts have been made for the Byzantine Fault Model. This naturally begs the question why so? One explanation is that for BFT systems centralization is in fact highly desirable as it simplifies the problem by pin-pointing a single point of accountability. Since historically, BFT systems have taken a rather niche role for applications that require "additional" safety, the principal design concern has always been maintaining consistency with efficiency and scalability being secondary concerns. Replication was traditionally done by a single authority and thus it is reasonable to assume a primary backup scheme, since neither fairness, nor total ordering were major concerns. As Byzantine Fault Tolerance moves into the mainstream with the popularity ascent of Blockchains these considerations are being revistited. For example, when operating a database as a consortium of mutually distrustful parties it is no longer desirable to grant priviledges to a leader and impose centralization for the sake of simplicity. Rather than trying to make a traditional BFT system more scalable we ask the question whether we can take the scalability lessons from Crash Failure settings and improve upon their robustness. While this is an attractive avenue, it is far from being straightforward. A natural way to scale a system is to move state and responsabilites to the clients. However, in a byzantine setting, giving up the comfort of a bounded and accountable set of replicas opens up the system to a wider set undesirable phenomena. Yet, this is exactly what we will do: Concretely, in this paper we will show how to design a BFT system that is almost entirely client driven, and all the while both safe and live.

// Unlike SMR that achieves agreement on a result by imposing a total order to evaluate the question, we just evaluate the result directly. Instead of operating multi consensus, I.e. a ledger where reaching consensus implies reaching consensus on every prior decision (like a blockchain), we just do binary consensus for each transaction in a seperate slot. Because its unecessary to reach agreement on this total order, we only need to respect an order to the set of conflicting transactions, a (implicit) partial order suffices. Transactions are still able to be totally ordered according to a serializable execution they correspond to by assigning a Timestamp and evaluation Isolation according to it.

Doing so has the following positive outcomes: - Indicus is robust to censorship and frontrunning as there exist no central authority (in SMR traditionally a leader) that decides what Transactions enter the system and decides on the ordering of Transactions - Indicus allows commutative/non-conflicting Transactions to both be executed and validated out of order, thus maximizing parallelism - Indicus minimizes the state, communication and computation load on replicas, as Clients serve as both execution hub and broadcast channel for their own Transactions. This avoids quadratic communication communication complexity in the normal case. (Theoretically always, but we keep some for practicality in the fallback) - As any Quorum system, Indicus is inherently load-balanced as there exists no leader bottleneck and all replicas have equal responsibility - In Indicus liveness is a client local property. Unlike SMR, where the entire system halts during view changes, Byzantine participants may stall system progress only for the objects their transactions touch. Hence, the system appears live to any non-conflicting Transaction

Since we do not sequence Transaction execution we require a concurrency control mechanism in order to maintain Isolation between Transactions. For this purpose, we design and implement a byzantine replicated MVTSO scheme, an aggressive version of Optimistic Concurrency control that empowers Read Transactions in the hope of reducing abort rates. Specifically, we allow to read old version, we allow to read uncommitted writes and we acquire read leases.

Additionally, we propose definitions for what it means to enforce an Isolation level in a transactional system with byzantine participants. These are general purpose formalizations and can serve as guideline for future byzantine Database systems.

In section X, we discuss Limitations of this approach.

2 Indicus

In the following we outline the Architecture and Protocols of Indicus, the first highly scalable (hopefully) database that tolerates both byzantine clients and replicas.

2.1 Architecture

In Indicus, Clients are more than external participants who propose Transactions to be executed. Instead, Clients are first class citizens that rejoice in the fair treatment of democracy and take part in everyday system activities. Of course, such priviledge comes at the cost of responsibilities, concretely, Indicus does not provide a continuously and mysteriously operating black box Transaction machine, but rather offers Clients the tools to operate the system itself. One the one hand, the simple paradigm of letting clients work for themselves is more scalable as replicas need to do less work and can service more clients. On the other hand, it naturally incentivises clients to be industrious, as they hold the keys to their own liveness. Clients that do not meet a certain standard for productivity may be excluded by the system; The joys of being permissioned.

Indicus follows a traditional optimistic concurrency control architecture as shown in Figure 1. Clients speculatively execute their Transactions, issuing remote reads when necessary and buffering writes locally. As Clients execute their own Transactions they need not declare their read/write keys or

values preemtively, but rather may conduct interactive Transactions, the most general Transaction model. Since execution is speculative and Clients are unaware of potential concurrency, they must validate their Transactions for Isolation correctness in order to be able to Commit. Intuitively, if there are no conflicting concurrent Transactions and a Client observed a consistent snapshot of the distributed Database state then it may commit, and otherwise it must abort or retry. Lastly, Transactions may span multiple shards, and hence the Commit/Abort decisions of each shard must be aggregated in a Two Phase Commit manner in order to finalize a safety appropriate result. Most notably, in Indicus not just exeuction is Client driven but the entire Transaction life cycle, thus maximizing scalability and fairness while putting each Client in charge of its own liveness. In the following we describe our Fault Model and define corresponding Transactional guarantees. We then outline the protocols for Execution, Validation and Writeback respectively.

2.2 Model and Definitions

2.2.1 System Model

We assume a partitioned Database where data objects are spread uniformly across shards (randomly or based on locality). For fault tolerance we assume that each Shard contains 5f+1 replicas where f is the number of (static) faulty or compromised replicas. A faulty replica may behave in Byzantine fashion, i.e. in addition to crashing or omitting messages it may send arbitrary messages and deviate from prescribed protocols in any way. We denote any participant (replica or client) that follows the protocol as *honest*, while faulty participants are dubbed *byzantine*. We assume there may exist a finite but unbounded number of byzantine Clients that may deviate from the protocol arbitrarily. We further assume a strong, but static adversary that can coordinate the faulty participants We do, however, assume the existance of sufficiently hard cryptographic primitives that allow for private/public key signatures and collision-resistant hashes that cannot be compromised by byzantine participants. We denote a signed message m, signed by principal p as $\langle m \rangle_{\sigma p}$.

We make no assumption on network synchrony in order to maintain safety, but in some cases may provide liveness only when the network is synchronous and messages are delayed by no more than a fixed but potentially unknown window. Unlike traditional State Machine Replication protocols in which the liveness of all Clients is correlated with the fate of the system (or often more specifically a leader), our system guarantees liveness not on a system basis, but on a per client basis. Concretely, we only guarantee liveness to clients that follow the protocol and conversely an honest client only loses liveness (even in an asynchronous setting) when it intertwines its fate with byzantine clients.

Application services may restrict the influence of Byzantine Clients on the Database state by authentication and enforcing access control. Beyond these measures, byzantine Clients that follow the protocol are indistinguishable from honest clients and may read and write at their leisure. While potential damage to the Database state cannot be avoided, it can be retraced by auditing the Transaction logs.

2.2.2 System properties

We offer Clients an interactive Transaction interface that implements ACID transactions. While all ACID guarantees may be violated by individual byzantine Replicas, the system maintains the illusion of an ACID compliant state to all honest Clients. Moreover, we guarantee to Clients that the Database is *byzantine Serializable* as defined below. Intuitively this Isolation level guarantees that all honest Clients experience the Database as if there serializable. In order to formally capture this we lay some ground work: [1]

Let $Op = \{r, w\} \times K \times V$ and $Dec = \{Commit, Abort\}$ be the sets of possible operations and decisions respectively, where K is the set of existing data items (keys) and V the range of possible values. A *request* $req \in (Op \cup Dec) \times C$ maps any such operation or decision to the issuing Client from set C. We denote with $Hon \subseteq C$ the subset of honest Clients.

History H. Informally, a H contains the operations (read/write) and decisions (commit/abort) of every transaction issued in the system. Formally, we define a *history* H as a finite sequence of requests.

We define a projection $H|_c$ as the subsequence of requests in H that were issued by Client c. A sequence of requests $s = req_i \dots req_{i+t}$ in $H|_c$ form a *Transaction* if req_i is the first request by Client c or $req_{i-1} \in Dec \times c$, and if $req_{i+t} \in Dec \times c$.

We further define:

Honest History H(P). Given protocol P, A *history* H is *honest* if it was generated by participants who all follow P. I.e. concretely, $H(P) \equiv H = H|_{Hon}$. and

Honest-View Equivalent. A *history* H is honest-view equivalent to a *history*H' if the Operations and Decisions of all honest Clients are the same and if the final writes are the same.

Byz-I Given a protocol P and an isolation level I: A history H is *byzantine-I* if there exists an honest history *H'* such that H is honest-view equivalent to H' and H' satisfies I.

Thus, informally a byzantine Isolation level states that the state that honest Clients experience must be explicable by an execution in which all participants were honest. This definition captures the requirements for any byzantine tolerant protocol that strives to maintain Isolation level I.

We further define some progress properties to limit the influence Byzantine Clients have on system throughput. These will be motiviating factors that we will return to in order

[1] Defs for commands etc in accordance with BFT DUr. cite

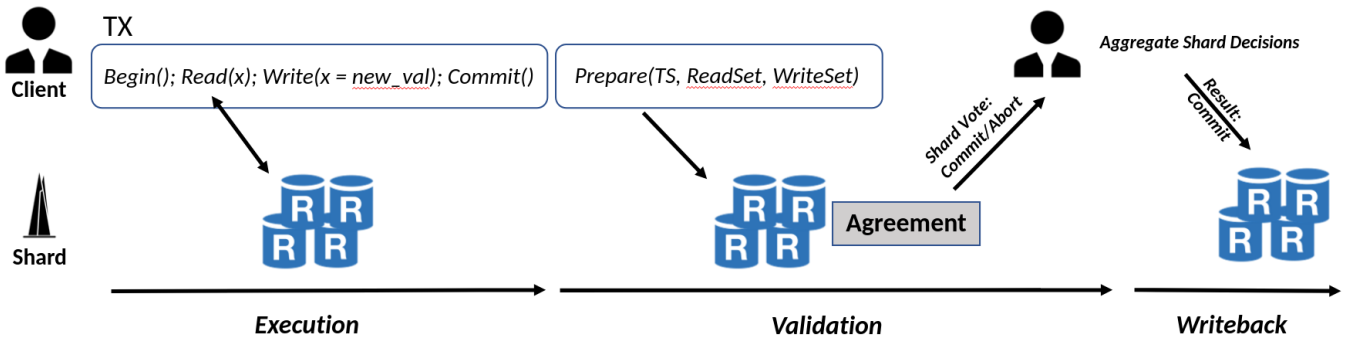

Figure 1: Transaction Lifecycle

to justify our choice of 5f+1 replicas as well as the use of an Optimistic Concurrency Control over a Locking based schema. We later outline an alternative design using 3f+1 replicas that does not maintain this property.

2.3 Execution

2.4 Validation

2.5 Concurrency Control

- Optimization: retries - Dependency resolution

2.6 Consistent logging.

2.7 Writeback and Multi-shard 2pc

- Optimization: Single shard logging

2.8 Failures

- Fallback: election, resolution, subtelties with mvtso, necessity even without dependencies.

2.9 Low Cost mode

3f+1 if not defending against byz colluders OCC if not worried about reads aborting

3 Implementation and Evaluation

4 Limitations

As is inherent to any Optimistic Execution and Concurrency Control, Indicus is vulnerable to highly congested workloads. When contention on select objects is high, concurrent execution of Transactions must yield the abort of some Transactions during Validation in order to maintain the Database Isolation guarantees. Note however, that when clients are in charge of execution, a pessimistic concurrency control solution such as two-phase-locking would incur an equal amount of deadlocks which would require resolution. The observation to make is that any system that conducts execution at the client application side speculates on concurrency. This however we stipulate, is unavoidable when trying to scale a system to the number of users rather than replica processing power. The traditonal ways to avoid the abort rate conundrum is to either restrict the transaction model, which in turn weakens the general applicability of the protocol, or to delegate execution to replicas and utilize State Machine Replication to serialize Transactions. SMR protocols with a single leader do not inquire any congestion based aborts as a single sequencer naturally eliminates concurrency. Indicus does not make these concessions in order to offer interactive Transactions and remain scalable. A workload that exhibits low commutativity and high contention should therefore refrain from adopting our system.

Similarly, as is the case in any transaction protocol, Indicus is vulnerable to ddos attacks by byzantine participants. A byzantine clients only opportunity at subverting progress for honest users is to artificially increase congestion. When such a client has unrestricted access control it may do so strategically iff it has control over the network. If it does not, it cannot reliably gain knowledge about concurrent transactions before they pass the validation step and must resort to flooding based attacks. Defense against such attacks is out of scope in our work, but is disincentivised as participants can be held accountable for their actions in a closed membership setting.

5 Related Work/Comparison

mention Hotstuff? A lot of recent effort has gone into designing high throughput and low latency databases that leverage synergies between transaction and replication layer to squeeze out any last performance. The recently proposed TAPIR transaction protocol leverages redundancy between transaction ordering and replication ordering to reduce total roundtrips, thus reducing latency in wide area networks. TAPIR shares several similarities with our system, most notably the absence

of a leader and the resulting unordered validation structure. TAPIR too, leverages optimistic concurrency control to allow for concurrency among commutative transactions. When congestion is high however, throughput and tail latency worsen as abort rate grows. Janus avoids transaction aborts by dynamically re-ordering conflicting transactions. This however, is made possible by assuming one-shot transactions, i.e. fixed read/write keys, and thus reduces the application generality. Another DB, Carousel similarly assumes a restricted Transaction model in order to parallelize execution and validation. While all of these databases offer low latency replication, they are fundamentally limited to tolerating crash-failures. This strong failure assumption makes them less secure and not suitable for storing mission-critical, financial or highly sensitive data. With the surge in Blockchain interest Byzantine Fault Tolerant (BFT) protocols are experiencing a second Spring. While originally developed to tolerate arbitrary bugs, these protocols find increasing importance in settings where participants are untrusted or malicious. Permissioned Blockchains, organized by a consortium of registered participants can use traditional BFT State Machine Replication (SMR) protocols in order to achieve agreement. Starting with PBFT, there have been numerous adaptations such as FaB or Generalized byzantine Paxos, and most notably Zyzzyva which leverages speculative execution and a semi-client driven protocol to reduce latency. SBFT modifies PBFT to scale to large replication degrees by utilizing collectors, threshold signatures and a fast path akin to Zyzzyva. Aardvark states the importance of robustness against byzantine failures and takes measures to increase the rate of leader rotation. Tendermint pushes this idea to the extreme by rotating leaders continuously, at the cost of assuming a synchronous network. Nevertheless, all protocols derived from the PBFT family suffer from the leader bottleneck as well as enforcing a total order even on commutative operations. BFT-Mir claims the leader proposing speed as the practical bottleneck in most implementations and improves upon this by allowing all replicas to act as proposers. Biblos achieves leaderless SMR by leveraging a non-skipping timestamp protocol. It furthermore allows for commutative Transactions to be executed in parallel at the cost of requiring read/write key sets to be known in advance. In order to preserve liveness however Bilbos falls back to a PBFT resolution. Q/U too, offers leaderless agreement via Qurorums for a limited read/write interface, but fails to terminate under contention. H/Q improves upon Q/U by adding PBFT fallback path under contention. Liskov et Al further explore byzantine Quorum protocols for a Read/Write interface, giving special attention to bounding the effects of byzantine clients. While the literature on BFT state machine replication is extensive, the efforts to offer a transactional interface for BFT is scarce. SMR itself can be utilized as a straw-man system to implement pre-defined Transactions (one-shot or stored procedures) by enforcing a common total order in which replicas will execute the transactions. HRDB offers a dedicated BFT Database, but assumes a trusted shepheard layer, thus not being truly BF resilient Byantium offers Snapshot Isolation for Transactions by re-purposing PBFT as atomic broadcast. It executes requests only at a primary and uses replicas to validate results in the total order defined by the SMR protocol. Augustus implements mini-transactions, a limited TX model that declares all operations before execution. It offers scalability via partitioning and achieves consistency within partitions by utilizing atomic broadcast. Whiile Augustus assumes an optimistic execution model that allows for aborts under concurrency, its follow up work Callinicos implements a locking scheme in order to avoid Transaction aborts. To do so efficiently it resorts to a limited transaction model that requires knowledge of read/write sets and otherwise locks an entire partition in order to guarantee mutual exclusion, thus voiding any concurrency. BFT Deferred update replication adopts an interactive OCC transaction model comparable to ours, allowing execution to be speculative at clients. However, it uses PBFT atomic broadcast to enforce SMR on validation, thus resorting to a leader, and totally ordering all requests. It moreover does not extend to multiple shards. Chainspace and Omniledger implement blockchain sharding by layering 2PC and atomic broadcast (within shards) for UTXO transactions. Rapidchain offers efficient sharding for a permissionless system.

[1]

6 Conclusion

Acknowledgments

The USENIX latex style is old and very tired, which is why there's no \acks command for you to use when acknowledging. Sorry.

Availability

USENIX program committees give extra points to submissions that are backed by artifacts that are publicly available. If you made your code or data available, it's worth mentioning this fact in a dedicated section.

References

[1] Ramakrishna Kotla, Lorenzo Alvisi, Mike Dahlin, Allen Clement, and Edmund Wong. Zyzzyva: speculative byzantine fault tolerance. *ACM SIGOPS Operating Systems Review*, 41(6):45–58, 2007.