How to Databasify a Blockchain: the Case of Hyperledger Fabric

Ankur Sharma  Felix Martin Schuhknecht  Divya Agrawal  Jens Dittrich

Big Data Analytics Group
Saarland Informatics Campus
https://bigdata.uni-saarland.de

November 1, 2018

Abstract

Within the last few years, a countless number of blockchain systems have emerged on the market, each one claiming to revolutionize the way of distributed transaction processing in one way or the other. Many blockchain features, such as byzantine fault tolerance (BFT), are indeed valuable additions in modern environments. However, despite all the hype around the technology, many of the challenges that blockchain systems have to face are fundamental transaction management problems. These are largely shared with traditional database systems, which have been around for decades already.

These similarities become especially visible for systems, that blur the lines between blockchain systems and classical database systems. A great example of this is Hyperledger Fabric, an open-source permissioned blockchain system under development by IBM. By having a relaxed view on BFT, the transaction pipeline of Fabric highly resembles the workflow of classical distributed databases systems.

This raises two questions: (1) Which conceptual similarities and differences do actually exist between a system such as Fabric and a classical distributed database system? (2) Is it possible to improve on the performance of Fabric by transitioning technology from the database world to blockchains and thus blurring the lines between these two types of systems even further? To tackle these questions, we first explore Fabric from the perspective of database research, where we observe weaknesses in the transaction pipeline. We then solve these issues by transitioning well-understood database concepts to Fabric, namely transaction reordering as well as early transaction abort. Our experimental evaluation shows that our improved version Fabric++ significantly increases the throughput of successful transactions over the vanilla version by up to a factor of 3x.

1 Introduction

Blockchains are one of the hottest topics in modern distributed transaction processing. However, from the perspective of database research, one could raise the question: what makes these systems so special over classical distributed databases, that have been out there for a long time already?

The answer lies in byzantine fault tolerance: while classical distributed database systems require a trusted set of participants, blockchain systems are able to deal with a certain amount of maliciously behaving nodes. This feature opens lots of new application fields such as transactions
between organizations, that do not fully trust each other. Unfortunately, ensuring BFT over all nodes of the network also heavily complicates transaction processing. If any node of the network is considered to be potentially malicious, a complex consensus mechanism is required to ensure the integrity of the system. This consensus mechanism assures, that a transaction can only commit, if a majority of the network agrees to it.

Of course, the expensiveness of the consensus mechanism has also been observed by the blockchain engineers. Therefore, some systems trade BFT with performance by simply assuming certain parties of the network to be trustworthy. A great example for this is Hyperledger Fabric [7], a popular open-source blockchain system introduced by IBM. In terms of BFT, it differs from other major players such as Bitcoin or Ethereum in two additional assumptions: First, Fabric assumes that the ordering service, which globally orders all transactions that go through the system, is trustworthy. Second, it allows for the forming of so called organizations. Within an organization, it is assumed that all peers trust each other. These two assumptions heavily simplify transaction processing. First of all, no complex consensus mechanism, such as PBFT [8], is necessary. Second, the trust within an organization allows for a distribution of work within it and enables parallelism, as not every peer in the organization has to execute every transaction.

With this relaxed view on BFT in mind, can we actually still consider Fabric a true blockchain system? A trustworthy ordering service, which globally arranges and schedules transactions, is a component that is present in classical distributed database systems as well. Further, the concept of an organization, in which all peers trust each other, is also present in distributed databases in its extreme form: all peers belong to a single organization. Besides of that, other core requirements of transaction management, such as ensuring transaction isolation or managing the data in a store, are essentially present one-to-one in both blockchains and database systems.

At the example of Fabric, it becomes obvious that conceptually the lines between blockchain systems and distributed database systems are rather blurry. We believe this blurriness should be seen as a chance for the database community: Due to all these conceptual similarities, it becomes possible to transition well-understood database technology to the world of blockchains, significantly enhancing this new technology.

The question remains which similarities can be exploited to transition database technology to Fabric and by how much can we improve on the state-of-the-art? To tackle this problem, we perform the following steps in this work:

1. To have a basis for the discussion, we first inspect the transaction flow of Hyperledger Fabric in the latest version 1.2 from a conceptual perspective. Fabric will serve as our case-study for the rest of the paper on how to "databasify" a blockchain system (Section 2).

2. Based on the analysis of the transaction flow in Fabric, we then inspect its components, that show the highest resemblance with those of database systems. We identify weaknesses in the implementation of Fabric of these components and describe, how database technology can be utilized to counter them. (Section 3).

3. We transition database technology to the transaction pipeline of Fabric. Precisely, we first improve on the ordering of transactions. By default, the system orders transactions arbitrarily after simulation, leading to unnecessary serialization conflicts. To counter this problem, we introduce an advanced transaction reordering mechanism, which aims at reducing the number of serialization conflicts between transactions within a block. This
mechanism significantly increases the number of valid transactions, that make it through
the system and therefore the overall throughput (Section 4.1).

4. Next, we advance the abort of transactions. By default, Fabric checks whether a transac-
tion is valid right before the commit. This late abort unnecessarily penalizes the system
by processing transactions, that have no chance to commit. To tackle this issue, we in-
troduce the concept of early abort to various stages of the pipeline. We identify invalid
transactions as early as possible and abort them, assuring that the pipeline is not throt-
tled by transactions that have no chance to commit eventually. A requirement for this
concept is a fine-grained concurrency control mechanism, by which we extend Fabric as
well (Section 4.2). These modifications significantly extend the vanilla Fabric, turning it
into what we call Fabric++.

5. We perform an extensive experimental evaluation of the optimizations of Fabric++ under
a custom blockchain benchmark simulating an asset transfer scenario. In total, we evaluate
the transactional throughput under 108 different configurations of Fabric and Fabric++
and show that we are able to significantly boost the performance over the vanilla version.
Additionally, we vary the number of channels and clients and show, that our optimizations
also have a positive impact on the scaling capabilities of the system (Section 5).

2 Hyperledger Fabric

Before diving into the conceptual similarities and differences between Fabric and distributed
database systems, we have to understand the workflow of Fabric. Let us describe in the following
section how it behaves in the latest version 1.2.

2.1 High-level Workflow

At its core, Fabric follows a simulate-order-validate-commit workflow, as shown in Figure 1:

(1) In the simulation phase, a client submits a transaction proposal to a specified subset of
the peers, called the endorsement peers or endorsers. The endorsers simulate the effects of the
transaction proposal against a local copy of their current state. Interestingly, none of the writes
become durable in the current state at this point. If the endorsers endorse the transaction
proposal, an actual transaction is formed from the execution result, that is then sent to the
ordering service (via the client).

(2) In the ordering phase, the ordering service establishes a global order among all received
transactions and distributes the ordered transactions at the granularity of blocks to all peers of
the network.

(3) In the validation phase, all peers individually validate the transactions within the received
blocks in terms of endorsement policy and serializability.

(4) In the commit phase, the blocks are appended to the local ledger and the changes made
by the valid transactions are applied to the current state.

Following these four phases assures that each honest peer stores the same transaction sequence.
2.2 Architecture

Fabric is a permissioned blockchain system, meaning all peers of the network are known at any point in time. Peers are grouped into organizations, which typically host them. For example, two companies trading with each other could each host an organization of 10 machines, forming a network of 20 peers. Within an organization, all peers trust each other.

Each peer runs a local instance of Fabric. This instance includes a copy of the ledger, containing the ordered sequence of all transactions that went through all four phases. This includes both valid and invalid transactions. Apart from the ledger, each peer also contains the current state in form of a state database. The current state can be seen as an optimization of the ledger: while the ledger simply contains the sequence of all processed transactions, the current state represents the state after the application of all valid transactions in the ledger to the initial state. Fabric implements the current state in form of a versioned key-value-store. For every key in the store, a pair of value and version-number is kept, where the version-number counts the number of changes that already happened to the value of this key.

Apart from the peers, which play an important role both in the simulation phase and the validation phase, there is a separate instance called the ordering service, which is the core component of the ordering phase and assumed to be trustworthy. Although it can be composed out of multiple machines for fault tolerance, it is a central service responsible for establishing a global order among all transactions.

2.3 Running Example

With the basic components of the architecture in mind, let us now discuss how transactions flow through the system. To do so, we present a simple running example in Figure 2 (simulation phase), Figure 3 (ordering phase), and Figure 4 (validation and commit phase), where two organizations A and B want to transfer money between each other.

Figure 1: High-level workflow of Fabric.

\(^1\)The version number is actually composed of transaction-ID and the block-ID, see Section 4.2.1 for details.
Each organization contributes two peers to the network. The balances of the organizations are stored by two variables $BalA$ and $BalB$, where $BalA$ stores the value 100 in its current version $v_3$ and $BalB$ stores 50 in version $v_2$. We can also see that the ledger already contains six transactions $T_1$ to $T_6$, where the four transactions $T_1$, $T_2$, $T_4$, and $T_6$ were valid ones and lead to the current state. The transactions $T_3$ and $T_5$ were invalid transactions. They are still stored in the ledger, although they did not pass the validation phase.

2.4 Simulation Phase

Transaction processing starts with the simulation phase in Figure 2. In step 1, a client proposes a transaction proposal (or short proposal) to the system. In our example, the proposal intends to transfer the amount of 30 from $BalA$ to $BalB$. The two involved operations $BalA-=30$ and $BalB+=30$ are expressed in a smart contract, an arbitrary program, that is bound to the proposal. Additionally to the smart contract, an endorsement policy must be specified. It determines which and/or how many peers have to endorse the proposal. In our example of money transfer between two organizations, a reasonable endorsement policy is to request endorsement from one peer of each organization — like two lawyers, preserving and defending the individual rights of their clients.

Therefore, in step 2, the proposal is sent to the two endorsement peers $A1$ and $B1$ according to the policy. These two peers now individually simulate the smart contract ($BalA-=30$, $BalB+=30$), that is bound to the proposal, against their local current state. Note that, as the name suggests, the simulation of the smart contract against the current state does not change the current state in any way. Instead, each endorsement peer builds an auxiliary read set $RS$ and a write set $WS$ during the simulation to keep track of all accesses that happen. In our case of money transfer over the amount of 30, the smart contract first reads the two current balances $BalA$ and $BalB$ along with their current version-numbers. Second, the smart contract updates the two balances according to the transferred amount, resulting in the new balances $BalA = 70$ and $BalB = 80$. Overall, this builds the following read and write set:

\[
RS = \{(BalA, v_3), (BalB, v_2)\} \quad WS = \{(BalA = 70, BalB = 80)\}
\]

\[2\] Smart contracts are typically called *chaincodes* in Fabric. However, as they do not conceptually differ from smart contracts in blockchain systems such as Ethereum, we stick to this term throughout the paper.
In this sense, the simulation of the smart contract is actually only a monitoring of the execution effects. The reason for performing only a simulation is that in this phase, we can not be sure yet whether this transaction will be allowed to commit eventually – this check will be performed later in the validation phase.

After the simulation of the smart contract on all endorsement peers, in step 3, the endorsement peers return their individually computed read and write sets to the client, that sent the transaction proposal. Additionally, they return a signature of their simulation, that will be relevant in the validation phase in Section 2.6. If all read sets and write sets match in step 4, the actual transaction (called \(T_7\) in the following) is formed from the results of the endorsement. This transaction \(T_7\) now contains the effects of the execution in form of the read and write set as well as all signatures and can be passed on to the ordering service.

### 2.5 Ordering Phase

As mentioned, the central component of the ordering phase is the ordering service, that we visualize in Figure 3. It receives all transactions, that made it through the simulation phase. Consequently, it receives in step 5 our transaction \(T_7\), that we followed through the simulation phase in Section 2.4. In step 6, we assume that it also receives two other transactions \(T_8\) and \(T_9\), that were endorsed in parallel to \(T_7\).

The ordering service now has the sole purpose of establishing a global order among the transactions. It treats the transactions in a black box fashion and does not inspect the transaction semantics, such as the read and write set, in any way. By default, it essentially arranges the transactions in the order in which they arrive, resulting in what we call for the rest of the paper the *arrival order*. In step 7, the ordering service now outputs the ordered stream of transactions in form of *blocks*, containing a certain number of transactions. Outputting whole blocks instead of individual transactions reduces the pressure on the network, as less communication overhead is produced.

Finally, the generated block is distributed to all four peers of the network to start the validation phase. Note that there is no guarantee that all peers receive a block at the same time, as the distribution happens partially from ordering service to peers directly as shown in step 8 and partially between the peers using a gossip protocol as shown in step 9. However, the service assures that all peers receive the blocks in the same order.

---

3They might not match due to non-determinism in the smart contract or due to malicious behavior of the endorsement peer(s).
2.6 Validation and Commit Phase

When a block arrives at a peer, the validation phase starts, visualized in Figure 4 for peer A1. The three remaining peers execute the same validation process. Overall, the validation phase has two purposes.

2.6.1 Endorsement Policy Evaluation

The first purpose is to validate the transactions in the block with respect to the endorsement policy. For example, it is possible that a malicious transaction was generated by a malicious client and a malicious peer in conspiracy to take advantage of the money transfer. Let us assume that transaction $T_8$ is such a malicious transaction and that the malicious client, which proposed $T_8$, works together with peer $A_2$, which is also malicious. Instead of using the legit write set $WS_{B_2} = \{BalA = 30, BalB = 120\}$ from $B_2$, the client creates a proposal with the write set $WS_{A_2} = \{BalA = 100, BalB = 120\}$, that it received from its collaborator $A_2$.

How is this transaction $T_8$ now detected in the validation phase? The key to this lies in the signatures $Sig_{A_2}$ and $Sig_{B_2}$, that the endorsement peers generate at the end of the simulation phase. The signature is computed over the read and write set, the executed smart contract, and the used endorsement policy. The client receives these cryptographically secure signatures and must pack them into the transaction along with the read and write set. The peers that validate the transaction recompute the signatures of all endorsement peers, that were responsible for transaction $T_8$ and compare the signatures with the received ones $Sig_{A_2}$ and $Sig_{B_2}$. In our example, in step 10, the peers detect that the signature of the honest peer $Sig_{B_2}$ does not match to the one they computed from the received write set and thus, would classify $T_8$ as invalid. $T_7$ and $T_9$, the remaining transactions in the block, are evaluated in parallel. Their signatures match the ones computed from the read and write set and therefore, these transactions are valid with respect to the endorsement policy.

2.6.2 Serializability Conflict Check

The second purpose of the validation is to analyze the transactions with respect to serializability conflicts, that can arise from the order of transactions. For every transaction, it must be checked whether the version-numbers of all keys in the read set match the version-numbers in the current state. Only if this is the case, a transaction operates on an up-to-date state. Considering our example, let us perform the serializability conflict check for the received block. $T_8$ is already marked as invalid as it did not pass the endorsement policy evaluation, so it is not checked again. $T_7$ passed the endorsement policy evaluation and is now tested for serialization conflicts in step 11. Its read set is $RS = \{(BalA, v_3), (BalB, v_2)\}$. The version numbers of $BalA$ and $BalB$ in the read set match the ones of the current state and therefore, $T_7$ is marked as valid. As a consequence, in step 12, the write set of $T_7$, namely $WS = \{BalA = 70, BalB = 80\}$ is written to the current state. This changes the current state to $BalA = (70, v_4)$ and $BalB = (80, v_3)$. Note that the version-numbers of the modified variables are incremented.

The next transaction to be checked is $T_9$ in step 13. Let us assume it also performs a money transfer and has the following read and write set:

$$RS = \{(BalA, v_3), (BalB, v_2)\} \quad WS = \{BalA = 0, BalB = 150\}$$
This transaction will not pass the conflict check, as it read \(BalA\) in version \(v_3\) and \(BalB\) in version \(v_2\), while the current state already contains \(BalA\) in version \(v_4\) and \(BalB\) in version \(v_3\). Therefore, it operated on outdated data and is marked as invalid. As a consequence, its write set is not applied to the current state and simply discarded.

Finally, after validating all transactions of the block, in step 14 the entire block is appended to the ledger along with the information about which transactions are valid or invalid.

### 3 Blurred Lines: Fabric vs Distributed Database Systems

As we now have an understanding of the workflow of Fabric, we are able to discuss its architecture in relation to distributed database systems. In particular, we are interested in aspects of Fabric, that are (a) conceptually shared with distributed database systems, but (b) have potential for the application of database technology.

#### 3.1 The Importance of Transaction Order

The first component we look at is the ordering mechanism. Such a component is also present in any distributed database system with transaction semantics and therefore a great candidate for transitioning database technology to Fabric.

As described in Section 2.5, Fabric relies on a single trustworthy ordering service for ordering transactions. Since Fabric simulates the smart contracts bound to proposals before performing the ordering, the order actually has an influence on the number of serialization conflicts between transactions. Again, this is a property shared with any parallel database system, that separates transaction execution from transaction commit.
In ordering transactions, various different strategies are possible: The simplest option is to arbitrarily order them, for instance in the order in which they arrive. While this arrival order is fast to establish, it can lead to serialization conflicts, that are potentially unnecessary. These conflicts increase the number of invalid transactions, which must be resubmitted by the client. Unfortunately, the vanilla Fabric follows exactly this naive strategy. This is caused by the design decision that the ordering service is not supposed to inspect the transaction semantics, such as the read and write set, in any way. Instead, it simply leaves the transactions in the order in which they arrive. This strategy can be problematic, as the example in Table 1 shows. In this example, four transactions are scheduled in the order in which they arrive, namely \( T_1 \Rightarrow T_2 \Rightarrow T_3 \Rightarrow T_4 \), where \( T_1 \) updates the key \( k_1 \) from version \( v_1 \) to \( v_2 \). Since the transactions \( T_2, T_3, \) and \( T_4 \) each read \( k_1 \) in version \( v_1 \) during their simulations, they have no chance to commit, as they operated on an outdated version of the value of \( k_1 \). They will be identified as invalid in the validation phase and the corresponding transaction proposals must be resubmitted by the client, resulting in a new round of simulation, ordering, and validation.

Table 1: For the order \( T_1 \Rightarrow T_2 \Rightarrow T_3 \Rightarrow T_4 \), only one out of four transactions is valid: \( T_2, T_3, \) and \( T_4 \) read the outdated version \( v_1 \) of key \( k_1 \), that has been updated by \( T_1 \) to \( v_2 \) before.

| Transaction | Read Set | Write Set | Is Valid? |
|-------------|----------|-----------|-----------|
| 1. \( T_1 \) | —        | \((k_1, v_1 \rightarrow v_2)\) | ✓         |
| 2. \( T_2 \) | \((k_1, v_1), (k_2, v_1)\) | \((k_2, v_1 \rightarrow v_2)\) | ×         |
| 3. \( T_3 \) | \((k_1, v_1), (k_3, v_1)\) | \((k_3, v_1 \rightarrow v_2)\) | ×         |
| 4. \( T_4 \) | \((k_1, v_1), (k_3, v_1)\) | \((k_4, v_1 \rightarrow v_2)\) | ×         |

Interestingly, for the four transactions from the previous example, there exists an order that is conflict free. In the schedule \( T_4 \Rightarrow T_2 \Rightarrow T_3 \Rightarrow T_1 \), as shown in Table 2 all four transactions are valid, as their read and write sets do not conflict with each other in this order.

Table 2: The order \( T_4 \Rightarrow T_2 \Rightarrow T_3 \Rightarrow T_1 \) results in all four transactions being valid.

| Transaction | Read Set | Write Set | Is Valid? |
|-------------|----------|-----------|-----------|
| 1. \( T_4 \) | \((k_1, v_1), (k_3, v_1)\) | \((k_4, v_1 \rightarrow v_2)\) | ✓         |
| 2. \( T_2 \) | \((k_1, v_1), (k_2, v_1)\) | \((k_2, v_1 \rightarrow v_2)\) | ✓         |
| 3. \( T_3 \) | \((k_1, v_1), (k_3, v_1)\) | \((k_3, v_1 \rightarrow v_2)\) | ✓         |
| 4. \( T_1 \) | —        | \((k_1, v_1 \rightarrow v_2)\) | ✓         |

This example shows that the vanilla orderer of Fabric misses a chance of removing unnecessary serialization conflicts. While this problem is new to the blockchain domain, as blockchains typically offer only a serial execution of transactions, within the database community, this problem is actually well known. There exist reordering mechanisms which aim at minimizing the number of serialization conflicts via a reordering of transactions \([21, 13, 20]\). However, in a database system, it is typically avoided to buffer a large number of incoming transactions before processing as low latency is mandatory. Thus, reordering is not always an option in such a setup. Fortunately, as blockchain systems buffer the incoming transactions anyways to group them into blocks, this gives us the opportunity to apply sophisticated transaction reordering mechanisms without introducing significant overhead.
We will add such a transaction reordering mechanism to Fabric in Section 4.1, which significantly enhances the number of valid transactions, that make it through the system.

3.2 On the Lifetime of Transactions

The second aspect we look at from a database perspective tackles the lifetime of transactions within the pipeline. In Fabric, every transaction that goes through the system is either classified as valid or as invalid with respect to the validation criteria. In the vanilla version, this classification happens in the validation phase right before the commit phase. A severe downside of this form of late abort is that a transaction, that violated the validation criteria already in an earlier phase, is still processed and distributed across all peers. This penalizes the whole system with unnecessary work, throttling the performance of valid transactions. Besides, this concept also delays the abort notification to the client.

We have to distinguish in which phase a violation happens. First, a violation can occur already in the simulation phase, in form of so called cross-block conflicts, meaning a transaction from a later block, which is currently in the simulation phase, conflicts with a valid transaction from an earlier block. Second, a violation can occur as well as in the ordering phase, in form of within-block conflicts between conflicting transactions in a single block.

Let us look at these two scenarios in isolation in Section 3.2.1 and Section 3.2.2 respectively.

3.2.1 Violation in the simulation phase (cross-block conflicts)

To understand the problem in the simulation phase, let us look at the following situation and how the vanilla version of Fabric handles it. Let us assume there are four transactions $T_1$, $T_2$, $T_3$, and $T_4$ that are currently in the ordering phase and that end up in a block of size four, which is shipped to all peers for validation. Before the validation of that block starts within a peer $P$, the smart contract of a transaction proposal $T_5$ starts its simulation in $P$. To do so, it acquires a read lock\(^4\) on the entire current state. While the simulation is running, the block has to wait for the validation, as it has to acquire an exclusive write lock on the current state. The problem in this situation is: if $T_1$, $T_2$, $T_3$, or $T_4$ write the value of a key, that is read by $T_5$, then $T_5$ simulates on stale data. Therefore, in the moment of the read, the transaction becomes virtually invalid. Still, in the vanilla version of Fabric, this stale read is not detected before the validation phase of $T_5$. Thus, $T_5$ would continue its simulation and go through the ordering phase, just to be invalidated in the very end.

3.2.2 Violation in the ordering phase (within-block conflicts)

Apart from conflicts across blocks, there can be conflicts between transactions within a block. These conflicts appear after putting the transactions into a particular order in the ordering phase. For instance, the example from Table 1 in Section 3.1 showed a schedule, where the three transactions $T_2$, $T_3$, and $T_4$ individually conflict with the previously scheduled transaction $T_1$ of the same block. Unfortunately, these conflicts are not detected within the orderer of the vanilla version of Fabric. The block containing $T_2$, $T_3$, and $T_4$ would be distributed across all peers of the network for validation, although 3/4 of transactions within the block are virtually invalid.

---

\(^4\)The read lock can be shared by multiple simulation phases, as they do not modify the current state.
As before, this originates from the design decision that the ordering service does not inspect transaction semantics.

The mentioned situations show that Fabric misses several chances to abort transactions right at the time of violation. In contrast to that, database systems are typically very eager in aborting transactions [10], as it decreases network traffic and saves computing resources. This concept of "cleaning" the pipeline as early as possible is called early abort in the context of databases, which apply this concept in various flavors. For instance, besides of the early abort of transactions, that violate certain criteria, database systems eliminate records from the query result set as early as possible by pushing down selection and projection operations in the query plan.

To overcome the mentioned problems, we will apply the concept of early abort at several stages of the transaction processing pipeline of Fabric. By this, we assure to utilize the available resources with meaningful work to the extend. We will detail this in Section 4.2.

4 Fabric++

We have outlined the problems of Fabric and how they relate to key problems known in the context of database systems. Let us now see precisely how we counter them. First, in Section 4.1, we introduce a transaction reordering mechanism, that aims at minimizing the number of unnecessary within-block conflicts. Second, in Section 4.2, we introduce early transaction abort to several stages of the Fabric pipeline. This also involves the introduction of a fine-grained concurrency control mechanism.

4.1 Transaction Reordering

When reordering a set of transactions $S$, multiple challenges must be faced. First, we have to identify which transactions of $S$ actually conflict with each other with respect to the actions they perform. Precisely, we have a conflict between two transactions $T_i$ and $T_j$ (denoted as $T_i \rightarrow T_j$), if $T_i$ writes to a key that is read by $T_j$. In this case, $T_i$ must be ordered after $T_j$ (denoted as $T_j \Rightarrow T_i$) to make the schedule serializable, as otherwise, the read of $T_j$ would be outdated. Unfortunately, the problem is typically more complex as cycles of conflicts can occur, such that simple reordering can not resolve the problem. For example, if we have the cycle of conflicts $T_i \rightarrow T_j \rightarrow T_k \rightarrow T_i$, there is no order of these three transactions that is serializable. Therefore, before reordering transactions, our mechanism must actually first remove certain transactions of $S$ to form a subset $S' \subseteq S$, from which a serializable schedule can be generated. Of course, a goal must be to remove as few transactions as possible. Finally, after computing $S'$, we can derive a concrete serializable schedule from the transactions in $S'$.

On a high-level, we have to carry out the steps as shown in the pseudo-code of Algorithm 5 to create a serializable schedule for a set of transactions $S$.

4.1.1 Example

To understand the principle and to discuss some of the implementation details, let us go through a concrete example. Let us assume we have a set $S$ of six transactions $T_0$ to $T_5$ to consider for
func reordering(Transaction[] S) {
    // Step 1: Inspect the read/write set of all transactions and build a conflict graph.
    Graph cg = buildConflictGraph(S)
    // Step 2: Within cg, we have to identify all occurring cycles. Divide cg into
    // strongly connected subgraphs using Tarjans algorithm [2] in divideIntoSubgraphs().
    Graph[] cg_subgraphs = divideIntoSubgraphs(cg)
    // Each strongly connected subgraph of cg with more than one node must contain at
    // least one cycle. We identify the cycles within the subgraphs using Johnsons
    // algorithm [1] in getAllCycles().
    Cycle[] cycles = emptyList()
    foreach subgraph in cg_subgraphs:
        if(subgraph.numNodes() > 1):
            cycles.add(subgraph.getAllCycles())
    // Step 3: To remove the cycles in cg, we have to remove conflicting transactions from
    // S. For each transaction of S, we count in how many cycles it occurs.
    MaxHeap transactions_in_cycles = emptyMaxHeap()
    foreach Cycle c in cycles:
        foreach Transaction t in c:
            if transactions_in_cycles.contains(t):
                transactions_in_cycles[t]++
            else:
                transactions_in_cycles[t] = 1
    // Step 4: Let us define S' as S. We now greedily remove the transaction from S' that
    // occurs in most cycles, until all cycles have been resolved.
    Transaction[] S' = S
    while not cycles.empty():
        Transaction t = transactions_in_cycles.popMax()
        S'.remove(t)
        foreach Cycle c in cycles:
            if c.contains(t):
                c.remove(t)
                cycles.remove(c)
        // Step 5: From S' we have to form the actual serializable schedule. We start by
        // building the (cycle-free) conflict graph of S'.
        Graph cg' = buildConflictGraph(S')
        // Compute schedule. We start at some node of the graph, that hasn't been visited yet.
        Transactions[] order = emptyList()
        Node startNode = cg'.getNextNode()
        while order.length() < cg'.numNodes():
            addNode = true
            if startNode.alreadyScheduled():
                startNode = cg'.getNextNode()
                continue
            // Traverse upwards to find a source
            foreach Node parentNode in startNode.parents():
                if not parentNode.alreadyScheduled():
                    startNode = parentNode
                    addNode = false
                    break
            // A source has been found, so schedule it and traverse downwards.
            if addNode:
                startNode.scheduled()
                order.append(startNode)
            foreach Node childNode in startNode.children():
                if not childNode.alreadyScheduled():
                    startNode = childNode
                    break
        // we invert the order to get the actual schedule
    return order.invert()
}

Figure 5: Pseudo code of the reordering mechanism.
reordering. These six transactions have read and write sets as shown in Table 3. In total, they access ten unique keys $K_0$ to $K_9$.

**Step (1):** Based on this information, we now have to generate the conflict graph of the transactions as done by the function `buildConflictGraph()` in line 3 of Algorithm 5. To do so in an efficient way, we interpret the rows of Table 3 as bit-vectors of length 10. Let us refer to them as $vec_r(T_i)$ for the reading accesses and $vec_w(T_i)$ for the writing accesses of a transaction $T_i$. For each transaction $T_i$, we now perform a bitwise $\&$-operation between $vec_r(T_i)$ and $vec_w(T_j)$ for all $j \neq i$. If the result of an $\&$-operation is not 0, we have identified a read-write conflict and create an edge in the conflict graph between the corresponding transactions. For example, for $T_0$ we have the reading accesses $vec_r(T_0) = 1100000000$ The bitwise $\&$-operation with $vec_w(T_1) = 1000000000$ leads to $vec_r(T_0) \& vec_w(T_1) = 1000000000$, which is not 0. This means $T_1$ writes a key that $T_0$ is reading and thus, we put a corresponding edge in the conflict graph.

As a result, we obtain the conflict graph $C(S)$ of our six transactions as shown in Figure 6.

![Figure 6: Conflict graph $C(S)$ of the transactions in $S$.](image)

Note that this algorithm has quadratic complexity on the number of transactions. Still, we apply it as the number of transactions to consider is very small in practice due to the limitation by the block size and therefore, the overhead is negligible.
Step (2): To identify the cycles, we apply Tarjan’s algorithm [16] in the function \textit{divideIntoSubgraphs()} in line 6 to identify all strongly connected subgraphs. In general, this can be done in linear time in $O(N + E)$ over the number of nodes $N$ and number of edges $E$ and results in the three subgraphs as shown in Figure 7.

Using Johnson’s algorithm [11], we then identify all cycles within the strongly connected subgraphs. Again, this step can be done in linear time in $O((N + E) \cdot (C + 1))$, where $C$ is the number of cycles. Therefore, if there are no cycles in the subgraphs, the overhead of this step is very small.

![Figure 7: The three strongly connected subgraphs of the conflict graph of Figure 6.](image)

We identify that the first subgraph (colored in green) consisting of $T_0$, $T_1$, and $T_3$ contains the two cycles $c_1 = T_0 \rightarrow T_3 \rightarrow T_0$ and $c_2 = T_0 \rightarrow T_3 \rightarrow T_1 \rightarrow T_0$. The second subgraph (colored in red) consisting of $T_2$ and $T_4$ contains the cycle $c_3 = T_2 \rightarrow T_4 \rightarrow T_2$. The third subgraph (colored in yellow) contains only one node and is thus cycle-free.

Step (3): From this information, we can build a table denoting for every transaction in which cycle it appears as shown in the lines 10 to 13 of Algorithm 5. Table 4 visualizes the result for our example. If a transaction $T_i$ is a part of a cycle $c_j$, the corresponding cell is set to 1, otherwise 0. The last row of the table sums up for every transaction in how many cycles it is contained in total.

| Cycle | $T_0$ | $T_1$ | $T_2$ | $T_3$ | $T_4$ | $T_5$ |
|-------|-------|-------|-------|-------|-------|-------|
| $c_1$ | 1     | 0     | 0     | 1     | 0     | 0     |
| $c_2$ | 1     | 1     | 0     | 1     | 0     | 0     |
| $c_3$ | 0     | 0     | 1     | 0     | 1     | 0     |
| **∑** | **2** | **1** | **1** | **2** | **1** | **0** |

Step (4): We now iteratively remove transactions, that participate in cycles, starting from the ones that appear in most cycles. The lines 25 to 32 of Algorithm 5 show the corresponding pseudo-code. As we can see, $T_0$ and $T_3$ both appear in two cycles, so we take care of them first. If we can choose between two transactions, such as $T_0$ and $T_3$, we pick the one with the smaller subscript. This assures that our algorithm is deterministic. We remove $T_0$, which clears all cycles in which $T_0$ appears, namely $c_1$ and $c_2$. The sum is updated accordingly, as we can see in Table 5.

The transactions $T_2$ and $T_4$ remain with a participation in cycle $c_3$ each. We remove $T_2$ which clears $c_3$ and thereby the last cycle. This results in the state of Table 6.
Table 5: Removing $T_0$ clears the cycles $c_1$ and $c_2$.

| Cycle | $T_0$ | $T_1$ | $T_2$ | $T_3$ | $T_4$ | $T_5$ |
|-------|-------|-------|-------|-------|-------|-------|
| $c_1$ | $X$   | $\theta$ | $\theta$ | $X$   | $\theta$ | $\theta$ |
| $c_2$ | $X$   | $X$   | $\theta$ | $X$   | $\theta$ | $\theta$ |
| $c_3$ | 0     | 0     | 1     | 0     | 1     | 0     |
| $\sum$| 0     | 0     | 1     | 0     | 1     | 0     |

Table 6: Removing $T_2$ clears the last cycle $c_3$.

| Cycle | $T_0$ | $T_1$ | $T_2$ | $T_3$ | $T_4$ | $T_5$ |
|-------|-------|-------|-------|-------|-------|-------|
| $c_1$ | $X$   | $\theta$ | $\theta$ | $X$   | $\theta$ | $\theta$ |
| $c_2$ | $X$   | $X$   | $\theta$ | $X$   | $\theta$ | $\theta$ |
| $c_3$ | $\theta$ | $\theta$ | $X$ | $\theta$ | $X$ | $\theta$ |
| $\sum$| 0     | 0     | 0     | 0     | 0     | 0     |

From this we now know that from the set $S' = \{T_1, T_3, T_4, T_5\}$ we can generate a serializable schedule, leading to the cycle-free conflict graph $C(S')$ (line 37) as shown in Figure 8.

**Step (5):** Generating the final schedule is essentially a repetitive execution of two parts until all nodes are scheduled: (a) the locating of the source node in the current subgraph (lines 43 to 51) and (b) the scheduling of all nodes that reachable from that source (lines 53 to 59).

![Figure 8: The cycle-free conflict graph $C(S')$, containing only the transactions $T_1$, $T_3$, $T_4$, and $T_5$.](image)

We start part (a) at the node of $C(S')$ representing the transaction with the smallest subscript, namely $T_1$. From this starting node, we have to find a source node, as sources have to be scheduled last. $T_1$ has two parents, namely $T_3$ and $T_4$, so it not a source. We follow the edge to $T_3$, which has not been visited yet but is also not a source, as it has $T_4$ as a parent as well. We follow the edge to $T_4$, which has not been visited yet and which is a source. Therefore, we can schedule $T_4$ safely at the last position in our schedule, to which we refer to as position 4. Now, part (b) starts as all nodes that are reachable from $T_4$ must be scheduled before it. $T_4$ has two children, namely $T_1$ and $T_3$. We follow the edge to $T_1$, which has not been scheduled yet. However, as $T_1$ has an incoming edge from $T_3$, we also can not directly schedule it. First, we visit $T_3$ and identify that it has a parent in form of $T_4$, the source at which we started. With this information, we know that $T_3$ must be scheduled at position 3 and $T_1$ must be scheduled at position 2. This ends part (b), as all reachable nodes have been scheduled. Next, we restart at the only remaining node $T_5$. As $T_5$ is not only a source but also a sink, we can schedule it.
instantly at position 1. This results in the final schedule $T_5 \Rightarrow T_1 \Rightarrow T_3 \Rightarrow T_4$, which is returned to the orderer.

Please note that our reordering mechanism is not guaranteed to abort a minimal number of transactions, as this would be a NP-hard problem. However, it offers a very lightweight way to generate a serializable schedule with a small number of aborts.

### 4.1.2 Batch Cutting

In the context of transaction reordering, we have to discuss and extend a mechanism within the ordering service, that we omitted for simplicity in the description of Fabric in Section 2, namely *batch cutting*. When the ordering service receives the transactions in form of a constant stream, it decides based on multiple criteria when to “cut” a batch of transactions to finalize it and to form the block. In the vanilla version, a batch is cut as soon as one of the following three conditions hold: (a) The batch contains a certain number of transactions. (b) The batch has reached a certain size in terms of bytes. (c) A certain amount of time has passed since the first transaction of this batch was received.

In Fabric++, we extend these criteria by one additional condition. We also cut the batch, if (d) the transactions within the batch access a certain number of unique keys. This condition ensures that the runtime of our reordering mechanism, in particular the time of step (1), remains bounded.

### 4.1.3 Micro-Benchmark

To analyze the effectiveness of our reordering mechanism, we first evaluate it in a stand-alone micro-benchmark in isolation of Fabric. For a given sequence of input transactions we compute the number of valid transactions for this particular sequence (called “arrival order” in the following plots) as well as for the sequence that is generated by our reordering mechanism (called

![Figure 9: Workload 1: Varying the number of conflicts.](image-url)
“reordered” in the following plots). Additionally, we measure the time to compute the reordered schedule. In Figure 9 we test a workload pattern with varying number of conflicts. For the interested reader, we provide a second micro-benchmark in the Appendix B on the effect of varying the length of the cycles (Figure 15) and see how well our reordering mechanism performs in comparison to the naive arrival order.

4.1.4 Micro-Benchmark 1: Interleave reads and writes to vary the number of conflicts

The first input sequence we test consists of two equal sized sub sequences, where one subsequence contains only transactions that perform writes (colored in red) and the other sequence only transactions that read (colored in blue). Each transaction performs only one operation (either read or write). Neither two writes nor two reads happen to the same key. For the example of $n = 6$ transactions, we start with the following sequence $S_1$:

$$S_1 = T[w(k_1)], T[w(k_2)], T[w(k_3)], T[r(k_1)], T[r(k_2)], T[r(k_3)]$$

To generate $S_i$, we move the last transaction of $S_{i-1}$ to the front, leading to the following sequences $S_2$, $S_3$, and $S_4$.

$$S_2 = T[r(k_3)], T[w(k_1)], T[w(k_2)], T[w(k_3)], T[r(k_1)], T[r(k_2)]$$
$$S_3 = T[r(k_2)], T[r(k_3)], T[w(k_1)], T[w(k_2)], T[w(k_3)], T[r(k_1)]$$
$$S_4 = T[r(k_1)], T[r(k_2)], T[r(k_3)], T[w(k_1)], T[w(k_2)], T[w(k_3)]$$

The more writing transactions happen before the corresponding reading transactions, the more conflicts happen. We want to find out whether our reordering mechanism can solve this problem.

Figure 9 shows the results for $n = 1024$ transactions. As we can see, our reordering mechanism is able to reorder the transactions for every input sequence in a way such that all transactions are valid. In contrast to that, the arrival order suffers under a lot of invalid reading transactions, if writing transactions happen before. We can also see that our reordering mechanism is computationally cheap: it takes only around 1 to 2 ms to rearrange the transactions on a Macbook Pro with Intel Core i7 running at 3.1 GHz.

4.2 Early Transaction Abort using Advanced Concurrency Control

The reordering mechanism previously described not only tries to minimize the number of unnecessary aborts, it also enables a form of early abort. Transactions, that are removed from $S$ because of their participation in conflict cycles can be aborted already in the ordering phase instead of later on the validation phase. This assures that less transactions are distributed across the network.

In the following, we want to push this concept of aborting transactions as early as possible in the pipeline to the limits. Additionally to early aborting transactions that occur in conflict cycles, we can integrate two more applications of early abort, as we will describe in Section 4.2.1 and Section 4.2.2. The first one is happening already in the simulation phase. Let us see in the following how this works.
4.2.1 Early Abort in the Simulation Phase

To realize early abort in the simulation phase, we first have to extend Fabric by a more fine-grained concurrency control mechanism, that allows for the parallel execution of simulation and validation phase within a peer. With such a mechanism at hand, we have the chance of identifying stale reads during the simulation already.

To understand the concept, let us consider the example from Section 3.2.1 again. With a fine-grained concurrency control mechanism, the block containing \( T_1, T_2, T_3, \) and \( T_4 \) would not have to pend for validation while the smart contract bound to the proposal \( T_5 \) is simulating. Instead, the four transactions would apply their updates in an atomic fashion while \( T_5 \) is simulating. As a consequence of this design, for every read \( T_5 \) performs, we can check whether the read value is still up-to-date. As soon as we detect a stale read, we can abort the simulation of the transaction proposal. Additionally, we directly notify the corresponding client about the abort, such that it can resubmit the proposal without delay.

Let us discuss in the following, how exactly our fine-grained concurrency control mechanism works and how we realize it in Fabric++. In the context of modern database systems, advanced concurrency control mechanisms are well established [14, 17, 18, 19, 12, 15]. Instead of locking the entire store, these techniques typically perform a fine-grained locking on the record level or even at the level of individual cells/values. As there is conceptually no difference between the store of a database system and the store used within the Fabric peers, similar techniques can be applied here.

As discussed in Section 2, Fabric implements its current state in form of a key-value store, which maps each individual key to a pair of value and version-number. The version-number is actually composed of the ID of the transaction, that performed the update, as well as the ID of the block that contains the transaction. In the original version of Fabric, the sole purpose of the version-numbers is to identify stale reads. In the validation phase, for every transaction we check whether the version-number of the read value still matches the one in the current state. We can go one step further and exploit the available version-numbers to implement a lock-free concurrency control mechanism protecting the current state. To do so, in Fabric++, we first remove the read-write lock, that was unnecessarily sequentializing simulation and validation phase. The version-number, that is maintained with each value, is sufficient to ensure the same transaction isolation semantics as the vanilla version. As no lock is acquired anymore, we need a mechanism to ensure that updates performed by the validation phase are not seen by simulation

\[ \text{Figure 10: Parallelization with early abort using our fine-grained concurrency control.} \]
phases running in parallel. To achieve this behavior, during simulation, we have to inspect the version-number of every read value and test whether it is still up-to-date.

Figure 10 visualizes this concept using a concrete example. At the start of the simulation phase, we first identify the block-ID of the last block that made it into the ledger. Let us refer to this block-ID as the last-block-ID. In our example, last-block-ID = 4. During the simulation of a smart contract bound to a transaction proposal $T_{exec}$, no read must encounter a version-number containing a block-ID higher than the last-block-ID. If it does see a higher block-ID it means that during the simulation phase, a validated transaction $T_{valid}$ in the validation phase modified a value in the read set of $T_{exec}$ and thus, the read set is outdated.

In our example, the read of $balA = 70$ in the simulation phase happens before the update of $balA$ to 50 in the validation phase. This is reflected by the version-number of $balA$, namely block-ID = 4. Therefore, this read is up-to-date and the simulation continues. In contrast to that, the read of $balB$ happens after the update of $balB$ to 100 in the validation phase. This is reflected by the version-number of $balB$, namely block-ID = 5. As 5 is higher than the last-block-ID = 4, we can directly classify $T_{exec}$ as invalid, as the transaction will not have a chance to pass the validation phase later on. Please note that the overall correctness of our lock-free mechanism is ensured by the atomic updates of the version-numbers.

4.2.2 Early Abort in the Ordering Phase

In addition to the early abort in the simulation phase, as explained in Section 4.2.1, we can transition a similar concept also to the ordering phase. As Fabric performs commits at the granularity of whole blocks, two transactions within the same block, that read the same key, must read the same version of that key. For example, let us consider two transactions $T_6$ and $T_7$, where $T_6$ is ordered before $T_7$ within the same block ($T_6 \Rightarrow T_7$). If $T_6$ read version $v_1$ of a key $k$ and $T_7$ read version $v_2$ of $k$ in their respective simulations, then $T_7$ is invalid. Such a version mismatch can happen, if between the simulations of $T_6$ and $T_7$ a change to the value of $k$ was committed by a valid transaction from a previous block. Therefore, as soon as we detect a version mismatch between transactions within the same block, we can early abort the latter transaction. Again, this strategy assures that only those transactions end up in a block, that have a realistic chance of commit.

5 Experimental Evaluation

In the previous section, we have extended and modified core components of Fabric in several ways, turning it into Fabric++. It is now time to evaluate the modifications in terms of effectiveness. Primarily, we are interested in the throughput of valid/successful and invalid/failed transactions, that make it through the system. Secondarily, we are interested on the influence of certain system configurations and the workload characteristics on the system.

5.1 Setup

Before starting with the actual experiments, let us discuss the setup. Our cluster consists of six identical servers, that are located within the same rack and connected via gigabit-ethernet.
Four machines serve as peers, one machine runs the ordering service, and one machine serves as the client, which fires transaction proposals. Each server consists of two quad-core Intel Xeon CPU E5-2407 (SandyBridge architecture) running at 2.2 GHz with 32KB of L1 cache, 256KB of L2 cache, and 10MB of a shared L3 cache. 24GB of DDR3 ram are attached to each of the two NUMA regions. The operating system used is a 64-bit Arch Linux with kernel version 4.17. Fabric is set up to use LevelDB as the current state database.

5.2 Benchmark Framework and Workload

In the database community, there exist numerous established benchmarks that can be used to test and to compare systems, such as TPC-C [5], TPC-H [6], or YCSB [1]. Unfortunately, since blockchains are still a relatively young field, there exist only very few benchmarks with standardized workloads.

At first, we looked at the Caliper [2] benchmarking suite which seemed like a natural candidate, as it is part of the Hyperledger project just like Fabric. It is compatible to Fabric 1.2, but comes with a few limitations: First, the framework provides only sample smart contracts and not a real benchmarking workload. Second, for certain metrics such as transactions per second or latency, it remains unspecified how they are actually measured. Third, it supports only a single channel. Apart from these limitations, other researcher have experienced incorrect behavior of Caliper in form of events, that were not properly registered. [4]. As a consequence, they released a fork of Caliper named Gauge [3] that claims to resolve these problems. Unfortunately, Gauge is not compatible with version 1.2 of Fabric right now. Next, we looked at Blockbench, which originates from a survey paper [9] on blockchain systems. While Blockbench actually provides some benchmarking workloads such as YCSB, again, it lacks the support for Fabric 1.2 and would need significant changes to make it compatible.

As a consequence of this journey, we decided to build our own benchmarking framework and to introduce a highly customizable workload. This allows us to fire transaction proposals at a specified rate from multiple clients in multiple channels. Our benchmark setup looks as follows: Initially, we create a certain number of accounts (10,000 accounts throughout this section, 20,000 accounts in Appendix C), each represented by a randomly generated account balance. Our workload is formed of a single smart contract, that reads and writes an adjustable number of account balances, simulating a typical asset transfer scenario between accounts. Among the accounts, there exist a certain number of hot accounts, that are involved in transactions more frequently than the remaining ones. By varying the number of read and write accesses per transaction, the probability of picking hot accounts, and the number of hot accounts, we are able to generate a wide range of different workloads.

In a single run, we fire a constant stream of transaction proposals, that are bound to our smart contract, for a certain amount of time at a certain firing rate. In the following, we test numerous different system and workload configurations to identify the impact of the system. In the individual experiments, we will detail the chosen configuration.

5.3 Transactional Throughput

We start our experimental evaluation by testing Fabric and Fabric++ under probably the most important criterium for a transaction processing system, namely the throughput of transactions.
Figure 11: Transactional Throughput of Fabric and Fabric++ under 108 different configurations.
We differentiate between successful and failed transactions: a good system should try to maximize the number of successful transactions while keeping the number of failed transactions as small as possible.

Table 7: Our 108 configurations for the evaluation of the transactional throughput as shown in Figure 11.

| Experiment Parameters                                             | Values          |
|-------------------------------------------------------------------|-----------------|
| Fired transaction proposals per second per client                 | 512             |
| Duration in which transaction proposals are fired                 | 90 sec          |
| Number of channels                                                | 1               |
| Number of clients per channel                                     | 4               |

| System Parameters                                                 | Values          |
|-------------------------------------------------------------------|-----------------|
| Maximum time to form a block                                      | 1 sec           |
| Maximum number of keys accessed per block                         | 16384           |
| Maximum size per block                                            | 2MB             |
| Maximum number of transactions per block (BS)                     | 256, 512, 1024   |

| Workload Parameters                                               | Values          |
|-------------------------------------------------------------------|-----------------|
| Number of account balances                                        | 10000           |
| Number of read & written balances per transaction (RW)            | 4, 8            |
| Probability for picking a hot account for reading (HR)            | 10%, 20%, 40%   |
| Probability for picking a hot account for writing (HW)            | 5%, 10%         |
| Number of hot account balances (HSS)                              | 1%, 2%, 4%      |

To measure this property, we fire a constant stream of transaction proposals for 90 seconds into a single channel using four clients. Each client fires at a rate of 512 proposals per second. This firing rate is sufficient to fully sustain the system in our setup. Table 7 shows the detailed configuration. To identify their impact on the throughput, we vary five important parameters: the maximum number of transactions per block (BS), the number of read balances and written balances per transaction (RW), the probability for picking a hot account for reading (HR) respectively for writing (HW), as well as the number of hot account balances (HSS). In total, we evaluate 108 different configurations in this experiment.

Figure 11 shows the results. First and foremost, we vary the maximum number of transactions per block (BS), as it has a large impact on the transaction processing in general and the ordering in particular. The results for Fabric and Fabric++ for BS=256 are presented in first and the second row, for BS=512 in the third and fourth row, and for BS=1024 in the fifth and sixth row, respectively. Along the columns, we vary the remaining four parameters RW, HR, HW, and HSS in a total of 36 configurations. For a single run, we show the transactional throughput (blue) that was achieved for each second of the 90 second run. This throughput is additionally split into successful transactions (green) and failed transactions (red).

To interpret the results, let us look at Figure 11 as a whole. We can see that Fabric++ significantly increases the throughput of successful transactions over Fabric for essentially all tested configurations. For vanilla Fabric, we can observe that under configurations accessing many accounts (RW=8), the number of failed transactions per second is actually significantly higher.
than the throughput of successful transactions. This problem is highly reduced by Fabric++, where the successful throughput is at least on par with the failed throughput, or even dominates it. The largest improvement of Fabric++ over Fabric in terms of successful transactions we observe is around factor 3x for the configuration BS=1024, RW=8, HR=40%, HW=10%, HSS=1%, which we also show in a zoomed-in version in Figure 14 of the appendix. We also observe a significant decrease in the throughput of the successful transactions with the increase in the hotness of the transactions. For large block-sizes (BS ∈ {512, 1024}), each block (bi) roughly updates every key in the hotset and a large fraction of coldset. This forces most of the transactions in block (bi+1) to abort because of read-write conflicts. So, we observe a pattern of blocks committing with alternating highly-successful and highly-failed transactions. In Fabric, most of the transactions are aborted due to this inter-block conflicts. In addition to this, due to a large block size, Fabric creates a large amount of within-block conflicts, which results in a large fraction of the total number of processed transactions to abort. In Fabric++, we observe a similar alternating behavior in terms of cross-block conflicts. However, since Fabric++ reorders the transactions within the block to remove the within-block conflicts, the number of successful transactions remain on-par with the number of failed transactions. We observe that the strength of Fabric++ lies in contended workload, where the hotness has temporal behavior. If, due to temporal behavior, hot reads, and updates end up in a same block, Fabric++ can possibly optimize the order of transactions to extract a largest set of transactions that have a chance to commit. In contrast to Fabric++, Fabric will behave similarly for temporal and non-temporal hotness in the workloads, forcing a large fraction of transactions to abort, even though they could commit.

Apart from the overall comparison of Fabric and Fabric++, we can analyze the influence of the parameters on the system. A larger block size generally results in a higher throughput. In the case of Fabric++, a larger block size also increases the reordering possibilities of our mechanism. Besides, we can see that a higher number of accesses per transactions results in more failed transactions.

### 5.4 Optimization Breakdown

In Section 5.3, we measured the throughput of Fabric++ with both optimizations activated. Let us now see at a sample configuration, how much the individual optimizations of reordering and early abort contribute to the improvement. Figure 12 shows the improvement breakdown for the configuration BS=1024, RW=8, HR=40%, HW=10%, HSS=1% in comparison to standard Fabric. While Fabric achieves only a throughput of around 100 successful transactions per second, activating one of our two optimization techniques alone improves this to around 150 transactions per second. In comparison to that, activating both techniques at the same time results in the highest throughput of successful transactions with around 220 transactions per second. This shows nicely how both techniques work together: Transactions, that are already early aborted in the simulation phase do not end up in a block in the ordering phase. As a consequence, only transactions, that have a realistic chance of being successful, are considered in the reordering process.

### 5.5 Scaling Channels and Clients

So far, in all experiments we used four clients to fire transactions into a single channel. Let us now vary the number of channels as well as the number of clients per channel to see the effect on the
throughput. Again, we use the configuration BS=1024, RW=8, HR=40%, HW=10%, HSS=1% and evaluate the average throughput of successful transactions for Fabric and Fabric++.

First, we vary the number of channels in Figure 13(a) from 1 to 8. Per channel, we use 2 clients to fire transaction proposals. We can see that when going from 1 channel to 4 channels, the throughput of both Fabric and Fabric++ significantly increases. Obviously, the additional mechanisms of Fabric++ do not harm the scaling with the number of channels. Only when using 8 channels, the throughput decreases again for both Fabric and Fabric++. This is simply the case because individual channels start competing for resources. This also increases the number of failed transactions: Scaling from 1 to 8 channels increases the number of failed transactions from 213 TPS to 837 TPS for Fabric and from 81 TPS to 704 TPS for Fabric++. Due to the competition for resources, individual simulations phase take longer and increase the chance of working on stale data.

(a) Varying the number of channels from 1 to 8. (b) Varying the number of clients per channel from 1 to 8. All clients fire their transaction proposals in a single channel.

Figure 13: The impact of the number of channels as well as the number of clients per channel on the throughput of successful transactions for the configuration BS=1024, RW=8, HR=40%, HW=10%, HSS=1%.
After varying the number of channels, let us now vary the number of clients per channel in Figure 13(b). We test 1, 2, 4, and 8 clients, where all clients fire their transaction proposals into a single channel. Here, the picture is a slightly different to the behavior when scaling channels. The throughput of Fabric increases very gently with the number of clients, and we see an improvement from around 60 to 105 successful transactions per seconds when going from 1 to 8 clients. For Fabric++, we see the highest throughput with around 205 successful transactions per second already for 2 clients. For 8 clients, the throughput drops by around factor 2 to the throughput of Fabric, clearly showing that the firing clients also compete for resources. This is also visible in an increase in failed transactions when going from 1 to 8 clients per channel, which increase from 86 TPS to 928 TPS for Fabric and from 20 TPS to 841 TPS for Fabric++.

6 Conclusion

In this work, we identified strong similarities of the transaction pipeline of contemporary blockchain systems at the case of Hyperledger Fabric and distributed database systems in general. We analyzed these similarities in detail and exploited them to transition mature techniques from the context of database systems to Fabric, namely transaction reordering to remove serialization conflicts as well as early abort of transactions, that have no chance to commit. In an extended experimental evaluation, where we tested 108 different configurations of workload and system, we showed that this improved version Fabric++ outperforms the vanilla Fabric in terms of throughput of successful transactions by up to factor 3x, while keeping the scaling capabilities intact.

Acknowledgments

We would like to thank Immanuel Haffner for helping us in setting up the Fabric cluster, running benchmarks, as well as profiling the internals of the system.

References

[1] https://github.com/brianfrankcooper/ycsb.
[2] https://github.com/hyperledger/caliper.
[3] https://github.com/persistentsystems/gauge.
[4] https://github.com/persistentsystems/gauge/blob/master/docs/caliper-changes.md.
[5] TPC-C On-Line Transaction Processing Benchmark http://www.tpc.org/tpcc/.
[6] TPC-H Decision Support Benchmark http://www.tpc.org/tpch/.
[7] E. Androulaki, A. Barger, V. Bortnikov, et al. Hyperledger fabric: a distributed operating system for permissioned blockchains. In EuroSys 2018, Porto, Portugal, April 23-26, pages 30:1–30:15, 2018.
[8] M. Castro and B. Liskov. Practical byzantine fault tolerance. In Third USENIX Symposium on Operating Systems Design and Implementation (OSDI), New Orleans, Louisiana, USA, February 22-25, pages 173–186, 1999.
[9] T. T. A. Dinh, R. Liu, M. Zhang, G. Chen, B. C. Ooi, and J. Wang. Untangling blockchain: A data processing view of blockchain systems. *IEEE Trans. Knowl. Data Eng.*, 30(7):1366–1385, 2018.

[10] Z. He and B. Hong. Impact of early abort mechanisms on lock-based software transactional memory. In *16th International Conference on High Performance Computing, HiPC 2009, December 16-19, 2009, Kochi, India, Proceedings*, pages 225–234, 2009.

[11] D. B. Johnson. Finding all the elementary circuits of a directed graph. *SIAM J. Comput.*, 4(1):77–84, 1975.

[12] P. Larson, S. Blanas, C. Diaconu, C. Freedman, J. M. Patel, and M. Zwilling. High-performance concurrency control mechanisms for main-memory databases. *PVLDB*, 5(4):298–309, 2011.

[13] G. Luo, J. F. Naughton, C. J. Ellmann, and M. Watzke. Transaction reordering. *Data Knowl. Eng.*, 69(1):29–49, 2010.

[14] T. Neumann, T. Mühlbauer, and A. Kemper. Fast serializable multi-version concurrency control for main-memory database systems. In *SIGMOD 2015, Melbourne, Victoria, Australia, May 31 - June 4*, pages 677–689, 2015.

[15] A. Sharma, F. M. Schuhknecht, and J. Dittrich. Accelerating analytical processing in MVCC using fine-granular high-frequency virtual snapshotting. In *SIGMOD 2018, Houston, TX, USA, June 10-15, 2018*, pages 245–258, 2018.

[16] R. E. Tarjan. Depth-first search and linear graph algorithms. *SIAM J. Comput.*, 1(2):146–160, 1972.

[17] T. Wang, R. Johnson, A. Fekete, and I. Pandis. Efficiently making (almost) any concurrency control mechanism serializable. *The VLDB Journal*, 26(4):537–562, Aug 2017.

[18] T. Wang and H. Kimura. Mostly-optimistic concurrency control for highly contended dynamic workloads on a thousand cores. *PVLDB*, 10(2):49–60, 2016.

[19] X. Yu, G. Bezerra, A. Pavlo, S. Devadas, and M. Stonebraker. Staring into the abyss: An evaluation of concurrency control with one thousand cores. *PVLDB*, 8(3):209–220, 2014.

[20] B. Zhang, B. Ravindran, and R. Palmieri. Reducing aborts in distributed transactional systems through dependency detection. In *Proceedings of the 2015 International Conference on Distributed Computing and Networking, ICDCN 2015, Goa, India, January 4-7, 2015*, pages 13:1–13:10, 2015.

[21] N. Zhou, X. Zhou, X. Zhang, et al. Reordering transaction execution to boost high-frequency trading applications. *Data Science and Engineering*, 2(4):301–315, 2017.

A Throughput Timeline

Figure 14 presents the detailed zoom-in of the run for configuration BS=1024, RW=8, HR=40%, HW=10%, HSS=1% for Fabric (Figure 14(a)) and Fabric++ (Figure 14(b)). We can see that
the throughput remains very stable over the run of 90 seconds. In the beginning, there is a small ramp-up phase visible, which is actually very interesting. For Fabric, the throughput of successful transactions directly starts very low with only 100 transactions per second. In contrast to that, for Fabric++, the initial throughput of successful transactions almost reaches 500 transactions per second with the number of failed transactions per second being 0. This shows that for the first block, our reordering mechanism manages to completely resolve all intra-block conflicts. After that, inter-block conflicts can arise which increase the number of failed transactions in any case.

![Fabric](a) Fabric

![Fabric++](b) Fabric++

Figure 14: Zoom-in of Figure [11] for the configuration BS=1024, RW=8, HR=40%, HW=10%, HSS=1%.

B Ordering Service Micro-Benchmark 2: Vary the length of cycles

In the following experiment, we want to analyze the impact of cycles on the arrival order and on our reordering mechanism. To do so, we again form a sequence of \( n \) transactions, that contains \( n/t \) cycles of size \( t \) transactions of the form

\[ T[r(k_0), w(k_0)], T[r(k_0), w(k_1)], T[r(k_1), w(k_2)], T[r(k_2), w(k_0)] \]

Again, we want to identify how many transactions are valid under the arrival order and when using our reordering mechanism. Figure 15 shows the results for 1024 transactions. For the arrival order, only half of transactions are valid, no matter of the cycle length. This is because aborting every second transaction breaks the cycles. In comparison to that, our reordering mechanism is able to achieve a high number of valid transactions, if the cycles are sufficiently long respectively, there are not too many cycles to cancel. Of course, our algorithm becomes more expensive with the length of the cycles to break. However, since extremely long cycles are very unlikely to occur in reality, the runtime of our mechanism will in general remain low in the ordering phase, as we will see in the full fledged evaluation later on.
Figure 15: Workload 2: Varying the length of the cycles.

C Extended Throughput Evaluation

Additionally to 10,000 account balances, as used in the previous experimental evaluation, we test Fabric and Fabric++ as well under 20,000 account balances and 2 read and write accesses per transaction (RW=2). We can observe that for RW=2, the number of successful transactions is significantly higher than the number of failed transactions due to less conflicts than for RW=4 or RW=8.

Figure 16: Transactional Throughput of Fabric and Fabric++ under 54 different configurations, where the number of account balances is 20,000.