Efficient Synchronization of Recursively Partitionable Data Structures

Technische Universität Berlin

Aljoscha Meyer

May 4, 2021

Efficient Synchronization of Recursively Partitionable Data Structures

Abstract

Given two nodes in a distributed system, each of them holding a data structure, one or both of them might need to update their local replica based on the data available at the other node. An efficient solution should avoid redundantly sending data to a node which already holds it.

We give conceptually simple yet asymptotically efficient probabilistic solutions based on recursively exchanging fingerprints for data structures of exponentially decreasing size, obtained by recursively partitioning the data structures. We apply the technique to sets, maps and radix trees. For data structures containing n items, this leads to $\mathcal{O}(log(n))$ round-trips. We give a scheme by which the fingerprints can be computed in $\mathcal{O}(log(n))$ time, based on an auxiliary data structure which requires $\mathcal{O}(n)$ space and which can be updated to reflect changes to the underlying data structure in $\mathcal{O}(log(n))$ time.

To minimize the number of round-trips, the technique requires up to $\mathcal{O}(n)$ space per synchronization session. It can be adapted to require only a bounded amount of memory, which is essential for robust, scalable implementations. While this increases the worst-case number of round-trips, it guarantees continuous progress, even in adversarial environments.

Contents

1 Introduction **4**
- 1.1 Motivating Examples . . . 4
- 1.2 Efficiency Criteria . . . 5
- 1.3 Recursively Comparing Fingerprints . . . 6
- 1.4 Thesis Outline . . . 6

2 Computing Fingerprints **8**

3 Set Reconciliation **10**
- 3.1 Recursive Set Reconciliation . . . 10
- 3.2 Proof of Correctness . . . 11
- 3.3 Complexity Analysis . . . 11
  - 3.3.1 Preliminary Observations . . . 12
  - 3.3.2 Communication Rounds . . . 12
  - 3.3.3 Communication Complexity . . . 13
  - 3.3.4 Computational Complexity . . . 13
- 3.4 Generic Optimizations . . . 14
  - 3.4.1 Detecting Singletons . . . 14
  - 3.4.2 Encodings . . . 14
  - 3.4.3 Branching Degree . . . 14
- 3.5 Reconciling Hash Graphs . . . 15

4 Bounded Memory Set Reconciliation **16**

5 Other Data Structures **17**
- 5.1 Higher-Dimensional Intervals . . . 17
- 5.2 Maps . . . 17
- 5.3 Tries . . . 17
- 5.4 Sequences . . . 17

6 Related Work **18**

7 Conclusion **19**

Chapter 1

Introduction

One of the problems that needs to be solved when designing a distributed system is how to efficiently synchronize data between nodes. Two nodes may each hold a particular set of data, and may then wish to exchange the ideally minimum amount of information until they both reach the same state. Typical ways in which this can happen are one node taking on the state of the other, or both nodes ending up with the union of information available between the two of them.

1.1 Motivating Examples

Distributed version control systems can be seen as an example of the latter case: users independently create new objects which describe changes to a directory, and when connecting to each other, they both fetch all new updates from the other in order to obtain a (more) complete version history. Regarded more abstractly, the two nodes compute the union of the sets of objects they store. A version control system might attempt to leverage structured information about those objects, such as a happened-before relation, but this does not lead to good worst-case guarantees. The set reconciliation protocol we give guarantees the exchange to only take a number of rounds logarithmic in the number of objects.

A different example are peer-to-peer publish-subscribe systems such as Secure Scuttlebutt [TLMT19]. A node in the system can subscribe to any number of topics, and nodes continuously synchronize all topics they share with other nodes they encounter on a randomized overlay network. Scuttlebutt achieves efficient synchronization by enforcing a linear happened-before relation between messages published to the same topic, i.e. each message is assigned a unique sequence number that is one greater than the sequence number of the previous message. When two nodes share interest in a topic, they exchange the greatest sequence number they have for this topic, whichever node sent the greater one then knows which messages the other is missing.

The price to pay for this efficient protocol is that concurrent publishing of new messages to the same topic is forbidden, since it would lead to different messages with the same sequence number, breaking correctness of the synchronization procedure. An unordered pubsub mechanism based on set reconciliation would be able to support concurrent publishing, the other design aspects such as the overlay network could be left unchanged.

An example from a less decentralized setting are incremental software updates.

A server might host a new version of an operating system, users running an old version want to efficiently download the changes. An almost identical problem is that of efficiently creating a backup of a file system to a server already holding an older backup. Both of these examples can abstractly be regarded as updating a map from file paths to file contents. Our protocol for mirroring maps could be used for determining which files need to be updated. The protocol allows synchronizing the actual files via an arbitrary nested synchronization protocol, e.g. rsync [TM+96].

1.2 Efficiency Criteria

There are a variety of criteria by which to evaluate a synchronization protocol. We exemplify them by the trivial synchronization protocol, which consists of both node immediately sending all their data to the other node.

Let n be the number of items held by node $\mathcal{A}$, and m the number of items held by node $\mathcal{B}$. To simplify things we assume for now that all items have the same size in bytes.

The most obvious efficiency criteria are the *total bandwidth* ($\mathcal{O}(n + m)$ bytes for the trivial protocol) and the number of *round-trips* ($1 \in \mathcal{O}(1)$ for the trivial protocol).

Efficiently using the network is not everything, the computational *time complexity per round-trip* must be feasible so that computers can actually run the protocol. It is lower-bounded by the amount of bytes sent in a given round, for the trivial protocol it is $\mathcal{O}(n + m)$.

Similarly, the *space complexity per round-trip* plays a relevant role, since computers have only a limited amount of memory. In particular, if an adversarial node can make a node run out of memory, the protocol can only be run in trusted environments. Even then, when non-malicious nodes of vastly differing computational capabilities interact (e.g. a microcontroller connecting to a server farm), "accidental denial of service attacks" can easily occur. Since the trivial protocol does not perform any actual computation, its space complexity per round-trip is in $\mathcal{O}(1)$.

The *space complexity per session* measures the amount of state nodes need to store across an entire synchronization session, in particular while idly waiting for a response.

In addition to the space required for per-round-trip computations and per session, an implementation of a protocol might need to store auxiliary information that is kept in sync with the data to be synchronized, in order to achieve sufficiently efficient time and space complexity per round-trip. Of interest is not only the *space for the auxiliary information*, but also its *update complexity* for keeping it synchronized with the underlying data.

A protocol might be asymmetric, with different resource usage for different nodes. If there is client and a clear server role, traditionally protocol designs aim to keep the resource usage of the server as low as possible, motivated by the assumption that many clients might concurrently connect to a single server, but a single client rarely connects to a prohibitive amount of servers at the same time.

Any protocol design has to settle on certain trade-offs between these different criteria, which will make it suitable for certain use cases, but unsuitable for others. We do believe that our designs occupy a useful place in the design space that is applicable to many relevant problems, such as those mentioned in the introduction.

A final, "soft" criterium is that of simplicity. While ultimately time and space complexities should guide adoption decisions, complicated designs are often a good indicator that the protocol will never see any deployment. Our designs require merely comparisons of (sums of) hashes, and the auxiliary data structure that enables efficient implementation is a simple balanced tree.

1.3 Recursively Comparing Fingerprints

We conclude the introduction with a brief sketch of a set reconciliation protocol (i.e. a protocol for computing the union of two sets on different machines) that exemplifies the core ideas. The protocol leverages the fact that sets can be partitioned into a number of smaller subsets. The protocol assumes that the sets contain elements from a universe on which there is a total order based on which intervals can be defined, and that nodes can compute fingerprints for any subset of the universe.

Suppose for example two nodes $\mathcal{A}$, $\mathcal{B}$ each hold a set of natural numbers. They can reconcile all numbers within an interval as follows: $\mathcal{A}$ computes a fingerprint over all the numbers it holds within the interval and then sends this fingerprint to $\mathcal{B}$, together with the interval boundaries. $\mathcal{B}$ then computes the fingerprint over all numbers it holds within that same interval. There are three possible cases:

- $\mathcal{B}$ computed the same fingerprint it received, then the interval has been fully reconciled and the protocol terminates.
- $\mathcal{B}$ has no numbers within the interval, $\mathcal{B}$ then notifies $\mathcal{A}$, $\mathcal{A}$ transmits all its numbers from the interval, and the interval has been fully reconciled.
- Otherwise, $\mathcal{B}$ splits the interval into two sub-intervals, such that $\mathcal{B}$ has a roughly equal number of numbers within each interval. $\mathcal{B}$ then initiates reconciliation for both of these intervals, the roles $\mathcal{A}$ and $\mathcal{B}$ reverse.

Crucially, in the last case, the two recursive protocol invocations can be performed in parallel. The number of parallel sessions increases exponentially, so the original interval is being reconciled in a number of rounds logarithmic in the greater number of items held by any node within that interval.

1.4 Thesis Outline

The remainder of this thesis fleshes out details and applies the same idea to some data structures, all of which share the property that they can be partitioned into smaller instances of the same data structure.

The viability of this approach hinges on the efficient computation of fingerprints, which is discussed and solved in chapter 2. We then give a thorough definition of the set conciliation protocol in chapter 3, and prove its correctness and its complexity guarantees. Chapter 4 gives a more concrete protocol that allows nodes to enforce limits on the amount of computational resources they spend, at the cost of increasing the number of roundtrips if these resource limits are reached. Chapter 5 shows how to apply the same basic ideas to k-d-trees, maps, tries and radix trees (TODO update as this cristallizes), and briefly discusses why it does not make sense to

apply it to arrays. Chapter 6 gives an overview of related work and justifies the chosen approach. We conclude in chapter 7.

Chapter 2

Computing Fingerprints

The protocols described in this thesis work by computing fingerprints of sets. This chapter defines and motivates a specific fingerprinting scheme that admits fast computation with small overhead for the storage and maintenance of auxiliary data structures.

- Merkle trees and why they don't cut it
- hashing into a group and adding things, using search trees for efficient computation
- why not monoids instead of groups
- fingerprint collisions (with and without help from malicious peers)
- miscellaneous (also nice for putting data structures into hash tables, intervals as primitive queries, progress over unreliable links)

TODO: move the following definitions to where they are needed

Definition 1. Let U be a set and $\preceq$ a binary relation on U. We call $\preceq$ a *linear order on* U if it satisfies three properties:

anti-symmetry: for all $x, y \in U$: if $x \preceq y$ and $y \preceq x$ then $x = y$

transitivity: for all $x, y, z \in U$: if $x \preceq y$ and $y \preceq z$ then $x \preceq z$

linearity: for all $x, y \in U$: $x \preceq y$ or $y \preceq x$

If $\preceq$ is a linear order, we write $x \prec y$ to denote that $x \preceq y$ and $x \neq y$.

Definition 2. Let U be a set, $\preceq$ a linear order on U, and $A \subseteq U$. A *binary search tree on* A is a rooted tree T with vertex set A such that for any inner vertex p with left child a and right child b: $a \prec p \prec b$.

Definition 3. Let $T = (V, E)$ be a binary search tree and $\varepsilon \in \mathbb{R}_{>0}$. We call T *ε-balanced* if $\mathit{height}(T) \leq \lceil \varepsilon \cdot \mathit{log}_2(|V|) \rceil$. Since the precise choice of ε will not matter for our complexity analyses, we will usually simply talk about *balanced* trees.

Definition 4. Let U be a set, $\oplus : U \times U \to U$, and $\mathbb{0} \in U$. We call $(U, \oplus, \mathbb{0})$ a *monoid* if it satisfies two properties:

associativity: for all $x, y, z \in U : (x \oplus y) \oplus z = x \oplus y \oplus z$

neutral element: for all $x \in U$:if $\mathbb{0} \oplus x = x = x \oplus \mathbb{0}$.

Definition 5. Let $(U, \oplus, \mathbb{0})$ be a monoid. We call it a *transitive monoid* if for all $x, z \in U$ there exists $y \in U$ such that $x \oplus y = z$.

Definition 6. Let $(U, \oplus, \mathbb{0})$ be a (transitive) monoid. We call it a *(transitive) group* if for all $x \in U$ there exists $y \in U$ such that $x \oplus y = \mathbb{0}$. This y is necessarily unique and denoted by $-x$. For $x, y \in U$ we write $x \ominus y$ as a shorthand for $x \oplus -y$.

Chapter 3

Set Reconciliation

In this chapter, we consider the set reconciliation protocol sketched in the introduction in greater detail. We define an unoptimized but simple version of the protocol in **??**, and we prove its correctness in section 3.2. Section 3.4 lists optimizations which eliminate unnecessary work from the protocol. We then define the proper set reconciliation protocol in **??** and do a complexity analysis in section 3.3. We conclude the chapter with an example application in section 3.5, briefly describing how the protocol can be applied to the synchronization of the hash graphs that arise e.g. in the context of distributed version control systems such as git [CS14].

3.1 Recursive Set Reconciliation

The set reconciliation protocol assumes that there is a set U, a linear order $\preceq$ on U, a node $\mathcal{X}_0$ locally holding some $X_0 \subseteq U$, and a node $\mathcal{X}_1$ locally holding $X_1 \subseteq U$. $\mathcal{X}_0$ and $\mathcal{X}_1$ exchange messages, a message consists of an arbitrary number of *interval fingerprints* and *interval item set*. An interval fingerprint is a triple $(x, y, fp([x, y)_{X_i}))$ for $x, y \in U$, an interval item set a triple (x, y, S) for $x, y \in U, S \subseteq [x, y)_{X_i}$.

Recall that $fp(A)$ denotes the fingerprint for $A \subseteq U$, and that $[x, y)_A := \{a \in A | x \preceq a \prec y\}$.

When a node $\mathcal{X}_i$ receives a message, it performs the following actions:

- For every interval item set (x, y, S) in the message, all items in S are added to the locally stored set X_i. The node then adds the interval item set $(x, y, [x, y)_{X_i} \setminus S)$ to the response, unless $[x, y)_{X_i} \setminus S = \emptyset$.
- For every interval fingerprint $(x, y, fp([x, y)_{X_j}))$ in the message, it does one of following:

  Case 1 (Equal Fingerprints). If $fp([x, y)_{X_j}) = fp([x, y)_{X_i})$, nothing happens.

  Case 2 (Recursion Anchor). The node may add the interval item set $(x, y, [x, y)_{X_i})$ to the response. If $|[x, y)_{X_j}| \leq 1$, it must do so.

  Case 3 (Recurse). Otherwise, the node selects $m_0 = x \prec m_1 \prec \ldots \prec m_k = y \in U$, $k \geq 2$ such that among all $[m_0, m_1)_{X_i}$ for $0 \leq l < k$ at least two intervals are non-empty. For all $0 \leq l < k$ it adds either the interval fingerprint $(m_l, m_{l+1}, fp([m_l, m_{l+1})_{X_i}))$ or the interval item set $(m_l, m_{l+1}, [m_l, m_{l+1})_{X_i})$ to the response.

- If the accumulated response is nonempty, it is sent to the other node. Otherwise, the protocol has terminated successfully.

To initiate reconciliation of an interval $[x, y)$, a node $\mathcal{X}_i$ sends a message containing solely the interval fingerprint $(x, y, fp([x, y)_{X_i}))$.

?? gives an example run of the protocol.

3.2 Proof of Correctness

We now prove the correctness of the protocol. The protocol is correct if for all $x, y \in U$ both nodes eventually hold $[x, y)_{X_i} \cup [x, y)_{X_j}$ after a node $\mathcal{X}_i$ has received a message pertaining to the interval $[x, y)$.

Case 1 (Interval Item Set). If the message contains the interval item set $(x, y, [x, y)_{X_j})$, then $\mathcal{X}_i$ adds all items to its set, resulting in $[x, y)_{X_i} \cup [x, y)_{X_j}$ as desired. The other node then receives $(x, y, [x, y)_{X_i} \setminus (x, y, [x, y)_{X_j}))$, ending up with $[x, y)_{X_j} \cup ((x, y, [x, y)_{X_i} \setminus (x, y, [x, y)_{X_j}))) = [x, y)_{X_i} \cup [x, y)_{X_j}$ as desired.

Case 2 (Interval Fingerprint). Otherwise, the message contains an interval fingerprint $(x, y, fp([x, y)_{X_j}))$.

Case 2.1 (Equal Fingerprints). If $fp([x, y)_{X_j}) = fp([x, y)_{X_i})$, the protocol terminates immediately and no changes are performed by any node. Assuming no fingerprint collision occurred, $[x, y)_{X_i} = [x, y)_{X_j} = [x, y)_{X_i} \cup [x, y)_{X_j}$ as desired.

Case 2.2 (Recursion Anchor). If $\mathcal{X}_i$ adds the interval item set $(x, y, [x, y)_{X_i})$, then case 1 applies when the other node receives the response, with the roles reversed.

Case 2.3 (Recurse). Let $count_i := |[x, y)_{X_i}|$ and $count_j := |[x, y)_{X_j}|$. $count_j \geq 2$, since otherwise $\mathcal{X}_j$ would have sent an item set for the interval. Similarly, $count_i \geq 2$, since we are not in case 2.2. Thus, $count_i + count_j \geq 4$, and the protocol has already been proven correct for all cases where $count_i + count_j < 4$.
We can thus finish the proof by induction on $count_i + count_j$, using the induction hypothesis that for all $x', y' \in U$ such that $|[x', y')_{X_i}| + |[x', y')_{X_j}| < count_i + count_j$ the protocol correctly reconciles $[x', y')_{X_i}$ and $[x', y')_{X_j}$.
$\mathcal{X}_i$ partitions the interval into $k \geq 2$ subintervals, of which at least two must be nonempty. Thus $|[m_l, m_{l+1})_{X_i}| < count_i$ for all $0 \leq l < k$. Furthermore, $[m_l, m_{l+1})_{X_j} \subseteq [x, y)_{X_j}$ and thus $|[m_l, m_{l+1})_{X_j}| \leq |[x, y)_{X_j}|$, so overall we have $|[m_l, m_{l+1})_{X_i}| + |[m_l, m_{l+1})_{X_j}| < count_i + count_j$ and can apply the induction hypothesis to conclude that every subinterval is correctly reconciled. Since the subintervals partition the original interval, the original interval is then correctly reconciled as well.

3.3 Complexity Analysis

The protocol gives nodes the freedom to respond to an interval fingerprint with an interval item set even if the interval fingerprint is arbitrarily large. For a meaningful complexity analysis we need to restrict the behavior of the node, a realistic modus operandi is for a node to send an interval item set whenever it holds a number of items less than or equal to some threshold $t \in \mathbb{N}, t \geq 1$ within the interval. Higher

choices for t reduce the number of roundtrips, but increase the probability that a items is being sent even though the other node already holds it.

A node is similarly given freedom over the number of subintervals into which to split an interval when recursing. We will assume a node always splits into at most $b \in \mathbb{N}, b \geq 2$ subintervals. As with t, higher numbers reduce the number of roundtrips at the cost of potentially sending items or fingerprints that did not need sending.

Because we want to analyze not only the worst-case complexity but also the complexity depending on the similarity between the two sets held by the participating nodes, we define some rather fine-grained instance size parameters: n_0 and n_1 denote the number of items held by $\mathcal{X}_0$ and $\mathcal{X}_1$ respectively. We let $n := n_0+n_1$, $n_{min} := min(n_0, n_1)$, $n_{max} := max(n_0, n_1)$, $n_\cap$, $n_\cup$ and $n_\triangle := |([x, y)_{X_0} \cup [x, y)_{X_1}) \setminus ([x, y)_{X_0} \cap [x, y)_{X_1})|$.

3.3.1 Preliminary Observations

The bit-level message complexity depends on the precise encoding of messages. When partitioning an interval into subintervals, it is not necessary to encode both boundaries of every interval since they overlap. When splitting into b subintervals, it suffices to transmit $b+1$ many boundaries (and of course the corresponding b fingerprints/item sets). If the receiving node remembers which boundaries it has sent in the last round, the two outer boundaries can also be omitted. We will refer to the protocol variant where nodes remember boundaries across communication rounds as the *stateful* protocol variant, and the variant where the outer boundaries of a sequence of subintervals needs to be transmitted as the *stateless* protocol variant.

A helpful observation for the following analysis is that the interval fingerprints that are being exchanged during a protocol run form a rooted tree where every vertex has at most b children. When a leaf of the tree is reached, an exchange of interval item sets follows. Equal fingerprints can also cut the tree short, but for the following worst-case analysis we will assume this does not occur.

Node $\mathcal{X}_i$ can branch at most $\lceil log_b(n_i) \rceil$ times, so the overall height of the tree is bounded by $2 \cdot \lceil log_b(n_{min}) \rceil$. The number of vertices of such a tree complete tree of height h is at most $\sum_{i=0}^{h} b^i = \frac{b^h-1}{b-1}$. For $h \leq 2 \cdot \lceil log_b(n_{min}) \rceil$, $\frac{b^h-1}{b-1} \leq 2 \cdot 2 \cdot n_{min} \leq 2n \in \mathcal{O}(n)$.

The parameter t determines when recursion is cut off, and thus influences the height of the tree. For $t = 1$, the protocol recurses as far as possible. For $t = b$, the last level of recursion is cut off, for $t = b^2$ the last two levels, and so on. Overall, the height of the tree is reduced by $\lfloor log_b(t) \rfloor$.

3.3.2 Communication Rounds

The number of communication rounds clearly corresponds to the height of the tree, plus 2 to account for the exchange of interval item sets, so the worst-case is $2 + 2 \cdot \lceil log_b(n_{min}) \rceil - \lfloor log_b(t) \rfloor \in \mathcal{O}(log_b(n))$. This number cannot be bounded by $n_\triangle$, as witnessed by problem instances where one node is missing exactly one item compared to the other node. In such an instance, $b-1$ branches in each recursion step result in equal fingerprints, but the one branch that does continue reaches the recursion anchor only after the full number of rounds. See fig. 3.1 for an example.

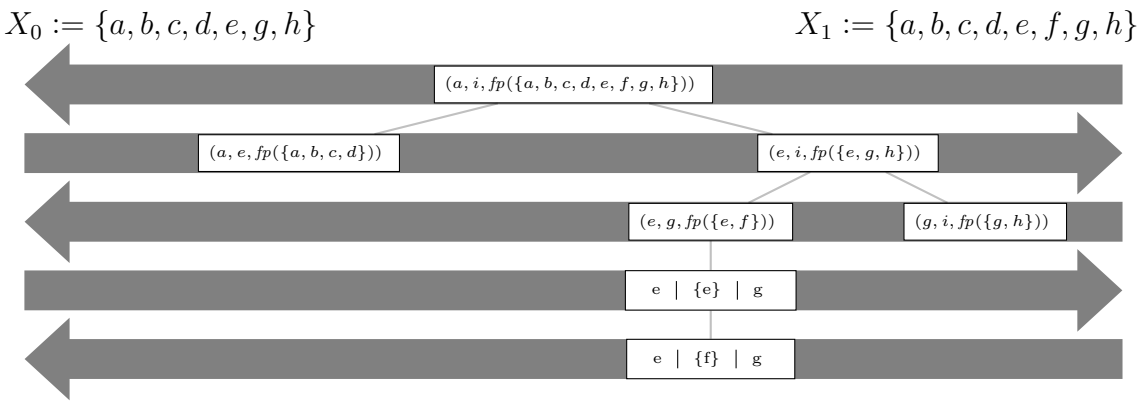

Figure 3.1: An example run of the protocol that takes the greatest possible number of rounds even though $n_{\triangle} = 1$. $b := 2, t := 1$.

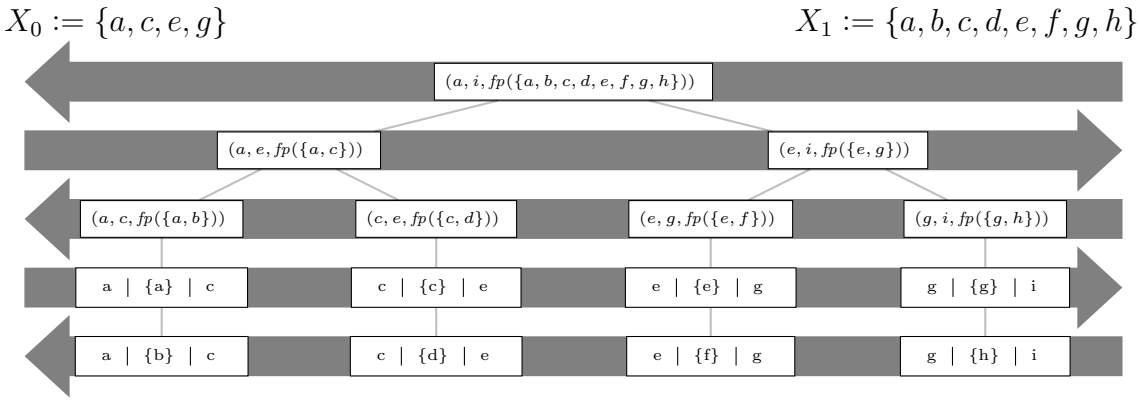

Figure 3.2: An example run of the protocol that requires transmitting the maximum amount of bytes. $b := 2, t := 1$.

3.3.3 Communication Complexity

The total number of bytes that needs to be transmitted during a protocol run is proportional to the number of vertices in the tree. A group of b vertices whose intervals partition the parent vertex require $(b-1) \cdot s_{item}$ bytes to encode the interval boundaries in the stateful protocol variant, $(b+1) \cdot s_{item}$ in the stateless variant. Each interval fingerprint adds another s_{fp} bytes for the fingerprint itself. The exchange of interval item sets consists in the worst case of exchanging every item using $\lceil \frac{n}{t} \rceil$ interval item sets, adding another $4 \cdot s_{item} + t \cdot s_{item}$ bytes in the stateless version (the first summand accounts for the two boundary items being sent by each node), or simply $t \cdot s_{item}$ bytes in the stateful version.

Overall, the complexity is TODO

Figure 3.2 shows a worst-case example, in which the tree of height $h := log_b(2 \cdot n_{min})$ has all $\frac{b^h - 1}{b-1}$ vertices.

3.3.4 Computational Complexity

TODO (write fingerprint chapter first?)

3.4 Generic Optimizations

While the simple formulation of the protocol is nicely suitable for a correctness proof, there are some cases that can be handled much more efficiently. We briefly discuss them and then present an optimized version of the protocol.

3.4.1 Detecting Singletons

When receiving an interval fingerprint $(x, y, fp([x, y)_{X_i}))$, a node can check whether $fp(\{x\}) = fp([x, y)_{X_i})$. If this is the case, $\mathcal{X}_i$ only stores x within the interval, so $\mathcal{X}_j$ can act as if it received the item x and the interval fingerprint $(l, y, \mathbb{0})$ where l is the least element in X_j such that $x \prec l \preceq y$.

3.4.2 Encodings

A first, simple encoding optimization is to encode empty intervals not via $\mathbb{0}$, but as a dedicated message part. This allows a more compact representation, which is appropriate since $\mathbb{0}$ should by far be the most frequently occurring fingerprint. We will designate an *empty interval fingerprint* as (x, y).

A second optimization that merely changes the encoding of messages consists of a more compact representation of "adjacent" intervals: Interval boundaries $[x_0, x_1), [x_1, x_2), \ldots, [x_{k-}$ can be encoded as a simple list $x_0, x_1, \ldots, x_{k-1}, x_k$. As long as the protocol only talks about intervals which partition the original interval, expressing k interval boundaries only consumes $1 + k$ space as opposed to the naïve $2k$.

We can in fact do even better: when a node receives some interval $[x, y)$ and splits it into $[x, m)$ and $[m, y)$, it merely needs to send m to the other node. If the other node can reconstruct which interval is being split at point m, then all the necessary information has been conveyed. Since any m falls into exactly one previously used interval, there can be no ambiguity. It thus suffices to send triples $(\mathit{fingerprint}_{\prec m}, m, \mathit{fingerprint}_{\succeq m})$. As a consequence, any particular $u \in U$ is transmitted at most one time during a protocol run. This optimization does however come at the cost of nodes needing to store interval boundaries across communication rounds.

3.4.3 Branching Degree

For the correctness of the protocol, large, populated intervals need to be partitioned and then recursively reconciled. It does however neither matter to partition the intervals into exactly two subintervals, nor that the intervals are evenly partitioned in the middle. Increasing the number of subintervals per recursion step reduces the total number of roundtrips at the cost of less efficient bandwidth usage, as will be seen in the complexity analysis in section 3.3.

In a similar vein, there is no inherent reason to initiate reconciliation by only sending a single fingerprint for the whole interval of interest, the initiating node and just as well partition the interval of interest and send fingerprints for each subinterval.

Splitting intervals into partitions of unequal sizes can be beneficial if missing items are not expected to be distributed uniformly at random across the whole order. Items might for example be ordered by the time at which they were created,

a long-running node would then expect a continuous stream of new items, but would rarely receive an unknown item from the far past. Interval selection could reflect this by e.g. partitioning the $\frac{1}{2}$ oldest items into the first subinterval, then the next $\frac{1}{4}$ into the second subinterval, the next $\frac{1}{8}$ into the third, and so on. If unknown old items are rare enough, the average size of intervals which require recursion is less than under a split into intervals of equal sizes.

3.5 Reconciling Hash Graphs

Chapter 4

Bounded Memory Set Reconciliation

- bounded memory - the need for backpressure
- credit-based backpressure
- bounded memory set reconciliation - conceptual
- bounded memory set reconciliation - protocol

Chapter 5

Other Data Structures

5.1 Higher-Dimensional Intervals

k-d-trees

5.2 Maps

two reconciliation sessions in parallel, one for the keys and one for the values, but both ordered by the keys

5.3 Tries

optimizing lexicographically ordered items

5.4 Sequences

Why we need content-based slicing, even recursively
    sequences as maps from rational numbers to items

Chapter 6

Related Work

- set reconciliation literature
- hash graph synchronization
- filesystem synchronization
- history-based synchronization

Chapter 7

Conclusion

TODO: conclude things

Work Plan

- by 05.05: basic set reconciliation chapter
- by 26.05: fingerprint chapter
- by 16.06: bounded-memory set reconciliation chapter
- by 07.07: other data structures chapter
- by 28.07: conclusion, coherence, polishing
- 15.08: self-inflicted soft deadline, unless adding more content

Possibly a chapter discussing more specifics that would occur when using set reconciliation as the core of an unordered p2p pubsub mechanism.

Bibliography

[CS14] Scott Chacon and Ben Straub. *Pro git*. Springer Nature, 2014.

[TLMT19] Dominic Tarr, Erick Lavoie, Aljoscha Meyer, and Christian Tschudin. Secure scuttlebutt: An identity-centric protocol for subjective and decentralized applications. In *Proceedings of the 6th ACM Conference on Information-Centric Networking*, pages 1–11, 2019.

[TM+96] Andrew Tridgell, Paul Mackerras, et al. The rsync algorithm. 1996.