From Rational Number Reconstruction to Set Reconciliation

Antoine Amarilli, Fabrice Ben Hamouda, Florian Bourse,
Robin Morisset, David Naccache, and Pablo Rauzy

École normale supérieure, Département d'informatique
45, rue d'Ulm, F-75230, Paris Cedex 05, France.
surname.name@ens.fr (except for fabrice.ben.hamouda@ens.fr)

Penser à voir si on garde ce titre.

~~Si cet abstract~~ vous convient, il faudrait en reprendre des bouts dans l'introduction.

Abstract. This work revisits *set reconciliation*, a problem consisting in synchronizing two multisets of fixed-size values while minimizing the amount of data transmitted. We propose a new number theoretic reconciliation protocol called "Divide & Factor" (D&F) which achieves optimal asymptotic transmission complexity like prior proposals. We then study the problem of synchronizing sets of variable-size files, and describe how constant-factor improvements can be achieved through the use of hashing with a carefully chosen hash size (balancing the quantity of data transferred and the risk of collisions). We show how this process can be applied to synchronize file hierarchies, taking into account the location of files. We describe `btrsync`, our open-source implementation of the protocol, and benchmark it against the popular software `rsync` to demonstrate that `btrsync` uses more CPU time but transmits less data.

Fabrice: j'écrirai quelque chose de plus positif comme: "`btrsync` transmits much less data at the expense of a small computationnal overhead."

~~Est-ce que~~ c'est vrai qu'on apporte quelque chose de nouveau dans le cas où c'est des fichiers de taille non fixe ? — Fabrice: je ne sais pas, il faudrait que l'on réanalyse la complexité propre de la dernière variante.

1 Introduction

Notations

Cette section est temporaire et ne sert que pour assurer des notations cohérentes le long du papier, notamment pour les indices:

- i, j: pour les fichiers/hashés h_i,F_i
- k: pour les rounds avec changement de modulo: $p_k, c_k, s_k, C_k, S_k, t_k, T_k$
- ℓ: pour les rounds liés aux traitements des collisions
- λ: pour une borne sur le k ou le ℓ

TODO: Corriger les n,n',η,η'.
TODO: faire commencer le i à 1 ou 0 partout
TODO: change t_0 because it is quite confusing !

Old Introduction

This work revisits *set reconciliation*, a problem consisting in synchronizing two multisets while minimizing the amount of data transmitted. Set reconciliation arises in many practical

situations, the most typical of which is certainly incremental backups performed over a slow network link.

Several efficient and elegant solutions are known to achieve set reconciliation of multisets containing atomic elements of a fixed size. For instance, [8] manages to perform set reconciliation using a bandwidth which is linear in the size of the symmetric difference of the multisets multiplied by the size of the elements, which is optimal in this setting. We refer the reader to [8,9,5] for more on this problem's history and its existing solutions.

However, in the case where the elements to be synchronized can be very large (e.g., files during a backup), we must use checksums to identify the differing files before transferring them, and the question of the size of the checksum to use is non-trivial. In this article, we propose a new reconciliation protocol called "Divide & Factor" (D&F) based on number theory. In terms of asymptotic transmission complexity, the proposed procedure reaches optimality as well. In addition, the new protocols offer a very interesting gamut of parameter trade-offs. We provide an analysis of the protocol's complexity in terms of transmission and computation, as well as a probabilistic analysis of the possible choices of checksum sizes; we also provide an implementation of the protocol and experimental results.

Fabrice@Antoine: Je ne suis pas tout à fait d'accord avec ce paragraphe, ton abstract est beaucoup mieux...

This paper is structured as follows: Section 2 presents "Divide & Factor", our set reconciliation protocol. Section 3 presents the transmission complexity of the protocol, introduces two transmission optimizations and analyzes them in detail. Section 4 analyzes the computational complexities of the proposed protocols. Section 5 explain how to use a set reconciliation algorithm to perform file synchronization. Section 6 presents our implementation `btrsync` and reports practical experiments and benchmarks against the popular software `rsync`.

New Introduction

File synchronization is the important practical problem consisting of retrieving a hierarchy of files on a remote host given some outdated or incomplete version of this hierarchy of files on the local machine. In many use cases, the bottleneck is bandwidth of the network link between the local machine and the remote host, and care must be taken to limit the quantity of data transferred over the link by using the existing copy of the set of files to the fullest possible extent. Popular file synchronization programs such as `rsync` use rolling checksums to skip remote file parts which match a file part at the same location on the local machine; however, they are usually unable to take advantage of the local files in subtle ways, like detecting that some large file is already present on the local machine but at a different location.

File synchronization is closely linked to the theoretical problem of *set reconciliation*: given two sets of fixed-size data items on different machines, determine the symmetric difference of the two sets while minimizing the amount of data transferred. The lower bound of the quantity of data to transfer is clearly the size of the symmetric difference

(i.e., the number of elements in the difference times the size of these elements), and some known algorithms achieve this bound [8]. We refer the reader to [8,9,5] (to quote a few references) for more on this problem's history and its existing solutions.

In this paper, we look at the problems of set reconciliation and file synchronization from a theoretical and practical perspective. Our contribution are as follows:

- We introduce "Divide & Factor", a new set reconciliation algorithm. D&F is based on number theory: it represents the items to synchronize as prime numbers, accumulates information about the symmetric difference in a series of rounds through the Chinese remainder theorem, and reconstitutes the result through the use of rational number reconstruction. The algorithm is described in Section 2.
- We show that D&F, like existing algorithms, has a transmission complexity which is linear in the size of the symmetric difference (Section 3). We study the computational complexity of D&F and present possible trade-offs between constant-factor transmission complexity and computational complexity through alternative design choices (Section 4).
- We explain how hash functions can be used with any set reconciliation algorithm to synchronize sets of elements which do not have a fixed size, such as file contents. We explain how the size of the hash function should be chosen to achieve a good tradeoff between the quantity of data to transfer and the risks of confusion. TODO have a section for this: also, isn't this D&F-specific?.
- We spell out in Section 5 how the previous construction can be extended to perform file synchronization, taking into account the location and metadata of files and managing situations such as file moves in an intelligent manner. We describe an algorithm to apply a set of move operations on a set of files in-place which avoids excessive use of temporary file locations.
- We present `btrsync`, our implementation of file synchronization through D&F, in Section 6, and benchmark it against `rsync`. The results show that `btrsync` has a higher computational complexity but transmits less data in most scenarios.

2 "Divide & Factor" Set Reconciliation

Section 2.2 presents a basic version of the proposed protocol. This basic version suffers from two limitations: it works only if the number of differences to reconcile is bound and it may fail leave the synchronized party in an erroneous state. Failure avoidance is overcome in section 2.3 and an extension to arbitrary numbers of differences is given in section 2.4.

2.1 Problem Definition and Notations

Oscar possesses an old version of a directory $\mathfrak{D}$ that he wishes to update. Neil has the new, up-to-date version $\mathfrak{D}'$: $\mathfrak{D}$ and $\mathfrak{D}'$ can differ both in their files and in their tree structures.

Oscar wishes to obtain $\mathfrak{D}'$ but *exchange as little data as possible* during the synchronization process.

To tackle this problem we separate the "*what*" from the "*where*" by considering files as a tuple of their location and content. In other words, we will first synchronize all the file contents and then move files to the adequate location. We consider that $\mathfrak{D}$ is a multiset of files which we denote as $\mathfrak{F} = \{F_0, \ldots, F_n\}$, and likewise represent $\mathfrak{D}'$ as $\mathfrak{F}' = \{F'_0, \ldots, F'_{n'}\}$.

Fabrice: NON, ce n'est pas vrai: on considère vraiment qu'un fichier est "path + content"

Let t_0 be the number of discrepancies between $\mathfrak{F}$ and $\mathfrak{F}'$ that Oscar wishes to learn, *i.e.* the symmetric difference of $\mathfrak{F}$ and $\mathfrak{F}'$:

$$t_0 = \#\mathfrak{F} + \#\mathfrak{F}' - 2\#\left(\mathfrak{F} \bigcap \mathfrak{F}'\right) = \#\left(\mathfrak{F} \bigcup \mathfrak{F}'\right) - \#\left(\mathfrak{F} \bigcap \mathfrak{F}'\right)$$

Fabrice: Qu'est-ce que $\mathfrak{D}$? Est-ce $\mathfrak{F}$? Dans tous les cas, il s'agit d'un set, vu la représentation indiquée

Given a file F, we denote by $\texttt{Hash}(F)$ its image by a collision-resistant hash function such as SHA-1. Let $\texttt{HashPrime}(F)$[1] be a function hashing files (uniformly) into primes smaller than 2^u for some $u \in \mathbb{N}$. Define the shorthand notations: $h_i = \texttt{HashPrime}(F_i)$ and $h'_i = \texttt{HashPrime}(F'_i)$.

Fabrice: `Hash` can be introduced latter when needed (section 2.3)

2.2 Description of the Basic Exchanges

The number of differences t_0 is unknown to Oscar and Neil. However, for the time being, we will assume that t_0 is smaller than some t and attempt to perform synchronization. If $t_0 \leq t$, synchronization will succeed; if $t_0 > t$ the parties will transmit more information later to complete the synchronization, as explained in section 2.4.

We generate a prime p such that:

$$2^{2ut} \leq p < 2^{2ut+1} \tag{1}$$

Given $\mathfrak{F}$, Oscar generates and sends to Neil the redundancy:

$$c = \prod_{F_i \in \mathfrak{F}} \texttt{HashPrime}(F_i) = \prod_{i=1}^{n} h_i \bmod p$$

Neil computes:

$$c' = \prod_{F'_i \in \mathfrak{F}'} \texttt{HashPrime}(F'_i) = \prod_{i=1}^{n'} h'_i \bmod p \quad \text{and} \quad s = \frac{c'}{c} \bmod p$$

Fabrice: Quel est l'intérêt de cette citation ???

Using [13] the integer s can be written as:

[1] The design of `HashPrime` is addressed in Appendix C.

$$s = \frac{a}{b} \bmod p \text{ where the } G_i \text{ denote files and } \begin{cases} a = \prod_{G_i \in \mathfrak{F}' \wedge G_i \notin \mathfrak{F}} \texttt{HashPrime}(G_i) \\ b = \prod_{G_i \notin \mathfrak{F}' \wedge G_i \in \mathfrak{F}} \texttt{HashPrime}(G_i) \end{cases}$$

Fabrice: je ne comprends pas le "where the G_i denote files..."

Note that if our assumption $t_0 \leq t$ is correct, $\mathfrak{F}$ and $\mathfrak{F}'$ differ by at most t elements and a and b are strictly less than 2^{ut}. The problem of recovering a and b from s efficiently is known as *Rational Number Reconstruction* [10,14]. theorem 1 (see [6]) guarantees that it can be solved in this setting. The following theorem is a slightly modified version of Theorem 1 in [6]:

Theorem 1. *Let $a, b \in \mathbb{Z}$ two co-prime integers such that $0 \leq a \leq A$ and $0 < b \leq B$. Let $p > 2AB$ be a prime and $s = ab^{-1} \bmod p$. Then a, b are uniquely defined given s and p, and can be recovered from A, B, s, p in polynomial time.*

Taking $A = B = 2^{ut} - 1$, Equation (1) implies that $AB < p$. Moreover, $0 \leq a \leq A$ and $0 < b \leq B$. Thus Oscar can recover a and b from s in polynomial time: a possible option is to use Gauss algorithm for finding the shortest vector in a bi-dimensional lattice [13]. By testing the divisibility of a and b by the h_i and the h'_i, Neil and Oscar can attempt to identify the discrepancies between $\mathfrak{F}$ and $\mathfrak{F}'$ and settle them.

Fabrice: certes, mais on utilise directement un Euclide étendu tronqué. Et pourquoi citer Vallée qui est un peu incompréhensible dans notre cas... TODO: check that in our program we ensure a and b co-prime !! otherwise, it may fail !!!!!

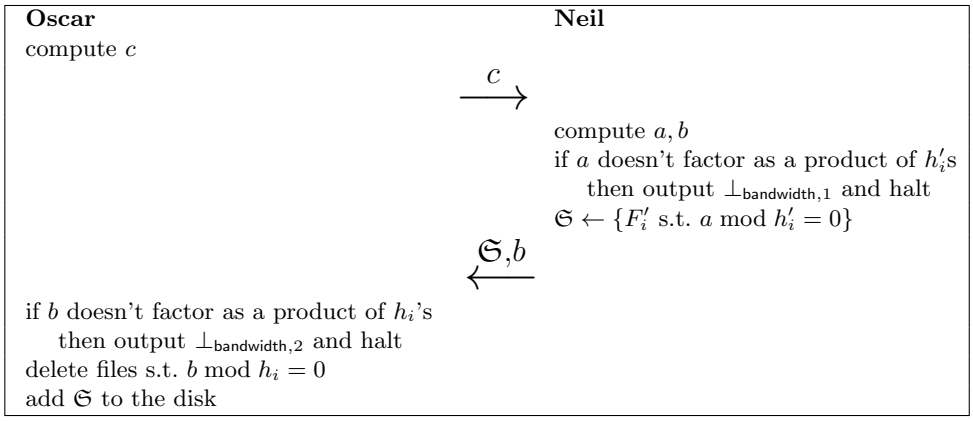

Fig. 1. Basic Protocol.

The formal description of the protocol is given in Figure 1. The "output $\perp_{\mathsf{bandwidth},\square}$ " protocol interruptions will:

– never occur if the assumption $t_0 \leq t$ holds.

– occur with high probability if $t_0 > t$. Indeed, for a potential $\perp_{\mathsf{bandwidth},1}$ to be overlooked, the ut-bit number a must perfectly factor over a set of n primes of size u. If we assume that a is "random", the probability γ that a is divisible by some h_i is essentially $\gamma \sim 1/h_i \sim 2^{-u}$, the probability that a is divisible by exactly t digests is:

$$\alpha = \binom{n}{t}\gamma^t(1-\gamma)^{n-t} \sim \binom{n}{t}2^{-ut}(1-2^{-u})^{n-t}$$

and the probability that the protocol does not terminate by a $\perp_{\mathsf{bandwidth},\square}$ when $t_0 > t$ is $\sim \alpha^2$.

However, we are not interested in the fact a is divisible by "exactly" t digests but the fact that a can be factorized over of the basis of h_i...

The very existence of $\perp_{\mathsf{bandwidth},\square}$'s is annoying for two reasons:

– A file synchronization procedure that works *only* for a limited number of differences is not really useful in practice. Thus, section 2.4 explains how to extend the protocol to perform the synchronization even when the number of differences t_0 exceeds the initial estimation t.
– If, by sheer bad luck, both $\perp_{\mathsf{bandwidth},\square}$'s went undetected (double accidental factorization) the Basic Protocol (Fig. 1) may leave Oscar in an inconsistent state.

Double accidental factorization is not only possible source of inconsistent states: as we did not specifically require `HashPrime` to be collision-resistant, the events

$$\perp_{\mathsf{collision},1} = \begin{cases} h'_i = h'_j \text{ for } i \neq j \\ \\ a \bmod h_i = 0 \end{cases} \quad \text{and/or} \quad \perp_{\mathsf{collision},2} = \begin{cases} h_i = h_j \text{ for } i \neq j \\ \\ b \bmod h'_i = 0 \end{cases}$$

will cause Neil to send wrong files in $\mathfrak{S}$ ($\perp_{\mathsf{collision},1}$) and/or have Oscar unduely delete files owned by Neil ($\perp_{\mathsf{collision},2}$).

Inconsistent states may hence stem from three events:

- accidental double factorization of a and/or b when $t_0 > t$ (probability α^2)
- $\perp_{\mathsf{collision},1}$ = collisions within the set $\{h'_i\}$
- $\perp_{\mathsf{collision},2}$ = collisions within the set $\{h_i\}$

Fabrice: ce qui est un peu tordu, c'est que les collisions qui nous intéressent sont les collisions entre $\mathfrak{S}$ et $\mathfrak{F} \cup \mathfrak{F}'$... il faut qu'on en discute.

Section 2.3 explains how protect the protocol from all inconsistent events at once.

2.3 Avoiding Inconsistency

The Basic Protocol of Figure 1 is fully deterministic. Hence if any sort of trouble occurs, repeating the protocol will be of no help. We modify the protocol as follows:

– Let $H \leftarrow \mathtt{Hash}(\mathfrak{F}')$
– Replace $\mathtt{HashPrime}(F)$ by a diversified $\hbar_k(F) = \mathtt{HashPrime}(k|F)$.
– Define the shorthand notations: $\hbar_{k,i} = \hbar_k(F_i)$ and $\hbar'_{k,i} = \hbar_k(F'_i)$.
– Let $\mathsf{StepProtocol}(k)$ denote the sub-protocol shown in Figure 2.
– Use the protocol of Figure 3 as a fully functional reconciliation protocol for $t_0 \leq t$.

Fabrice: This notation (hash of a set) is not defined... and it may be useful to recall the definition of `Hash` *here (collision-resistant) if the reader has forgotten it...*

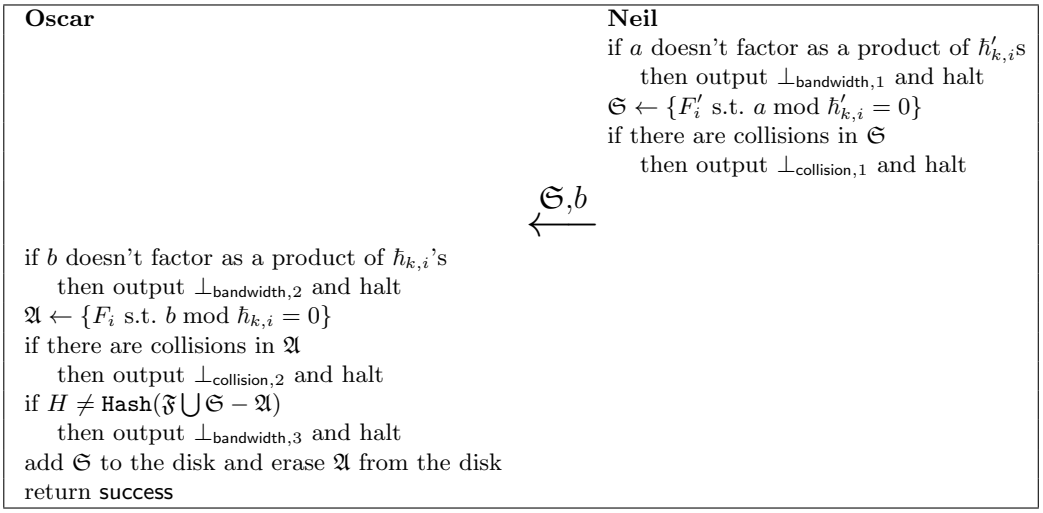

Fig. 2. StepProtocol(k).

Note: To avoid transmitting the (potentially very voluminous) $\mathfrak{S}$ during StepProtocol before knowing if one of the errors $\perp_{\mathsf{bandwidth},2}$, $\perp_{\mathsf{bandwidth},3}$, $\perp_{\mathsf{collision},2}$ will occur, Neil may transmit

$$\mathfrak{S}' = \{\mathtt{Hash}(F'_i),\ F'_i \in \mathfrak{S}\}$$

instead of $\mathfrak{S}$ and send $\mathfrak{S}$ only after successfully passing the $\perp_{\mathsf{bandwidth},3}$ test. The definition of H must be changed accordingly to

$$H = \mathtt{Hash}(\{\mathtt{Hash}(F'_i),\ F'_i \in \mathfrak{F}'\})$$

Fabrice: en même temps, le Hash d'un ensemble, que l'on n'a pas définit, a tout intérêt à déjà être de cette forme, sinon, on a des problèmes pour avoir une concaténation "propre"

2.4 Handling a High Number of Differences

To extend the protocol to an arbitrary t_0, Oscar and Neil agree on an infinite set of primes $p_1, p_2, \ldots$ As long as the protocol fails with a $\perp_{\mathsf{bandwidth},\square}$ status, Neil and Oscar redo the protocol with a new p_ℓ and Neil will keep accumulating information about the difference between $\mathfrak{F}$ and $\mathfrak{F}'$ as shown in Appendix A. Each of this repetition is called a round. Note that no information is lost and that the transmitted modular knowledge about the difference adds up until it reaches a threshold sufficient to reconcile $\mathfrak{F}$ and $\mathfrak{F}'$.

More precisely, let us suppose $2^{2ut_k} \leq p_\ell < 2^{2ut_k+1}$. Let us write $P_k = p_1 \ldots p_k$ and $T_k = u(t_1 + \ldots t_k)$. After receiving the redundancies $c_1, \ldots, c_k$ corresponding to $p_1, \ldots, p_k$, Neil has as many information as if Oscar had transmitted a redundancy C_k corresponding to the modulo P_k, and can compute $S_k = C'_k/C_k$ from $s_k = c'_k/c_k$ and S_{k-1} using the CRT

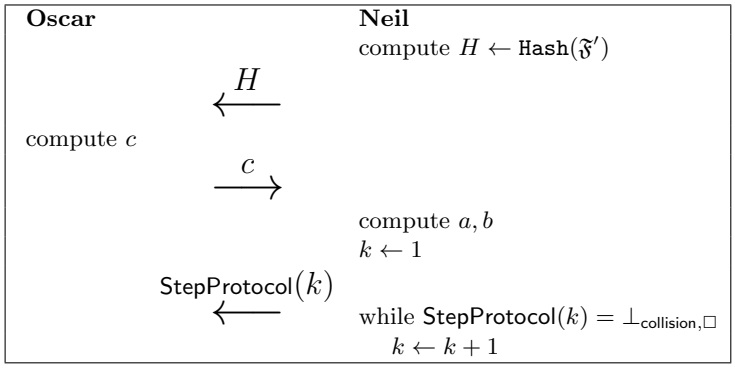

Fig. 3. Fully Functional Protocol for $t_0 \leq t$.

(TODO ref ?). Therefore, the number λ of rounds used is the minimum number k such that $T_k \geq t_0$. If $t_1 = t_2 = \cdots = t$, then $\lambda = \lceil t/t_0 \rceil$.

All $\perp$ treatments were removed from Appendix A for the sake of clarity (these can be very easily added by modifying Appendix A *mutatis mutandis*). In essence, the rules are: add information modulo a new p_ℓ whenever the protocol fails with a $\perp_{\mathsf{bandwidth},\square}$ and increment k whenever the protocol fails with a $\perp_{\mathsf{collision},\square}$.

Fabrice: plus clair si on le met dans une deuxième figure en appendix quand même, je pense...

A typical execution sequence is thus expected to be something like:

$$\perp_{\mathsf{bandwidth},1}, \perp_{\mathsf{bandwidth},1}, \perp_{\mathsf{bandwidth},1}, \perp_{\mathsf{bandwidth},1}, \perp_{\mathsf{collision},1}, \perp_{\mathsf{collision},1}, \mathsf{success}$$

3 Transmission Complexity

This section explores two strategies for reducing the size of p and hence improving transmission by *constant factors* (from an asymptotic communication standpoint, improvements cannot be expected as the protocol already transmits information proportional to t_0, the difference to settle). Excluding the core information $\mathfrak{S}$ and assuming that no $\perp_{\mathsf{collision},\square}$ events occurred, the transmission complexity of the protocol of Appendix A is:

$$\lambda \log(\max_{k=1}^{\lambda} c_k) + \log b \;\leq\; \lambda \log(\max_{k=1}^{\lambda} p_k) + \frac{1}{2} \log \prod_{k=1}^{\lambda} p_k \;\leq\; 3\lambda(ut_0 + 1) = O(\lambda u t_0) = O(ut),$$

As we have no control over t, decreasing u is the main natural optimization option. We will get back to this later on in this paper (section 3.2).

Fabrice: maybe deal with the case when doubling ? and precise that we deal with constant t ???

3.1 Probabilistic Decoding: Reducing p

Generate a prime p about twice shorter than the p recommended in section 2.2, namely:

$$2^{ut} < p \leq 2^{ut+1} \tag{2}$$

Let $\eta = \max(n, n')$. The new redundancy c is calculated as previously and is hence also approximately twice smaller. Namely:

$$s = \frac{a}{b} \bmod p \text{ and } \begin{cases} a = \prod_{G_i \in \mathfrak{F}' \wedge G_i \notin \mathfrak{F}} \texttt{HashPrime}(G_i) \\ \\ b = \prod_{G_i \notin \mathfrak{F}' \wedge G_i \in \mathfrak{F}} \texttt{HashPrime}(G_i) \end{cases}$$

and since there are at most t differences, we must have:

$$ab \leq 2^{ut} \tag{3}$$

By opposition to section 2.2 we do not have a fixed bound for a and b anymore; Equation (3) only provides a bound for the *product* ab. Therefore, we define a sequence of $t+1$ couples of bounds:

$$(A_i, B_i) = \left(2^{ui}, 2^{u(t-i)}\right) \forall i \in \{0, \ldots, t\}$$

Equations (2) and (3) imply that there must exist at least one index i such that $0 \leq a \leq A_i$ and $0 < b \leq B_i$. Then using Theorem 1, since $A_i B_i = 2^{ut} < p$, given (A_i, B_i, p, s) one can recover (a, b), and hence the difference between $\mathfrak{F}$ and $\mathfrak{F}'$.

The problem is that (unlike section 2.2) we have no guarantee that such an (a, b) is unique. Namely, we could (in theory) stumble over an $(a', b') \neq (a, b)$ satisfying (3) for some index $i' \neq i$. We conjecture that such failures happen with negligible probability (that we do not try to estimate here) when u is large enough, but this makes the modified protocol heuristic only. If failures never occur, this variant will roughly halve the quantity of transmitted bits with respect to section 2.2.

Fabrice: "heuristic" ⟶ not really, because of the final hash verification, though we cannot compute exactly the complexity...

3.2 The File Laundry: Reducing u

What happens if we brutally shorten u in the basic Divide & Factor protocol? As expected by the birthday paradox, we should start seeing collisions. Let us analyze the statistics governing the appearance of collisions.

Consider `HashPrime` as a random function from $\{0,1\}^*$ to $\{0, \ldots, 2^u - 1\}$. Let X_i be the random variable:

$$X_i = \begin{cases} 1 & \text{if file } F_i \text{ collides with another file.} \\ \\ 0 & \text{otherwise.} \end{cases}$$

Clearly, we have $\Pr[X_i = 1] \leq \frac{\eta - 1}{2^u}$. The average number of colliding files is hence:

$$\mathbb{E}\left[\sum_{i=0}^{\eta-1} X_i\right] \leq \sum_{i=0}^{\eta-1} \frac{\eta - 1}{2^u} = \frac{\eta(\eta - 1)}{2^u}$$

Fabrice: maybe say it in another way, since we already do not suppose HashPrime is collision-resistant

Fabrice: What is η exactly ? n, n', $n + n'$, $|\mathfrak{F} \cup \mathfrak{F}'|$? I prefer the last one actually, i.e., the total number of files...

For instance, for $\eta = 10^6$ files and 32-bit digests, the expected number of colliding files is less than 233.

However, it is important to note that a collision can only yield a *false positive*, and never a *false negative*. In other words, while a collision may obliviate a difference[2] a collision will never create a nonexistent difference *ex nihilo*.

Fabrice: difference or discrepancy ?

Thus, it suffices to replace $\texttt{HashPrime}(F)$ by a diversified $\hbar_\ell(F) = \texttt{HashPrime}(\ell|F)$ to quickly filter-out file differences by repeating the protocol for $\ell = 1, 2, \ldots$ At each iteration the parties will detect new files and new deletions, fix these and "launder" again the remaining multisets.

Fabrice: already said in Section 2.3... and maybe it is better to say that, in this first analysis, we suppose we do not filter out files, because this only improves the algo...

Assume that the diversified $\hbar_\ell(F)$'s are random and independent. To understand why the probability that a stubborn file persists colliding decreases exponentially with the number of iterations λ, assume that η remains invariant between iterations and define the following random variables:

$$X_i^\ell = \begin{cases} 1 & \text{if file } F_i \text{ collides with another file during iteration } \ell. \\ 0 & \text{otherwise.} \end{cases}$$

$$Y_i = \textstyle\prod_{\ell=1}^{\lambda} X_i^\ell = \begin{cases} 1 & \text{if file } F_i \text{ collides with another file during all the } \lambda \text{ first protocol iterations.} \\ 0 & \text{otherwise.} \end{cases}$$

Fabrice je ne vois pas la difference entre X_i^ℓ dans cette section et Z_i^ℓ dans la suivante. Peux-tu preciser STP

By independence, we have:

$$\Pr\left[\, Y_i = 1 \,\right] = \prod_{\ell=1}^{\lambda} \Pr\left[\, X_i^\ell = 1 \,\right] = \Pr\left[\, X_i^1 = 1 \,\right] \ldots \Pr\left[\, X_i^\lambda = 1 \,\right] \leq \left(\frac{\eta - 1}{2^u}\right)^\lambda$$

Therefore the average number of colliding files is:

$$\mathbb{E}\left[\sum_{i=0}^{\eta-1} Y_i\right] \leq \sum_{i=0}^{\eta-1} \left(\frac{\eta - 1}{2^u}\right)^\lambda = \eta \left(\frac{\eta - 1}{2^u}\right)^\lambda$$

And the probability that at least one false positive will survive k rounds is:

$$\epsilon_k \leq \eta \left(\frac{\eta - 1}{2^u}\right)^\lambda$$

For the previously considered instance[3] we get $\epsilon_2 \leq 5.43\%$ and $\epsilon_3 \leq 2 \cdot 10^{-3}\%$.

[2] *e.g.* make the parties blind to the difference between `index.htm` and `iexplore.exe`.

[3] $\eta = 10^6$,$u = 32$.

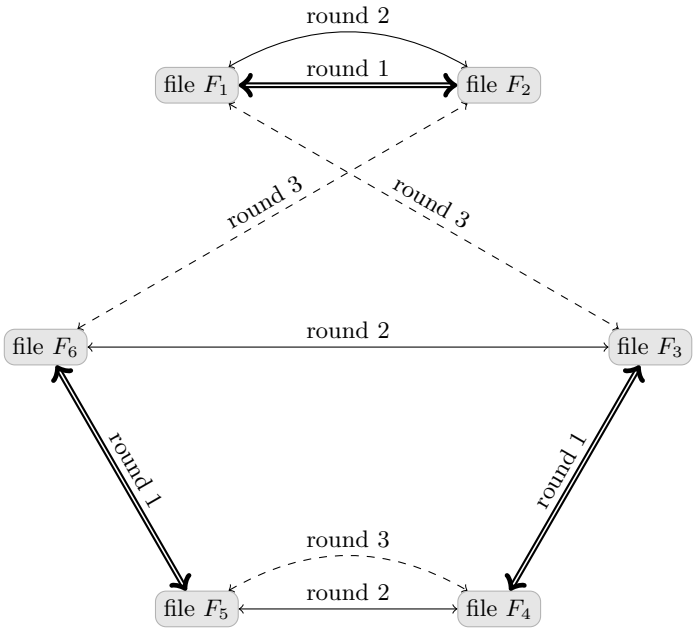

Fig. 4. Illustration of three masquerade balls. Each protocol round is materialized by a different type of arrow. Arrows denotes collisions.

A more refined (but somewhat technical) analysis. As mentioned previously, the parties can remove the files confirmed as different during iteration k and work during iteration $k+1$ only with common and colliding files. Now, the only collisions that can fool round k, are the collisions of file-pairs (F_i, F_j) such that F_i and F_j have both already collided during *all the previous iterations*[4]. We call such collisions "masquerade balls" (*cf.* Figure 4). Define the two random variables:

$$Z_i^\ell = \begin{cases} 1 & \text{if } F_i \text{ participated in masquerade balls during all the } \ell \\ & \text{first protocol iterations.} \\ 0 & \text{otherwise.} \end{cases}$$

$$X_{i,j}^\ell = \begin{cases} 1 & \text{if files } F_i \text{ and } F_j \text{ collide during iteration } \ell. \\ 0 & \text{otherwise.} \end{cases}$$

[4] Note that we do not require that F_i and F_j repeatedly collide *which each other*. *e.g.* we may witness during the first round $\hbar_1(F_1) = \hbar_1(F_2)$, $\hbar_1(F_3) = \hbar_1(F_4)$ and $\hbar_1(F_5) = \hbar_1(F_6)$ while during the second round $\hbar_2(F_1) = \hbar_2(F_2)$, $\hbar_2(F_3) = \hbar_1(F_6)$ and $\hbar_2(F_5) = \hbar_2(F_4)$ as shown in Figure 4.

Set $Z_i^0 = 1$ and write $p_\ell = \Pr\left[Z_i^\ell = 1 \text{ and } Z_j^\ell = 1 \right]$ for all ℓ and $i \neq j$. For $k \geq 1$, we have:

$$\begin{aligned}
\Pr\left[Z_i^\lambda = 1 \right] &= \Pr\left[\exists j \neq i,\, X_{i,j}^\lambda = 1,\, Z_i^{\lambda-1} = 1 \text{ and } Z_j^{\ell-1} = 1 \right] \\
&\leq \sum_{j=0, j\neq i}^{\eta-1} \Pr\left[X_{i,j}^\lambda = 1 \right] \Pr\left[Z_i^{\lambda-1} = 1 \text{ and } Z_j^{\lambda-1} = 1 \right] \\
&\leq \frac{\eta - 1}{2^u} p_{\lambda-1}
\end{aligned}$$

Fabrice: maybe put this in appendix and add a few comments...

Furthermore $p_0 = 1$ and

$$\begin{aligned}
p_\ell &= \Pr\left[X_{0,1}^\ell = 1,\, Z_0^\ell = 1 \text{ and } Z_1^\ell = 1 \right] + \Pr\left[X_{0,1}^\ell = 0,\, Z_0^\ell = 1 \text{ and } Z_1^\ell = 1 \right] \\
&\leq \Pr\left[X_{0,1}^\ell = 1,\, Z_0^{\ell-1} = 1 \text{ and } Z_1^{\ell-1} = 1 \right] \\
&\quad + \sum_{i\geq 2, j\geq 2} \Pr\left[X_{0,i}^\ell = 1,\, X_{1,j}^\ell = 1,\, Z_0^{\ell-1} = 1 \text{ and } Z_1^{\ell-1} = 1 \right] \\
&= \Pr\left[X_0^\ell = X_1^\ell \right] \Pr\left[Z_0^{\ell-1} = 1 \text{ and } Z_1^{\ell-1} = 1 \right] \\
&\quad + \sum_{i\geq 2, j\geq 2} \Pr\left[X_{0,i}^\ell = 1 \right] \Pr\left[X_{1,j}^\ell = 1 \right] \Pr\left[Z_0^{\ell-1} = 1 \text{ and } Z_1^{\ell-1} = 1 \right] \\
&\leq \frac{1}{2^u} p_{\ell-1} + \frac{(\eta-2)^2}{2^{2u}} p_{\ell-1} = p_{\ell-1} \left(\frac{1}{2^u} + \frac{(\eta-2)^2}{2^{2u}} \right)
\end{aligned}$$

hence:

$$p_\ell \leq \left(\frac{1}{2^u} + \frac{(\eta-2)^2}{2^{2u}} \right)^\ell,$$

and

$$\Pr\left[Z_i^\lambda = 1 \right] \leq \frac{\eta - 1}{2^u} \left(\frac{1}{2^u} + \frac{(\eta-2)^2}{2^{2u}} \right)^{\lambda-1}$$

And finally, the survival probability of at least one false positive after k iterations satisfies:

$$\epsilon'_\lambda \leq \frac{\eta(\eta-1)}{2^u} \left(\frac{1}{2^u} + \frac{(\eta-2)^2}{2^{2u}} \right)^{\lambda-1}$$

For $(\eta = 10^6, u = 32, \lambda = 2)$, we get $\epsilon'_2 \leq 0.013\%$.

How to select u? For the sake of simplicity, we consider $t = t_1 = t_2 = \ldots$. For a fixed λ, ϵ'_λ decreases as u grows. For a fixed u, ϵ'_λ also decreases as λ grows. Transmission, however, grows with both u (bigger digests) and k (more iterations). We write for the sake of clarity: $\epsilon'_\lambda = \epsilon'_{\lambda,u,\eta}$.

Fix η. Note that the number of bits transmitted per iteration ($\simeq 3ut$), is proportional to u. This yields an expected transmission complexity bound $T_{u,\eta}$ such that:

$$T_{u,\eta} \propto u \sum_{\lambda=1}^{\infty} \lambda \cdot \epsilon'_{\lambda,u,\eta} = \frac{u\eta\,(\eta-1)}{2^u} \sum_{\lambda=1}^{\infty} \lambda \left(\frac{1}{2^u} + \frac{(\eta-2)^2}{2^{2u}} \right)^{\lambda-1} = \frac{u\eta\,(\eta-1)\,8^u}{\left(2^u - 4^u + (\eta-2)^2\right)^2}$$

Dropping the proportionality factor $\eta\,(\eta-1)$, neglecting $2^u \ll 2^{2u}$ and approximating $(\eta-2) \simeq \eta$, we can optimize the function:

$$\phi_\eta(u) = \frac{u \cdot 8^u}{(4^u - \eta^2)^2}$$

$\phi_{10^6}(u)$ admits an optimum for $u = 19$.

Note: The previous analysis is incomplete because of the following approximations:

- We consider u-bit prime digests while u-bit strings contain only about $2^u/u$ primes.
- We used a fixed u in all rounds. Nothing forbids using a different u_ℓ at each iteration, or even fine-tuning the u_ℓ's adaptively as a function of the laundry's effect on the progressively reconciliated multisets..
- Our analysis treats t as a constant, but large t values increase p and hence the number of potential files detected as different per iteration - an effect disregarded *supra*.

A different approach is to optimize t and u experimentally, *e.g.* using the open source D&F program `btrsync` developed by the authors (*cf.* section 6).

3.3 How to Stop a Probabilistic Washing Machine?

Remplacer ℓ par λ

We now combine both optimizations and assume that ℓ laundry rounds are necessary for completing some given reconciliation task using a half-sized p. By opposition to section 2.2, confirming correct protocol termination is now non-trivial.

We say that a *round failure* occurs whenever a round results in an $(a', b') \neq (a, b)$ satisfying Equation (3). Let the round failure probability be some function $\zeta(u)$ (that we did not estimate). If u is kept small (for efficiency reasons), the probability $(1-\zeta(u))^\ell$ that the protocol will properly terminate may dangerously drift away from one.

Fabrice: oups, pas sûr de tout comprendre ici, il faudrait que l'on en rediscute... En particulier, le CRT n'a pas d'erreur a priori, sauf erreur de ma part.

If v of $\ell + v$ rounds fail, Oscar needs to solve a problem called *Chinese Remaindering With Errors* [2]:

Problem 1. (Chinese Remaindering With Errors Problem: CRWEP*).* Given as input integers v, B and $\ell + v$ points $(s_1, p_1), \ldots, (s_{\ell+v}, p_{\ell+v}) \in \mathbb{N}^2$ where the p_i's are coprime, output all numbers $0 \leq s < B$ such that $s \equiv s_i \bmod p_i$ for at least ℓ values of i.

~~*La définition de $\zeta(u)$ est assez gratuite vu que ça n'intervient pas dans la suite.*~~

We refer the reader to [2] for more on this problem, which is beyond our scope. Boneh [3] provides a polynomial-time algorithm for solving the CRWEP under certain conditions satisfied in our setting.

But how can we confirm the solution? As mentioned in section 2.3, Neil will send to Oscar $H = \texttt{Hash}(\mathfrak{F}')$ as the interaction starts. As long as Oscar's CRWEP resolution will not yield a state matching H, the parties will continue the interaction.

4 Computational Complexity

In this section, we are interested in computing the computational complexity of our protocol, when there is no collision, to simplify the analysis. We first present some variants of the protocol we described above, and then we analyse the complexity of all these variants and propose some algorithmic optimizations to speed up the file reconciliation. A summary of all costs can be found in Table 1.

4.1 Basic Complexity

Let $\mu(l)$ be the time required to multiply two l-bit numbers[5]. For naive (*i.e.* convolutive) algorithms $\mu(l) = O(l^2)$, but using FFT multiplication [11], $\mu(l) = \tilde{O}(l)$. FFT is experimentally faster than convolutive methods starting at $l \sim 10^6$. The modular division of two l-bit numbers and the reduction of $2l$-bit number modulo a l-bit number are also known to cost $\tilde{O}(\mu(l))$ [4]. Indeed, in packages such as `gmp`, division and modular reduction run in $\tilde{O}(l)$, for sufficiently large l.

As proven in TODO APPENDIX, the complexity of `HashPrime` is $u^2\mu(u)$. Hence, we have the costs depicted in the third column of Table 1.

4.2 Adapting p_k

Using p_k ut_k-bit primes is not very practical, because generating a big prime number is slow, and storing a list of such primes require to be able to bound the number of rounds, and to fix all the parameters (u and $t_1, t_2, \dots$). That is why, in this section we show that we can adapt p_k to be able to generate them easily and also, to speed up the computations with a constant factor.

Let $\texttt{Prime}[i]$ denote the i-th prime[6]. Besides conditions on size, the *only* property required from p is to be co-prime with all the h_i's and all the h'_i's. We can hence consider the following variants:

[5] We assume that $\forall l, l', \mu(l + l') \geq \mu(l) + \mu(l')$.
[6] with $\texttt{Prime}[1] = 2$

Variant 1: Smooth p_k:

$$p_k = \prod_{j=r_k}^{r_{k+1}-1} \texttt{Prime}[j],$$

where the bounds r_k are chosen to ensure that each p_k has the proper size. Generating such a prime is much faster than generating a big prime p_k.

Variant 2: $p_k = \texttt{Prime}[k]^{r_k}$ where the exponents r_k are chosen to ensure that each p_k has the proper size. This variant is even faster than the previous one, but require to choose h_k bigger than all $\texttt{Prime}[k]^{r_k}$.

Variant 3: $p_k = P_k = 2^{ut_k}$. In this case, $C_k = \prod_{i=1}^{n} h_i \bmod p_k$, $c_1 = C_1$ and $c_k = (C_k - C_{k-1})/p_{k-1}$, *i.e.*, c_k is the slice of bits $ut_{k-1} \dots ut_k - 1$ of C_k (with $t_0 = 0$), which we write $c_k = C_k[ut_{k-1} \dots ut_k]$. Using Algorithm 1 for the computation of c_k, this variant offers substantial constant-factor accelerations compared to previous variants, because it removes the need of all modulo operations and of CRT re-combination when multiple p_k are needed, since C_k is just the binary concatenation of c_k and C_{k-1} !

Let us explain the main ideas of Algorithm 1. Let $X_i = \prod_{j=1}^{n} h_j$ ($X_0 = 1$), $X_{i,k} = X_i[ut_{k-1} \dots ut_k]$ and let $D_{i,k}$ be the u upper-bits of the product of $Y_{i,k} = X_i[0 \dots ut_k]$ and h_i, *i.e.*, $D_{i,k} = (Y_{i,k} \times h_i)[ut_k \dots u(t_k+1)]$ ($D_{i,0} = 0$ and $D_{0,k} = 0$). Since $X_{i+1} = X_i \times h_{i+1}$, we have, for $k \geq 0$, $i \geq 0$:

$$\begin{aligned} D_{i+1,k+1} \times 2^{ut_{k+1}} + Y_{i+1,k+1} &= Y_{i,k+1} \times h_{i+1} \\ &= (X_{i,k+1} \times 2^{ut_k} + Y_{i,k}) \times h_{i+1} \\ &= X_{i,k+1} \times 2^{ut_k} \times h_{i+1} + Y_{i,k} \times h_{i+1} \\ &= X_{i,k+1} \times h_{i+1} \times 2^{ut_k} + (D_{i,k} \times 2^{ut_k} + \dots). \end{aligned}$$

Therefore, if we only consider bits $[ut_k \dots u(t_{k+1}+1)]$, for $k \geq 1$,$i \geq 0$:

$$D_{i+1,k+1} \times 2^{u(t_{k+1}} + X_{i+1,k+1} = X_{i,k+1} \times h_{i+1} + D_{i,k}.$$

Since $c_k = X_{n,k}$, the algorithm is correct.

We remark we only need to store $D_{k,i}$ and $D_{k+1,i}$ during round k (for all i). So the space complexity is $O(nu)$.

4.3 Algorithmic Optimizations using Product Trees

The non-overwhelming (but nonetheless important) complexities of the computations of (c, c') and of the factorizations can be even reduced to $\tilde{O}(\frac{n}{t_k}\mu(ut_k))$ and $\tilde{O}(\frac{n}{T_k}\mu(uT_k))$

Algorithm 1 Computation of c_k for $p_k = 2^{ut_k}$

Require: k, the set h_i, $(D_{k,i})$ as explained in the article
Ensure: $c_{k+1} = \prod_{i=1}^{n} h_i \bmod p_{k+1}$, (D_{k+1}) as explained in the article
1: **if** $k = 0$ **then**
2: $X \leftarrow 1$
3: **else**
4: $X \leftarrow 0$
5: **for** $i = 0, \ldots, n-1$ **do**
6: $Z \leftarrow X \times h_{i+1}$
7: $D_{i+1,k+1} \leftarrow Z[u(t_{k+1} - t_k) \ldots u(t_{k+1} - t_k + 1)]$
8: $X \leftarrow Z[0 \ldots u(t_{k+1} - t_k)]$
9: $c_{k+1} \leftarrow X$

using product trees. Therefore, the cost drops to $\tilde{O}(nu)$ with FFT [11]. To simplify the presentation, we suppose there is only one round, and $p = p_1$, $t = t_1 = T_1$, and we assume that $t = 2^\tau$ is a power of two dividing n. We also only focus on the case p prime, for the sake of simplicity. But the algorithm can be easily adapt to the other variants.

The idea is the following: group h_i's by subsets of t elements and compute the product of each such subset in $\mathbb{N}$.

$$H_j = \prod_{i=jt}^{jt+t-1} h_i \in \mathbb{N}$$

Each H_j can be computed in $\tilde{O}(\mu(ut))$ using the standard product tree method described in Algorithm 2 (for $j = 0$) illustrated in Figure 5. Thus, all these $\frac{n}{t}$ products can be computed in $\tilde{O}(\frac{n}{t}\mu(ut))$. We can then compute c by multiplying the H_j modulo p, which costs $\tilde{O}(\frac{n}{t}\mu(ut))$.

Algorithm 2 Product Tree Algorithm

Require: the set h_i
Ensure: $\pi = \pi_1 = \prod_{i=0}^{t-1} h_i$, and π_i for $i \in \{1, \ldots, 2t-1\}$ as in Figure 5
1: $\pi \leftarrow$ array of size t
2: **function** PRODTREE(i,start,end)
3: **if** start = end **then**
4: **return** 1
5: **else if** start + 1 = end **then**
6: **return** h_{start}
7: **else**
8: mid $\leftarrow \lfloor \frac{\text{start+end}}{2} \rfloor$
9: $\pi_{2i} \leftarrow$PRODTREE($2i$,start,mid)
10: $\pi_{2i+1} \leftarrow$PRODTREE($2i+1$,mid,end)
11: **return** $\pi_{2i} \times \pi_{2i+1}$
12: $\pi_1 \leftarrow$PRODTREE$(1, 0, t)$

What is the caveat?

The same technique applies to factorization[7], but with a slight *caveat*.

After computing the tree product, we can compute the residues of a modulo H_0. Then we can compute the residues of a mod H_0 modulo the two children π_2 and π_3 of $H_0 = \pi_1$ in the product tree (depicted in Figure 5), and so on. Intuitively, we descend the product tree doing modulo reduction. At the end (*i.e.*, as we reach the leaves), we obtain the residues of a modulo each of the h_i ($i \in \{0, \ldots, t-1\}$). This is described in Algorithm 6 and illustrated in Figure 6. We can use the same method for the tree product associated to any H_j, and the residues of a modulo each of the h_i ($i \in \{jt, \ldots, jt+t-1\}$) for any j, *i.e.*, a modulo each of the h_i for any i. Complexity is $\tilde{O}(\mu(ut))$ for each j, which amounts to a total complexity of $\tilde{O}(\frac{n}{t}\mu(ut))$.

Algorithm 3 Division Using a Product Tree

Require: $a \in \mathbb{N}$, π the product tree of Algorithm 2
Ensure: $A[i] = a \bmod \pi_i$ for $i \in \{1, \ldots, 2t-1\}$, computed as in Figure 3

```
A ← array of size t
function modTree(i)
    if i < 2t then
        A[i] ← A[⌊i/2⌋] mod π_i
        modTree(2i)
        modTree(2i + 1)
A[1] ← a mod π_1
modTree(2)
modTree(3)
```

4.4 Summary

Summary of costs is depicted in Table 1.

4.5 Remove Hashing to Prime Numbers

In most cases, the most costly operation is the hashing to prime number. That is why, a further optimization can consist in removing the need to hash to prime numbers by hashing to any integer coprime with all the p_i. The problem is that, even if there are no collision and a and b are correctly recovered, the fact that h_i divides a does not mean F_i should be a file in $\mathcal{S}$, because, for example, we can have $a = 150$, $h_1 = 10$, $h_2 = 15$, $h_3 = 6$, and $a = h_1 \times h_2$, but h_3 divides a. Therefore, we need a slightly more complex method for the factorization and a careful probability analysis to ensure this case does not come too often. This is left for future work.

Remarks Fabrice:

[7] We explain the process with a, this is applicable *ne variatur* to b as well.

Entity	Computation	Complexity expressed in $\tilde{O}$ of			
		Basic algo.		Opt. algo.	
		p_i prime	$p_i = 2^{ut_i}$	p_i prime	$p_i = 2^{ut_i}$
Both	computation of h_i and h'_i	$nu^2\mu(u)$			
	for round i				
Both	compute redundancies c_i and c'_i	$n \cdot \mu(ut_i)$		$\frac{n}{t_i} \cdot \mu(ut_i)$	
Neil	compute $s_i = c'_i/c_i$[a] or $S_i = C'_i/C_i$	$\mu(uT_i)$			
Neil	compute S_i from S_{i-1} and s_i (CRT)[a]	$\mu(uT_i)$	n/a	$\mu(uT_i)$	n/a
Neil	find a_i, b_i such that $S_i = a_i/b_i \bmod p_i$	$\mu(uT_i)^3$			
Neil	factor a_i	$n \cdot \mu(uT_i)$		$\frac{n}{T_i} \cdot \mu(uT_i)$	
	last round				
Oscar	factor b_i	$n \cdot \mu(ut_i)$		$\frac{n}{t_i} \cdot \mu(ut_i)$	
	Overwhelming complexity	TODO	TODO	TODO	TODO

[a] only for p_i prime or equivalent TODO;
[b] only for $p_i = 2^{ut_i}$.
[c] using advanced algorihms in [10,14] — naive extended GCD leads $(uT_i)^2$.

Table 1. Global Protocol Complexity

- we should separate move resolution algo from implementation
- we should implement the non-prime version stuff... but difficult to find the correct subset of hi which factorizes...

5 From Set Reconciliation to Files Synchronization

The previous sections presented how to perform set reconciliation on a set of fixed-size hashes, and how to choose the hash size when it is used to represent the content of files. This suggests a simple protocol to synchronize a set of files: first synchronize the set of hashes as described, then exchange the content of the files corresponding to the hash values that were found to be in the symmetric difference.

However, in practice, we do not wish to synchronize file *sets*, but file *hierarchies*: we are not just interested in the *content* of the files but in their *metadata*. The most important metadata is the file's path (i.e., its name and its location in the filesystem), though other kinds of metadata exist (e.g., modification time, owner, permissions). In many use cases, the file metadata can change without changes in the file content: for instance, files can be *moved* to a different location. When performing reconciliation, we must be aware of this fact, and reflect file moves without re-transferring the contents of the moved files.

In this section, we present how to solve this *file synchronization* problem, using our set reconciliation program as a building block.

5.1 General Principle

To perform file synchronization, Oscar and Neil will compute the hash of the contents of each of their file (which we will denote as the *content hash*). They will then compute the hash of the concatenation of the content hash and file metadata (which we will call the *mixed hash*). Oscar and Neil then perform set reconciliation on the set of mixed hashes.

Once the reconciliation has completed, all mixed hashes common to Oscar and Neil represent files which have the same content and metadata in Oscar and in Neil. Neil now sends to Oscar the content hashes and metadata of the files whose mixed hash appears in Neil but not in Oscar. Oscar is now aware of the metadata and content hash of all of Neil's files that do not exist in Oscar with the same content and metadata (we will call them the *missing files*).

Oscar now looks at the list of content hashes of the missing files. For some of these hashes, Oscar may already have a file with the same content hash, only with the wrong metadata. For others, Oscar may not have any file with the same content hash. In the first case, Oscar can recreate Neil's file by altering the metadata, without retransferring the file contents. In the second case, Oscar needs to retrieve the full file contents from Neil. To complete the file synchronization, we first deal with all of Oscar's files which fall in the first case: this is a bit tricky because "altering the metadata" is not obvious to perform in-place when it involves file moves, and is the focus of Section 5.2. Once this has been performed, we transfer all missing files: this is explained in Section 5.3.

5.2 Moving Existing Files

To reproduce the structure of Oscar on Neil's disk, we need to perform a sequence of file moves. Sadly, it is not straightforward to apply the moves, because, if we take a file to move, its destination might be blocked, either because a file already exists (we want to move a to b, but b already exists), or because a folder cannot be created (we want to move a to b/c, but b already exists as a file and not as a folder). Note that for a move operation $a \to b$, there is at most one file blocking the location b: we will call it the *blocker*.

If the blocker is absent on Oscar, then we can just delete the blocker. However, if a blocker exists, then we might need to move it somewhere else before we solve the move we are interested in. This move itself might have a blocker, and so on. It seems that we just need to continue until we reach a move which has no blocker or whose blocker can be deleted, but we can get caught in a cycle: if we must move a to b, b to c and c to a, then we will not be able to perform the operations without using a temporary location.

How can we perform the moves? A simple way would be to move each file to a unique temporary location and then rearrange files to our liking: however, this performs many unnecessary moves and could lead to problems if the program is interrupted. We can do something more clever by performing a decomposition in Strongly Connected Components (SCC) of the *move graph* (with one vertex per file and one edge per move operation going

from to the file to its blocker or to its destination if no blocker exists). The computation of the SCC decomposition is simplified by the observation that because two files being moved to the same destination must be equal, we can only keep one arbitrary in-edge per node, and look at the graph pruned in this fashion: its nodes have in-degree at most one, so the strongly connected components are either single nodes or cycles. Once the SCC decomposition is known, the moves can be applied by applying each SCC in a bottom-up fashion, an SCC's moves being solved either trivially (for single files) or using one intermediate location (for cycles).

The detailed algorithm is implemented as two mutually recursive functions and presented as Algorithm 4.

5.3 Transferring Missing Files

Once all moves have been applied, Oscar's hierarchy contains all of its files which also exist on Neil, and they have been put at the correct location. The only thing that remains is to transfer the contents of Neil's files that do not exist in Oscar's hierarchy and create those files at the right position. To do so, a simple solution is to use `rsync` to synchronize explicitly the correct files on Neil to the same location in Oscar's hierarchy, using the fact that Oscar is now aware of all of Neil's files and their location.

6 Implementation

We implemented and benchmarked the Divide & Factor protocol described in the previous sections. The implementation is called `btrsync`, its source code is available from [1].

6.1 Organization

Fabrice: Finish that stuff and put it in appendix ?

The program is composed of three parts: shell script, python program cmd, python program... TODO

The Shell Script sets up two instances of the Python program on Oscar and Neil and establishes a bidirectional communication channel between them using two Unix pipes between their standard inputs and outputs.

The Python Program uses `gmp` to perform all the number theory operations and performs the actual synchronization. It proceeds in two phases:

Algorithm 4 Perform Moves

Require: $\mathfrak{D}$ is a dictionary where $\mathfrak{D}[f]$ denotes the intended destinations of f

```
1: M ← []
2: T ← []
3: for f in 𝔇's keys do
4:     M[f] ← not_done
5: function UNBLOCK_COPY(f, t)
6:     if t is blocked by some b then
7:         if b is not in 𝔇's keys then
8:             unlink(b)                          ▷ We don't need b
9:         else
10:            RESOLVE(b)                         ▷ Take care of b and make it go away
11:    if T[f] was set then
12:        f ← T[f]
13:    copy(f, d)
14: function RESOLVE(f)
15:    if M[f] = done then
16:        return                                 ▷ Already managed by another in-edge
17:    if M[f] = doing then
18:        T[f] ← mktemp()
19:        move(f, T[f])
20:        M[f] ← done
21:        return                                 ▷ We found a loop, moved f out of the way
22:    M[f] ← doing
23:    for d ∈ 𝔇[f] do
24:        if d ≠ f then
25:            unblock_copy(f, d)                 ▷ Perform all the moves
26:    if f ∉ 𝔇[f] and T[f] was not set then
27:        unlink(f)
28:    if T[f] was set then
29:        unlink(T[f])
30: for f in 𝔇's keys do
31:    RESOLVE(f)
```

6.2 What is Implemented ?

TODO: write this beautyfully

- variant 2 of adapting p_i (power of small primes)
- hash to prime (remove use of hash to prime, not supported yet, though it WOULD be VERY interesting)
- $t_i = t' \times 2^i$ with t' a constant, i.e., doubling t at each round, for complexity reasons...

Small difference with theory:

- move reconciliation: An optimization implemented by `btrsync` over the algorithm described here is to move files instead of copying them and then remove the original file. Moves are faster than copies on most filesystems as the OS does not need to copy the actual file contents to perform moves. TODO: Fabrice@Antoine, que voualis-tu dire ici ? J'ai juste copié-collé ce que tu avais écris
- hash to prime: slightly different, I should correct this and use the fast version described in appendix if I have time (Fabrice)

The only metadata managed by `btrsync` is the file path (name and location). Other kinds of metadata (modification date, owner, permissions) are not implemented, though it would be easy to do so.

6.3 Experimental Comparison to `rsync`

We compared `rsync`[8] and our Divide & Factor implementation (called `btrsync`) under the following experimental conditions:

Fabrice: Can't we move this in appendix ? I'm not sure it's really essential. We can replace it by a sentence as: "We call `rsync` with carefully chosen options in order to do a fair comparison (cf Appendix)"

Test Directories: The directories used for transmission and time comparisons are described in Table 3.

Command-Line Options: `rsync` was called with the following options, for the reasons below:

▶ `--delete` to delete existing files on Oscar which do not exist on Neil like `btrsync` does.

▶ `-I` to ensure that `rsync` did not cheat by looking at file modification times (which `btrsync` does not do).

▶ `--chmod="a=rx,u+w"` in an attempt to disable the transfer of file permissions (which `btrsync` does not transfer). Although these settings ensure that `rsync` does not need to

[8] `rsync` version 3.0.9, used both as a competitor to benchmark against and as an underlying call in our own code.

transfer permissions, verbose logging suggests that it does transfer them anyway, so `rsync` must lose a few bytes per file as compared to `btrsync` for this reason.

▶ `-v` Transmission accounting was performed by calling `rsync` with the `-v` flag (which reports the number of sent and received bytes). For `btrsync` we added a piece of code counting the amount of data transmitted during `btrsync`'s own negotiations.

Network Configuration: Experiments were performed without any network transfer, by synchronizing two folders on the same host. Hence, time measurements should mostly represent the CPU cost of the synchronization.

Results: Results are given in Table 2. In general, `btrsync` spent more time than `rsync` on computation (especially when the number of files is large, which is typically seen in the experiments involving `synthetic`). Transmission results, however, turn out to be favorable to `btrsync`.

In the trivial experiments where either Oscar or Neil have no data at all, `rsync` outperforms `btrsync`. This is especially visible when Neil has no data: `rsync` immediately notices that there is nothing to transfer, but `btrsync` engages in information transfers to determine the symmetric difference.

On non-trivial tasks, however, `btrsync` outperforms `rsync`. This is the case of the `synthetic` datasets, where `btrsync` does not have to transfer information about all unmodified files, and even more so in the case where there are no modifications at all. For Firefox source code datasets, `btrsync` saves a very small amount of bandwidth, presumably because of unmodified files. For the `btrsync` source code dataset, we notice that `btrsync`, unlike `rsync`, was able to detect the move and avoid retransferring the moved folder.

7 Conclusion and Further Improvements

Vérifier les claims de cette liste, et en parler dès l'intro.

The main contributions of our work are:

- We present the novel "Divide & Factor" protocol for set reconciliation, which is based on number theory and is optimal with respect to transfer size.
- We study the problem of set reconciliation of directories of files. We discuss the optimal size of message digests in this setting, as well as a move resolution algorithm to reproduce a directory structure.
- We present `btrsync`, an open source implementation of the "Divide & Factor" protocol.
- We demonstrate the usability of this implementation through benchmarks on synthetic and real-world tasks, and show that `btrsync` exchanges less data than the popular software `rsync`.
- The optimizations presented in this paper apply to [] as well.

TODO résoudre ce point.

Many fine questions of the probabilistic discussions in the paper are left as future work. Another further line of research would be to pursue development of `btrsync` to make it suitable for end users.

Be more specific!

Future work: divide large files in smaller blocks.

8 Acknowledgment

The authors acknowledge Guillain Potron for his early involvement in this research work.

9 ToDo

- @Fabrice: Pour éviter le cas empty → source trop gros, on pourrait imaginer l'astuce suivante: si jamais Neil la taille de c est plus petite que la taille du produit des nombres premiers $p_1...p_n$ utilisés, Neil envoie un message pour l'indiquer, et on arrête là le protocole. Et Oscar peut directement factoriser ce nombre envoyé...
- @Fabrice: il faut discuter de la taille de la taille maximale des "petits" premiers utilisés pour les variantes de p et montrer que cela n'enlève pas trop d'entropie pour les h_i. Encore une fois, je m'en occupe la semaine prochaine si besoin. ..... OK on s'en moque (Fabrice)
- Refaire une dernière fois les expériences, vu que Fabrice a significativement amélioré les perfs.
- Faire clarifier par Fabrice l'histoire du doublement... Oui, il faudrait expliciter un peu
- Fabrice: comparer nos perfs avec http://ipsit.bu.edu/programs/reconcile/ ? ou pas ?

References

1. Amarilli, A., Ben Hamouda, F., Bourse, F., Morisset, R., Naccache, D., Rauzy, P.: `https://github.com/RobinMorisset/Btrsync`
2. Bleichenbacher, D., Nguyen, P.Q.: Noisy polynomial interpolation and noisy chinese remaindering. In: Advances in Cryptology – Proceedings of Eurocrypt 2000. pp. 53–69. Springer-Verlag (2000)
3. Boneh, D.: Finding smooth integers in short intervals using crt decoding. In: Proceedings of the 32nd Annual ACM Symposium on Theory of Computing. pp. 265–272 (2000)
4. Burnikel, C., Ziegler, J., Stadtwald, I., D-Saarbrucken: Fast recursive division (1998)
5. Eppstein, D., Goodrich, M., Uyeda, F., Varghese, G.: What's the difference?: efficient set reconciliation without prior context. In: ACM SIGCOMM Computer Communication Review. vol. 41, pp. 218–229. ACM (2011)
6. Fouque, P.A., Stern, J., Wackers, J.G.: Cryptocomputing with rationals. In: Blaze, M. (ed.) Financial Cryptography. Lecture Notes in Computer Science, vol. 2357, pp. 136–146. Springer (2002)
7. Hofmann, M., Pierce, B.C., Wagner, D.: Edit lenses. In: Field, J., Hicks, M. (eds.) POPL. pp. 495–508. ACM (2012)
8. Minsky, Y., Trachtenberg, A.: Scalable set reconciliation. In: 40th Annual Allerton Conference on Communications, Control and Computing (October 2002), a full version can be downloaded from `http://ipsit.bu.edu/documents/BUTR2002-01.ps`

9. Minsky, Y., Trachtenberg, A., Zippel, R.: Set reconciliation with nearly optimal communication complexity. Information Theory, IEEE Transactions on 49(9), 2213–2218 (2003)
10. Pan, V., Wang, X.: On rational number reconstruction and approximation. SIAM Journal on Computing 33, 502 (2004)
11. Schönhage, A., Strassen, V.: Schnelle multiplikation grosser zahlen. Computing 7(3), 281–292 (1971)
12. Tridgell, A.: Efficient algorithms for sorting and synchronization. Ph.D. thesis, PhD thesis, The Australian National University (1999)
13. Vallée, B.: Gauss' algorithm revisited. J. Algorithms 12(4), 556–572 (1991)
14. Wang, X., Pan, V.: Acceleration of Euclidean algorithm and rational number reconstruction. SIAM Journal on Computing 32(2), 548 (2003)

A Extended Protocol

First phase during which Neil amasses modular information on the difference

Oscar		**Neil**
		start the protocol with p_1
	$\xrightarrow{c_1}$	
		computes a, b using p_1
		if a factors properly then go to Final Phase
		else perform the protocol with p_2
	$\xrightarrow{c_2}$	
		computes $c \bmod p_1p_2 = \mathrm{CRT}_{p_1,p_2}(c_1, c_2)$
		computes a, b using p_1p_2
		if a factors properly then go to Final Phase
		else perform the protocol with p_3
	$\xrightarrow{c_3}$	
		computes $c \bmod p_1p_2p_3 = \mathrm{CRT}_{p_1,p_2,p_3}(c_1, c_2, c_3)$
		computes a, b using $p_1p_2p_3$
		if a factors properly then go to Final Phase
		else perform the protocol with p_4
	$\vdots$	

Final Phase

		Let $\mathfrak{S} = \{F'_i \text{ s.t. } a \bmod h'_i = 0\}$
	$\xleftarrow{\mathfrak{S},b}$	
deletes files s.t. $b \bmod h_i = 0$		
adds $\mathfrak{S}$ to the disk		

Note that parties do not need to store the p_i's in full. Indeed, the p_is could be subsequent primes sharing their most significant bits. This reduces storage per prime to a very small additive constant $\cong \ln(p_i) \cong \ln(2^{2tu+2}) \cong 1.39(tu+1)$ of about $\log_2(tu)$ bits.

B Power of Two Protocol

In this variant Oscar computes c in $\mathbb{N}$:

$$c = \prod_{F_i \in \mathfrak{F}} \texttt{HashPrime}(F_i) = \prod_{i=1}^{n} h_i \in \mathbb{N}$$

and considers $c = \bar{c}_{n-1}|\ldots|\bar{c}_2|\bar{c}_0$ as the concatenation of n successive u-bit strings. Again, we omit the treatment of $\perp$s for the sake of clarity.

First phase during which Neil amasses modular information on the difference		
Oscar		**Neil**
computes $c \in \mathbb{N}$		
	$\xrightarrow{\bar{c}_0}$	
		computes a, b modulo 2^u if a factors properly then go to Final Phase else request next chunk $\bar{c}_1$
	$\xrightarrow{\bar{c}_1}$	
		construct $c \bmod 2^{2u} = \bar{c}_1\|\bar{c}_0$ computes a, b modulo 2^{2u} if a factors properly then go to Final Phase else request next chunk $\bar{c}_2$
	$\xrightarrow{\bar{c}_2}$	
		construct $c \bmod 2^{3u} = \bar{c}_2\|\bar{c}_1\|\bar{c}_0$ computes a, b modulo 2^{3u} if a factors properly then go to Final Phase else request next chunk $\bar{c}_3$
	$\vdots$ (for $2t$ rounds)	

Final Phase		
		Let $\mathfrak{S} = \{F'_i$ s.t. $a \bmod 2^{2tu} = 0\}$
	$\xleftarrow{\mathfrak{S},b}$	
deletes files s.t. $b \bmod 2^{2tu} = 0$ adds $\mathfrak{S}$ to the disk		

C Hashing Into Primes

Hashing into primes is frequently needed in cryptography. A recommended implementation of $\texttt{HashPrime}(F)$ is given in Algorithm 6. If u is large enough (*e.g.* 160) one might sacrifice uniformity to avoid repeated file hashings by defining $\texttt{HashPrime}(F) = \texttt{NextPrime}(\texttt{Hash}(F))$. Yet another acceleration (that further destroys uniformity) consists in replacing `NextPrime` by Algorithm 5 where $\alpha = 2 \times 3 \times 5 \times \cdots \times \texttt{Prime}[d]$ is the product of the first primes until some rank d.

analyse ci-contre

Fabrice: Cela accélère un peu, car il y a environ $\frac{n}{\log(n)\varphi(\alpha)}$ nombres premiers $\leq n$ et congrus à 1 modulo α, contre $\frac{n}{\log(n)}$ nombres premiers $\leq n$ (voir `http://fr.wikipedia.org/wiki/Th%C3%A9or%C3%A8me_de_la_progression_arithm%C3%A9tique#Version_quantitative`). Dans l'algo 5, h est donc premier avec proba $\frac{\frac{n}{\log(n)\varphi(a)}}{\frac{n}{\alpha}} = \frac{\alpha}{\log(n)\varphi(\alpha)}$, tandis que dans l'algo 6, h est premier avec proba $\frac{1}{\log(n)}$. On a alors une accélération d'environ 10 si on prend $d = 60$.

Algorithm 5 Fast Nonuniform Hashing Into Primes

1: $h = \alpha \left\lfloor \frac{\texttt{Hash}(F)}{\alpha} \right\rfloor + 1$
2: **while** h is composite **do**
3: $\quad h = h - \alpha$
4: **return** h

Algorithm 6 Possible Implementation of $\texttt{HashPrime}(F)$

1: $i = 0$
2: **repeat**
3: $\quad h = 2 \cdot \texttt{Hash}(F|i) + 1$
4: $\quad i = i + 1$
5: **until** h is prime
6: **return** h

Antoine sur quelle plate-formes ont ete obtenus les resultats experimentaux? quelles vitesses de processeur etc?

Entities and Datasets		Transmission (Bytes)						Time (s)	
Neil's $\mathfrak{F}'$	Oscar's $\mathfrak{F}$	$\texttt{TX}_{\texttt{rs}}$	$\texttt{RX}_{\texttt{rs}}$	$\texttt{TX}_{\texttt{bt}}$	$\texttt{RX}_{\texttt{bt}}$	$\delta_{\texttt{rs}} - \delta_{\texttt{bt}}$	$\frac{\delta_{\texttt{bt}}}{\delta_{\texttt{rs}}}$	$\texttt{time}_{\texttt{rs}}$	$\texttt{time}_{\texttt{bt}}$
source	empty	778311	1614	779517	10140	9732	1.0	0.1	0.4
empty	source	24	12	11927	5952	17843	496.6	0.1	0.4
empty	empty	24	12	19	30	13	1.4	0.0	0.3
synthetic	synthetic_shuffled	54799	19012	7308	3417	-63086	0.1	0.2	1.5
synthetic_shuffled	synthetic	54407	18822	6822	3042	-63365	0.1	0.2	0.8
synthetic	synthetic	54799	19012	327	30	-73454	0.0	0.1	0.7
firefox-13.0.1	firefox-13.0	40998350	1187	39604079	3305	-1392153	1.0	1.5	10.2
source_moved	source	778176	1473	2757	1966	-774926	0.0	0.1	0.6

Table 2. Experimental results. `rs` and `bt` subscripts respectively denote `rsync` and `btrsync`. The two first columns indicate the datasets. Synchronization is performed *from* Neil *to* Oscar. `RX` and `TX` denote the quantity of received and sent bytes and $\delta_{\square} = \texttt{TX}_{\square} + \texttt{RX}_{\square}$. $\delta_{\texttt{rs}} - \delta_{\texttt{bt}}$ and $\delta_{\texttt{bt}}/\delta_{\texttt{rs}}$ express the absolute and the relative differences in transmission between the two programs. The last two columns show timing results.

Directory	Description
synthetic	A directory containing 1000 very small files containing the numbers $1, 2, \ldots, 1000$.
synthetic_shuffled	`synthetic` with: 10 deleted files 10 renamed files 10 modified files
source	A snapshot of `btrsync`'s own source tree
source_moved	`source` with one big folder (a few megabits) renamed.
firefox-13.0	The source archive of Mozilla Firefox 13.0.
firefox-13.0.1	The source archive of Mozilla Firefox 13.0.1
empty	An empty folder.

Table 3. Test Directories.

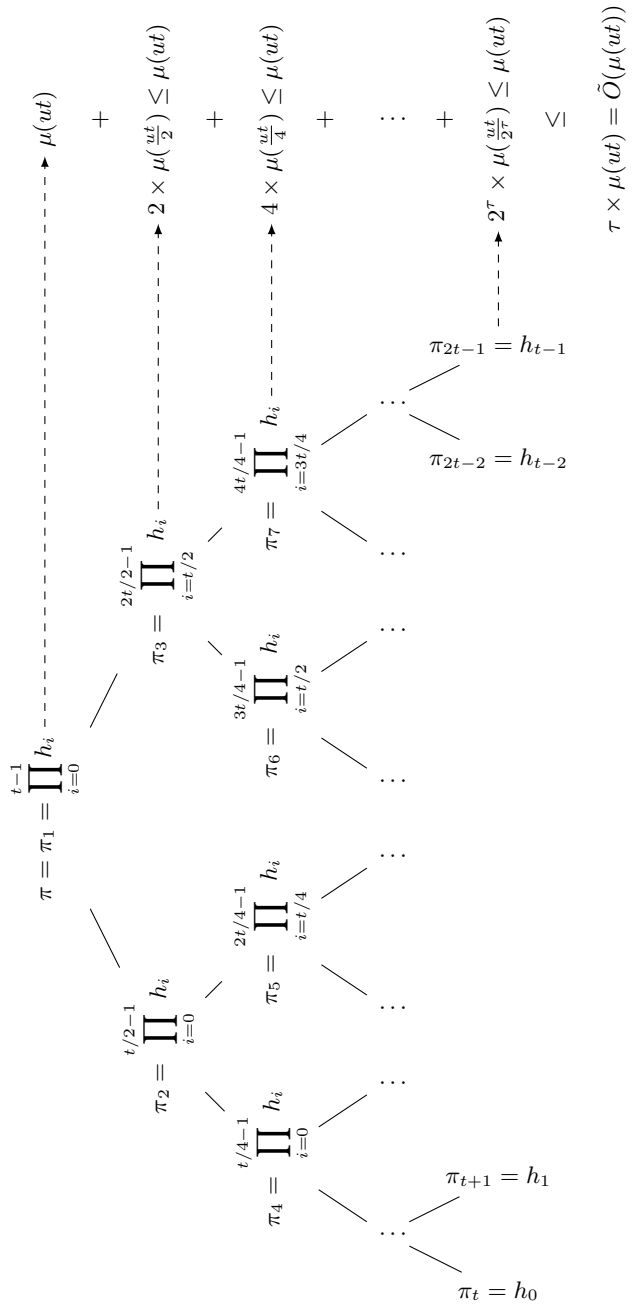

Fig. 5. Product Tree

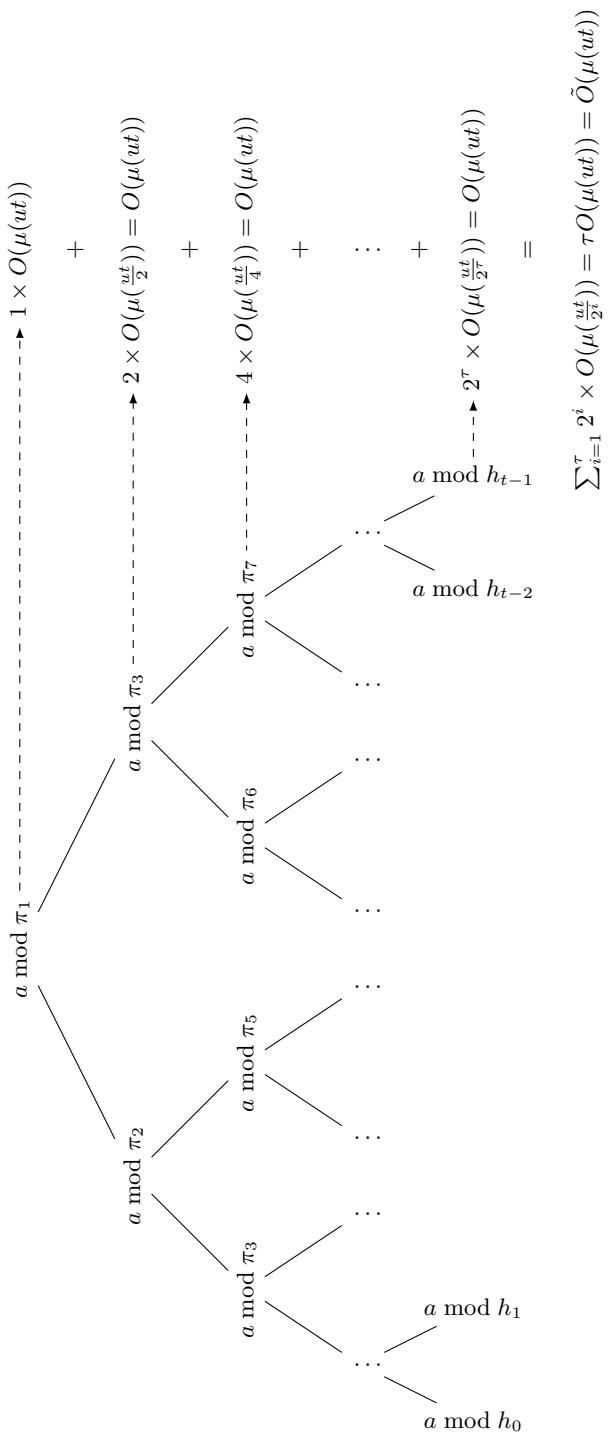

Fig. 6. Modular Reduction From Product Tree