Egalitarian Computing

Alex Biryukov and Dmitry Khovratovich, ***University of Luxembourg***

https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/biryukov

This paper is included in the Proceedings of the 25th USENIX Security Symposium

August 10–12, 2016 • Austin, TX

ISBN 978-1-931971-32-4

Open access to the Proceedings of the 25th USENIX Security Symposium is sponsored by USENIX

Egalitarian computing

Alex Biryukov
University of Luxembourg
alex.biryukov@uni.lu

Dmitry Khovratovich
University of Luxembourg
khovratovich@gmail.com

Abstract

In this paper we explore several contexts where an adversary has an upper hand over the defender by using special hardware in an attack. These include password processing, hard-drive protection, cryptocurrency mining, resource sharing, code obfuscation, etc.

We suggest memory-hard computing as a generic paradigm, where every task is amalgamated with a certain procedure requiring intensive access to RAM both in terms of size and (very importantly) bandwidth, so that transferring the computation to GPU, FPGA, and even ASIC brings little or no cost reduction. Cryptographic schemes that run in this framework become *egalitarian* in the sense that both users and attackers are equal in the price-performance ratio conditions.

Based on existing schemes like Argon2 and the recent generalized-birthday proof-of-work, we suggest a generic framework and two new schemes:

- MTP, a memory-hard Proof-of-Work based on the memory-hard function with fast verification and short proofs. It can be also used for memory-hard time-lock puzzles.
- MHE, the concept of memory-hard encryption, which utilizes available RAM to strengthen the encryption for the low-entropy keys (allowing to bring back 6 letter passwords).

Keywords: MTP, MHE, Argon2, memory-hard, asymmetric, proof-of-work, botnets, encryption, time-lock puzzles.

1 Introduction

1.1 Motivation

Historically attackers have had more resources than defenders, which is still mostly true. Whether it is secret key recovery or document forgery, the attackers are ready to spend tremendous amount of computing power to achieve the goal. In some settings it is possible to make most attacks infeasible by simply setting the key length to 128 bits and higher. In other settings the secret is limited and the best the defender can do is to increase the time needed for the attack, but not to render the attack impossible.

Passwords, typically stored in a hashed form, are a classical example. As people tend to choose passwords of very low entropy, the security designers added unique salts and then increased the number of hash iterations. In response the attackers switched to dedicated hardware for password cracking, so that the price of single password recovery dropped dramatically, sometimes by a few orders of magnitude.

A similar situation occurred in other contexts. The Bitcoin cryptocurrency relies on continuous preimage search for the SHA-256 hash function, which is much cheaper on custom ASICs, consuming up to 30,000 times less energy per solution than most efficient x86 laptops [2]. Eventually, the original concept of an egalitarian cryptocurrency [25] vanished with the emergence of huge and centralized mining pools.

Related problems include password-based key derivation for hard-drive encryption, where the data confidentiality directly depends on the password entropy, and where offline attack is exceptionally easy once the drive is stolen. Similar situation arise in the resource sharing and spam countermeasures. In the latter it is proposed that every user presents a certain proof (often called proof-of-work), which should be too expensive for spammers to generate on a large scale. Yet another setting is that of code obfuscation, in which powerful reverse-engineering/de-compilation tools can be used in order to lift the proprietary code or secrets embedded in the software.

1.2 Egalitarian computing

Our idea is to remedy the disparity between ordinary users and adversaries/cheaters, where latter could use botnets, GPU, FPGA, ASICs to get an advantage and run a cheaper attack. We call it *egalitarian computing* as it should establish the same price for a single computation unit on all platforms, so that the defender's hardware is optimal both for attack and defence. Equipped with egalitarian crypto schemes, defenders may hope to become to be on par with the most powerful attackers.

The key element of our approach is large (in size) and intensive (in bandwidth) use of RAM as a widely available and rather cheap unit for most defenders. In turn, RAM is rather expensive on FPGA and ASIC[1], and slow on GPU, at least compared to memoryless computing tasks. All our schemes use a lot of memory and a lot of bandwidth — almost as much as possible.

We suggest a single framework for this concept and concrete schemes with an unique combination of features.

In the future, adoption of our concept could allow a homogenization of computing resources, a simplified security analysis, and relaxed security requirements. When all attackers use the same hardware as defenders, automated large-scale attacks are no longer possible. Shorter keys, shorter passwords, faster and more transparent schemes may come back to use.

Related work The idea of extensive memory use in the context of spam countermeasures dates back at least to 2003 [5, 13] and was later refined in [15]. Fast memory-intensive hash functions were proposed first by Percival in 2009 [27] an later among the submissions of the Password Hashing Competition. Memory-intensive proofs-of-work have been studied both in theory [16] and practice [6, 32].

Paper structure We describe the goals of our concept and give a high level overview in Section 2. Then we describe existing applications where this approach is implicitly used: password hashing and cryptocurrency proofs of work (Section 3). We present our own progress-free Proof-of-Work MTP with fast verification, which can also serve as a memory-hard time-lock puzzle, in Section 4. The last Section 5 is devoted to the novel concept of memory-hard encryption, where we present our scheme MHE aimed to increase the security of password-based disk encryption.

[1] The memory effect on ASICs can be illustrated as follows. A compact 50-nm DRAM implementation [17] takes 500 mm^2 per GB, which is equivalent to about 15000 10 MHz SHA-256 cores in the best Bitcoin 40-nm ASICs [1] and is comparable to a CPU size. Therefore, an algorithm requiring 1 GB for 1 minute would have the same AT cost as an algorithm requiring 2^{42} hash function calls, whereas the latter can not finish on a PC even in 1 day. In other words, the use of memory can increase the AT cost by a factor of 1000 and more at the same time cost for the desktop user.

2 Egalitarian computing as framework

2.1 Goal

Our goal is to alter a certain function $\mathscr{H}$ in order to maximize its computational cost on the most efficient architecture – ASICs, while keeping the running time on the native architecture (typically x86) the same. We ignore the design costs due to nontransparent prices, but instead estimate the running costs by measuring the time-area product [8, 31]. On ASICs the memory size M translates into certain area A. The ASIC running time T is determined by the length of the longest computational chain and by the ASIC memory latency.

Suppose that an attacker wants to compute $\mathscr{H}$ using only a fraction αM of memory for some $\alpha < 1$. Using some tradeoff specific to $\mathscr{H}$, he has to spend $C(\alpha)$ times as much computation and his running time increases by the factor $D(\alpha)$ (here $C(\alpha)$ may exceed $D(\alpha)$ as the attacker can parallelize the computation). In order to fit the increased computation into time, the attacker has to place $\frac{C(\alpha)}{D(\alpha)}$ additional cores on chip. Therefore, the time-area product changes from AT_1 to AT_α as

$$AT_\alpha = A \cdot (\alpha + \frac{\beta C(\alpha)}{D(\alpha)}) T \cdot D(\alpha) = \\ = AT_1(\alpha D(\alpha) + C(\alpha)\beta), \quad (1)$$

where β is the fraction of the original memory occupied by a single computing core. If the tradeoff requires significant communication between the computing cores, the memory bandwidth limit Bw_{max} may also increase the running time. In practice we will have $D(\alpha) \geq C(\alpha) \cdot Bw/Bw_{max}$, where Bw is the bandwidth for $\alpha = 1$.

Definition 1 *We call function $\mathscr{F}$* memory-hard *(w.r.t. M) if any algorithm $\mathscr{A}$ that computes $\mathscr{H}$ using αM memory has the computation-space tradeoff $C(\alpha)$ where $C()$ is at least a superlinear function of $1/\alpha$.*

It is known [19] that any function whose computation is interpreted as a directed acyclic graph with T vertices of constant in-degree, can be computed using $O(\frac{T}{\log T})$ space, where the constant in $O()$ depends on the degree. However, for concrete hash functions very few tradeoff strategies have been published, for example [9].

2.2 Framework

Our idea is to combine a certain computation $\mathscr{H}$ with a memory-hard function $\mathscr{F}$. This can be done by modifying $\mathscr{H}$ using input from $\mathscr{F}$ (*amalgamation*) or by transforming its code to an equivalent one (*obfuscation*).

The **amalgamation** is used as follows. The execution of $\mathscr{H}$ is typically a sequence of smaller steps $H_i, i < T$, which take the output V_{i-1} from the previous step and produce the next output V_i. For our purpose we need another primitive, a memory-hard function $\mathscr{F}$, which fills the memory with some blocks $X[i], i < T$. We suggest combining $\mathscr{H}$ with $\mathscr{F}$, for example like:

$$\mathscr{H}' = H'_T \circ H'_{T-1} \circ \cdots \circ H'_1,$$

where

$$H'_i(V_{i-1}) = H(V_{i-1} \oplus X[i-1]).$$

Depending on the application, we may also modify $X[i]$ as a function of V_{i-1} so that it is impossible to precompute $\mathscr{F}$. The idea is that any computation of $\mathscr{H}'$ should use T blocks of memory, and if someone wants to use less, the memory-hardness property would impose computational penalties on him. This approach will also work well for any code that uses nonces or randomness produced by PRNG. PRNG could then be replaced by (or intermixed with the output of) F.

The **obfuscation** principle works as follows. Consider a compiler producing an assembly code for some function $\mathscr{H}$. We make it to run a memory-hard function $\mathscr{F}$ on a user-supplied input I (password) and produce certain number of memory blocks. For each if-statement of the form

if x **then** A
else B

the compiler computes a *memory-hard bit* b_i which is extracted from the block $X[i]$ (the index can also depend on the statement for randomization) and alters the statement as

if $x \oplus b_i$ **then** A
else B

for $b_i = 0$ and

if $x \oplus b_i$ **then** B
else A

for $b_i = 1$. This guarantees that the program will have to run $\mathscr{F}$ at least once (the bits b_i can be cached if this if-statement is used multiple times, ex. in a loop).

Accessing the memory block from a random memory location for each conditional statement in practice would slow down the program too much, so compiler can perform a tradeoff depending on the length of the program, the number of conditional statements in it and according to a tunable degree of required memory-hardness for a program. Memory-hard bits could be mixed into opaque predicates or other code obfuscation constructs like code-flattening logic.

We note that in order for a program to run correctly, the user needs to supply correct password for $\mathscr{F}$, even though the source code of the program is public. A smart decompiler, however, when supplied with the password, can obtain clean version of the program by running $\mathscr{F}$ only once.

Our schemes described in the further text use the amalgamation principle only, so we leave the research directions in obfuscation for future work.

3 Egalitarian computing in applications

In this section we outline several applications, where memory-hard functions are actively used in order to achieve egalitarian computing.

3.1 Password hashing with a memory-hard function

The typical setting for the password hashing is as follows. A user selects a password P and submit it to the authentication server with his identifier U. The server hashes P and unique salt S with some function $\mathscr{F}$, and stores $(U, S, F(P,S))$ in the password file. The common threat is the password file theft, so that an attacker can try the passwords from his dictionary file D and check if any of them yields the stolen hash. The unique S ensures that the hashes are tried one-by-one.

Following massive password cracking attacks that use special hardware [23, 30], the security community initiated the Password Hashing Competition [3] to select the hash function that withstands the most powerful adversaries. The Argon2 hash function [10] has been recently selected as the winner. We stress that the use of memory-hard function for password hashing does not make the dictionary attacks infeasible, but it makes them much more expensive in terms of the single trial cost.

Definition and properties of Argon2 We use Argon2 in our new schemes described in Sections 4 and 5. Here we outline the key elements of the Argon2 design that are used in our scheme. For more details and their rationale we refer the reader to [10].

Argon2 takes P, S, and possibly some additional data U as inputs. It is parametrized by the memory size M, number of iterations t, and the available parallelism l. It fills M blocks of memory $X[1], X[2], \ldots, X[M]$ (1 KB each) and then overwrites them $(t-1)$ times. Each block $X[i]$ is generated using internal compression function F, which takes $X[i-1]$ and $X[\phi(i)]$ as inputs. For $t = 1$ this works as follows, where H is a cryptographic hash

function (Blake2b).

$$\begin{aligned} &X[1] = H(P,S); \\ &X[i] = F(X[i-1], X[\phi(i)]),\ i > 1; \\ &Out \rightarrow H(X[M]). \end{aligned} \tag{2}$$

The indexing function $\phi(i)$ is defined separately for each of two versions of Argon2: 2d and 2i. The Argon2d version, which we use, compute it as a function of the previous block $X[i-1]$.

The authors proved [10] that all the blocks are generated distinct assuming certain collision-resistant-like properties of F. They also reported the performance of 0.7 cpb on the Haswell CPU with 4 threads, and 1.6 cpb with 1 thread.

Tradeoff security of Argon2 Using the tradeoff algorithm published in [9], the authors report the values $C(\alpha)$ and $D(\alpha)$ up to $\alpha = 1/7$ with $t = 1$. It appears that $C(\alpha)$ is exponential in α, whereas $D(\alpha)$ is linear.

α	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{7}$
$C(\alpha)$	1.5	4	20.2	344	4660	2^{18}
$D(\alpha)$	1.5	2.8	5.5	10.3	17	27

Table 1: Time and computation penalties for the ranking tradeoff attack for Argon2d.

3.2 Proofs of work

A *proof-of-work scheme* is a challenge-response protocol, where one party (Prover) has to prove (maybe probabilistically) that it has performed a certain amount of computation following a request from another party (Verifier). It typically relies on a computational problem where a solution is assumed to have fixed cost, such as the preimage search in the Bitcoin protocol and other cryptocurrencies. Other applications may include spam protection, where a proof-of-work is a certificate that is easy to produce for ordinary sender, but hard to generate in large quantities given a botnet (or more sophisticated platform).

The proof-of-work algorithm must have a few properties to be suitable for cryptocurrencies:

- It must be *amortization-free*, i.e. producing q outputs for $\mathscr{B}$ should be q times as expensive;
- The solution must be *short* enough and verified quickly using *little memory* in order to prevent DoS attacks on the verifier.
- The time-space tradeoffs must be *steep* to prevent any price-performance reduction.
- The time and memory parameters must be *tunable independently* to sustain constant mining rate.
- To avoid a clever prover getting advantage over the others the advertised algorithm must be the most efficient algorithm to date (*optimization-freeness*).
- The algorithm must be *progress-free* to prevent centralization: the mining process must be stochastic so that the probability to find a solution grows steadily with time and is non-zero for small time periods.
- Parallelized implementations must be limited by the memory bandwidth.

As demonstrated in [11], almost any hard problem can be turned into a proof-of-work, even though it is difficult to fulfill all these properties. The well-known hard and NP-complete problems are natural candidates, since the best algorithms for them run in (sub)exponential time, whereas the verification is polynomial. The proof-of-work scheme Equihash [11] is built on the *generalized-birthday*, or k-XOR, problem, which looks for a set of n-bit strings that XOR to zero. The best existing algorithm is due to Wagner [34]. This problem is particularly interesting, as the time-space tradeoff steepness can be adjusted by changing k, which does not hold, e.g., in hard knapsacks.

Drawbacks of existing PoW We briefly discuss existing alternatives here. The first PoW schemes by Dwork and Naor [14] were just computational problems with fast verification such as the square root computation, which do not require large memory explicitly. The simplest scheme of this kind is Hashcash [7], where a partial preimage to a cryptographic hash function is found (the so called *difficulty test*). Large memory comes into play in [13], where a random array is shared between the prover and the verifier thus allowing only large-memory verifiers. This condition was relaxed in [15], where superconcentrators [28] are used to generate the array, but the verifier must still hold large memory in the initialization phase. Superconcentrators were later used in the Proof-of-Space construction [16], which allows fast verification. However, the scheme [16] if combined with the difficulty test is vulnerable to cheating (see Section 4.4 for more details) and thus can not be converted to a progress-free PoW. We note that the superconcentrators make both [15] and [16] very slow.

Ad-hoc but faster schemes started with scrypt [27], but fast verification is possible only with rather low

amount of memory. Using more memory (say, using Argon2 [10]) with a difficulty test but verifying only a subset of memory is prone to cheating as well (Section 4.4).

The scheme [11] is quite promising, but the reference implementation reported is quite slow, as it takes about 30 seconds to get a proof that certifies the memory allocation of 500 MB. As a result, the algorithm is not truly progress-free: the probability that the solution is found within the first few seconds is actually zero. It can be argued that this would stimulate centralization among the miners. In addition, the memory parameter does not have sufficient granularity and there is no correlation between the allocated memory and the minimal time needed to find the proof.

Finally, we mention schemes Momentum [21] and Cuckoo cycle [32], which provide fast verification due to their combinatorial nature. They rely on the memory requirements for the collision search (Momentum) or graph cycle finding (Cuckoo). However, Momentum is vulnerable to a sublinear time-space tradeoff [11], whereas the first version of the Cuckoo scheme was recently broken in [6].

We summarize the properties of the existing proof-of-work constructions in Table 2. The AT cost is estimated for the parameters that enable 1-second generation time on a PC.

4 MTP: Proofs of work and time-lock puzzles based on memory-hard function

In this section we present a novel proof-of-work algorithm MTP (for Merkle Tree Proof) with fast verification, which in particular solves the progress-free problem of [11]. Our approach is based on the memory-hard function, and the concrete proposal involves Argon2.

Since fast memory-hard functions $\mathscr{F}$ such as Argon2 perform a lengthy chain of computations, but do not solve any NP-like problem, it is not fast to verify that Y is the output of F. Checking some specific (say, last) blocks does not help, as explained in detail in the further text. We thus have to design a scheme that lower bounds the time-area product for the attacker, even if he computes a slightly modified function.

4.1 Description of MTP

Consider a memory-hard function $\mathscr{F}$ that satisfies Equation (2) (for instance, Argon2) with a single pass over the memory producing T blocks and a cryptographic hash function H (possibly used in $\mathscr{F}$). We propose the following non-interactive protocol for the Prover (Figure 1) in Algorithm 1, where L and d are security parameters. The average number of calls to F is $T + 2^d L$.

Algorithm 1 MTP: Merkle-tree based Proof-of-Work. Prover's algorithm

Input: Challenge I, parameters L, d.

1. Compute $\mathscr{F}(I)$ and store its T blocks $X[1]$, $X[2]$, …, $X[T]$ in the memory.
2. Compute the root Φ of the Merkle hash tree (see Appendix A).
3. Select nonce N.
4. Compute $Y_0 = H(\Phi, N)$ where G is a cryptographic hash function.
5. For $1 \le j \le L$:
$$i_j = Y_{j-1} \pmod{T};$$
$$Y_j = H(Y_{j-1}, X[i_j]).$$
6. If Y_L has d trailing zeros, then $(\Phi, N, \mathscr{Z})$ is the proof-of-work, where $\mathscr{Z}$ is the opening of $2L$ blocks $\{X[i_j - 1], X[\phi(i_j)]\}$. Otherwise go to Step 3.

Output: Proof $(\Phi, N, \mathscr{Z})$.

The verifier, equipped with $\mathscr{F}$ and H, runs Algorithm 2.

Algorithm 2 MTP: Verifier's algorithm

Input: Proof $(\Phi, N, \mathscr{Z})$, parameters L, d.

1. Compute $Y_0 = H(\Phi, N)$.
2. Verify all block openings using Φ.
3. Compute from $\mathscr{Z}$ for $1 \le j \le L$:
$$X[i_j] = F(X[i_j - 1], X[\phi(i_j)]);$$
$$Y_j = G(Y_{j-1}, X[i_j]).$$
4. Check whether Y_L has t trailing zeros.

Output: Yes/No.

4.2 Cheating strategies

Let the computation-space tradeoff for $\mathscr{H}$ and the default memory value T be given by functions $C(\alpha)$ and $D(\alpha)$ (Section 2).

Memory savings Suppose that a cheating prover wants to reduce the AT cost by using αT memory for some $\alpha < 1$. First, he computes $\mathscr{F}(I)$ and Φ, making $C(\alpha)T$

Scheme	AT cost	Speed	Verification		Tradeoff	Paral-sm	Progress
			Fast	M/less			-free
Dwork-Naor I [14]	Low	High	Yes	Yes	Memoryless	Yes	Yes
Dwork-Naor II [13]	High	Low	Yes	No	Memoryless	Constr.	Yes
Dwork-Naor III [15]	Medium	Low	Yes	No	Exponential	Constr.	Yes
Hashcash/Bitcoin [7]	Low	High	Yes	Yes	Memoryless	Yes	Yes
Pr.-of-Space [16]+Diff.test	High	Low	Yes	Yes	Exponential	No	No
Litecoin	Medium	High	Yes	Yes	Linear	No	Yes
Argon2-1GB + Diff.test	High	High	No	No	Exponential	No	Yes
Momentum [21]	Medium	High	Yes	Yes	Attack [11, 33]	Yes	Yes
Cuckoo cycle [32]	Medium [6]	Medium	Yes	Yes	Linear [6]	Yes	Yes
Equihash [11]	High	Medium	Yes	Yes	Exponential	Constr.	Yes
MTP	High	High	Yes	Yes	Exponential	Constr.	Yes

Table 2: Review of existing proofs of work. Litecoin utilizes scrypt with 128KB of RAM followed by the difficulty test). M/less – memoryless; constr. – constrained.

calls to F. Then for each N he has to get or recompute L blocks using only αT stored blocks. The complexity of this step is equal to the complexity of recomputing random L blocks during the first computation of $\mathscr{F}$. A random block is recomputed by a tree of average size $C(\alpha)$ and depth $D(\alpha)$. Therefore, to compute the proof-of-work, a memory-saving prover has to make $C(\alpha)(T + 2^d L)$ calls to F, so his amount of work grows by $C(\alpha)$.

Block modification The second cheating strategy is to compute a different function $\mathscr{F}' \neq \mathscr{F}$. More precisely, the cheater produces some blocks $X[i']$ (which we call *inconsistent* as in [16]) not as specified by Equation (2) (e.g. by simply computing $X[i'] = H(i')$). In contrast to the verifiable computation approach, our protocol allows a certain number of inconsistent blocks. Suppose that the number of inconsistent blocks is εT, then the chance that no inconsistent block is detected by L opened blocks is

$$\gamma = (1-\varepsilon)^L.$$

Therefore, the probability for a proof-of-work with εM inconsistent blocks to pass the opening test is γ. In other words, the cheater's time is increased by the factor $1/\gamma$. We note that it does not make sense to alter the blocks after the Merkle tree computation, as any modified block would fail the opening test.

Overall cheating penalties Let us accumulate the two cheating strategies into one. Suppose that a cheater stores αT blocks and additionally allows εT inconsistent blocks. Then he makes at least

$$\frac{C(\alpha+\varepsilon)(T+2^d L)}{\gamma} \tag{3}$$

calls to F. The concrete values are determined by the penalty function $C()$, which depends on $\mathscr{F}$.

4.3 Parallelism

Both honest prover and cheater can parallelize the computation for 2^t different nonces. However, the latency of cheater's computation will be higher, since each block generates a recomputation tree of average depth $D(\alpha + \varepsilon)$.

4.4 Why simpler approach does not work: grinding attack

Now we can explain in more details why the composition of $\mathscr{F}$ and the difficulty test is not a good proof-of-work even if some internal blocks of $\mathscr{H}$ are opened. Suppose that the proof is accepted if $H(X[T])$ has certain number d of trailing zeros. One would expect that a prover has to try 2^d distinct I on average and thus call $\mathscr{F}$ 2^t times to find a solution. However, a cheating prover can simply try 2^d values for $X[T]$ and find one that passes the test in just 2^d calls to H. Although $X[T]$ is now inconsistent, it is unlikely to be selected among L blocks to open, so the cheater escapes detection easily. Additionally checking $X[T]$ would not resolve the problem since a cheater

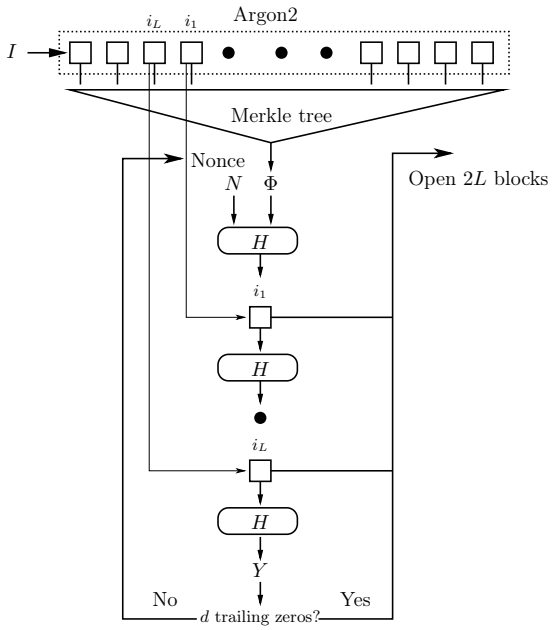

Figure 1: MTP: Merkle-tree based Proof-of-Work with light verification.

would then modify the previous block, or $X[\phi(T)]$, or an earlier block and then propagate the changes. A single inconsistent block is just too difficult to catch[2].

4.5 MTP-Argon2

As a concrete application, we suggest a cryptocurrency proof-of-work based on Argon2d with 4 parallel lanes. We aim to make this PoW unattractive for botnets, so we suggest using 2 GB of RAM, which is very noticeable (and thus would likely alarm the user), while being bearable for the regular user, who consciously decided to use his desktop for mining. On our 1.8 GHz machine a single call to 2-GB Argon2d runs in 0.5 seconds, but the Merkle tree computation is more expensive, as we have to hash 2 GB of data splitted into 1 KB blocks. We suggest using Blake2b for H, as it is already used in Argon2d, but restrict to 128-bit output, so that the total running time is about 3 seconds. In this case a single opening has $16 \cdot 21$ bytes of hashes, or 1.3 KB in total.

We suggest $L = 70$, so that the entire proof consists of 140 blocks and their openings, or 180 KB in total. Let us figure out the cheating advantage. The $C()$ and $D()$ functions are given in Table 1). Assuming certain ratio between the area needed to implement Blake2b and the area needed for DRAM, we get the following lower bound on the ASIC-equipped cheater.

Proposition 1 *For $L = 70$ and 2 GB of RAM the time-area product can be reduced by the factor of 12 at most, assuming that each Blake2b core occupies an equivalent of 2^{16} bytes.*

Proof. Assuming that each core occupies 2^{16} bytes, we obtain $\beta = 2^{-15}$ in terms of Equation (1). Since the cheater has the success chance $\gamma = (1-\varepsilon)^L$, Equation (1) is modified as follows:

$$AT_\alpha = AT_1 \frac{\alpha D(\alpha+\varepsilon) + C(\alpha+\varepsilon)/2^{15}}{(1-\varepsilon)^L}. \qquad (4)$$

Consider three options:

- $\alpha, \varepsilon < 1/12$. Then $C(\alpha+\varepsilon) \geq 4660$ (Table 1) and we have
$$AT_\alpha \geq AT_1 \cdot \frac{4660}{32768} \geq AT_1 \cdot 0.12.$$
- $\alpha < 1/12$, $1/6 \geq \varepsilon > 1/12$. Then $C(\alpha+\varepsilon) \geq 20$, $(1-\varepsilon)^L > 1/441$, and we have
$$AT_\alpha \geq AT_1 \cdot \frac{20 \cdot 441}{32768} \geq AT_1 \cdot 0.27.$$
- $\alpha < 1/12$, $1/6 \leq \varepsilon$. Then $(1-\varepsilon)^L > 2^{15}$, and the time-area product increases.
- $\alpha > 1/12$. Then $AT_\alpha \geq AT_1 \cdot 1/12$.

This ends the proof.

We conclude that a cheater can gain at most 12x-advantage, whereas he can still be detected in the future by memory-rich verifiers. Tradeoffs are also not helpful when implementing this Proof-of-Work on ASIC. Altogether, our proposal should reduce the relative efficiency of potential ASIC mining rigs and allow more egalitarian mining process. Even if someone decides to use large botnets (10,000 machines and more), all the botnets machines would have to use the same 2 GB of memory, otherwise they would suffer large penalty. We note that if $\varepsilon = 0$, i.e. the prover is honest, then his maximal advantage is $\max \frac{1}{\alpha D(\alpha)} \leq 2$.

4.6 MTP as a tool for time-lock puzzles and timestamping

The paradigm of inherently sequential computation was developed by [12] in the application to CPU benchmarking and [29] for timestamping, i.e. to certify that the document was generated certain amount of time in the past. Rivest et al. suggested *time-lock puzzles* for this purpose.

[2] We have not seen any formal treatment of this attack in the literature, but it appears to be known in the community. It is mentioned in [26] and [4].

In our context, a time-lock puzzle solution is a proof-of-work that has lower bound on the running time assuming unlimited parallelism.

The verifier in [20, 29] selects a prime product $N = pq$ and asks the prover to compute the exponent $2^{2^D} \pmod N$ fpr some $D \approx N$. It is conjectured that the prover who does not know the factors can not exponentiate faster than do D consecutive squarings. In turn, the verifier can verify the solution by computing the exponent 2^D modulo $\phi(N)$, which takes $\log(D)$ time. So far the conjecture has not been refuted, but the scheme inherently requires a secret held by the verifier, and thus is not suitable for proofs-of-work without secrets, as in cryptocurrencies.

Time-lock puzzles without secrets were suggested by Mahmoody et al. [22]. Their construction is a graph of hash computations, which is based on depth-robust graphs similarly to [16]. The puzzle is a deterministic graph such that removing any constant fraction of nodes keeps its depth above the constant fraction of the original one (so the parallel computation time is lower bounded). A Merkle tree is put atop of it with its root determining a small number of nodes to open. Therefore, a cheater who wants to compute the graph in less time has to subvert too many nodes and is likely to be caught. As [16], the construction by Mahmoody et al., if combined with the difficulty filter, is subject to the grinding attack described above.

The MTP-Argon2 construction can be viewed as a time-lock puzzle and an improvement over these schemes. First, the difficulty filter is explicitly based on the grinding attack, which makes it a legitimate way to solve the puzzle. Secondly, it is much faster due to high speed of Argon2d. The time-lock property comes from the fact that the computation chain can not be parallelized as the graph structure is not known before the computation.

Suppose that MTP-Argon2 is parallelized by the additional factor of R so that each core computes a chain of length about T/R. Let core j compute j-th (out of R) chain, chronologically. Then bu step i each core has computed i blocks and has not computed $T/R-i$ blocks, so the probability that core j requests a block that has not been computed is

$$\frac{(j-1)(T/R-i)}{(j-1)T/R+i} \leq \frac{(j-1)(T/R-i)}{jT/R}.$$

Summing by all i, we obtain that core j misses at least $\frac{T(1-1/j)}{2R}$, so the total fraction of inconsistent blocks is about $0.5 - \frac{\ln R}{2R}$. Therefore, ε quickly approaches 0.5, which is easily detectable. We thus conclude that a parallel implementation of MTP-Argon2 is likely to fail the Merkle tree verification.

5 Memory-hard encryption on low-entropy keys

5.1 Motivation

In this section we approach standard encryption from the memory-hardness perspective. A typical approach to hard-drive encryption is to derive the master key from the user password and then use it to encrypt chunks of data in a certain mode of operation such as XTS [24]. The major threat, as to other password-based security schemes, are low-entropy passwords. An attacker, who gets access to the hard drive encrypted with such password, can determine the correct key and then decrypt within short time.

A countermeasure could be to use a memory-hard function for the key derivation, so that the trial keys can be produced only on memory-rich machines. However, the trial decryption could still be performed on special memoryless hardware given these keys. We suggest a more robust scheme which covers this type of adversaries and eventually requires that the entire attack code have permanent access to large memory.

5.2 Requirements

We assume the following setting, which is inspired by typical disk-encryption applications. The data consists of multiple chunks $Q \in \mathscr{Q}$, which can be encrypted and decrypted independently. The only secret that is available to the encryption scheme $\mathscr{E}$ is the user-input password $P \in \mathscr{P}$, which has sufficiently low entropy to be memorized (e.g., 6 lowercase symbols). The encryption syntax is then as follows:

$$\mathscr{E}: \mathscr{P} \times \mathscr{S} \times \mathscr{Q} \to \mathscr{C},$$

where $S \in \mathscr{S}$ is associated data, which may contain salt, encryption nonce or IV, chunk identifier, time, and other secondary input; and $C \in \mathscr{C}$ is ciphertext. S serves both to simplify ciphertext identification (as it is public) and to ensure certain cryptographic properties. For instance, unique salt or nonce prevents repetition of ciphertexts for identical plaintexts. We note that in some settings due to storage restriction the latter requirement can be dropped. Decryption then is naturally defined and we omit its formal syntax.

In our proposal we do not restrict the chunk size. Even though it can be defined for chunks as small as disk sectors, the resistance to cracking attacks will be higher for larger chunks, up to a megabyte long.

A typical attack setting is as follows. An attacker obtains the encrypted data via some malicious channel or installs malware and then tries different passwords to decrypt it. For the sake of simplicity, we assume that the

plaintext contains sufficient redundancy so that a successful guess can be identified easily. Therefore, the adversary tries D passwords from his dictionary $\mathscr{D} \subset \mathscr{P}$. Let T be the time needed for the fastest decryption operation that provides partial knowledge of plaintext sufficient to discard or remember the password, and A_0 be the chip area needed to implement this operation. Then the total amount of work performed by the adversary is

$$W = D \cdot T \cdot A_0.$$

At the same time, the time to encrypt T' for a typical user should not be far larger than T. Our goal is to maximize W with keeping T' the same or smaller.

The memory-hard functions seem to serve perfectly for the purpose of maximizing W. However, it remains unclear how to combine such function $\mathscr{F}$ with $\mathscr{E}$ to get *memory-hard encryption* (MHE).

Now we formulate some additional features that should be desirable for such a scheme:

- The user should be able to choose the requested memory size A independently of the chunk length $|Q|$. Whereas the chunk length can be primarily determined by the CPU cache size, desirable processing speed, or the hard drive properties, the memory size determines the scheme's resistance to cracking attacks.

- The memory can be allocated independently for each chunk or reused. In the former case the user can not allocate too much memory as the massive decryption would be too expensive. However, for the amounts of memory comparable to the chunk size the memory-hard decryption should take roughly as much as memoryless decryption. If the allocated memory is reused for distinct chunks, much more memory can be allocated as the allocation time can be amortized. However, the decryption latency would be quite high. We present both options in the further text.

- Full ciphertext must be processed to decrypt a single byte. This property clearly makes T larger since the adversary would have to process an entire chunk to check the password. At the same time, for disk encryption it should be fine to decrypt in the "all-or-nothing" fashion, as the decryption time would still be smaller than the user could wait.

- Encryption should be done in one pass over data. It might sound desirable that the decryption should be done in one pass too. However, this would contradict the previous requirement. Indeed, if the decryption can be done in one pass, then the first bytes of the plaintext can be determined without the last bytes of the ciphertext[3].

- Apart from the memory parameter, the total time needed to allocate this memory should be tunable too. It might happen that the application does not have sufficient memory but does have time. In this case, the adversary can be slowed down by making several passes over the memory during its initialization (the memory-hard function that we consider support this feature).

Our next and final requirement comes from adversary's side. When the malware is used, the incoming network connection and memory for this malware can be limited. Thus, it would be ideal for the attacker if the memory-intensive part can be delegated to large machines under attacker's control, such as botnets. If we just derived the secret-key K for encryption as the output of the memory-hard hash function $\mathscr{F}$, this would be exactly this case. An adversary would then run $\mathscr{F}$ for dictionary D on his own machine, produce the set $\mathscr{K}$ of keys, and supply them to malware (recall that due to low entropy there would be only a handful of these keys). Thus the final requirement should be the following:

- During decryption, it should be impossible to delegate the entire memory-hard computation to the external device without accessing the ciphertext. Therefore, there could be no memory-hard precomputation.

5.3 Our scheme

Our scheme is based on a recent proposal by Zaverucha [35], who addresses similar properties in the scheme based on Rivest's All-or-Nothing transform (ANT). However, the scheme in [35] does not use an external memory-hard function, which makes it memory requirements inevitably bound to the chunk size. Small chunks but large memory is impossible in [35].

Our proposal is again based on the All-or-Nothing transformation, though we expect that similar properties can be obtained with deterministic authenticated encryption scheme as a core primitive. The chunk length q (measured in blocks using by $\mathscr{F}$) and memory size $M \geq q$ are the parameters as well as some blockcipher E (possibly AES). First, we outline the scheme where the memory is allocated separately for each chunk. The reader may also refer to Figure 2.

The underlying idea is to use both the header and the body blocks to produce the ciphertext. In tun, to recompute the body blocks both the ciphertext and the header must be available during trial decryption.

[3]The similar argument is made for the online authenticated ciphers

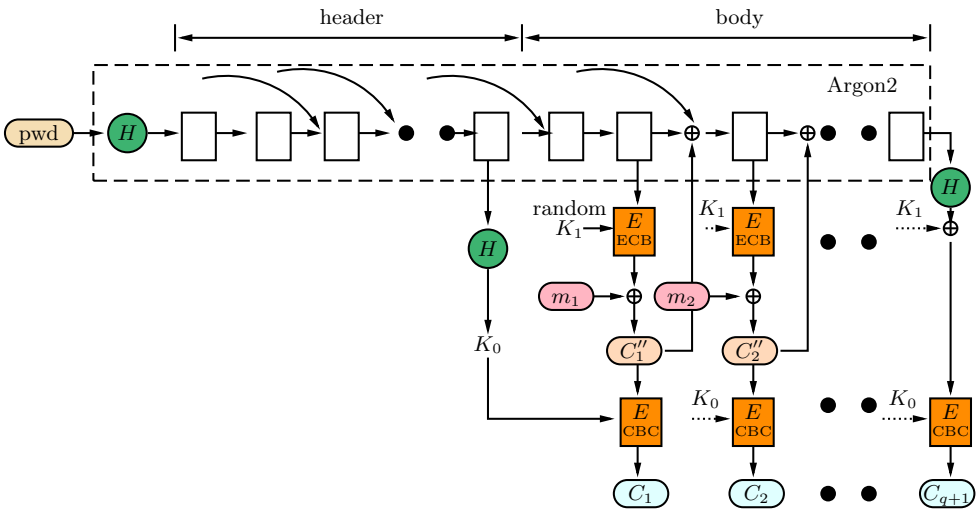

Figure 2: MHE: Disk encryption using memory-hard function Argon2.

The version of the MHE scheme which allocates the same memory for multiple chunks is very similar. The S input is ignored at the beginning, so that the header memory blocks do not depend on the data. Instead, we set $K_0 = H(X_0, S)$, so that the body blocks are affected by S and M, and thus are different for every chunk. In this case the body blocks have to be stored separately and should not overwrite the header blocks for $t > 1$.

Let us verify that the scheme in Algorithm 3 satisfies the properties we listed earlier:

- The allocated memory size M can be chosen independently of the chunk length q (as long as $M > q$).
- The body memory blocks are allocated and processed for each chunk independently. In addition, the header blocks are also processed independently for each chunk in the single-chunk version.
- In order to decrypt a single byte of the ciphertext, an adversary would have to obtain K_1, which can be done only by running $\mathscr{F}$ up to the final block, which requires all C_i'', which are in turn must be derived from the ciphertext blocks.
- Encryption needs one pass over data, and decryption needs two passes over data.
- The total time needed to allocate and fill the header is tunable.
- The computation of the body memory blocks during decryption can not be delegated, as it requires knowledge both of the header and the ciphertext. It might be possible to generate the header on an external machine, but then random access to its blocks to decrypt the ciphertext is required.

We note that properties 1, 5, and 6 are not present in [35].

Security First, we address traditional CPA security. We do not outline the full proof here, just the basic steps. We assume that the adversary does not have access to the internals of Argon2, and that blockcipher E is a secure PRF. Next, we assume collision-resistance of the compression function F used in $\mathscr{F}$. Given that, we prove that all the memory blocks are distinct, which yields the CPA security for C'. From the latter we deduce the CPA security for the final ciphertext. We note that in the case when the collision-resistance of F can not be guaranteed, we may additionally require that X_i undergo hashing by a cryptographic hash function H' before encryption, so that the plaintext blocks are still distinct. All these properties hold up to the birthday bound of the blockcipher.

Next, we figure out the tradeoff security. The genuine decrypting user is supposed to spend M memory blocks for $\mathscr{F}$ and q memory blocks to store the plaintext and intermediate variables (if the ciphertext can be overwritten, then these q blocks are not needed). Suppose that an adversary wants to use αM memory for header and body. Then each missing block, if asked during decryption, must be recomputed making $C(\alpha)$ calls to F. The best such strategy for Argon2, described in [9], yields $C(\alpha)$ that grows exponentially in $1/\alpha$. For example, using $1/5$ of memory, an adversary would have to make 344 times as many calls to F, which makes a memory-reducing encryption cracking inefficient even on special hardware.

in [18].

Algorithm 3 Memory-hard encryption with independent memory allocation (for each chunk).

Input: Password P, memory size M, associated data S, chunk Q, number of iterations t, memory-hard function $\mathscr{F}$ (preferably Argon2), blockcipher E, cryptographic hash function H (e.g. SHA-3).

1. Run $\mathscr{F}$ on (P,S) with input parameters M and t but fill only $M-q$ blocks (the *header*) in the last iteration. Let X_0 be the last memory block produced by $\mathscr{F}$.
2. Produce $K_0 = H(X_0)$ — the first session key.
3. Generate a random session key K_1.
4. Generate the remaining blocks $X_1, X_2, \ldots, X_q$ (*body*) for $\mathscr{F}$ as follows. We assume that each chunk M consists of smaller blocks $m_1, m_2, \ldots, m_q$ of length equal to the block size of $\mathscr{F}$. For each $i \geq 1$:
   - Encrypt X_{i-1} by E in the ECB mode under K_1 and get the intermediate ciphertext block C'_i.
   - Add the chunk data: $C''_i = C'_i \oplus m_i$.
   - Encrypt C''_i under K_0 in the CBC mode and produce the final ciphertext block C_i.
   - Modify the memory: $X_{i-1} \leftarrow X_{i-1} \oplus C''_i$.
   - Generate the block X_i according to the specification of $\mathscr{F}$. In Argon2, the modified X_{i-1} and some another block $X[\phi(X_{i-1})]$ would be used.
5. After the entire chunk is encrypted, encrypt also the key K_1 :

$$C_{t+1} = E_{K_0}(H(X_t) \oplus K_1).$$

Output: $C_1, \ldots, C_{t+1}$.

Performance We suggest taking $l = 4$ in Argon2 in order to fill the header faster using multiple cores, which reportedly takes 0.7 cpb (about the speed of AES-GCM and AES-XTS). The body has to be filled sequentially as the encryption process is sequential. As AES-CBC is about 1.3 cpb, and we use two of it, the body phase should run at about 4 cpb. In a concrete setting, suppose that we tolerate 0.1 second decryption time (about 300 Mcycles) for the 1-MB chunk. Then we can take the header as large as 256 MB, as it would be processed in 170 Mcycles + 4 Mcycles for the body phase.

6 Conclusion

We have introduced the new paradigm of egalitarian computing, which suggests amalgamating arbitrary computation with a memory-hard function to enhance the security against off-line adversaries equipped with powerful tools (in particular with optimized hardware). We have reviewed password hashing and proofs of work as applications where such schemes are already in use or are planned to be used. We then introduce two more schemes in this framework. The first one is MTP, the progress-free proof-of-work scheme with fast verification based on the memory-hard function Argon2, the winner of the Password Hashing Competition. The second scheme pioneers the memory-hard encryption — the security enhancement for password-based disk encryption, also based on Argon2.

References

[1] Avalon asic's 40nm chip to bring hashing boost for less power, 2014. `http://www.coindesk.com/avalon-asics-40nm-//chip-bring-hashing-boost-less-power/`.

[2] Bitcoin: Mining hardware comparison, 2014. available at `https://en.bitcoin.it/wiki/Mining_hardware_comparison`. We compare 2^{32} hashes per joule on the best ASICs with 2^{17} hashes per joule on the most efficient x86-laptops.

[3] Password Hashing Competition, 2015. `https://password-hashing.net/`.

[4] 2016. Andrew Miller, Bram Cohen, private communication.

[5] ABADI, M., BURROWS, M., AND WOBBER, T. Moderately hard, memory-bound functions. In *NDSS'03* (2003), The Internet Society.

[6] ANDERSEN, D. A public review of cuckoo cycle. `http://www.cs.cmu.edu/~dga/crypto/cuckoo/analysis.pdf`, 2014.

[7] BACK, A. Hashcash – a denial of service counter-measure, 2002. available at `http://www.hashcash.org/papers/hashcash.pdf`.

[8] BERNSTEIN, D. J., AND LANGE, T. Non-uniform cracks in the concrete: The power of free precomputation. In *ASIACRYPT'13* (2013), vol. 8270 of *Lecture Notes in Computer Science*, Springer, pp. 321–340.

[9] BIRYUKOV, A., AND KHOVRATOVICH, D. Tradeoff cryptanalysis of memory-hard functions. In *Asiacrypt'15* (2015). available at http://eprint.iacr.org/2015/227.

[10] BIRYUKOV, A., AND KHOVRATOVICH, D. Argon2: new generation of memory-hard functions for password hashing and other applications. In *Euro S&P'16* (2016). available at https://www.cryptolux.org/images/0/0d/Argon2.pdf.

[11] BIRYUKOV, A., AND KHOVRATOVICH, D. Equihash: Asymmetric proof-of-work based on the generalized birthday problem. In *NDSS'16* (2016). available at https://eprint.iacr.org/2015/946.pdf.

[12] CAI, J., LIPTON, R. J., SEDGEWICK, R., AND YAO, A. C. Towards uncheatable benchmarks. In *Structure in Complexity Theory Conference* (1993), IEEE Computer Society, pp. 2–11.

[13] DWORK, C., GOLDBERG, A., AND NAOR, M. On memory-bound functions for fighting spam. In *CRYPTO'03* (2003), vol. 2729 of *Lecture Notes in Computer Science*, Springer, pp. 426–444.

[14] DWORK, C., AND NAOR, M. Pricing via processing or combatting junk mail. In *CRYPTO'92* (1992), vol. 740 of *Lecture Notes in Computer Science*, Springer, pp. 139–147.

[15] DWORK, C., NAOR, M., AND WEE, H. Pebbling and proofs of work. In *CRYPTO'05* (2005), vol. 3621 of *Lecture Notes in Computer Science*, Springer, pp. 37–54.

[16] DZIEMBOWSKI, S., FAUST, S., KOLMOGOROV, V., AND PIETRZAK, K. Proofs of space. In *CRYPTO'15* (2015), R. Gennaro and M. Robshaw, Eds., vol. 9216 of *Lecture Notes in Computer Science*, Springer, pp. 585–605.

[17] GIRIDHAR, B., CIESLAK, M., DUGGAL, D., DRESLINSKI, R. G., CHEN, H., PATTI, R., HOLD, B., CHAKRABARTI, C., MUDGE, T. N., AND BLAAUW, D. Exploring DRAM organizations for energy-efficient and resilient exascale memories. In *International Conference for High Performance Computing, Networking, Storage and Analysis 2013* (2013), ACM, pp. 23–35.

[18] HOANG, V. T., REYHANITABAR, R., ROGAWAY, P., AND VIZÁR, D. Online authenticated-encryption and its nonce-reuse misuse-resistance. In *CRYPTO'15* (2015), R. Gennaro and M. Robshaw, Eds., vol. 9215 of *Lecture Notes in Computer Science*, Springer, pp. 493–517.

[19] HOPCROFT, J. E., PAUL, W. J., AND VALIANT, L. G. On time versus space. *J. ACM 24*, 2 (1977), 332–337.

[20] JERSCHOW, Y. I., AND MAUVE, M. Offline submission with RSA time-lock puzzles. In *CIT* (2010), IEEE Computer Society, pp. 1058–1064.

[21] LORIMER, D. Momentum – a memory-hard proof-of-work via finding birthday collisions, 2014. available at http://www.hashcash.org/papers/momentum.pdf.

[22] MAHMOODY, M., MORAN, T., AND VADHAN, S. P. Publicly verifiable proofs of sequential work. In *ITCS* (2013), ACM, pp. 373–388.

[23] MALVONI, K. Energy-efficient bcrypt cracking, 2014. Passwords'14 conference, available at http://www.openwall.com/presentations/Passwords14-Energy-Efficient-Cracking/.

[24] MARTIN, L. Xts: A mode of aes for encrypting hard disks. *IEEE Security & Privacy*, 3 (2010), 68–69.

[25] NAKAMOTO, S. Bitcoin: A peer-to-peer electronic cash system. http://www.bitcoin.org/bitcoin.pdf.

[26] PARK, S., PIETRZAK, K., ALWEN, J., FUCHSBAUER, G., AND GAZI, P. Spacecoin: A cryptocurrency based on proofs of space. *IACR Cryptology ePrint Archive 2015* (2015), 528.

[27] PERCIVAL, C. Stronger key derivation via sequential memory-hard functions. http://www.tarsnap.com/scrypt/scrypt.pdf.

[28] PIPPENGER, N. Superconcentrators. *SIAM J. Comput. 6*, 2 (1977), 298–304.

[29] RIVEST, R. L., SHAMIR, A., AND WAGNER, D. A. Time-lock puzzles and timed-release crypto. https://people.csail.mit.edu/rivest/pubs/RSW96.pdf.

[30] SPRENGERS, M., AND BATINA, L. Speeding up GPU-based password cracking. In *SHARCS'12* (2012). available at http://2012.sharcs.org/record.pdf.

[31] THOMPSON, C. D. Area-time complexity for VLSI. In *STOC'79* (1979), ACM, pp. 81–88.

[32] TROMP, J. Cuckoo cycle: a memory bound graph-theoretic proof-of-work. Cryptology ePrint Archive, Report 2014/059, 2014. available at http://eprint.iacr.org/2014/059, project webpage https://github.com/tromp/cuckoo.

[33] VAN OORSCHOT, P. C., AND WIENER, M. J. Parallel collision search with cryptanalytic applications. *J. Cryptology 12*, 1 (1999), 1–28.

[34] WAGNER, D. A generalized birthday problem. In *CRYPTO'02* (2002), vol. 2442 of *Lecture Notes in Computer Science*, Springer, pp. 288–303.

[35] ZAVERUCHA, G. Stronger password-based encryption using all-or-nothing transforms. available at http://research.microsoft.com/pubs/252097/pbe.pdf.

A Merkle hash trees

We use Merkle hash trees in the following form. A prover P commits to T blocks $X[1], X[2], \ldots, X[T]$ by computing the hash tree where the blocks $X[i]$ are at leaves at depth $\log T$ and nodes compute hashes of their branches. For instance, for $T = 4$ and hash function G prover P computes and publishes

$$\Phi = G(G(X[1], X[2]), G(X[3], X[4])).$$

Prover stores all blocks and all intermediate hashes. In order to prove that he knows, say, $X[5]$ for $T = 8$, (or to *open* it) he discloses the hashes needed to reconstruct the path from $X[5]$ to Φ:

$$\begin{aligned} \text{open}(X[5]) = (&X[5], X[6], g_{78} = G(X[7], X[8]), \\ &g_{1234} = G(G(X[1], X[2]), G(X[3], X[4])), \Phi), \end{aligned}$$

so that the verifier can make all the computations. If G is collision-resistant, it is hard to open any block in more than one possible way.