

AntCrypt^{*}

Proposal for the Password Hashing Competition

March 31, 2014

Markus Dürmuth, Ralf Zimmermann

Horst Görtz Institute for IT Security
Ruhr University Bochum
Universitätsstr. 150
44780 Bochum, Germany

* The name “AntCrypt” was chosen because the core of our construction resembles an anthill: in both, a huge quantity of small workers carry out tiny tasks in apparent chaos, however, in reality this “chaos” is orchestrated so that all results come together and form the final result.

1 Introduction

Arguably the biggest threat to password hashing schemes stems from GPUs, FPGAs, and ASICs, who provide enormous computing power which can speed up verification of a batch of passwords (e.g., in an offline guessing attack). Common constructions for password hashes use, at their core, two different methods to limit speed-up of verification operations.

- First, aggressively iterated constructions proportionally increase the computation times for verification on all platforms. The (well-understood) problem with constructions solely relying on iterated constructions is that they are typically quite fast when implemented on GPUs and FPGAs, as they can be parallelized very well.
- Frequent memory access (e.g., memory-hardness and similar ideas) are intended to slow down implementations on hardware basically utilizing memory bandwidth and memory latency. Large memory requirements (such as scrypt) will force the attacker to access main memory (on GPUs), while moderate memory usage (such as bcrypt) leaves the attacker with a trade-off between using a large number of registers and thus voiding memory access, or using fewer registers but accessing global memory.

One concern with bcrypt is that the size of the memory used in the computation is fixed to 4 kByte and cannot be changed, and that 4kByte is potentially not enough memory to effectively thwart efficient implementations on FPGAs. With scrypt, one concern is that the huge memory requirements are problematic if deployed on servers handling frequent login requests, and make the server susceptible to denial-of-service attacks. Another potential concern is that memory access is “relatively rare” in the sense that there is one hash function computation between two memory access operations.

In our proposal, we opt for a middle-ground, which seems to offer the best benefit of both worlds:

1. *Memory usage:* We use *moderate amounts of memory*, tunable with a parameter from 256 Bytes upwards, where a reasonable choice seems to be around 32 kBytes. We ensure very frequent access to all regions of the memory, similar to bcrypt and different from scrypt, to avoid previously mentioned potential problems.

In addition, our construction makes use of a (to the best of our knowledge novel) idea that aims to slow down implementations on GPUs and FPGAs/ASICs specifically.

2. *Control-flow divergence:* Our code will frequently branch depending on the current state (and thus ultimately on the password), to (i) avoid good parallelization on GPUs, and (ii) increase the size of implementations (and thus increase the cost and decrease the throughput) on FPGAs/ASICs.

2 The Key-Derivation Scheme

In this section, we will provide a description of our construction and comment on the design choices. A more detailed discussion will follow in Section 3.

2.1 Parameters and Main Data Structure

Unless stated otherwise, all data types are 32-bit words. As the prototype of the PHS function provides two cost parameters, we derive the internal parameters from them as follows:

- `state_bytes` defines the amount of memory used for the state in bytes and is defined as

$$\text{state_bytes} = 2^{\text{m_cost}+8}.$$

Analogously, `state_words` defines the number of 32-bit words the state contains.

- `inner_rounds` defines the number of iterations for the inner loop, iterating over all `state_words` state positions. We require a minimum of at least two inner rounds as follows:

$$\text{inner_rounds} = \max\left(\left\lfloor \frac{\text{m_cost}}{16} \right\rfloor, 2\right)$$

- `outer_rounds` defines the number of iterations for the outer loop. We require at least one outer round and define it as follows:

$$\text{outer_rounds} = \max(\text{t_cost}, 1)$$

The *primary data structure* is a memory buffer `buf = [prefix, memory]`. The `prefix` can be used as a to generate different hash values from the buffer. If not stated otherwise, we refer to `buf` as the `memory` without the `prefix`.

In the algorithm, we use two such buffers: `state` of size `state_words+1` 32-bit words, as well as a `rehash` buffer of `16+1` words. The size of `rehash` is equivalent the output length of the *primary hash function* + 1 word as a prefix. These buffers are accessed either on byte level as bytes `[0, 1, ..., (state_bytes + 3)]` or as words `[[{3, 2, 1, 0}, {7, 6, 5, 4}, ..., {..., (state_bytes + 3)}]`.

2.2 Algorithm Definition

Algorithm 1 describes the basic structure of the derivation function. In the following, we will describe the main functions `init`, `update_entropy`, `update_state` and `compute_output` in more detail.

Algorithm 1 Pseudocode of AntCrypt

Require: $t_cost > 0$, $m_cost > 0$, $outlen > 0$, $salt$, pw ,

Ensure: key

```
1: init( $salt$ ,  $pw$ )                                {Initialize  $state$ }
2: for  $i = 0$  to  $outer\_rounds$  do
3:    $update\_entropy()$                              {Distribute entropy over the state}
4:   # The following loop is referred to as  $update\_state()$ 
5:   for  $j = 0$  to  $inner\_rounds$  do
6:      $int\_update\_state()$                          {Waste time operating on  $state$ }
7:   end for
8: end for
9:  $compute\_output()$                                {Final output transformation}
```

Description of $init()$: The initialization function $init()$ fills the empty state memory with its initial content and is implemented in the reference implementation as the function $phs_init()$. Please note this function addresses the memory byte-by-byte, not as 32-bit words.

1. The $salt$ is copied to the beginning of the $state$ memory. It is interpreted byte-wise, i.e., $state[0] = salt[0]$, $state[1] = salt[1]$, \dots . We use a fixed size for the salt (16 bytes as suggested in the proposal), and do not see a reason for supporting variable sized salts: 16 bytes (128 bit) should offer sufficient security against appropriate attacks and we do not need to add any separator or length into the buffer when using a fixed length.
2. The password is appended to the array after the salt, also stored byte-per-byte. The maximum length of a password accepted is $state_bytes - 16 - 1$ bytes to leave enough space for the salt, the password and an end-identifier. The minimal state supported consists of 256 bytes. Thus, a 128 byte password as required will always be accepted.
3. A “password-end” marker $0x80$ is appended directly after the password and the remaining space is filled with $0x00$.

Description of $update_entropy()$ The function $update_entropy$ (which maps to $phs_upd_entropy()$ in the reference implementation) uses a hash function, hashing the entire $state$. As common hash functions have a much smaller output compared to $state$ – e.g., 128 bit for MD5 or 512 bit for SHA-512 – we need to extend these constructions to adapt for the larger output size. In the implementation, we use the $rehash$ buffer and its prefix to derive the new state.

Similar constructions are well-known in the cryptographic literature, and in the random oracle model it is easy to prove that the resulting function constitutes a secure hash function. Basically, we compute

$$h := H(state),$$

and then

$$s_i = H(i \parallel h)$$

and forming the next state as

$$s_0, s_1, \dots, s_k.$$

The “intermediate” hash value h is also used in the `compute_output()` function for an additional feature.

Description of `update_state()` The function `update_state()` accesses the buffer `state` `inner_rounds` \times `state_words` times and aims at wasting CPU cycles and efficiently slow down parallel computation on different platforms. In the reference implementation, this function is implemented as the function `phs_upd_state()`.

Algorithm 2 Pseudocode of `update_state()`

```
1: for  $i = 0$  to inner_rounds do
2:   for  $j = 0$  to state_words do
3:     res = (state[ $j$ ] ROR  $i$ )
4:     tgt_addr = res % state_words
5:     reset idx permutation
6:     for  $j = 0$  to #F do
7:       choose unused idx by evaluating res
8:       res =  $F_{idx}(\text{res})$ 
9:     end for
10:    state[tgt_addr] = state[tgt_addr] XOR res
11:  end for
12: end for
```

Algorithm 2 describes the update algorithm. We use a set of functions $F_i(x)$, where `#F` is the number of functions and use a calling sequence of these functions, where every function is called exactly once. After all `#F` functions process the data, the word at the target address is updated by using a bit-wise XOR.

Please note that the sequence is not pre-defined, but depends on the value `res` (and thus the initial value `state`[j]). Thus, in theory, all `#F!` sequences are possible.

The currently implemented functions (defined in `phc.h`) are given in Table 1. Please note that the functions are currently being evaluated and may be tweaked later.

Description of `compute_output()` The function `compute_output()` uses the `state` memory after the last outer round to generate the derived key material. It is implemented as `phs_gen_output()` in the reference implementation.

It consists of two steps, depending on the requested output length. If the output length is less or equal to 512 bit, i.e., the output length of SHA-512, only the first step is necessary.

```

/* integer operations */
#define F00(X) ( (X) + 0x01234567 )
#define F01(X) ( (X) * 0x89ABCDEF )

/* bit operations */
#define F02(X) ( (X) >> 3 )
#define F03(X) ( ROTR((X), 7) )
#define F04(X) ( (X) ^ 0x01234567 )
#define F05(X) ( (X) & 0xFEFEFEFE )
#define F06(X) ( (X) | 0x02020202 )

/* floating point operations */
#define F07(X) ( (uint32_t) ( 2147483648.L \
    * sin (((double) X)/1000000000.L ) ) )
#define F08(X) ( (uint32_t) ( 2147483648.L \
    * cos (((double) X)/1000000000.L ) ) )
#define F09(X) ( (uint32_t) ( 2147483648.L \
    * tan (((double) X)/5000000000.L ) ) )

/* 1/x: [1,2] -> [0.5, 1] (bijective) */
#define F10(X) ( (uint32_t) ( (double) ( 2 * 4294967296.L \
    * ( 1 / (1.5 + (double) X / 4294967296.L ) ) - 0.75 ) ) )

```

Table 1. List of the functions F_i used in `update_state()`.

First, we generate the intermediate hash, which would be generated during the next call to `update_entropy`. It basically is identical to the first step of `update_entropy()`, i.e., we apply the hash function to the entire state:

$$h := H(\text{state}).$$

Depending on the desired output length, we use up to 64 byte from h , addressing the buffer byte-wise and starting with byte 0.

In case more than 64 byte were requested, we use the `prefix` for the `state`, initialized with 1, and hash the full `state` including the `prefix` to derive a new intermediate value

$$h' := H(i \parallel \text{state}).$$

We use the same function used in `update_entropy()` to derive a new `state` from h' , overwriting the previous `state` and append up to `state_bytes` bytes to the output. This procedure can be repeated up to $2^{32} - 1$ times, effectively producing more than 2^{40+m_cost} bytes of key material.

This construction has another advantage: By storing the “intermediate” value h as final output, we are able to recompute the “next” state. This means that we can “resume” the computation of the state from a previously stored hash value, i.e., we can retroactively increase the hardness with respect to an increased parameter `t_cost` (cf. Section 3 for more details).

3 Design choices and remarks

Next, we comment on some of the design choices that underly our construction.

3.1 Implementation

One of our main intentions was to keep the overall structure and design as simple as possible, as this facilitates analysis and implementations. This also means we omitted some features from the implementation that are easy to add for a future (reworked) version. For the same reason we omitted most optimizations of the implementation and provide a rather straightforward implementation which is presumably easy to analyze. If selected for the second round we would provide an optimized version. The overall structure is very simple, with a clear distinction between the “cryptographically hard” step (`update_entropy()`), where we use established cryptographic primitives, and the “computationally hard” step (`update_state()`), where we are relatively free to do arbitrary computations that achieve our goals.

Some features that can easily be added (and will be added in future versions):

Parallelism There is a very easy modification to make the computation parallelizable for the honest server that computes the hash. Instead of processing each cell individually when computing the `update_state()`, we can read several (for example 16) consecutive cells, compute their output in parallel, and then write back simultaneously. (As we XOR the result on the target cell the order of writing does not matter.) This provides sufficient parallelism for the honest server, while not being advantageous for the attacker, as these parallel threads are still diverging.

Extending hardness Without further modifications, the above construction allows the legitimate server to increase the hardness of an existing hash without knowledge of the password, within certain constraints. It is necessary that the intermediate hash is stored in its entirety, i.e., the output has at least 512 bit. Furthermore, only the `t_cost` parameter can be increased (i.e., internally the `outer_rounds` parameter), the `m_cost` parameter needs to be fixed. Increasing the strength is very straightforward (and we will make code for doing so available in the near future). As the final step `compute_output()` is equivalent to the first part of the `update_entropy()` step (for an output length of 512 bit), we can simply resume the computation from this step on by first completing the second half, i.e., populating the entire `state` buffer from this value and then resuming with the iterations, adding so many iterations that the wanted iteration count is met, and finalizing with the final hashing.

3.2 Divergence and choice of the functions F_x

The specific choice of the functions F_i used in the construction depends on a number of factors, including the attacker’s compute architecture. We are still

evaluating different choices for these functions, so the currently selected functions are likely to change in future versions; any comments are appreciated.

Some important considerations are the following: If the functions take too long to compute, then an attacker can potentially queue them up to compute the same ones in parallel, thwarting the divergence of the threats. However, if they are too fast to evaluate, then the “overhead” imposed by the computations in the inner loop that are not part of the F_i ’s, e.g., computing the permutation, reduces the effectiveness of the divergence. (As computing the permutation incurs some substantial overhead, we consider using just a random sequence of indices; choosing a permutation, however, has the desirable property to rule out a number of timing side channel attacks as discussed later in this text.)

The overlap between different functions F_i , i.e., the potential to execute them in parallel, should be minimized; we attempted to achieve this by choosing functions with distinct assembler instructions, additionally ensuring that they are not easily transformable into each other. (Note the absence of the “bitwise invert” function, which can be expressed with an XOR.)

On using floating point operations We are aware that using floating point operations in such constructions is unusual, but we believe that they are helpful in minimizing the overlap, and they are also quite costly to implement on FPGAs and ASICs. We avoid rounding errors by converting back each result to an integer, thus being able to control any potential rounding error. But again, the specific choice of the F_x is still somewhat experimental, and we might opt to remove floating point instructions if they incur problems with portability.

4 Security

4.1 Cryptographic security

Our construction inherits its cryptographic strength quite directly from the security of the underlying hash function. We describe our construction using SHA-512, but it can be easily substituted with any other hash function with sufficiently large state/output size. We have selected SHA-512 as it is a widely accepted design which has proven security over several years, and implementations are easily available in common libraries. In fact, it should be straight-forward to prove (in the random oracle model) that, provided that the functions F_x are permutations, or at least behave “sufficiently random”, then the overall construction behaves like a random function.

In general, constructing secure hash functions is a delicate matter, and large efforts have gone into the design of such functions. Therefore, we feel that it is mandatory to rely on well-established constructions to achieve cryptographic security instead of attempting to use home-made constructions. One of the beauties of our construction is that we separate the task of providing *cryptographic strength* from the task of *slowing down verification*, (cryptographic strength is largely realized by the re-hashing done in `update_entropy`, putting minimal

requirements on `update_state` only, while the slow-down is largely realized in `update_state`).

The only thing that is required to really inherit these properties is that the applying the step `update_state` does not lose too much entropy. However, by our construction, applying the sequence of the F_x to the current state is always a permutation, as we XOR the output to the target value. (This is very similar to the well-known Feistel structure, which also always is a permutation for arbitrary round functions.) And if one application of the sequence is a permutation, then by repeating this argument, the entire function `update_state` constitutes a permutation.

4.2 Speed up

The intended use of function `update_state` is to slow down the computation of the password hash, thus this is the critical place to look for optimizations.

On CPUs On CPUs, we believe that only minor optimizations can be done. The pseudo-random nature of the order of applying the functions F_x means that there is very little (constant) structure that can be exploited for optimizations. Note that, when looking retroactively at one particular run, there will be structure that can potentially be exploited, however, as the structure changes for each application of the permuted chain of F_x 's such structure needs to be detected during runtime. As the functions F_x are very short (ranging from a single assembler instruction up to a few), we believe that code for detecting and exploiting such structure would most likely slow down the execution more than it helps in speeding up.

Also note that we plan to consider other functions F_x in the future, and we hope to be able to provide a more formal argument regarding the potential speed up in future versions of this document.

On GPUs On GPUs, these random permutations will lead to a substantial amount of branch divergence, which means that the parallel executions of the hash function for a parallel brute force attempt (running for different passwords) will have divergent control flow. For “ideal” functions F_x with no overlap, no overhead outside the F_x , and ideally random selection, we would expect a slow-down equal to the number of functions, i.e., by a factor of 16. (Here “slowdown” is comparing the runtime for the case with convergent execution, e.g., all threads hashing the same password, with the runtime for divergent threads, e.g., when hashing different passwords in each thread.) In practice, there is overhead, e.g., caused by the final XOR and the computation of the permutation, and the functions have not entirely disjoint assembler instructions (e.g., we need some re-scaling of the values for the floating point instructions), so these ideal goals will likely not be met.

On FPGAs/ASICs While FPGAs and ASICs are very dangerous in terms of efficient implementation of brute-force attacks, the construction was chosen to render dedicated hardware attacks almost useless.

While many of the functions are easily mapped to hardware, floating point operations come at a high price. We analyzed the available cores for Xilinx Spartan 6, Virtex 6 and Artix 7 devices^{**}. The CORDIC-core offers sine and square-root with 8 to 48 bit operands. For 32-bit operation, the minimum area is 3664 LUTs and 3588 FFs for sine (Spartan-6) and 975 LUTs and 1202 FFs for square-root (Artix-7) and has a latency of more than 32 clock cycles.

The floating-point core provides addition/subtraction, division, square-root and multiplication with configurable latency (time-area tradeoff) and may use available DSP cores, and the area consumption is heavily dependent on these configurations.

The use of more than one FPU function will significantly increase the area and latency of the generation on FPGAs. In addition, to support all possible sequences of the F_i functions, the complexity of the routing will increase dramatically: Every output needs to be routed to every other function as input. Thus, every function needs a large multiplexer, increasing the routing delay and increasing the critical path.

The second limiting factor is the memory usage, as fast memory cores are available but limited in size and number. To implement a 64 kByte state (`m_cost` = 8) will already use about 29 18k-BRAMs on Xilinx FPGAs. Thus, the memory area will become a limiting factor even with medium state sizes.

In practice, we think that using FPGAs or producing dedicated ASICs will not be the first choice for an attacker, as the construction is by design very cumbersome to implement and artificially adds latency, enforces complex routing and needs area-consuming FPU arithmetic.

4.3 Side channel attacks

Storing the password in memory The password is written to the `state` in the beginning and immediately overwritten by the output of the hash function. No copy needs to be stored beyond the initialization of the `state` memory. This should effectively prevent reading the password from memory.

(Cache) timing attacks The different functions F_x have usually, depending on the platform, different execution times. This could lead to timing attacks or to cache timing attacks. However, as we always use the same functions, just in differently permuted order, the time between memory access is constant (assuming that each operation runs in time independent of the data). In other words, the execution time between memory access is constant, thus no information is leaked. Then also the overall running time is constant, and no timing leak exists.

Other side channel attacks More involved side channel attacks, such as power consumption and electromagnetic emanation, are outside the scope of our consideration, as they depend on the specific architecture the code is running on. Also, they do not seem appropriate for the case considered, as an adversary that

^{**} cf. Xilinx DS858 and Xilinx DS335 specification

has physical access to a machine verifying the correct password typically has easier attacks at hand.

5 Final remarks

5.1 Statements

We ensure that we have not inserted, and are not aware of, any deliberately hidden weaknesses in the scheme described above.

The scheme is and will remain available worldwide on a royalty free basis, and we are unaware of any patent or patent application that cover the use or implementation of the submitted algorithm.

6 Test vectors

In this section, we will provide several testvectors. Please note that the requested list of $2^8 \times 2^8$ password/seed combinations for meaningful cost factors will take a very long time to generate. To do this, please compile the included source code and run the program `phc_tv`.

The format of the output is similar to previously used formats:

```
$<seed>$<t_cost>:<m_cost>$<hash>
```

where the cost parameters are represented as two-digit numbers and the seed and hash are in hex-representation.

The output listed in Tables 2, 3, 4 and 5 was generated by the `phc_demo` program, which is also provided as source code.

