Groth16 Implementation Specification in BitVM2

Fiamma [1]

August 2024

Abstract

In this paper, we present our methodology for implementing Groth16 verification within the Bitcoin framework, utilizing the BitVM2 repository. We extend our gratitude to contributors such as Robin, Lukas, Weikeng, and the Zerosync team for their invaluable contributions to this project. The objectives of this paper are threefold:

- Peer Review: We invite the BitVM2 community to review our design and implementation, providing feedback to enhance its robustness and efficiency.
- Knowledge Dissemination: We aim to provide a detailed exposition on the integration of BitVM2 with Groth16 to enlighten further developers and researchers about its operational mechanics.
- Community Collaboration: By sharing our findings, we seek to catalyze collaborative development efforts within the community to expedite the safe production readiness of BitVM2.

Through this publication, we hope to contribute significantly to the evolution and adoption of BitVM2, fostering a safer and more efficient deployment in production environments.

Contents

1 Basic Data

This section outlines key details about Fiamma to date and will be updated as needed.

- ZK algorithm: Groth16 [9]
- Script size: 2.605447015 G
- Subscript number: 994
- Max subscript szie: 3.710443 M
- Min subscript szie: 0.562906 M
- Average subscript size: 2.621174 MB
- Max input size: 1152 Bytes
- Max output size: 576 Bytes
- Signature: Winternitz
- Hash: Blake3

2 Basic Theory

This section outlines essential basic theories used in our implementation.

- Groth16 [9]: The algorithm that we are implementing currently
- On Proving Pairing [8]: an efficient solution for implementing ZKP on Bitcoin combined with BitVM2 [3]
- Non-deterministic computation: an efficient way to reduce the script size in terms of verification point
- Affine module: an efficient way to decrease the script size in terms of computation
- Montgomery reduction and Karatsuba multiplication: an efficient way to reduce the script size of big integer multiplication

2.1 Groth16 verification program

The verification progress of Groth16[9] proceeds as follows:

$$0/1 \Leftarrow Vrf(R, \sigma, a_1, ..., a_l, \pi) : Parse\pi = ([A]_1, [C]_1, [B]_2) \in G_1^2, G_2 \tag{1}$$

Accept the proof if and only if

$$[A]_1 \cdot [B]_2 = [\alpha_1] \cdot [\beta]_2 + \sum_0^l a_i [\frac{\beta\mu_i(x) + \alpha v_i(x) + \omega_i(x)}{\gamma}]_1 \cdot [\gamma]_2 + [C]_1 \cdot [\delta]_2 \tag{2}$$

Let

$$[msm]_1 = \sum_0^l a_i [\frac{\beta\mu_i(x) + \alpha v_i(x) + \omega_i(x)}{\gamma}]_1 \tag{3}$$

It should be noted that $a_0 = 1$

2.2 On proving pairing

This is an efficient way to prove correctness of [8], the algorithm shows in page 25: Algorithm 9: Multi Miller loop with embedded c exponentiation.

If we integrate this algorithm into Groth16, the entire process would be structured as follows:

$P_1 = [msm]_1; Q_1 = -[\gamma]_2$

$P_2 = [C]_1; Q_2 = -[\delta]_2$
$P_3 = [\alpha]_1; Q_3 = -[\beta]_2$
$P_4 = [A]_1; Q_4 = [B]_2$
Q_4 is non-fixed, Q_1, Q_2, and Q_3 is fixed.

Input: $A = [(P_1, Q_1), (P_2, Q_2), (P_3, Q_3), (P_4, Q_4)], c, c^{-1} \in F_{q^k}, s \in F_{q^3}, P_{Q_4} \leftarrow \mathcal{L}(Q_4)$

Output: $1\ if \prod_{i=0}^{n} e(P_i, Q_i) = 1$

```
assert $c \cdot c^{-1} = 1$
$f \leftarrow c^{(} - 1), lc \leftarrow 0$
Initialize array $T$ such that $T[4] = Q_4$
for $i$ = $L - 2$ to 0 do
    $f$ = $f^2$
    for j = 1 to n do
        $l \leftarrow P_{Q_4}[lc]$
        $f = f \cdot l.evaluate(P_4)$
        $Q_4$ is not fixed then
            $T \leftarrow T[4]$
            assert $l.is_{t}angent(T)$
            $T[4] = l.double(T)$
        end
        if $bit^2 == 1$ then
            $f = f \cdot c^{-1}$ if $bit$ == 1 else $f \cdot c$ end
            $l \leftarrow P_{Q_4}[lc + 1]$
            $f = f \cdot l.evaluate(P_4)$
            $Q_4$ is not fixed then
                $Q^{'} = Q_4$ if $bit$ == 1 else $-Q_4$
            $T \leftarrow T[4]$
            assert $l.is_{l}ine(T, Q^{'})$
            $T[4] = l.add(T, Q^{'})$
            end
        end
    end
    $lc = lc + 2$
    $f \leftarrow f \cdot s \cdot (c^{-1})^q \cdot (c^{-1})^{q^2} \cdot (c^{-1})^{q^3}$
    for j = 0 to n do
        $l_{1..3} \leftarrow (P_{Q_j}[lc + i])_{i=0}^{2}$
        $f \leftarrow f \cdot l_1.evaluate(P_j) \cdot l_2.evaluate(P_j) \cdot l_3.evaluate(P_j)$
        $Q_4$ is not fixed then
            $Q_1 \leftarrow \pi_p(Q), Q_12 \leftarrow \pi_p(Q_1), Q_3 \leftarrow \pi_p(Q_2)$
            $T \leftarrow T[4]$
            assert $l_1.is_line(T, Q_1); T \leftarrow T + Q_1$
            assert $l_2.is_line(T, -Q_2); T \leftarrow T - Q_1$
            assert $l_3.is_line(T, Q_3)$
        end
    end
end
return $f == 1?$
```

2.3 Non-deterministic computation

We reference this concept from cairo-vm [5]. Simply put, the prover may do additional work that is not part of the proven computation.

Consider the task of computing a square root of a number x as part of a larger function. The deterministic approach is to use some square-root algorithm to compute $y = \sqrt{x}$ which means that we need to include its execution trace in our script. In contrast, the nondeterministic approach is much more efficient: the prover computes the square-root y using the same algorithm but does not include this computation in the script. Instead, the script only need to prove that is that y2 = x, which can be accomplished using a single SQUARE script gadget.

We adapt this concept with double and add operations based on G_1 and G_2

2.3.1 Double in G_1 and G_2

- Show that the pair (λ, μ), indeed defines a tangent through T, showing that $y_1 - \lambda x_1 - mu = 0$ and $2\lambda y_1 = 3x_1^2$. This step is dominated by $2\tilde{m}$ and one $\tilde{s}$
- Compute λ^2 which is simple one $\tilde{s}$
- Compute $x_3 = \lambda^2 - 2x_1$ and $2\lambda y_3 = -\mu - \lambda x_3$ which is dominated by computing λx_3

2.3.2 Add in G_1 and G_2

- Show that the pair (λ, μ), indeed defines a tangent through T, showing that $y_1 - \lambda x_1 - mu = 0$ and $y_2 - \lambda x_2 - mu = 0$. This step is dominated by $2\tilde{m}$
- Compute λ^2 which is simple one $\tilde{s}$
- Compute $x_3 = \lambda^2 - x_1 - x_2$ and $2\lambda y_3 = -\mu - \lambda x_3$ which is dominated by computing λx_3

2.4 Popular coordinate systems

Pairings can be computed over elliptic curves represented in any coordinate system, but popular choices have been homogeneous projective and affine coordinates, depending on the ratio between inversion and multiplication.

2.4.1 Homogeneous projective coordinates

The choice of projective coordinates has proven particularly advantageous at the 128-bit security level for single pairing computation, due to the typically high inversion/multiplication ratio in this setting. The tangent line evaluated at $P = (x_P, y_P)$ can be computed with the following formula:

$$g_{2\phi(T)}(P) = -2YZy_p + 3X^2x_pw + (3b^{'}Z^2 - Y^2)w^3 \tag{4}$$

2.4.2 Affine coordinates

The choice of affine coordinates has proven more useful at higher security levels and embedding degrees, primarily due to the norm map's role in simplifying the computation of inverses at higher extensions. The main advantages of affine coordinates lie in their ease of implementation and the straightforward format of the line functions. These features enable faster accumulation within the Miller loop, particularly when additional sparsity is exploited.

If $T = (x_1, y_1)$ is a point in $E^t(F_{p2})$, one can compute the point $2T := T + T$ with the following formula:

$$(1/y_p)g_{2\phi(T)}(P) = 1 + (-x_p/y_p)\lambda w + (1/y_p)(\lambda x_1 - y_1)w^3 \tag{5}$$

You can read this paper [12] for more details on this topic. We extend our heartfelt thanks to all contributors of the PR [10] in BitVM2 [3].

2.5 Big integer multiplication

2.5.1 Montgomery reduction

Montgomery reduction [6], also known as REDC, is an algorithm that simultaneously computes the product by R' and reduces modulo N more quickly than the naïve method. Unlike conventional modular reduction which focuses on making the number smaller than N, Montgomery reduction aims at making the number more divisible by R. It achieves this by adding a sophisticatedly chosen small multiple of N to cancel the residue modulo R. Dividing the result by R yields a much smaller number. This number is so much smaller that it is close enough to the reduction modulo N, and computing the reduction modulo N requires only a final conditional subtraction. Because all computations are done using only reduction and divisions with respect to R, not N, the algorithm runs faster than straightforward modular reduction by division.

2.5.2 Karatsuba multiplication

Karatsuba multiplication [7] is a divide-and-conquer algorithm for (non-modular) multiplication that reduces the computational cost for n-bit integers from $O(n^2)$ in classical multiplication to $O(n^{\log_2^3})$ which is $O(n^{1.58 \cdots})$.

When applied to modular exponentiation with n-bit numbers, including the exponent, the cost is reduced from $O(n^3)$ to $O(n^{1+log_2^3})$ which is $O(n^{2.58 \cdots})$. One of several methods to leverage the benefits of Karatsuba multiplication during modular reduction is to pre-compute the (non-modular) inverse of the modulus to slightly more than n bits, achievable at a cost of $O(n^2)$, this is considered negligible in the context of O, using classical algorithms.

Karatsuba multiplication can be used in conjunction with Montgomery reduction. We highly appreciate the work [11] of Robin and the Zerosync team, which has reduced for multiplication to approximately 55% of the native version.

3 Script

3.1 Split basics

This section outlines essential basic knowledge used in our implementation.

3.1.1 Signature hash type

A SIGNHASH [1] flag is used to indicate which part of the transaction is signed. The mechanism provides a flexibility in constructing transactions. There are in total 6 different flag combinations that can be added to a digital signature in a transaction.

- 0x01 = SIGHASH_ALL : Sign all inputs and outputs
- 0x02 = SIGHASH_NONE : Sign all inputs and no output
- 0x03 = SIGHASH_SINGLE : Sign all inputs and the output with the same index
- 0x81 = SIGHASH_ANYONECANPAY | SIGHASH_ALL : Sign its own input and all outputs
- 0x82 = SIGHASH_ANYONECANPAY | SIGHASH_NONE : Sign its own input and no output
- 0x83 = SIGHASH_ANYONECANPAY | SIGHASH_SINGLE : Sign its own input and the output with the same index

3.1.2 Transaction structure

The transaction structure of Bitcoin is illustrated in the following table.

Field	Size	Description
Version	4 bytes	The version number for the transaction. Used to enable new features
Maker	1 bytes	Used to indicate a segwit transaction. Must be 00
Flag	1 bytes	Used to indicate a segwit transaction. Must be 01 or greater
Input Count	Variable	Indicates the number of inputs
Input-TXID	32 bytes	The TXID of the transaction containing the output you want to spend
Input-VOUT	4 bytes	The index number of the output you want to spend
Input-ScriptSig Size	Variable	The size in bytes of the upcoming ScriptSig
Input-ScriptSig	Variable	The unlocking code for the output you want to spend
Input-Sequencer	4 bytes	Set whether the transaction can be replaced or when it can be mined
Output Count	Variable	Indicates the number of outputs
Output-Amount	8 bytes	The value of the output in satoshis
Output-ScriptPubKey Size	Variable	The size in bytes of the upcoming ScriptPubKey
Output-ScriptPubKey	Variable bytes	The locking code for this output
Witness-Stack Items	Variable	The number of items to be pushed on to the stack as part of the unlocking code.
Witness-Stack Items-Size	Variable	The size of the upcoming stack item
Witness-Stack Items-Item	Variable	The data to be pushed on to the stack
Locktime	4 bytes	Set a time or height after which the transaction can be mined

The blue part means it will be stored in the segwit part. Anyone could check more details in [13]

3.1.3 Block size calculation

Understanding the calculation of Bitcoin block size following the Taproot upgrade is essential.

Block size The following illustration depicts the block size calculation:

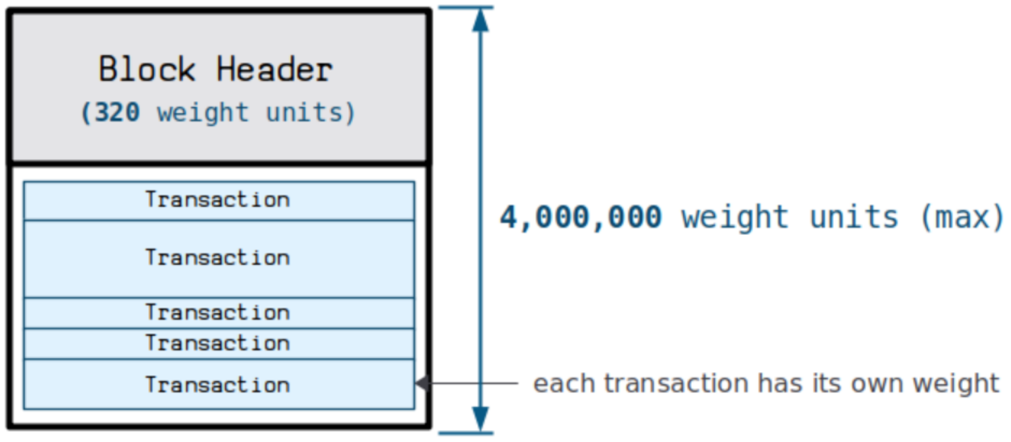


Figure 1: Block size

Transaction size The transaction size calculation is illustrated below:

For more detailed information, please refer to [14]

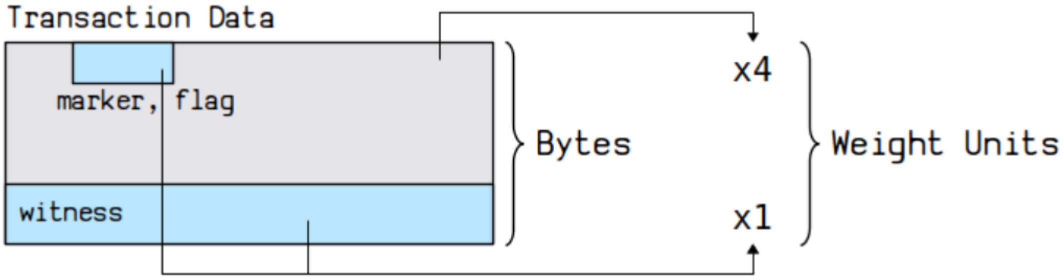


Figure 2: Transaction size

Script chunk limitation Based on the design of BitVM2 [3], We aim for each script chunk to be packed into one block as a transaction. So, the transaction size could not exceed 4,000,000 - 320 = 3,999,680 weight units [14].

A disputed transaction, characterized by 1 input and 2 outputs, typically has an average non-witness data size of approximately 464 weight units. Consequently, the witness size limitation for such transactions is calculated as 3,999,680 - 464 = 3,999,216 weight units.

As outlined in BitVM2 [3], the disputed transaction needs the signature of Committee, and the signature type is SIGNHASH_SINGLE. Let's assume that the number of Committee is 7 and the size of each schnorr signature is 65 bytes (64 bytes for SIGNHASH_ALL)

So the limitation will be 3,999,216 - 7 * 65 - 8(stack item size) = 3,998,753 weight units.

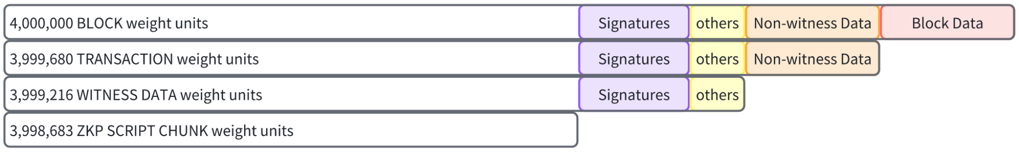

4,000,000 BLOCK weight units
Signatures
others
Non-witness Data
Block Data
3,999,680 TRANSACTION weight units
Signatures
others
Non-witness Data
3,999,216 WITNESS DATA weight units
Signatures
others
3,998,683 ZKP SCRIPT CHUNK weight units

Figure 3: ZKP script chunk limitation

As for the structure of subscript, the normal flow should be as follows:

- check the hash [4] is consistent with the input and output
- check the Winternitz [15] signature
- execute the subscript

So, if we take these factors into account, the subscript structure should be like the following picture

It should be noted that: we need to reduce the maxmial input and putput size of subscript less than 1024 to invoke at most 2 rounds blake3 [4] hash progress

3.1.4 Limitations

Several constraints must be considered:

- Max script size: 4MB
- Max stack depth: 1000 (main stack and alternate stack combined);
- Max stack item size: 520 bytes
- Max input size for arithmetic opcodes: 32-bit words
- Max input bytes for Blake3: 512 Bytes

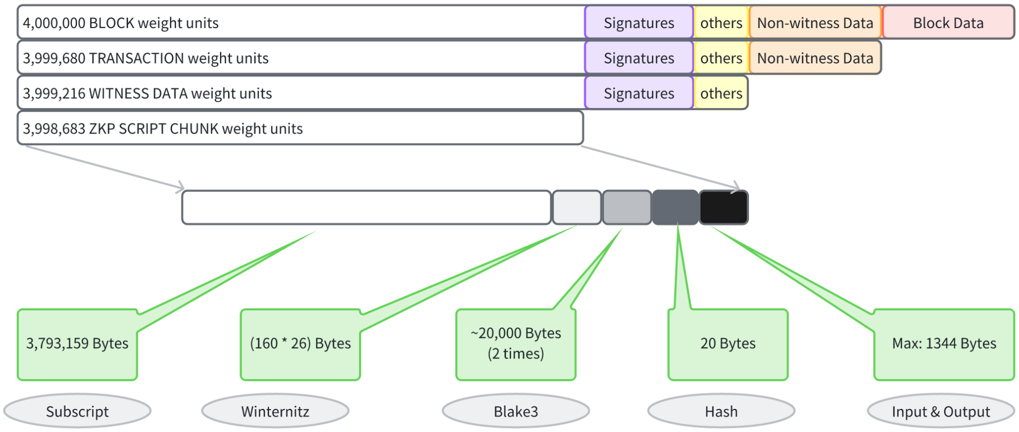


Figure 4: structure of subscript

3.2 Split pricinple

In this section, we will primarily discuss why we chose a manual method for splitting the ZKP verification script.

3.2.1 Automated

Automating the splitting of the entire computation is appealing. Ideally, the whole process shoule be similar to the following figure:

The overall flow shoule be as follows:

- The program will be complied into a set of customized gadgets first;
- Each gadget will correspond with a script gadget;
- The accmulator will begin to split all gadgets;
- Becasue each script gadget has a fixed size, these gadgets will be spilt as one chunk when their accumulated size is almost equal to 4M;
- Node A executes the script program locally to generate input and output for each chunk;
- All the input and output locate in stack, so Node A has to commit all the value in the stack;

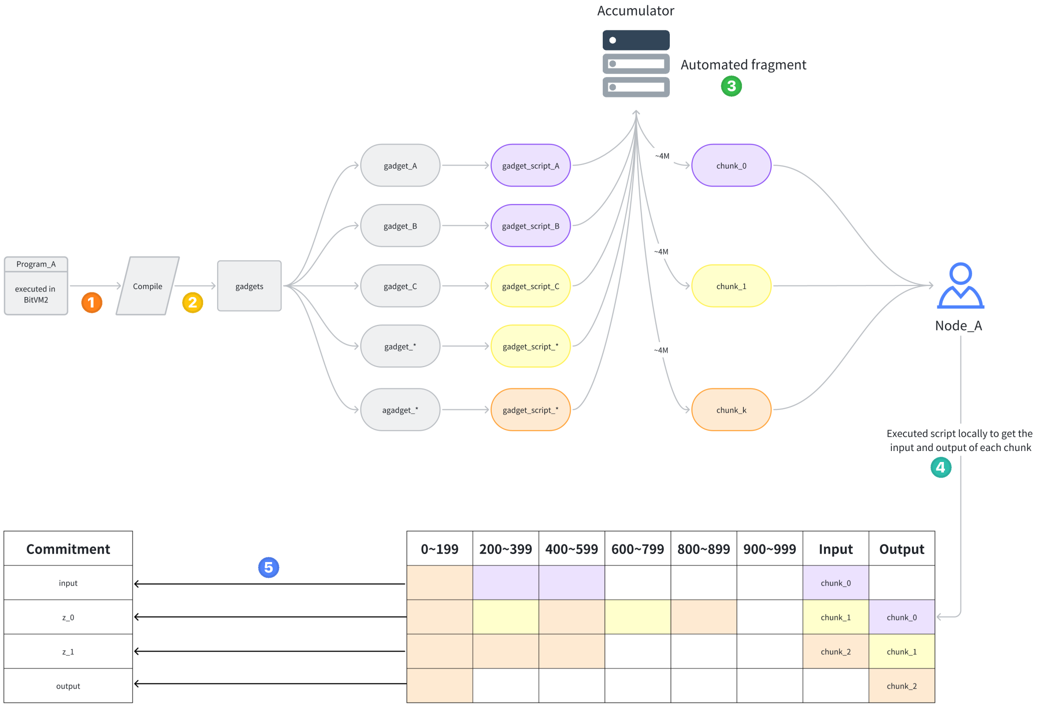
It is important to note that stack depth is another factor that should be considered. In the automated approach, there are several constraints:

- It is easy to exceed the stack depth limitation;
- It commits many values which will not be used in the current chunk;
- It must implememnt enough gadgets to support any computation, which means achieving turing completeness;
- Executing a large scirpt program is much slower;
- It increases the costs when verifying the expected input and output on-chain;
- The logic of each chunk is unreadable;

However, automated fragmentation has its advantages as well, such as generating the minimal number of script chunks. But since we do not upload all chunks to the Bitcoin network, the number of chunks is less of a concern unless their size becomes excessively large.

3.2.2 Manual

Why do we select a manual way to split the entire program?



Commitment
input
z_0
z_1
output

0~199	200~399	400~599	600~799	800~899	900~999	Input	Output
🟧	🟪	🟪				chunk_0	
🟧	🟨	🟧	🟨	🟧		chunk_1	chunk_0
🟧	🟧	🟧				chunk_2	chunk_1
🟧							chunk_2

Figure 5: automated fragment

- We avoid exceeding the stack depth limitation by only placing necessary data for the current chunk on the stack;
- We commit only the data used in the current chunk;
- We just need to implement gadgets to support ZKP verification as any computaion could generate a ZK proof;
- We use a Rust program to generate the input and output for each chunk;
- This approach minimizes costs when verifying the expected input and output on-chain;
- The logic of each chunk is readable.

While this approach may generate more chunks, as mentioned earlier, only one script chunk is executed on Bitcoin, so this is acceptable.

The overall flow proceeds as follows:

- We first concurrently implement the Rust and script versions of the Groth16 ZKP verification;
- The rust version includes the witness generation of each chunk;
- The script version includes all split chunks;
- We ensure each chunk meets the size and depth constraints;
- Node A executes the Rust program locally to generate inputs and outputs for each chunk;
- Node A commits all the inputs and outputs.

3.3 Split data

We present the results directly on how we split the script whose size exceeds the 4M limitation, as demonstrated in **??**, there are only some operations of F_{q12} that need to be split after we optimize the operations for G_1 and G_2

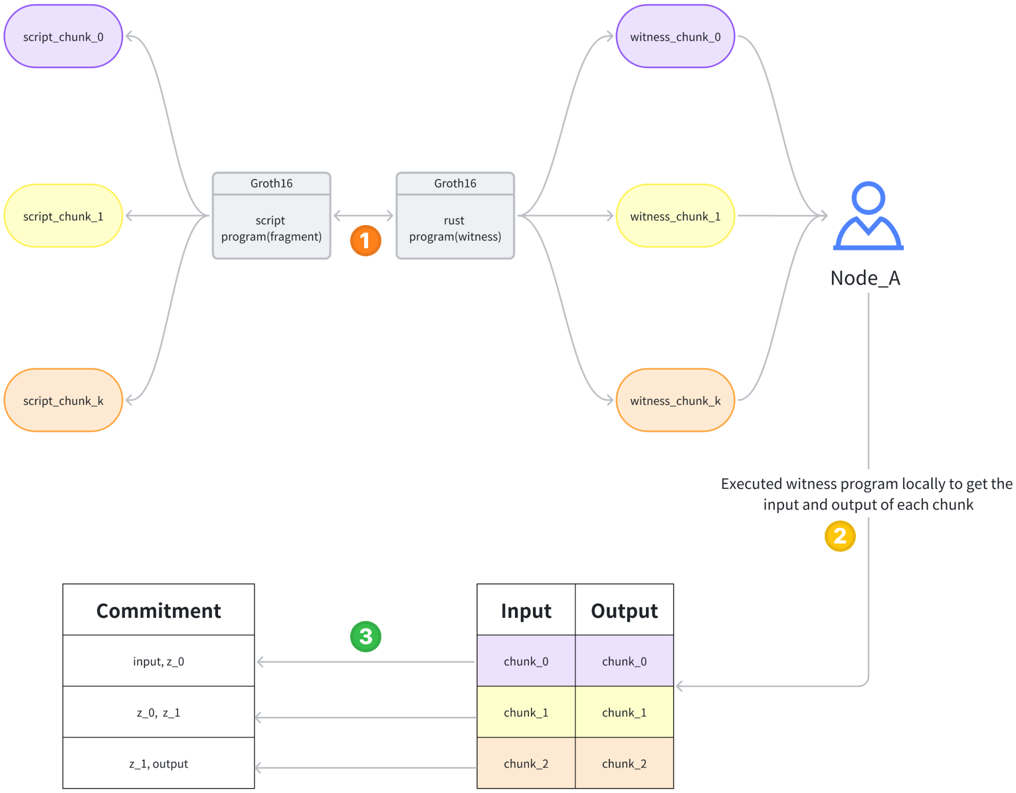


Figure 6: manual fragmentation

operator type	script size	max depth	exceed 4M?
$F_{q12} : a * b$	11,641,775 bytes	545	yes
$F_{q12} : \ frobenius\_map(1)$	4,541,887 bytes	-	yes
$F_{q12} : \ mul\_by\_034$	9,810,459 bytes	-	yes
$F_{q12} : \ ell\_by\_constant$	9,525,050 bytes	383	yes

We will demonstrate how we manually split these four large scripts one by one, striving to satisfy the following properties concurrently:

- Ensuring that size and stack depth limitations are not exceeded;
- Minimizing the size of inputs and outputs;
- Making the logic of each chunk as clear and readable as possible;

3.3.1 Split code

We will give an example to explain how the split process works based on $Fn \ \ double\_ell\_affine()$, there have 2 types of code:

- Split_code: it will generate all subscripts;
- Witness_code: it will generate all inputs and outputs for all subscripts

Split_pseudocode

```
pub fn split_double_ell_affine() -> Vec<Script> {
        let mut res = vec![];

        // [p.c0, p.c1, c3, c4, lamda, mu, Q.x, Q.y, 1]
        // [p.c0, p.c1, c3, c4, x3, y3]
        // [final_c0] | [y3, x3]
        // [final_c0, x3, y3]
        res.push(script! {

            { Pairing::double_line_g2_optimization() }

            { Fq2::toaltstack() }
            { Fq2::toaltstack() }

            // [p.c0, p.c1, c3, c4,]
            // compute b = p.c1 * (c3, c4)
            { Fq6::mul_by_01() }
            // [p, c3, c4, b]

            // [p.c0, b]
            { Fq12::mul_fq6_by_nonresidue() }
            // [p.c0, b * beta]

            // compute final c0 = a + beta * b
            { Fq6::add(6, 0) }

            { Fq2::fromaltstack() }
            { Fq2::fromaltstack() }

        });

        // [c0, c3, c4, b, a, p.c0, p.c1, lamda, p.x, -mu, p.y]
        // [c0, c3, c4, b, a, p.c0, p.c1, p.y, -mu, lamda * p.x]
        // [c0, c3, c4, b, a, p.c0, p.c1, lamda * p.x, p.y * -mu]
        // [c0, c3, c4, b, a, p.c0 + p.c1]
        // [c3, c4, b, a, e, c0]
        // [c4, b, a, e, c0, c3]
        // [c4, b, a, e, c0 + c3]
        // [b, a, e, c0 + c3, c4]
        // [b, a, e]
        // [e, a + b]
        // [final_c1]
        res.push(script! {

            { Fq2::mul_by_fq(1, 0) }
            { Fq2::copy(29) }
            { Fq2::equalverify() }
            { Fq2::mul_by_fq(1, 0) }
            { Fq2::copy(28) }
            { Fq2::equalverify() }
            { Fq6::add(6, 0) }
            { Fq2::roll(22) }
            { Fq2::roll(22) }
            { Fq2::add(2, 0) }
            { Fq2::roll(20) }
            { Fq6::mul_by_01() }
            { Fq6::add(12, 6) }
            { Fq6::sub(6, 0) }
        });
        unsafe {
            FQ12_SPLIT_DOUBLE_ELL_COUNTER += 1;
        }
        res
    }
```

Witness_code

```
    pub fn witness_add_ell_affine(
        f: &Fq12,
        p1: &ark_G1Affine,
        affi_t4: &G2Affine,
        affi_q4: &G2Affine,
    ) -> (Fq12, Vec<ScriptContext>, G2Affine) {
        let mut contexts = vec![];
        //////////////////// script-1 ///////////////
        // non-fixed (non-constant part) P4
        let (lambda, mu, x3, y3) = Self::line_add_g2(affi_t4, affi_q4);

        let affi_t4_new = G2Affine::new(x3, y3);

        let mut p = ark_G1Affine::new(p1.x, p1.y);

        p.x = -p1.x / p1.y;
        p.y = Fq::from(1) / p1.y;
        //////////////////// script-1 ///////////////
        let mut context = ScriptContext::default();

        let c0 = ark_bn254::Fq2::from(1);

        let mut c3 = lambda.clone();
        c3.mul_assign_by_fp(&p.x);

        let mut c4 = -mu.clone();
        c4.mul_assign_by_fp(&p.y);

        let mut b = f.c1 as Fq6;
        b.mul_by_01(&c3, &c4);

        let mut a = f.c0 as Fq6;
        a.mul_by_fp2(&c0);
        let mut beta_b = b;
        <Fq12Config as Fp12Config>::mul_fp6_by_nonresidue_in_place(&mut beta_b);

        let final_c0 = a + beta_b;

        // fill inputs for subscript - 1
        context
            .inputs
            .extend(f.c0.to_base_prime_field_elements().collect::<Vec<Fq>>());
        ...
        // fill outputs for subscript - 1
        context
            .outputs
            .extend(final_c0.to_base_prime_field_elements().collect::<Vec<Fq>>());
        ...

        //////////////////// script-2 ///////////////
        let mut context = ScriptContext::default();
        // fill inputs for subscript - 2
        context
            .inputs
            .extend(c0.to_base_prime_field_elements().collect::<Vec<Fq>>());
        ...

        let e = f.c0 + f.c1;
        let mut e = e;
        let c0_c1 = c0 + c3;
        e.mul_by_01(&c0_c1, &c4);

        let a_b = a + b;
        let final_c1 = e - a_b;

        // fill outputs for subscript - 2
        context
            .outputs
            .extend(final_c1.to_base_prime_field_elements().collect::<Vec<Fq>>());
        contexts.push(context);

        (Fq12::new(final_c0, final_c1), contexts, affi_t4_new)
    }
```

3.3.2 Split result

We present the split result directly in the following table:

operator typr	subscript	script size	max depth	exceed 4M?
$F_{q12} : a \ * \ b$		6,727,971 bytes	545	yes
	chunk0	3,085,202 bytes	545	no
	chunk1	3,642,594 bytes	545	no
$F_{q12} : \ a \ * \ a$		4,495,790 bytes	-	yes
	chunk0	2,246,763 bytes	545	no
	chunk1	2,249,027 bytes	545	no
$F_{q12} : \ doube_ell$		6,416,951 bytes	-	yes
	chunk0	3,708,344 bytes	545	no
	chunk1	2,708,607 bytes	545	no
$F_{q12} : \ add_ell$		6,415,463 bytes	-	yes
	chunk0	3,706,856 bytes	545	no
	chunk1	2,708,607 bytes	545	no
$F_{q12} : \ ell\_by\_constant$		4,714,973 bytes	383	yes
	chunk0	2,568,521 bytes	545	no
	chunk1	2,145,297 bytes	545	no

Note: we combine Fq12_ell and double_g2, add_g2 operation to get a smaller subscript number

3.4 Benchmark data

This section primarily provides benchmark data for various operators used in the Groth16 verification process.

3.4.1 Operator script size origin

We will first present some initial benchmark data from our current implementation, including:

- Double and Add operators in G_1 group;
- Double and Add operators in G_2 group;
- Field operators in extension field;

G1 group

operator type	script size	max depth	exceed 4M?
$2 \cdot g_1$	1,752,916 bytes	131	no
$g_1 \cdot g_1^{'}$	3,997,319 bytes	< 1000	no

G2 group

operator type	script size	max depth	exceed 4M?
$2 \cdot g_2$	7,019891 bytes	815	yes
$g_2 \cdot g_2^{'}$	9,270,854 bytes	293	yes

Field

operator type	script size	max depth	exceed 4M?
$F_{q12} : a + b$	6,644 bytes	220	no
$F_{q12} : 2 * a$	6,793 bytes	217	no
$F_{q12} : a * b$	11,641,775 bytes	545	yes
$F_{q12} : a * a$	7,772,080 bytes	545	yes
$F_{q12} :\ mul\_fq6\_by\_nonresidue$	4,923 bytes	146	no
$F_{q12} :\ frobenius\_map(1)$	4,541,887 bytes	-	yes
$F_{q12} :\ frobenius\_map(2)$	2,224,363 bytes	-	yes
$F_{q12} :\ mul\_by\_034$	9,810,459 bytes	-	yes
$F_{q12} :\ ell\_by\_constant$	9,525,050 bytes	383	yes
$F_{q6} :\ a * b$	3,873,847 bytes	275	no
$F_{q6} :\ frobenius\_map(1)$	1,518,206 bytes	-	no
$F_{q6} :\ frobenius\_map(2)$	598,274 bytes	-	no
$F_{q6} :\ mul\_by\_01\_with\_1\_constant$	3,280,529 bytes	221	no
$F_{q6} :\ mul\_by\_fp2\_constant$	1,520,337 bytes	101	no
$F_{q6} : mul\_by\_01$	3,769,633 bytes	-	no
$F_{q6} :\ mul\_by\_fp2$	2,252,362 bytes	167	no
$F_{q2} :\ a * b$	750,883 bytes	113	no

3.4.2 Operator script size optimization

The fewer script chunks, the better. Therefore, before we split the large operators, we aim to optimize them first. We will present our new data initially, followed by an explanation of the principle.

- Double and Add operators in G_1 group;
- Double and Add operators in G_2 group;
- Field operators in extension field;

G1 group

operator type	script size	optimized script size	exceed 4M?
$2 \cdot g_1$	1,752,916 bytes	699,519 bytes	no
$g_1 \cdot g_1^{'}$	3,997,319 bytes	562,445 bytes	no

G2 group

operator type	script size	optimized script size	exceed 4M?
$2 \cdot g_2$	7,019,891 bytes	1,847,059 bytes	yes
$g_2 \cdot g_2^{'}$	9,270,854 bytes	1,563,501 bytes	yes

Field

operator type	script size	optimized script size	exceed 4M?
$F_{q12} : a + b$	6,644 bytes	6,644 bytes	no
$F_{q12} : 2 * a$	6,793 bytes	6,793 bytes	no
$F_{q12} : a * b$	11,641,775 bytes	**6,727,971** bytes	yes
$F_{q12} : a * a$	7,772,080 bytes	**4,495,790** bytes	yes
$F_{q12} : mul\_fq6\_by\_nonresidue$	4,923 bytes	4,923 bytes	no
$F_{q12} : frobenius\_map(1)$	4,541,887 bytes	2,879,118 bytes	no
$F_{q12} : frobenius\_map(2)$	2,224,363 bytes	2,791,858 bytes	no
$F_{q12} : mul\_by\_034$	9,810,459 bytes	**4,289,505** bytes	yes
$F_{q12} : ell\_by\_constant$	9,525,050 bytes	**4,714,973** bytes	yes
$F_{q6} : a * b$	3,873,847 bytes	2,236,303 bytes	no
$F_{q6} : frobenius\_map(1)$	1,518,206 bytes	933,606 bytes	no
$F_{q6} : frobenius\_map(2)$	598,274 bytes	958,404 bytes	no
$F_{q6} : mul\_by\_01\_with\_1\_constant$	3,280,529 bytes	1,925,215 bytes	no
$F_{q6} : mul\_by\_fp2\_constant$	1,520,337 bytes	960,147 bytes	no
$F_{q6} : mul\_by\_01$	3,769,633 bytes	2,136,209 bytes	no
$F_{q6} : mul\_by\_fp2$	2,252,362 bytes	1,273,623 bytes	no
$F_{q2} : a * b$	750,883 bytes	424,433 bytes	no

4 Automated slashing

As Fiamma is a POS chain powered by Babylon, we refer to 2 important security properties proposed in the Babylon litepaper [2].

- **Fully slashable PoS security**: Whenever there is a safety violation, the Bitcoin stake is guaranteed to be slashed
- **Staker security**: Each Bitcoin staker is guaranteed to be able to withdraw its funds, or unbond, as long as the staker follows the PoS protocol honestly

Fiamma is a ZKP verification chain based on BitVM2. The core feature provides by BitVM2 is make the POS chain is slashable which satisfy the 1st property. We also provide another way to make the whole protocol satisfy the 2nd property as well, It's not the key point in this paper, we will give additional clarification in other ways.

4.1 BitVM2

We will give the overall design and then also give some problems we also need to solved and optimized in the future.

4.1.1 Overall design

The overall design of BitVM2 in Fiamma is as follows:

Now, we will give clear clarification on it with 3 paragraphs

Roles We intrduce the roles firstly:

- S: The staker, The guys who stake BTC as a POS node in Fiamma
- C: The Committee, which is made up of important partners of Fiamma
- Challenger: The guy who is willing to kick-off the challenge by paying 1 BTC
- user_b: The guy who kick off the dispute transaction on Bitcoin

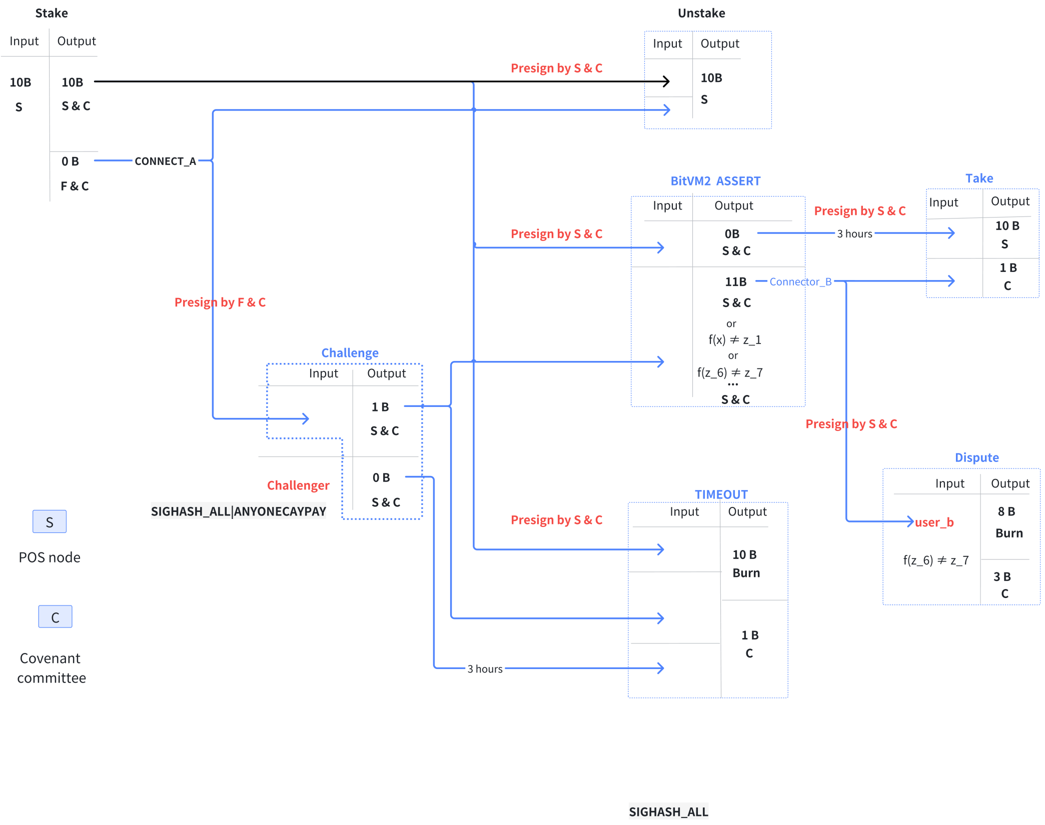


Figure 7: bitvm2 protocol

Transactions We intrduce the function of each transaction:

- Stake transaction: Somebody stakes e.g 5 BTC as a POS node of Fiamma by issuing a Stake transaction
- Unstake transaction: S could take the staking asset at anytime by issuing a Unstake transaction
- Challenge transaction: Challenger could kick off the Challenge transaction by paying 1 BTC, It should be noted that Challenger is permissionless
- Assert transaction: Generate a taproot address and then tranfer the staking BTC to taproot address, anyone could spend this taproot address if he can provide a valid witness of any leaf
- Take transaction: Take the staking BTC back to S if nobody can spend the taproot address in a fixed time, the Challenger lose 1 BTC
- Dispute transaction: Anyone could kick off this transaction if he can spend the taproot address, the S will be slashed
- Timeout transaction: If challenge happens, S still could be slashed if the Assert transaction does not be kicked off in a fixed time

Overall flow We will give the overall flow of the whole protocol as the following picture:

- S stake e.g 5 BTC as a POS node of Fiamma
- S pre-sign 6 transactions with Committee at the initial phase (We should stop S provides service if he doesn't sign these 6 transactions)
- S provides ZKP verification service in Fiamma chain and store the inputs, outputs, and all immediate vaules in POS chain as well as the Winternitz signature for these values;
- At the phase, the proof state is soft finality which means that it's secured by POS network until Now

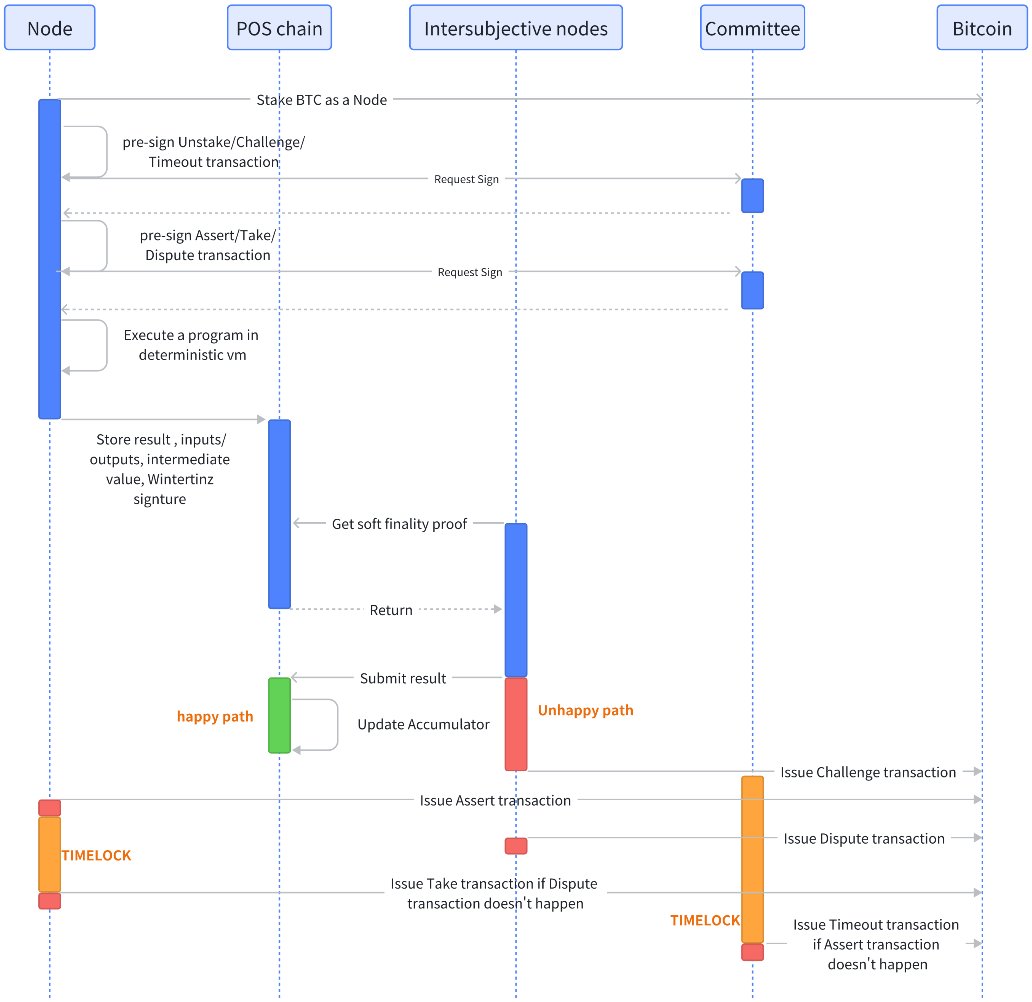

Node
POS chain
Intersubjective nodes
Committee
Bitcoin
Stake BTC as a Node
pre-sign Unstake/Challenge/ Timeout transaction
Request Sign
pre-sign Assert/Take/ Dispute transaction
Request Sign
Execute a program in deterministic vm
Store result , inputs/ outputs, intermediate value, Wintertinz signture
Get soft finality proof
Return
Submit result
happy path
Update Accumulator
Unhappy path
Issue Challenge transaction
Issue Assert transaction
TIMELOCK
Issue Dispute transaction
Issue Take transaction if Dispute transaction doesn't happen
TIMELOCK
Issue Timeout transaction if Assert transaction doesn't happen

Figure 8: bitvm2 simple flow

- The intersubjective nodes in Fiamma will actively to monitor the POS chain and access to the proof which its state is soft
- When the proof get enough number of confirmation, e.g exceed the threshold, the proof state move to hard finality
- Only if one of intersubjective nodes is honest, he will find S give a wrong claim, then he kick the Challenge transaction
- S broadcast the Assert transaction
- Anyone (user_b) will kick off the Dispute transactions, normally, user_b is the Challenger
- Then S is slashed
- If the Dispute transaction doesn't be broadcast after a fixed time, S take asset back by issuing the Take transactions

4.1.2 Core features

Our protocol have several nice properties as follows:

- Trustless for asset security: The asset is safe and there doesn't introduce any additional assumption
- Trust-minimized for protocol: if S is malicious, S will be slashed only if one of intersubjective nodes is honest
- Decentralized for network: As intersubjective nodes is permissionless, so the protocol is decentralized enough and also safe enough

- Nice economic game: S and intersubjective nodes wouldn't love to be malicious becasue of the punish machanism. nodes would love to challenge if S is malicious because of the potential reward

4.1.3 To be improved and completed

The protocol we describe before is an ideal case, there still have a few points to be moted, improved, and completed

- BTC slashing: Because Babylon only support double sign fault, the POS node have to stake additional BTC to correspond to validity fault. It will be improved when Fiamma help Babylon integrate BitVM2 part;
- Taproot address generation: The subscript number of MSM is dynamic as the coefficient is different for different proof
- Reward delivery: It cannot know ahead who will be the Challenger at initial phase, so the Committee will responsible for this part
- Security: The protocol is definitely safe and works only if S and C won't be malicious at the same time
- Staking for intersubjective nodes: The intersubjective nodes have to stake some asset on Fiamma or Bitcoin in case they are out of service or malicious
- On-chain cost for challenge process: As we don't need to store the data and commitments on Bitcoin, we should keep reduce the size of all subscripts to save the challenge cost

5 Summary

Bibliography

[1] **all signhash types**. URL: https://wiki.bitcoinsv.io/index.php/SIGHASH_flags.

[2] **Babylon litepaper**. URL: https://docs.babylonchain.io/assets/files/btc_staking_litepaper-32bfea0c243773f0bfac63e148387aef.pdf.

[3] **BitVM2**. URL: https://bitvm.org/bitvm2.

[4] **blake3**. URL: https://github.com/BLAKE3-team/BLAKE3-specs/blob/master/blake3.pdf.

[5] **CairoVM**. URL: https://eprint.iacr.org/2021/1063.pdf.

[6] **Modular Multiplication Without Trial Division**. URL: https://www.ams.org/journals/mcom/1985-44-170/S0025-5718-1985-0777282-X/S0025-5718-1985-0777282-X.pdf.

[7] **Multidigit Multiplication for Mathematicians**. URL: https://cr.yp.to/papers/m3.pdf.

[8] **On Proving Pairings**. URL: https://eprint.iacr.org/2024/640.pdf.

[9] **On the Size of Pairing-based Non-interactive Arguments**. URL: https://eprint.iacr.org/2016/260.pdf.

[10] **PR**. URL: https://github.com/BitVM/BitVM/commit/d40e501cbcffb25a631b18ae6e7381e29f88a0fa.

[11] **PR**. URL: https://github.com/BitVM/BitVM/pull/75/commits/1467368e68649224325fb089319e818a26d7bdb2.

[12] **The Realm of the Pairings**. URL: https://eprint.iacr.org/2013/722.pdf.

[13] **Transaction constructure**. URL: https://learnmeabitcoin.com/technical/transaction/.

[14] **Transaction size**. URL: https://learnmeabitcoin.com/technical/transaction/size/.

[15] **witernitz signature**. URL: https://toc.cryptobook.us/book.pdf.