Bitcoin proof costs, from header chains to funded disputes

Red Team Gladiatorics / Research paper / 1–20 November 2026

Abstract

How much does it cost to prove a Bitcoin payment, and how much more does it cost to give a challenger an enforceable objection? We follow a 940,000-satoshi mainnet payment from a recursive header proof to a Bitcoin Core regtest dispute. We compare that experiment with November's BitVM2 Nova work, the BitVM3 Nova garbled verifier, and Babylon's published BABE benchmarks. For each measurement we identify the statement, the machine, and the work included in the timer.

The native header prover produced small proofs. Two 40-header testnet4 results took 242.21 s and 1,304.83 s under different concurrency profiles. The difficulty for BitVM2 appeared downstream, when its verifier had to authenticate a linear multiscalar multiplication. A program running in the SP1 virtual machine verified the Nova header proof and checked that the payment appeared in the chain. Garbling the full SP1 proof verifier produced 785,442,492 ciphertexts. In the trace dispute, the operator recorded the verifier's execution so a challenger could contest any of its 3,460,139,945 checks. [1, 2, 3, 4]

On the same 11,600-sat Testnet4 payment with 78 candidate garblings generated and 10 retained, BABE setup took 287.69 s and BitVM3 Nova trace dispute setup took 12,324.83 s. Their public setups occupied 284,321,992 B and 1,087,002,760 B. The largest private-regtest paths consumed 25,982 vB for the valid-proof defense and 2,664,913 vB for a false-wire refutation, excluding funding. Section 12 compares both with the 7/3 pilot. [8, 9, 10]

Interpreting the comparison

Question	Evidence used here
Did a prover establish the claimed statement?	A verified proof, identified inputs, and an explicit description of what the circuit checks.
How large was the circuit?	R1CS constraints, zkVM instructions, and Boolean gates are reported in their native units. Only like units share an axis.
What does the challenger have to retain and do?	Published bytes, optional local indexes, independent replay, authenticated openings, and evaluation have separate measurements.
What reached Bitcoin?	A private regtest spending valueless coins, with policy admission, mined witnesses and virtual sizes. The historical payment supplies the fixture.

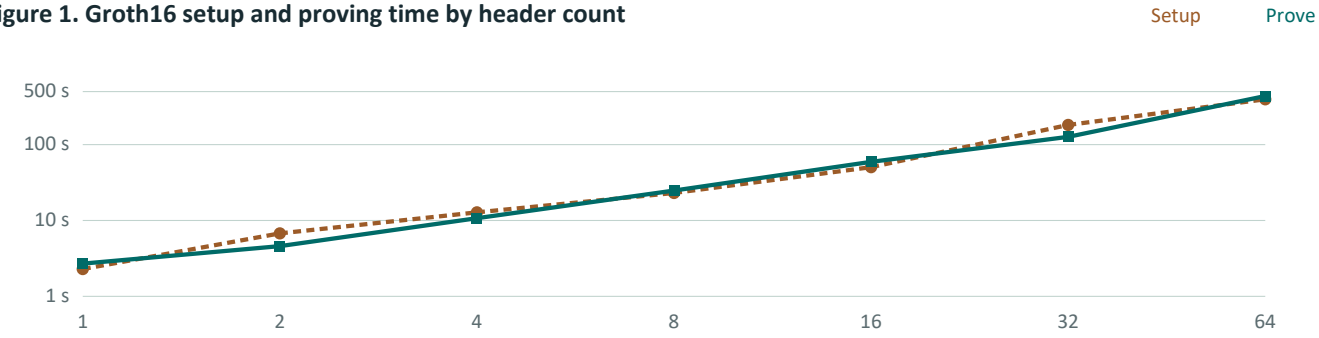
The research question concerns verification and dispute costs for a known payment. Private-input confidentiality and production custody require separate evidence. A "proof" below identifies a particular proof system and statement.

1. Native proving on a laptop

On 4 November the Groth16 gym ran BN254 circuits on a Ryzen AI 9 365 with 20 threads and 31.1 GB RAM under Windows. Every proof in the recorded sweep verified, and each check rejected a changed public input. A compressed proof occupied 128 bytes, the verification key 488 bytes, and verification took 2–5 ms. These small outputs coexisted with sizable proving keys and several minutes of proving. [5]

Header count	Constraints	Setup	Prove	Peak RSS
1	121,782	2.3 s	2.7 s	0.33 GB
4	486,438	12.8 s	10.7 s	1.3 GB
16	1,945,062	50 s	59 s	4.8 GB
64	7,779,558	396 s	437 s	10.4 GB

Figure 1. Groth16 setup and proving time by header count



Header count doubles at each tick. Both axes use logarithmic spacing.

The dedicated SPV circuit used a 123-byte transaction and a 13-level Merkle branch. It needed 1,796,669 constraints, 46.02 s of setup and 45.04 s of proving. It verified in 3 ms. The fixture had a fixed transaction layout. The header sweep omitted retarget, median-time-past and version rules, and used seeded benchmark keys with known toxic waste. It provides a reference for a pairing-based prover on this laptop. Section 3 measures a payment proof that also authenticates the header chain. [5]

2. BitVM2 Nova and the IPA opening

The earlier BitVM2 feasibility discussion asked why a compact Nova proof could still be expensive to verify through Bitcoin Script. We checked its numbers against the prover report, the reconstructed key dimensions, and the measured arithmetic gadget. Here k means headers per fold. Each timing below includes coordinated folding, export and cold verification of two independent proofs. [1, 2, 3]

Testnet4 workload	Shards × threads	Wall time	Compressed bytes per proof	Peak export job commit
k=10, 40 headers, sequential	1 active × 2	1,304.83 s	12,976 B	2,284.18 MiB
k=10, 40 headers, concurrent	2 active × 4	242.21 s	12,976 B	2,265.48 MiB
k=20, 80 headers, concurrent	2 active × 4	544.76 s	13,288 B	4,289.67 MiB

Each shard joined the next at its full state. The k=20 run also rebuilt its verifier key during cold verification. The 5.39× ratio between the two k=10 observations includes changes in concurrency, threads and cache conditions. The memory column gives the largest individual child job, in committed memory. [1]

Measured components and their extrapolation

Item	Result	Evidence
Retained k=5 Nova key, primary / secondary	524,288 / 16,384	Padded R1CS constraints and IPA generators per side
Checking the origin of 78 compressed commitments	887,723,899 / 727,322,403 B	Measured aggregate Script bytes for the Pallas / Vesta profiles
One four-bit-window MSM term	4,296 constraints	Measured slope over 1, 2, 4 and 8 terms
Primary MSM at 524,288 terms	2,252,341,248 constraints	Product of the measured slope and generator count
Script schedule across both engines	646,467,120,106,606 B	Projection for 10,506,798 additions, about 588 TiB

The inner-product argument (IPA) requires a weighted sum over the commitment generators. This multiscalar multiplication (MSM) introduces authenticated intermediate values wherever the implementation splits it into Script fragments. The 588 TiB figure sums the modeled scripts across one specific schedule. A challenger would spend a selected dispute path. The November 7 handoff paused this verifier route. Its native proving measurements remain valid reference points. [2, 3, 14]

3. Proving a six-confirmation mainnet payment

The public statement binds the mainnet header chain from genesis through height 975,072 and output 0 of transaction a652ab0c…0045 in block 975,067. The output pays 940,000 sats to P2WPKH, with six confirmations under the proven tip. Nova folds the headers. An SP1 guest verifies the two Nova segments and the payment branch. The garbled SP1 verifier checks the final proof gate by gate. [4, 6]

Measured stage	Work	Time or artifact
Header folding	975,072 k=1 steps, two shards, ten threads each	137,435 s on the slower shard (38.2 h)
Spartan/IPA export	Two 12,596-byte header proof bundles	44.5 s and 31.1 s
Native verification	One check per header bundle	7.8 s and 7.0 s
Guest execution	34,055,622,753 SP1 instructions	1,283 s, 7.5 GB peak RSS
Application proving	13,290 shard tasks, 4,430 recursion tasks	Verified compressed proof, 1 h 54 min for the final resumed attempt
Final wrapping	Shrink-wrap, then fixed-schedule BLAKE3 wrap	112 s and 72 s
Full garbled verification	785,442,492 ciphertexts × 16 B	12,567,079,872 B per garbling

The final GPU attempt reused 9,645 memoized shard tasks and proved 3,645 fresh. Earlier attempts ended under resource pressure, so the available records do not establish a cold end-to-end GPU time. The historical garble-and-evaluate time was 167.1 s locally and 261.6 s in a later CI run. The new shared-host replay accepted the payment in 1,867.0 s. [4, 6, 7]

New RTX 5080 proving sample

The 19 November sample checked the guest ELF and all three private input digests against the successful run, used an empty memo, and stopped after a five-minute sampling interval. Three fresh SP1 shard tasks completed in 101.093 s, 140.553 s, 181.440 s. Their latencies overlap. The process ran for 307.4 s including shutdown. Shared device use peaked at 15,900 MiB. This run exercises the application prover. A rate extrapolated from these startup tasks would mix startup latency, concurrency and unrelated GPU load. [8]

4. Challenging any step of the garbled verifier

The BitVM3 Nova trace dispute in PR #26 uses the mainnet payment proof from section 3. The operator runs the full SP1 verifier and numbers each input, gate result and terminal output check. Records bind the gate's truth table and source wires. Three indexed SHA-256 trees commit to program records, initial inputs and published wire values. The challenger independently compiles the same verifier and derives the expected program and input roots from the pinned payment fixture. [8, 9]

Trace item	Exact count or size
Input wires, including two constants	8,635,082
Boolean gates, including free gates	3,451,504,862
Wire values	3,460,139,944
Selectable checks, including the terminal output	3,460,139,945
Program / value / initial-input authentication depth	32 / 32 / 24
Encoded dispute input	39,188 bits

The operator flips one committed wire value in a copy of the accepted trace. The payment proof and the canonical program stay fixed. After funding, the challenger re-executes the verifier, compares the published values and authenticates the first disagreement. For a gate it supplies the program record, both operands and the claimed output, with their paths. The garbled predicate evaluates whether those authenticated values disobey the committed truth table. An honest opening takes the delayed operator payout path. [8, 9]

How the Bitcoin selection works

The parties pre-sign a single chain before broadcasting funding. Each bit has two challenger-token hashlocks. The mined choice reveals one token per bit. Its hash selects the corresponding garbler-label hashlock in the response script. The same pre-signed transaction IDs support choices at all 3.46 billion indices. One funded session selects one index. A regression selects alternative bit vectors across a frame boundary. The funded Testnet4 run exercised one chosen refutation (section 9). [8]

The harness uses Bitcoin Core 31.1 on an isolated regtest and runs both roles on one host. The terminal transaction spends test coins according to the local dispute. Network data availability, independent service operation, reorg recovery and an application withdrawal or DLC execution transaction are outside this measurement. [8, 9]

5. Setup, storage and challenger work

The new run used a Ryzen AI 9 365 laptop, WSL2, eight threads for trace hashing and Nova folding, and seven candidate instances with three retained. The archive stage streamed all gates into a random-access representation. The 16,384-leaf pages retain their roots, with lower Merkle levels rebuilt on demand. The full protocol used real garbling and a Nova/Spartan certificate for the selected cut-and-choose partition. [8]

Phase	Measured wall time
Generate the canonical trace and index	214.84 s
Build all three commitments, including copy and checks	1,714.19 s
Generate and commit all seven candidate garblings	187.21 s

Stored or generated material	Bytes
Reusable gate records and page offsets	13,035,787,758 B
Packed wire values, one vector	432,517,493 B
Sparse Merkle roots, all three trees	16,809,984 B
Ciphertext, one local check instance	107,852,400 B
Generated / retained ciphertext, seven / three	754,966,800 / 323,557,200 B

The challenger derives program and input pins independently before participating. The cold independent pass took 1,457.21 s. In the disposable-label refutation preflight, selection after funding took 246.66 s. Section 9 measures retained-circuit evaluation for the Testnet4 payment.

The 13.03 GB program archive is an index over the verifier and can be regenerated. The published value vector is session-specific. A deployment that delivers that vector, distributes a precomputed index, or rebuilds it on demand shifts cost between roles. This run reports the copies and the logical categories explicitly so those designs can be priced separately.

6. Settlement and Babylon BABE

The graph encodes 39,188 label bits in groups of up to 440, yielding 90 choice frames and 90 response frames. Core mined both a refutation and a delayed payout in separate regtest preflights. Both used disposable garbler labels with the real trace and transaction graph. Section 9 reports the funded Testnet4-fixture run using a verified Nova certificate. [8]

Trace dispute transaction graph on Core 31.1	Measured vB
Choice frames / response frames	906,204 / 1,758,535
Refutation / delayed payout terminal	158 / 146
Mined refutation / mined delayed-payout dispute	2,664,897 / 2,664,885
Wallet funding, separately	165
Largest individual frame	19,743

The 2,664,897-vB maximum covers permitted timeout prefixes and terminal leaves and equals the mined refutation. Pre-signed templates occupied 8,608,896 B. Babylon’s published honest-setup path totals 4,533 + 13,079 + 149 = 17,761 vB. The trace dispute graph used 150.0× that footprint. [8, 10]

Babylon's BABE verifies a Groth16 proof using witness encryption for a linear pairing relation and a garbled scalar-multiplication circuit. Its single-instance measurement is 22.16 MiB retained setup data, 174.90 ms setup and 126.53 ms decryption. In its 307-generated, six-retained configuration, the paper reports 14 s of cut-and-choose setup and 0.86 s of evaluation. Soldering those six instances adds 969.37 s in the paper's SP1 benchmark. Its setup machine was a Ryzen 7 7840U. Those experiments use a different proof format, number of labels and security setting. [10, 11]

For the seven-generated, three-retained pilot, the chance of retaining a particular three bad circuits while opening four good ones is 1/35, about 2.86%. Babylon selects its published parameter sets for at most 2^{-40} under that event. Sections 8–10 use a shared Testnet4 payment and report local settlement for both designs. Sections 11 and 12 compare the pilot with Babylon’s 78/10 setting on that payment. [8, 10]

7. A local BABE result on Testnet4

We proved the Testnet4 genesis header using the repository's Groth16 header circuit. The witness is BIP 94's 80-byte header. Seven public field elements bind the previous hash, end hash, compact difficulty and expanded target. The circuit checks the double SHA-256 hash and proof of work. This fixture covers one header at a supplied target. The payment proof also establishes that the transaction appears in a block with six confirmations. [5, 16]

We used GOAT's fixed-input garbled scalar circuit and its sparse decomposable randomized encoding (DRE) adaptor. The harness supplies the paper's fixed-statement witness-encryption equations. We ran two independent Groth16 setups and proofs, each with five fresh BABE instances. The prover received the selected labels, evaluated the blinded first proof point, and decrypted with the remaining proof elements. All ten trials recovered their 32-byte secrets. Altering the last proof point broke recovery in every trial. Changing a public input failed native verification in both runs. [10, 15]

Operation or artifact	Observed on the Ryzen AI 9 365
Groth16 application circuit	121,782 constraints, 7 public field inputs
Groth16 key setup, runs 1 / 2	4.312 / 4.272 s
Groth16 prove, runs 1 / 2	3.131 / 2.962 s
Native verification, prepared key	4.16 / 1.95 ms, 128 B per proof
Boolean topology expansion, once per run	2.366 / 1.580 s
Boolean gates, total / with ciphertext	3,307,431 / 740,116
BABE off-chain instance setup	491.93 ms (394.39–618.58)
Evaluation and pairing decryption	814.76 ms (678.67–1,030.50)
Pairing decryption within evaluation	2.84 ms (1.68–5.93)
Gate ciphertexts / serialized DRE adaptor	11,841,856 / 16,577,064 B
Combined circuit and adaptor / WE ciphertext	28,418,920 B / 96 B

Instance times are medians and full ranges across ten trials on the shared WSL2 laptop, with eight Rayon threads available to Groth16. Each evaluation includes the Boolean witness pass, gate-label evaluation, adaptor decoding, point reconstruction and pairing decryption. Storage is 27.10 MiB for ciphertexts and adaptor, plus 96 B of setup ciphertext. The 512 active input labels occupy 8,192 B. Per-label Bitcoin hashlocks and transaction signing, cut-and-choose and soldering are outside the timers.

The local medians for evaluation and decryption were 731 ms in the first run and 857 ms in the replication. Babylon reports 126.53 ms on its own implementation and hardware. Differences in circuit representation, serialization and evaluator code prevent attributing the gap to the CPU or the Bitcoin network. [10, 15]

8. BABE and BitVM3 Nova on one Testnet4 payment

Both routes authenticate output 0 of transaction 11c18d7c…b301aa0e, paying 11,600 sat in block 161,011. Six linked headers end at 161,016. The SP1 payment program checks the transaction, output, Merkle branch, header links, and a Nova header proof from an agreed checkpoint at 160,000. We independently decoded the transaction, folded its Merkle path and matched the 128-byte payment record in both routes' 1,292-byte public statement. Its SHA-256 digest begins 220205b4 and ends 5719fc43.

The common RTX 5080 run produced a verified compressed SP1 proof in 19,702.20 s (5 h 28 min). It proved 6,156 shards and replayed 2,186, then proved 2,716 recursion tasks and replayed 65. Both outer routes reuse the same shrink proof. Charging the shared GPU time to only one side would distort the comparison.

Measured stage	BABE fixed input	BitVM3 Nova full garbled SP1 verifier
Shared Nova → SP1 → shrink	Identical input	Identical input
Outer wrap setup / prove	282.56 s / 1,496.35 s	8.75 s / 70.71 s (earlier)
Groth16 incl. load, CPU / GPU	353.33 s / 389.02 s	BLAKE3 proof used directly
Groth16 after load, CPU / GPU	152.36 s / 124.66 s	—
Proof preparation from shrink[1]	2,173.46 s	80.52 s
Garbler setup	306.92 ms	844.50 s
Circuit storage	28,418,920 B incl. adaptor	12,567,079,872 B ciphertext
Boolean gates requiring ciphertext	740,116	785,442,492
Active labels for one opening	8,192 B	138,161,280 B
Release generation	0.007 ms (selection only)	38.77 s
Garbled evaluation, including outcome	497.77 ms (median of five)	517.81 s (one run)

For this payment, the full garbled verifier emitted 442.2 times the bytes of the BABE circuit and adaptor and published 552,645,120 B of input commitments. BABE used an additional 96 B of witness-encryption material. Its standard Groth16 circuit and proving key occupied 2,437,991,441 and 5,862,173,061 B on disk. The R1CS has 15,972,262 constraints. Groth16 ran once per backend with the same key and witness. The GPU process ran first and peaked at 15.76 GiB host RAM. The CPU process peaked at 16.28 GiB and shared the OS file cache. The BABE medians use the GPU proof. The garbled verifier streamed files from the Windows volume under WSL. [1] Sum of timed wrap phases plus the GPU Groth16 call where required. [9, 15, 17]

For BABE we proved the standard SP1 BN254 wrap, generated a Groth16 proof with the official v6.1.0 circuit artifacts, and verified it against the payment program's verification key and public bytes before supplying it to the garbled scalar circuit. Five fresh honest-setup instances each had to recover their secret and reject an altered proof. On the full garbled verifier route the evaluator authenticated a release of labels for the selected proof and checked the result against the garbler's output commitments. Both final evaluations ran on the CPU.

9. The BitVM3 Nova trace dispute

The trace dispute lets the challenger identify a single incorrect step in the SP1 verifier's recorded execution. The operator commits the trace and prepares seven local-check garblings. Nova folds the translation checks across the three retained instances. We timed the full-width extension separately from the materialized full-verifier run on the preceding page. Adding both columns together would charge two alternative verification routes.

Trace protocol stage	Matched payment
Verifier replay / gate records	262.84 s / 13,034,102,446 B
Selectable checks / packed value bytes	3,460,139,945 / 432,517,493 B
5080 program / input / value commitment	273.26 s / 0.15 s / 17.11 s
Tree build, all phases / sparse tree roots	308.06 s / 16,809,984 B
Seven candidate garblings	70.79 s
Nova folds / primary / secondary constraints	19,594 / 122,856 / 10,349
Nova public parameters / compression key setup	5.52 s / 0.44 s
Nova folds / final Spartan compression	5,254.47 s / 15.45 s
Compressed Nova proof / verifier key	11,688 / 52,884,552 B
Public setup ready, elapsed / serialized bytes	5,438.58 s / 327,384,008 B
Verify public setup and opened instances	44.93 s
Independent program and initial-input pin	983.45 s
Wallet and pre-signatures / template storage	0.30 s / 8,608,896 B
Post-funding selection	60.13 s
Retained garbled evaluation	4.32 s
Worst dispute, without / with this funding	2,664,897 / 2,665,062 vB

The challenger selected wire 3,000,000,000 after funding and recovered the refutation secret. Bitcoin Core mined 182 transactions on private regtest. The refutation used 2,664,897 vB plus 165 vB for funding. The underlying payment came from Testnet4.

The RTX 5080 produced the shared SP1 proof, accelerated the measured Groth16 prover and computed the operator's trace commitments. The challenger independently replayed the verifier and computed its program and input commitments on CPU. We measured fixed-input BABE with honest setup. Section 11 measures the additional BABE setup work under the same seven-candidate setting as the trace dispute.

10. Funding the BABE dispute

We funded the valid 11,600-sat Testnet4 payment claim on private regtest using the saved RTX 5080 Groth16 proof. A fresh fixed-input BABE instance committed both label hashes for each of its 512 proof-point bits and the SHA-256 hash of the 32-byte plaintext secret. We pre-signed the graph before broadcasting funding. The prover decoded the confirmed assertion, the challenger published the matching active labels, and the prover evaluated the circuit with the 512 labels read back from the mined challenge. Its decrypted 32-byte secret unlocked WronglyChallenged. [8, 10, 15]

Private regtest, same payment	BABE (vB)	BitVM3 Nova trace dispute (vB)
Funding transaction	208	165
Maximum path through final settlement, excluding funding	25,899	2,664,897
Maximum path including this funding	26,107	2,665,062

Bitcoin Core mined all three BABE outcomes from the same funding on separate forks. An unchallenged Assert and Withdraw used 8,978 vB. An unanswered challenge ended with NoWithdraw at 25,728 vB. The largest path comprised Assert (8,716 vB), ChallengeAssert (16,767 vB), WronglyChallenged (154 vB) and Withdraw (262 vB). A fresh Core node independently validated each fork from genesis. Core rejected a wrong active label, a wrong secret and both premature timelocked spends.

The two settlement columns use the same historical payment but exercise different objections. BABE defends a valid proof. The trace challenger refutes an incorrect wire value planted in the verifier's execution record. The BABE implementation uses uncompressed proof-point coordinates, 512 single-use Lamport tokens and two Taproot challenge inputs of 256 bits each. Both inputs commit to the same transaction outputs. Babylon reports 17,761 vB for its own three-transaction encoding. The corresponding three local transactions through secret recovery used 25,637 vB. [9, 10]

All totals sum the individually rounded virtual sizes of canonical signed transactions. The maximum covers the fixed graph and its 64-byte default Schnorr signatures. Both experiments use private regtest coins. Network fees, time waiting for public confirmations and alternative encodings are outside the measurement.

11. The seven-candidate pilot

We repeated the funded BABE experiment on the same payment and GPU-produced Groth16 proof. Both stacks committed seven candidates before the partition draw. In the BABE run a separate process recorded that commitment and supplied the random nonce. Each stack opened four candidates and kept three. For a garbler that commits exactly three bad candidates before a uniform draw, the probability of retaining all three is 1 in 35 (2.86%). This target describes the partition event. Neither the trial nor that probability establishes overall protocol soundness. [8, 9, 10]

Measured stage	BABE	BitVM3 Nova trace dispute
Build and commit seven candidates	2.23 s	70.79 s
Nova steps / compressed certificate	256 / 11,688 B	19,594 / 11,688 B
Nova folding / compression	55.59 s / 14.72 s	5,254.47 s / 15.45 s
Check opened candidates and public setup	1.49 s	44.93 s
Total setup through durable write	86.05 s	5,438.58 s
Public setup on disk	85,319,232 B	327,384,008 B
Worst path excluding funding	25,917 vB	2,664,897 vB
Worst path including this funding	26,125 vB	2,665,062 vB

Both runs used the trace dispute's Nova/Spartan circuit to prove that one set of Bitcoin input labels translates correctly into the other two retained garblings. BABE bound 512 proof-point bits. The trace dispute bound 39,188 bits to authenticate the selected verifier step. The BABE verifier regenerated the four opened candidates from their seeds, checked the retained payloads against the original commitments and verified the compressed certificate. Only the four opened seeds enter the public setup.

All three BABE evaluations recovered their distinct secrets from one mined label release in 0.860 s. Core admitted each alternative secret spend at the same weight. The mined answer and withdrawal were the largest path at 25,917 vB before funding. The unanswered branch used 25,728 vB and the unchallenged withdrawal used 8,978 vB. A fresh verifier checked the public setup against the on-chain locks and fresh Core nodes replayed all three forks from genesis.

Setup time starts with candidate generation and ends after verification and writing the public setup. The table excludes preparation of the shared payment proof, graph signing and funding. Serialized setup includes retained garblings, commitments, opened seeds and the certificate. The reusable verifier key and the archive of the verifier's execution are outside those byte totals. Sections 8 and 9 report the respective upstream costs. Each timing is one wall-clock observation on the shared laptop.

12. BABE and BitVM3 Nova at the 40-bit target

BABE answers a challenge to a valid payment proof by revealing a secret. In the BitVM3 Nova trace dispute, the challenger identifies a verifier step and shows that its recorded result is wrong. We tested both at the 78-generated, 10-retained setting from Table 2 of Babylon's BABE paper. A uniform draw opens 68 candidates after the garbler commits to all 78. If exactly ten candidates are bad, just one of the 1,258,315,963,905 possible retained sets contains all ten. The probability of that event is about 7.95×10^{-13}, below 2^{-40}. Both benchmarks used this partition and a Nova proof soldering nine garblings to one base. The bound assumes a fresh, unbiased challenge following commitment. The experiment measures costs for honest setup participants. [8, 10]

Protocol and split	Setup (s)	Public setup (B)	Worst path (vB)	With funding (vB)
BABE 7/3	86.05	85,319,232	25,917	26,125
BABE 78/10	287.69	284,321,992	25,982	26,190
BitVM3 Nova 7/3	5,438.58	327,384,008	2,664,897	2,665,062
BitVM3 Nova 78/10	12,324.83	1,087,002,760	2,664,913	2,665,078

Setup starts with candidate generation and includes verifying the openings and certificate, then writing the public setup. The trace dispute setup also writes the public label commitments for its transaction driver before stopping the clock. Storage is the length of the serialized setup. The reusable key, upstream payment proof and the archive of the verifier's execution are accounted separately. Each timing is one observation on the shared host. The RTX 5080 supplied the saved proof and rebuilt the commitments to the verifier's execution trace. Both soldering provers used six Rayon threads on CPU.

At 78/10, BABE folded 256 steps in 106.42 s. The trace dispute folded 19,594 in 9,536.56 s. Each step had 156,365 primary constraints. Both serialized certificates occupied 12,224 B and the reusable verifier key occupied 89,237,904 B. Checking the public setup, including the 68 openings, took 39.02 s for BABE and 1,318.94 s for the trace dispute. The trace dispute operator discarded candidate payloads after committing to them and regenerated the ten retained candidates during setup.

Bitcoin Core admitted each of the ten BABE answers. Fresh nodes replayed all three settlement branches. The largest branch defended the valid payment through withdrawal. The trace challenger selected wire 3,000,000,000 after funding in 221.70 s. Retained evaluation took 4.16 s for BABE and 53.52 s for the trace dispute. Bitcoin Core mined the refutation using one of the deepest leaves in the spending tree. That path reached the maximum size permitted by the graph. Moving from 7/3 to 78/10 added 65 vB to BABE and 16 vB to the trace dispute before funding.

13. Script validity and unspent status

The payment inclusion proof establishes that a confirmed transaction created the output. The Script execution proof checks whether a particular Bitcoin Script succeeds. The unspent-output proof follows outputs created within a chosen block window and checks whether they remain unspent at its end. The distinction matters when a benchmark is cited as evidence for a balance or withdrawal. [4, 12, 13]

Script execution in Nova

The Script execution engine reproduced 2,756 applicable Core vectors for legacy Script, P2SH and witness v0. In 1,808 valid cases, all 29,121 tested circuit steps matched the engine's next state. The suite skipped 2,239 Taproot cases. The P2SH multisig example took 25.8 s to fold 31 steps and 0.65 s to verify the recursive proof. Spartan/IPA compression then took 34.0 s, produced 11,480 B and verified in 0.84 s. Signature and locktime claims have separate checks in this design. [12]

Windowed UTXO state

Unspent outputs, blocks 975,067–975,072	Measurement
Uniform step circuit / augmented primary	51,404 / 70,617 R1CS constraints
Native trace generation	1,743 s
Nova folding	224,790 steps across eight concurrent segments
Longest segment / aggregate throughput	35,839 s (9.96 h) / 6.27 steps per second
Compressed segments	8 × 13,800 B = 110,400 B
Verify eight segments, chaining and the payment membership	1.47 s, after 3.61 s to rebuild the key

The sealed map held 74,593 leaves. The mainnet payment output from section 3 was unspent at the window tip, and it was the only window output paying that script. The circuit rehashes transaction bodies and links their Merkle roots to the headers. Its assume-valid model inherits block validity and its window defines the balance being reported. A full-chain UTXO-set or wallet balance would require broader coverage. [13]

Conclusion

At the 40-bit partition target, the BitVM3 Nova trace dispute took 42.8 times as long to set up, stored 3.82 times the public bytes and used 102.6 times the block space of the local BABE implementation. The two paths settle different objections on the same payment. BABE releases 512 proof-point labels once. The trace challenger selects one verifier step after funding and authenticates it with a 39,188-bit opening.

Sources and reproduction

The November 19–20 experiments cover the payment inclusion proof and its trace dispute, the GPU shard sample and the Testnet4 BABE benchmarks. Source 10 supplies Babylon's reference measurements and cut-and-choose parameters. Times are wall-clock unless stated. MB/GB use powers of 1,000, MiB/GiB/TiB use powers of 1,024. Bitcoin transaction virtual size rounds weight divided by four upward.

1. Native Nova proving coordinator and Windows measurements.

2. BitVM2 verifier dimensions, measured Script components and IPA projection.

3. BitVM2 four-bit-window multiplication measurements and projection (M26a).

4. Header-chain and payment inclusion proof, statement and historical timings (M16).

5. Groth16 gym, raw November 4 MSI benchmark records.

6. SP1 CUDA prover, verified payment proof and cache accounting.

7. Full garbled SP1 verifier (M14), continuous integration run.

8. November 19 funded trace dispute, public receipt and reproduction instructions. See the local evidence package and the trace dispute notes in the source branch.

9. Verifier trace dispute and earlier funded pilot (PR #26).

10. Garg et al., BABE, January 2026, section 9, tables 1, 2 and 5.

11. Babylon Trustless Bitcoin Vaults, public testnet protocol parameters.

12. Bitcoin Script execution proof and differential checks (M17).

13. Unspent-output proof for a six-block window (M18).

14. November 7 BitVM2 engineering handoff, scope and measured commitment-origin checks (M23).

15. GOAT Network, fixed-input scalar circuit and sparse DRE adaptor used in the local benchmark.

16. BIP 94, Testnet4 genesis header and network parameters.

17. Succinct, SP1 6.9.0 standard wrap and Groth16 proving path.

Historical baseline a183eef. The local experiment extends PR #26 base b404ea4. The matched Testnet4 evidence package contains the public proof files, fixture checker, raw timing receipts and output commitments. Section 6 reports the separately measured mainnet-fixture regtest preflight. In the matched evidence ZIP, the challenge folder holds the public Nova setup, the selected opening and the raw regtest transactions and blocks. The m16-check setup, opening and chain commands replay the respective checks. The babe-settlement folder contains the raw BABE branches and receipt. replay_settlement.py submits each fork to a fresh Core 31.1 node. The babe-cut-and-choose folder contains the seven-candidate run, public setup, challenge pins and Core branches. The replay command uses the public setup verifier before submitting the blocks. The separate 78/10 evidence archive includes the compact comparison, public setups, pins and mined data. Build both verifiers with the babylon-40 feature to replay it.