HEKATON: Horizontally-Scalable zkSNARKs via Proof Aggregation

Michael Rosenberg
micro@umd.edu
UMD

Tushar Mopuri
tmopuri@upenn.edu
UPenn

Hossein Hafezi
h.hafezi@nyu.edu
NYU

Ian Miers
imiers@umd.edu
UMD

Pratyush Mishra
prat@upenn.edu
UPenn

July 26, 2024

Abstract

Zero-knowledge Succinct Non-interactive ARGuments of Knowledge (zkSNARKs) allow a prover to convince a verifier of the correct execution of a large computation in private and easily-verifiable manner. These properties make zkSNARKs a powerful tool for adding accountability, scalability, and privacy to numerous systems such as blockchains and verifiable key directories. Unfortunately, existing zkSNARKs are unable to scale to large computations due to time and space complexity requirements for the prover algorithm. As a result, they cannot handle real-world instances of the aforementioned applications.

In this work, we introduce HEKATON, a zkSNARK that overcomes these barriers and can efficiently handle arbitrarily large computations. We construct HEKATON via a new “distribute-and-aggregate” framework that breaks up large computations into small chunks, proves these chunks in parallel in a distributed system, and then aggregates the resulting chunk proofs into a single succinct proof. Underlying this framework is a new technique for efficiently handling data that is shared between chunks that we believe could be of independent interest.

We implement a distributed prover for HEKATON, and evaluate its performance on a compute cluster. Our experiments show that HEKATON achieves strong horizontal scalability (proving time decreases linearly as we increase the number of nodes in the cluster), and is able to prove large computations quickly: it can prove computations of size $2^{35}$ gates in under an hour, which is much faster than prior work.

Finally, we also apply HEKATON to two applications of real-world interest: proofs of batched insertion for a verifiable key directory and proving correctness of RAM computations. In both cases, HEKATON is able to scale to handle realistic workloads with better efficiency than prior work.
1 Introduction

Zero-knowledge Succinct Non-interactive ARGuments of Knowledge (zkSNARKs) are cryptographic proofs that allow a prover to convince a verifier that, given a function $F$ and public input $x$, there is a private witness $w$ such that $F(x, w) = 1$. zkSNARKs hide all information about $w$, and are small and easy to verify regardless of the complexity of $F$. Recent efficient constructions of zkSNARKs [GGPR13; Gro16; GWC19; CHMMW20; Set20] have enabled a range of applications and industrial deployments that rely on zkSNARKs to improve efficiency and privacy characteristics. However, zkSNARKs cannot currently scale to prove useful computations on realistic problem sizes. Indeed, many proposed applications of zkSNARKs, such as verifiable key transparency [TFZBT22], proofs of program execution or vulnerability [BCGTV13; ZGKPP18], or machine learning inference [LXZ21], are limited to toy problem sizes due to zkSNARK scalability limitations.

In more detail, in typical zkSNARKs, to prove the correctness of a computation $F$ on inputs $(x, w)$, we must first express $F$ as an arithmetic circuit $C_F$. The size of the latter is often much larger than the description of $F$, leading to two scalability issues:

- **Poor parallelization:** zkSNARK provers perform a number of expensive operations whose cost grows linearly with the circuit size $|C_F|$. Unfortunately, these operations have diminishing parallelizability, particularly for real implementations that must account for inter-process communication costs between cores, processors, and even compute clusters.

- **Large memory overheads** The space complexity of the prover also tends to scale linearly with $|C_F|$. As a result, memory often ends up being the key bottleneck for proving complex computations: while one can always wait longer for a proof, one cannot always add more RAM to the prover’s machine.

These problems are not merely asymptotic, but lead to high concrete costs even for relatively simple computations. For instance, a circuit for proving the multiplication of two $700 \times 700$ matrices requires 685 million gates. Prior work [WZCPS18] reports that using even a 64-threaded machine to prove this circuit requires hundreds of minutes and a prohibitive $1.7TB$ of RAM.

Moreover, these problems are exacerbated for many exciting SNARKs applications which, frequently, are RAM Programs (e.g., proofs of vulnerability, zkRollups, etc.). This poses two challenges. First, representing a RAM program as a bare circuit requires all branches be taken and loops be unrolled, drastically increasing circuit size. Second, circuits do not, natively, provide memory access and the methods for providing memory either offer high overhead or, as we will see, place constraints how we can address the space and time complexity.

**Distributed proving: a path forward?** A promising approach to scale zkSNARKs up to large circuits is to distribute the proving algorithm across a set of workers in a way that ensures that the proof computation is parallelized across the workers, and which ensures that each worker’s local memory requirements are low. In this work, we revisit this approach and design HEKATON, a new horizontally-scalable zkSNARK whose prover algorithm can be distributed over large compute clusters much more efficiently than all prior work [WZCPS18; Xie+22; LXZSZ24]. We detail the technical contributions underlying HEKATON next.

1.1 Our contributions

HEKATON is the result of several contributions:

1. **Scalable proving via divide-and-aggregate.** We distill a generic framework for constructing scalable zkSNARKs that we call *divide-and-aggregate*, or DNA for short. Our framework proceeds as follows:
to prove the execution of a function $F$ represented as a circuit $C_F$, partition $C_F$ into smaller subcircuits $C_1, \ldots, C_n$, and have each cluster machine use a simpler “inner” zkSNARK to prove the satisfaction of an individual subcircuit separately. Then, invoke a proof aggregation protocol \cite{BMMTV21; GMN22; ABST23} to relatively cheaply aggregate these subcircuit proofs into a single succinctly-verifiable proof.

Instantiating the foregoing blueprint requires addressing two challenges: aggregating proofs for different subcircuits, and sharing wires between subcircuits.

(2) **Multi-circuit aggregation.** We generalize prior proof aggregation schemes \cite{BMMTV21; GMN22} to efficiently aggregate proofs for different circuits. This allows our framework to handle computations with arbitrary subcircuit structure. In contrast, schemes implicit in prior work \cite{BMMTV21} are only able to handle uniform circuits that just repeat the same subcircuit many times.

(3) **Shared wires via efficient global memory.** All known aggregation schemes (including ours) are only able to achieve succinct verification when proofs for neighbouring subcircuits share a small constant amount of data. However, this is incompatible with our goal of designing a proof system for arbitrary circuits where subcircuits might share many wires.

We overcome this challenge via a new low-overhead technique for providing subcircuits with efficient access to a global memory bank. This allows subcircuits to share wire values by accessing them from the memory bank, rather than directly passing them between subcircuits. To efficiently prove correctness of memory accesses, we design a new memory-checking circuit that can itself be partitioned into subcircuits that share just a constant number of wires. The resulting workflow is illustrated in Fig. 1.

On a technical level, our approach extends recent work on permutation-based memory checking techniques \cite{ZGKPP18; BCGJM18} that uses commit-carrying SNARKs by extending our aggregation scheme into one that supports a commit-carrying mode.

(4) **Efficient instantiation and implementation.** We instantiate our DNA zkSNARK blueprint by choosing Mirage \cite{KPPS20} as the inner zkSNARK, and designing a new “commit-carrying” aggregation scheme for Mirage. We call the resulting system HEKATON.

We implement HEKATON in a Rust library that supports distributed proving. Our implementation also provides a novel framework for writing partitioned circuits that enables us to minimize communication (in distributed mode) and memory requirements (in both modes). We provide details in Section 7.
Applications and evaluation. As noted above, our evaluation of HEKATON demonstrates that it can prove computations that are orders-of-magnitude larger than prior work, while requiring a fraction of the time. We provide details in Section 8. To showcase the improvements HEKATON offers over state of the art provers, we implement two real world applications. In the first application, we use HEKATON to prove the correctness of batched updates to a Verifiable Key Directory [TFZBT22], whose configuration, for the first time, matches that of deployed systems, i.e., using a sparse Merkle Tree and SHA256 as a hash function. In the second application, we adapt our memory-checking techniques to build a proof for RAM execution that achieves a throughput of 50kHz. Our system is able to prove a large execution trace of $2^{25}$ instructions via a circuit of $2^{35}$ constraints.

1.2 Related work

Lookup tables. Our global memory technique is reminiscent of recent work on lookup arguments [GW20; ZBKMNS22]. Indeed, one could consider constructing a large lookup table containing the shared wire values. However, a key difference is that most lookup arguments assume that the commitment to the table is constructed honestly in an offline phase (often over public values), and hence optimize their proving algorithm for this regime. In our setting, the “table” is for secret wire values unique to the witness and therefore must be constructed online during the proving phase by the untrusted prover. To use a lookup argument, we would first need to check that the table commitment was generated honestly, and would need to design a proving algorithm optimized for this regime.

Partitioning circuits. Assuming sufficient resources, HEKATON’s latency is determined by the size of the largest of the partitioned subcircuit and the number of shared wires between the subcircuits. The latter costs only 13 constraints per shared wire, thus partitioning the circuit into equal-sized subcircuits leads to good performance since in practice, these partitioned circuits have a very small number of shared wires. We leave the problem of optimal circuit partitioning to future work, but note that HEKATON is compatible with prior automatic circuit partitioning schemes [Cos+15; San+23].

1.2.1 Distributed zkSNARKs

Like HEKATON, all prior works on distributed zkSNARKs use a coordinator which gets as input the circuit $C$, the public input $x$, and the witness $w$, and is responsible for distributing these to the worker nodes, who in turn jointly compute the zkSNARK proof. Existing protocols differ in how these workers perform the latter computation. We provide an asymptotic comparison between the above systems in Table 1, and focus below on qualitative differences.

<table>
<thead>
<tr>
<th>system</th>
<th>supported computation</th>
<th>proof size</th>
<th>per-worker time</th>
<th>total comm.</th>
<th>verifier time</th>
<th>SRS size</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIZK [WZCPS18]</td>
<td>arbitrary</td>
<td>$O(1)$</td>
<td>$O(</td>
<td>C</td>
<td>/n)$</td>
<td>$O(</td>
</tr>
<tr>
<td>DeViro [Xie+22]</td>
<td>data-parallel</td>
<td>$O(n)$</td>
<td>$O(</td>
<td>C</td>
<td>/n)$</td>
<td>$O(</td>
</tr>
<tr>
<td>Pianist [LXZSZ24]</td>
<td>arbitrary</td>
<td>$O(1)$</td>
<td>$O(</td>
<td>B</td>
<td>)$</td>
<td>$O(1)$</td>
</tr>
<tr>
<td>HEKATON</td>
<td>arbitrary</td>
<td>$O(\log n)$</td>
<td>$O(</td>
<td>B</td>
<td>)$</td>
<td>$O(1)$</td>
</tr>
</tbody>
</table>

Table 1: Comparison of distributed zkSNARKs. Here $C$ is the circuit being proved, $n$ is the number of subcircuits, $k$ is the number of unique subcircuits, and $B$ is the largest unique subcircuit.

DIZK [WZCPS18] initiated the study of distributed algorithms for zkSNARK provers, focusing on the zkSNARK in [Gro16] (though the techniques are applicable to other proof systems as well). In more detail,
the core contributions of DIZK are the design and implementation of efficient distributed algorithms for the core operations performed in zkSNARK provers, namely FFTs and multi-scalar multiplications. The primary drawback of DIZK is its need for a linear amount of inter-worker communication. In practice, this greatly increases the latency of proving: the per-gate cost is $10 \times$ worse than local proving. In contrast, HEKATON requires only constant inter-worker communication, and is able to achieve per-gate costs that are very similar to local proving.

DeVirgo [Xie+22] is a SNARK with a distributed prover that focuses on supporting only data-parallel computations, i.e., the circuit being proved consists of repeated copies of a single subcircuit. DeVirgo’s prover requires the primary node to perform cryptographic work that scales with the size of the subcircuit, and requires linear inter-worker communication. In contrast, as noted above, HEKATON supports arbitrary computations, requires only constant inter-worker communication, and ensures that the cryptographic work performed by the primary node scales only with the number of workers.

Pianist [LXZSZ24] is a very recent work that designs a distributed proving algorithm for the Plonk zk-SNARK [GWC19]. At a high level, Pianist relies on the elegant observation that using bivariate polynomial commitments allows one to decompose Plonk’s global permutation check, which is used for circuit wiring correctness, into local per-worker permutation checks. The resulting protocol, produces constant proof size and verifier time, whereas the latter costs scale logarithmically for HEKATON.\footnote{Note that one can generically reduce HEKATON’s proof sizes via depth-1 recursive proof composition [Cos+15; BCGMMW20].}

However, Pianist, as instantiated, requires an SRS whose size scales linearly with circuit size. To be precise, the SRS for the bivariate polynomial commitment in Pianist depends on the degree of the variables. The degree of the first variable corresponds to subcircuit size, and that of the other to the number of workers. As a result, Pianist’s SRS size is $O(|C|)$. In contrast, because HEKATON’s SRS consists of the SRS(es) for subcircuits and a small SRS for aggregation, its SRS size is dominated by the number of unique subcircuits. For many circuits of interest (e.g., RAM programs), the number and size of the unique subcircuits is much smaller than the total circuit size, leading to substantial SRS size savings for HEKATON. We provide a thorough experimental comparison of HEKATON and Pianist in Section 8.2.

Mangrove [NDCTB24] is a concurrent theoretical work that uses similar commit-and-prove-based permutation-checking techniques as us. Unlike us, however, Mangrove only reports estimated performance numbers, and does not provide a full implementation or evaluation. Furthermore, applying Mangrove’s techniques to distributed proving would lead to a prover that requires linear inter-worker communication.

Distributed proving based via recursive proofs. A promising idea for distributed proving is to use recursive verification, where the system uses recursive proofs to aggregate subcircuit proofs. The idea would be to replace the custom aggregation scheme in HEKATON with a system that verifies batches of subcircuit proofs recursively in a tree-like manner. However, this approach has several drawbacks.

First, even assuming state-of-the-art folding-based techniques [BCLMS21; KST22] that have reduced the cost of recursive verification, a back-of-the-envelope calculation shows that such an approach would have over $2 \times$-worse aggregation time than HEKATON. This is even when we assume that the aggregation step is also distributed; without this assumption, the recursive verification approach would have much worse aggregation time.

Second, even a state-of-the-art recursive aggregation scheme, would still need to support shared wires between subcircuits, and as discussed in Section 2.1.1, existing approaches for this would incur much higher overhead than HEKATON. Adapting the memory-checking techniques in HEKATON to a recursive verification setting is a non-trivial research problem, and indeed formed a key component of concurrent work [NDCTB24].
Finally, scalable implementation of these approaches would be challenging due to the need for complex multi-round communication and straggler management.
2 Techniques

We construct our divide-and-aggregate zkSNARK via the blueprint outlined in Section 1.1, which we recall next. To prove a circuit $C$, (a) partition $C$ into subcircuits $(C_1, \ldots, C_n)$, (b) replace shared wires with accesses to a global memory, (c) construct augmented subcircuits $(C'_1, \ldots, C'_n)$ that additionally check memory accesses, (d) prove each $C'_i$ using via an inner zkSNARK (denoted ARG), and (e) aggregate the resulting proofs using an aggregation scheme for ARG proofs (denoted Agg).

In the rest of this section, we will expand on each of these steps, focusing on our novel memory checker (Section 2.1) and our new commit-carrying aggregation scheme (Section 2.2) that supports this memory checker.

2.1 Partition-friendly memory checking

To check the consistency of these memory accesses, one can design a “memory-checker” circuit $M$ using standard memory checking techniques [BEGKN91]. However, integrating this into our blueprint requires that the memory checker $M$ can be partitioned into $n$ subcircuits $M_1, \ldots, M_n$, such that (a) $M_i$ checks the memory accesses made by $C_i$, and (b) $M_i$ and $M_{i+1}$ share just a constant number of wires. Given such an $M$, we can obtain a DNA zkSNARK for arbitrary circuits by invoking our blueprint on augmented subcircuits $C'_i$ that invoke $C_i$ and $M_i$ together. Let us thus focus on constructing such a partitioning-friendly memory checker.

2.1.1 Limitations of existing memory checkers

**Attempt 1: Online memory checkers.** In online memory checking [BEGKN91], memory is committed to via a Merkle tree and read/write operations consist of checking/updating the Merkle path for that location. This requires sharing only a single wire value (the Merkle root) between subcircuits, and is used in all prior divide-and-aggregate SNARKs [BCTV14a; CTV15]. However, online memory checking creates both asymptotic and concrete bottlenecks.

Asymptotically, Merkle path checking imposes a logarithmic overhead: if $s_i$ is the number of shared wires in the $i$-th subcircuit $C_i$, then checking $C_i$’s memory accesses requires $s_i \cdot O(\log s)$ hash function invocations, where $s = \sum_i s_i$ is the size of the global memory. Since in the worst case the number of shared wires can be as large as the circuit size $|C_i|$, the size of the memory checker subcircuit $M_i$, and hence that of the augmented subcircuit $C'_i$, can be as large as $O(|C_i| \log |C_i|)$. Concretely, even with zkSNARK-friendly [AABDS20; GKRRS21] hash functions that require $\sim 300$ gates, each memory access costs $300 \log |s|$ gates, which results in unacceptable slowdowns in practice.

**Attempt 2: Memory trace checkers.** A second approach [BCGT13] verifies memory operations by recording them in a memory trace, and performing cheap checks on the trace entries. In more detail, a memory trace logs each memory operation as $(\text{subcircuit-number}, \text{op} = \text{read/write}, \text{addr}, \text{value})$. The checker then considers two versions of this trace: one sorted by address (denoted $A$), and another sorted by subcircuit number (denoted $T$), and performs both local and global checks on these traces. As we will show later, the local checks are easily partitioned across subcircuits, but the global check requires verifying that the two traces contain the same entries, and are hence permutations of each other. Efficiently implementing this “permutation check” in a manner amenable to partitioning is a challenge, as we explain next.

**Permutation checking.** State-of-the-art permutation checking techniques [ZGKPP18; BCGJM18; KPPS20] require the optimal $O(s)$ gates to check permutations between $s$-sized traces. They achieve this low cost by relying on randomized Reed–Solomon fingerprinting [Lip89; Lip90], which performs the following randomized check to ensure that two traces $T := (t_1, \ldots, t_s)$ and $A := (a_1, \ldots, a_s)$ are permutations:
1. construct trace polynomials $T(X) := \prod_{i=1}^{n}(X - t_i)$ and $A(X) := \prod_{i=1}^{n}(X - a_i)$;
2. sample a random field element $r \leftarrow \mathbb{F}$; and
3. check that $T(r) = A(r)$.

This check can be realized quite efficiently via a circuit that witnesses $T$, $A$, and $r$, computes the products $\prod_{i=1}^{n}(r - t_i)$ and $\prod_{i=1}^{n}(r - a_i)$, and checks that these are equal. We now show that this computation can also be partitioned easily across subcircuits in a way that ensures that each subcircuit only pays for the accesses it makes to the memory.

### 2.1.2 Partitioning polynomial evaluation

Our partitioning strategy relies on the simple but crucial observation that the trace polynomials can be evaluated incrementally. In more detail, notice that for every $j \in [s]$, if we are given the running products $\tau := \prod_{i=1}^{j}(r - t_i)$ and $\alpha := \prod_{i=1}^{j}(r - a_i)$, and the $j + 1$-th trace entries $t_{j+1}$ and $a_{j+1}$, we can easily compute the next running products $\tau' := \tau \cdot (r - t_{j+1})$ and $\alpha' := \alpha \cdot (r - a_{j+1})$. We can then iterate this process to eventually compute $T(r)$ and $A(r)$. We leverage this observation to construct memory-checking subcircuits $M_1, \ldots, M_n$ as follows.

**Notation.** Denote by $s_i$ the number of memory accesses in the $i$-th subcircuit $C_i$, by $k_i = \sum_{j=1}^{i-1}s_j$ the total number of accesses in the previous subcircuits $C_1, \ldots, C_{i-1}$, and finally by $S_i$ the set $\{k_i + 1, \ldots, k_i + s_i\}$.

Let $T = (t_1, \ldots, t_s)$ and $A = (a_1, \ldots, a_s)$ be the memory traces as defined in Section 2.1.1. We split these traces up into $n$ subtraces, one for each subcircuit, as follows. The $i$-th subtrace of $T$ is defined to contain those entries access by $C_i$, i.e., $T_i := \{t_j \mid j \in S_i\}$. The $i$-th subtrace of $A$ is defined analogously as $A_i := \{a_j \mid j \in S_i\}$. 2 Finally, denote by $\tau_i$ and $\alpha_i$ the running products up to (but excluding) the $i$-th subcircuit, i.e., $\tau_i := \prod_{j \in [i]}(r - t_j)$ and $\alpha_i := \prod_{j \in [i]}(r - a_j)$.

**Construction.** We are now ready to describe how our construction of the memory-checking subcircuits $M_i$.

- $M_i$ receives as public input the random challenge $r$, the last entries $t$ and $a$ of $T_{i-1}$ and $A_{i-1}$ respectively, as well as the running products $\tau_i$ and $\alpha_i$.
- The public output of $M_i$ consists of the last entries $t'$ and $a'$ of $T_i$ and $A_i$ respectively, as well as the updated running evaluations $\tau_{i+1}$ and $\alpha_{i+1}$.
- $M_i$ receives as additional witness the subtraces $T_i$ and $A_i$.
- $M_i$ performs the following computations:
  - $M_i$ replaces each memory gate with with a ‘read’ from the value in the corresponding entry in $T_i$.
  - $M_i$ checks that entries in $T_i$ all have the subcircuit number set to $i$, and that $A_i$ is sorted by address.
  - $M_i$ checks that consecutive entries in $A_i$ are consistent: if they have the same address, then they must contain the same value.
  - $M_i$ computes the new running products $\tau_{i+1} := \tau_i \cdot \prod_{j \in S_i}(r - t_j)$ and $\alpha_{i+1} := \alpha_i \cdot \prod_{j \in S_i}(r - a_j)$

This approach, illustrated in Fig. 2, achieves the desired performance: the $i$-th memory-checking circuit processes exactly an $s_i/s$-fraction of the traces in just $O(s_i)$ gates, as required, and the hidden constants are great: just 13 gates per shared wire. We provide formal details and analyses of this construction in Section 4.2.

We note that our full construction in Section 4.2 incorporates the optimizations mentioned in Section 2.4 to further simplify and reduce the cost of memory-checking.

---

2 Note that $A_i$ can be entirely disjoint from $T_i$, i.e., it might contain no entries corresponding to $C_i$’s memory accesses.
2.2 Aggregating heterogeneous commit-carrying zkSNARKs

A question left open by the description of the memory-checker construction in Section 2.1 is that of how to derive the challenge $r$ for the trace polynomial evaluation. An approach taken by prior work [ZGKPP18; KPPS20] is to derive the random challenge $r$ by hashing (via a random oracle $\rho$) commitments to the traces. This requires the SNARK proof for each subcircuit to additionally ensure that the witnessed traces are consistent with their claimed commitments. Prior work achieves this consistency check efficiently by using special commit-carrying zkSNARKs (cc-zkSNARK) [CFQ19] where the commitment scheme is co-designed with the SNARK so that the contents of the commitment are “natively” available as witnesses in the proven circuit.

We briefly recall their high-level strategy here, and then explain why it does not work in our setting. First, the prover commits to the traces, obtaining commitments $tcm$ and $acm$; then, it derives the challenge $r := \rho(tcm, acm)$; and finally, it invokes the zkSNARK’s proving algorithm on the augmented subcircuits $C'_i$ to prove that the committed traces are valid memory traces. Unfortunately, the natural extension to our setting where the prover derives the challenge $r$ by hashing the individual subcircuit commitments fails to meet our goals because the proof is no longer succinct.

**Our approach: commit-carrying aggregation schemes.** We resolve this issue by defining a new notion of commit-carrying aggregation schemes that support aggregating not only the proofs the underlying inner cc-zkSNARK, but also their commitments. Our notion also naturally supports multi-circuit aggregation that allows the prover to aggregate proofs for multiple circuits into a single combined proof.

With this new tool in hand, we can augment our blueprint construction as follows. Given an inner cc-zkSNARK ARG and a commit-carrying aggregation scheme Agg for ARG, to prove a partitioned circuit $C = (C_1, \ldots, C_n)$,

1. construct augmented subcircuits $(C'_1, \ldots, C'_n)$ and memory subtraces $(T_1, \ldots, T_n)$ and $(A_1, \ldots, A_n)$;
2. compute trace commitments $(tcm_1, \ldots, tcm_n)$ and $(acm_1, \ldots, acm_n)$ using ARG’s commitment scheme;
3. aggregate these with Agg to obtain succinct commitments $(tcm, acm)$, and derive $r := \rho(tcm, acm)$;
4. prove each circuit $C'_i$ with ARG and aggregate the resulting inner proofs with Agg.

This approach achieves succinctness while preserving soundness. We provide details of our construction in Section 6.

2.3 Our aggregation scheme for Mirage

To instantiate the foregoing blueprint, we choose the Mirage cc-zkSNARK [KPPS20] as our inner zkSNARK. Mirage is a commit-carrying variant of the Groth16 zkSNARK [Gro16], and inherits the succinctness and prover efficiency of the latter. In more detail, a Mirage proof, like a Groth16 one, consists of three group elements $(A, B, C)$, while the commitment is another group element $D$. The verifying key, like that of Groth16,
contains elements $\alpha \in G_1$ and $\beta, \delta, \in G_2$, as well as an extra element $\eta \in G_2$. To check a proof $\pi = (A, B, C)$ with respect to a commitment $D$, the verifier checks that the following pairing equation is satisfied.

\[ e(A, B) = e(\alpha, \beta) e(C, \delta) e(D, \eta) \]

**Background for Mirage.** To construct an aggregation scheme for Mirage, we follow existing work on aggregation schemes for Groth16 [BMMTV21; GMN22] by relying on inner-pairing product argument systems [BCCGP16; BBBPWM18] to enable proving more general bilinear products. Two variants are relevant for our purposes: “MIPPs”, which prove the inner product $\sum a_i \cdot B_i$ for (committed) vectors $a \in \mathbb{F}_n^m, B \in G_2^m$, and “TIPPs”, which prove the pairing product $\prod_i e(r_i \cdot A_i, B_i)$ for a scalar $r$ and (committed) vectors $A \in G_1^n, B \in G_2^n$.

To construct an aggregation scheme for Mirage that proves the correctness of a batch of Mirage proofs $(A, B, C)$ with respect to commitments $D$, we will use the above ingredients to prove the following randomized pairing check with respect to a random challenge $r$:

\[ \prod_i e(r_i \cdot A_i, B_i) = e(\alpha, \beta) \sum_i \prod_i e(r_i \cdot C_i, \delta) \prod_i e(r_i \cdot D_i, \eta) \]

To aggregate these proofs, the aggregation prover first commits to the $A, B, C, D$ using a “structure-preserving” commitment scheme [AFGHO16], hashes these to obtain $r$, and then proves the above equation using MIPPs and TIPPs. Clearly, the left hand side can be proven using a TIPP, while the remaining terms can be proven using MIPPs. Unfortunately, we are not done yet, as our setting requires us to aggregate proofs for multiple circuits simultaneously.

**Multi-circuit aggregation.** When aggregating multiple circuits, the $\delta$ and $\eta$ components of the verifying key differ across circuits, meaning our randomized check changes to the following:

\[ \prod_i e(r_i \cdot A_i, B_i) = e(\alpha, \beta) \sum_i \prod_i e(r_i \cdot C_i, \delta_i) \prod_i e(r_i \cdot D_i, \eta_i) \]

While all the pairing checks can now be proven using TIPPs, preserving succinctness requires care. To see this, let us inspect the two kinds of TIPPs performed in the above equation. Notice that in the TIPP for the left-hand-side check $(\prod_i e(r_i \cdot A_i, B_i))$, both arguments are prover-supplied, and hence it is fine for the prover to provide commitments to $A$ and $B$. The TIPPs on the right-hand side, however, involve one prover-supplied argument ($C$ and $D$) and one verifier-supplied argument ($\delta$ and $\eta$). While the prover can provide commitments to $C$ and $D$, it would be unsound for it to do the same for $\delta$ and $\eta$, as it could commit to arbitrary elements that cause the checks to pass.

Therefore, the verifier must obtain these commitments itself. The straightforward solution of having the verifier compute them itself would violate succinctness. To resolve this issue, we leverage the fact that the subcircuit structure is known at setup time, which in turn means that the vectors $\delta$ and $\eta$ are also known at setup time. This means that commitments to the latter can be computed then and included in the verifying key, thus preserving succinctness.

**Reducing the number of TIPPs.** The aggregation scheme proposed so far is sound and relatively efficient. However, it requires the prover to prove multiple TIPPs, worsening prover complexity and proof size. Because aggregation is performed on a single machine (and is not parallelized), improving its prover complexity is important. We do so by devising a method for batching multiple TIPP instances together into a single TIPP instance, and providing a single TIPP proof for the latter.

**Our aggregation scheme for Mirage.** The sketch above omits details, including how we handle public inputs. We provide these details in Section 5.
2.4 Optimizations

The foregoing discussions omit a number of optimizations that we developed to improve the efficiency of our construction. We describe these next.

Read-only memory. Notice that we can assume that the memory is read-only, as the prover can initialize the memory with values of all the shared wires, and subcircuits which access these wires can simply check that the memory contains the correct value. This optimization greatly reduces the concrete cost of the local checks performed on the memory traces.

Reducing SRS size. The description in the foregoing sections obscures the fact that, as described, each augmented subcircuit $C'_i$ is unique, even if the topology of the underlying subcircuit $C_i$ is shared with other subcircuits. This is because we need to embed into $C'_i$ information about (a) its circuit number $i$ (so that it can check the trace entries correspond to its own memory access), and (b) the addresses it reads from (so that it reads the appropriate shared wires). Because of this, each subcircuit becomes unique, leading to a blowup in SRS size when we instantiate HEKATON with the Mirage cc-zkSNARK [KPPS20] as the latter has circuit-specific setup (we would need to generate a separate SRS for each subcircuit).

To resolve this issue, we observe that the aforementioned information (circuit number and addresses) is not dependent on the prover’s witness, and can be committed to via Mirage’s commitment scheme in a preprocessing phase. We provide details in Section 6.3.

Homogenizing public inputs via Merkle trees. Recall from Section 2.1.2 that each memory-checking subcircuit $M_i$ receives as public input the triple $i_{in} = (r, (t, a), (\tau_{k_i}, \alpha_{k_i}))$, and outputs the pair $o_{out} = ((t', a'), (\tau_{k_{i+1}}, \alpha_{k_{i+1}}))$. This means that each $M_i$’s public input is necessarily different. Existing constructions of aggregation schemes [BMMTV21] can handle such heterogeneous public inputs, but incur prover complexity and proof size overheads. Instead, we propose to homogenize the public inputs of all the subcircuits by careful and selective use of online memory checking: we commit to $(i_{in}, o_{out})$ in the $i$-th leaf of a Merkle tree, and use the corresponding path to verifiably access the $i$-th leaf in the $i$-th subcircuit. This allows us to eschew the complex public input handling mechanisms of prior aggregation schemes.

Batch setup optimization. We perform a common setup for all circuits simultaneously. This helps us improve efficiency by reducing the number of TIPPs by 1.
3 Preliminaries

Indexed relations. An indexed relation $\mathcal{R}$ is a set of triples $(i, x, w)$ where \( i \) is the index, \( x \) is the instance, and \( w \) is the witness; the corresponding indexed language $\mathcal{L}(\mathcal{R})$ is the set of pairs $(i, x)$ for which there exists a witness \( w \) such that $(i, x, w) \in \mathcal{R}$. An indexed relation is said to be committable if the witness can be split into two parts, $w = (w_1, w_2)$.

Definition 3.1 (circuit satisfiability). The committable indexed relation CSAT is the set of triples $(i, x, w) = ((\mathbb{F}, \ell, m, n, C), x, (w_1, w_2))$ where $\mathbb{F}$ is a finite field, $\ell, m,$ and $n$ are natural numbers, $x \in \mathbb{F}^\ell$, $w_1 \in \mathbb{F}^m$, and $w_2 \in \mathbb{F}^n$ are vectors, and $C : \mathbb{F}^{\ell + m + n} \rightarrow \mathbb{F}$ is an arithmetic circuit over $\mathbb{F}$ such that $C(x, w_1, w_2) = 0$.

To obtain the usual notion of circuit satisfiability we can set $m = 0$ and $w_1 = \perp$.

Security parameters. For simplicity of notation, we assume that all public parameters have length at least $\lambda$, so that algorithms which receive such parameters can run in time $\text{poly}(\lambda)$.

Random oracles. We denote by $\mathcal{U}(\lambda)$ the set of all functions that map $\{0,1\}^\ast$ to $\{0,1\}^{\lambda}$. A random oracle with security parameter $\lambda$ is a function $\rho : \{0,1\}^\ast \rightarrow \{0,1\}^{\lambda}$ sampled uniformly at random from $\mathcal{U}(\lambda)$.

Bilinear groups. We denote groups by $\mathbb{G}$. A bilinear function $e : \mathbb{G}_1 \times \mathbb{G}_2 \rightarrow \mathbb{G}_T$ is a type-3 bilinear pairing if there is no efficiently computable group homomorphism from $\mathbb{G}_2$ to $\mathbb{G}_1$. $e$ is degenerate if there is a non-identity $G \in \mathbb{G}_1$ such that $e(G, H) = 1$ for all $H \in \mathbb{G}_2$. We use additive notation for $\mathbb{G}_1$ and $\mathbb{G}_2$, and multiplicative notation for $\mathbb{G}_T$. As shorthand we sometimes write $[a]_1$ for $a \cdot G$ and $[b]_2$ for $b \cdot H$.

3.1 Commitment schemes

A commitment scheme $\text{CM} = (\text{Setup}, \text{Commit})$ over a universe of message spaces $\{\mathcal{M}_i\}_{i \in \mathbb{N}}$ enables a party to generate a (perfectly) hiding and (computationally) binding commitment to a given message $m \in \mathcal{M}$.

- **Setup**: on input a security parameter and a description of the message space $\mathcal{M}$, $\text{CM}.\text{Setup}$ samples a commitment key $ck$.
- **Commitment**: on input public parameters $ck$, message $m \in \mathcal{M}$, and randomness $r$, $\text{CM}.\text{Commit}$ outputs a commitment $C$ to $m$.

3.2 Commit-carrying zkSNARKs

A tuple of algorithms $\text{ARG} = (\mathcal{G}, \mathcal{C}, \mathcal{P}, \mathcal{V})$ is a commit-carrying succinct non-interactive argument of knowledge (cc-SNARK) [CFQ19] in the random oracle model (ROM) for a committable indexed relation $\mathcal{R}$ if it satisfies the following syntax and properties:

- **Setup**: On input a security parameter $\lambda$ and a set of indices $[i_1]_i = 1, n$, $\mathcal{G}$ outputs corresponding proving keys $[\text{ipk}_i]_i = 1, n$, commitment keys $[\text{ick}_i]_i = 1, n$, and verification keys $[\text{ivk}_i]_i = 1, n$. When $n = 1$, we omit indices.
- **Commitment**: On input a commitment key $\text{ick}$, a message $w_1$, and commitment randomness $r$, $\mathcal{C}$ outputs a commitment $\text{cm}$.
- **Proving**: On input a proving key $\text{ipk}$, an instance $x$, a witness $w = (w_1, w_2)$, and commitment randomness $r$, $\mathcal{P}$ outputs a proof $\pi$.
- **Verifying**: On input a verification key $\text{ivk}$, an instance $x$, a commitment $\text{cm}$, and a proof $\pi$, $\mathcal{V}$ outputs a bit indicating whether $\pi$ is a valid proof.

Throughout, we assume without loss of generality that the proving key $\text{ipk}$ contains $\text{ick}$, $\text{ivk}$ and $i$. These algorithms must satisfy the following properties:
Multi-instance completeness. For every set of indices \([i_j]_{j=1}^n\) and every efficient adversary \(A\), the following probability is 1.

\[
\Pr \left[ \begin{array}{c}
\forall i \in [n]: \\
(i_i, x_i, w_i) \in R
\end{array} \right] \\
\begin{array}{c}
\forall i \in [n]: \\
(i_i, x_i, w_i) \in R
\end{array}
\] \\
\begin{array}{c}
\forall i \in [n]: \\
\forall i \in [n]: \\
\forall i \in [n]:
\end{array}
\]

\[
\begin{array}{c}
\forall i \in [n]: \\
\forall i \in [n]: \\
\forall i \in [n]:
\end{array}
\] \\
\begin{array}{c}
\forall i \in [n]: \\
\forall i \in [n]: \\
\forall i \in [n]:
\end{array}
\]

Multi-instance knowledge soundness. For every efficient adversary \(\hat{\mathcal{P}}\) and auxiliary input distribution \(\mathcal{D}\), there exists an efficient extractor \(\mathcal{E}_{\hat{\mathcal{P}}}\) such that for every set of indices \([i_j]_{j=1}^n\), the following probability is negligible:

\[
\Pr \left[ \begin{array}{c}
\exists i \in [n]: \\
(i_i, x_i, w_i) \notin R
\end{array} \right] \\
\begin{array}{c}
\exists i \in [n]: \\
(i_i, x_i, w_i) \notin R
\end{array}
\] \\
\begin{array}{c}
\exists i \in [n]: \\
\forall i \in [n]: \\
\forall i \in [n]:
\end{array}
\]

\[
\begin{array}{c}
\forall i \in [n]: \\
\forall i \in [n]: \\
\forall i \in [n]:
\end{array}
\] \\
\begin{array}{c}
\forall i \in [n]: \\
\forall i \in [n]: \\
\forall i \in [n]:
\end{array}
\]

Multi-instance zero-knowledge. There exists an efficient simulator \(S = (S_1, S_2)\) such that for every efficient stateful adversary \(\hat{V} = (\hat{V}_1, \hat{V}_2, \hat{V}_3)\), the following probabilities are negligibly close:

\[
\Pr \left[ \begin{array}{c}
\forall i \in [n]: \\
(i_i, x_i, w_i) \in R
\end{array} \right] \\
\begin{array}{c}
\forall i \in [n]: \\
(i_i, x_i, w_i) \in R
\end{array}
\] \\
\begin{array}{c}
\forall i \in [n]: \\
\forall i \in [n]: \\
\forall i \in [n]:
\end{array}
\]

\[
\begin{array}{c}
\forall i \in [n]: \\
\forall i \in [n]: \\
\forall i \in [n]:
\end{array}
\] \\
\begin{array}{c}
\forall i \in [n]: \\
\forall i \in [n]: \\
\forall i \in [n]:
\end{array}
\]

In the foregoing, \(\rho[\mu]\) is the function that, on input \(x\), equals \(\mu(x)\) if \(\mu\) is defined on \(x\), or \(\rho(x)\) otherwise. This definition uses explicitly-programmable random oracles [BR93]. Note that we can recover the definition of a standard zkSNARK by setting the commitment algorithm \(C\) to be a no-op and considering \(n = 1\).

3.3 Aggregation schemes

Let \(\text{ARG} = (G, C, \mathcal{P}, \nu)\) be a ccSNARK for CSAT. Then, at a high level, an aggregation scheme for ARG is a ccSNARK that proves the validity of a batch of ARG proofs (for possibly different circuits). Formally, an aggregation scheme for ARG is a ccSNARK for the committable indexed relation \(R_{\text{Agg}}\) defined below.

**Definition 3.2 (aggregation relation).** The committable indexed relation \(R_{\text{Agg}}\) is the set of triples \((i, x_i, (w_1, w_2)) = ([ivk_i]_{i=1}^n, x_i, ([cm_i]_{i=1}^n, [\pi_i]_{i=1}^n))\) where, for each \(i \in [n]\),
• ivk_i is an honestly-generated verification key under ARG for some index i for CSAT,
• x is a valid instance for CSAT with respect to i (i.e., (i, x) ∈ L(CSAT)),
• cm_i is a commitment, and
• π_i is a valid proof under ivk_i for x, i.e., V(ivk_i, x, cm_i, π_i) = 1.
4 Partitioning circuits via memory checking

We now describe our transformation that partitions a circuit $C$ into augmented subcircuits via the memory-checking infrastructure described in Section 2.1. Our transformation can be decomposed into two steps. First, in Section 4.2, we show how to augment a partitioned circuit $C$ into a ‘ROM’-circuit where wires between subcircuits are replaced with memory accesses. Then, in Section 4.3, we show how to check these memory accesses with a memory checker, and how to split the checks performed by the latter across subcircuits.

The resulting transformation, which we denote by $f = (f_1, f_{w_1}, f_{w_2})$, maps a partitioned circuit instance $(i, x, w) \in k$-CSAT (Definition 4.3) to a batch of CSAT instances $[(i_1, x', (w_{i,1}, w_{i,2}))]_{i=1}^k$ such that $x'$ is of the form $(1, r, \alpha, \beta)$, where $(\alpha, \beta)$ are field elements, and $r$ is the root of a Merkle tree. The reduction satisfies the following lemma.

**Lemma 4.1.** There exists an efficient transformation $f = (f_1, f_{w_1}, f_{w_2})$ satisfying the following properties:

- **Completeness:** For all $\alpha, \beta \in F$, if $(i, x, w) \in k$-CSAT, then $x$ is the 0-th leaf of $r_t$ and $(i_1, x', (w_{i,1}, w_{i,2})) \in CSAT$ for all $i \in [k]$.
- **Knowledge soundness:** Let $CM = (\text{Setup}, \text{Commit})$ be a commitment scheme. There exists an efficient extractor $E$ such that for every $i$, efficient adversary $A$ and auxiliary distribution $D$, the following probability is negligible:

$$
\Pr \left[ \begin{array}{l}
(\hat{i}, \hat{x}, \hat{w}) \not\in k$-CSAT \\
\forall i \in [k]: \ \ (\hat{i}, \hat{x}', (w_{i,1}, w_{i,2})) \in CSAT \\
\text{and} \\
x \text{is the 0-th leaf of } r_t
\end{array} \right] = \frac{\rho \leftarrow U(\lambda)}{\rho}
$$

We now describe the components of the transformation $f$, beginning with some notation.

**4.1 Notation**

We begin by defining a notion of graph and circuit partitions.

**Definition 4.2.** A labelled $k$-partition $V$ of a graph $G = (V, E)$ is a list $\{V_1, \ldots, V_k\}$ that partitions the vertex set $V$. Specifically, $V$ is a $k$-partition of $G$ if and only if the sets in $V$ are non-empty and mutually disjoint, and the union of these sets equals $V$.

In the following, let $G = (V, E)$ be a directed acyclic graph, and $V$ a $k$-partition of $G$. The cut-set of two vertex subsets $V_1, V_2 \subseteq V$ is defined as the set of edges originating in $V_1$ and terminating in $V_2$. That is, $\text{cut}(V_1, V_2) := \{(u, v) \in E \mid u \in V_1, v \in V_2\}$. The reduced cut-set of $V_1, V_2$ (denoted $d$-cut$(V_1, V_2)$) is obtained from $\text{cut}(V_1, V_2)$ by removing all edges with the same source vertex except the lexicographically first one. That is, for all $e = (u, v) \in \text{cut}(V_1, V_2)$, $e \in d$-cut$(V_1, V_2)$ if and only if there is no $v' \in V_2$ such that $(u, v') \in \text{cut}(V_1, V_2)$.

We call an arithmetic circuit $C$ partitioned if there exists a $k$-partition $C = \{C_1, \ldots, C_k\}$ of the underlying graph. Each $C_i \in C$ is called a subcircuit of $C$. We identify circuits and subcircuits with their underlying graphs. For example, we will call the graph $k$-partition underlying a partitioned circuit $C$ a $k$-partition of the circuit itself. This allows us to define the following relation.


**Definition 4.3** (k-partitioned circuit satisfiability). The committable indexed relation $k$-CSAT is the set of triples $((i, x, w) = ((i_{\text{CSAT}}, \mathcal{C}), x_{\text{CSAT}}, w_{\text{CSAT}}))$ where $(i_{\text{CSAT}}, x_{\text{CSAT}}, w_{\text{CSAT}}) \in \text{CSAT}$ and $\mathcal{C} = \{C_1, \ldots, C_k\}$ is a $k$-partition of $\mathcal{C}$.

Notice that the set of wires shared between two subcircuits $C_i, C_j \in \mathcal{C}$ is exactly the reduced cut-set $\text{d-cut}(C_i, C_j)$. We will denote by $S$ the set of all shared wires, i.e., $S = \bigcup_{i \neq j} \text{d-cut}(C_i, C_j)$.

### 4.2 Eliminating shared wires with ROM circuits

We introduce a new circuit model: circuits with read-only access to a memory bank. We call such circuits ROM circuits.

**Definition 4.4.** A ROM circuit over the field $\mathbb{F}$ is an arithmetic circuit over $\mathbb{F}$ equipped with access to a memory $M$, which is an array of $\mathbb{F}$ elements that is indexed by the elements of $\mathbb{F}$. A ROM circuit, in addition to the standard addition and multiplication gates, contains a read gate that takes as input an index $i$ and outputs the value $M[i]$.

Throughout this section, we will use $s$ to denote the size of a memory bank $M$.

**Definition 4.5** (k-partitioned ROM circuit satisfiability). The indexed relation $k$-RCSAT is the set of all triples $(i, x, w) = ((\mathbb{F}, k, \ell, n, s, C, C), x, (w, M))$ where $\mathbb{F}$ is a finite field, $\ell, n$, and $s$ are natural numbers, $x \in \mathbb{F}^\ell$ and $w \in \mathbb{F}^n$ are vectors over $\mathbb{F}$, $M$ is a memory bank of size $s$ over $\mathbb{F}$, and $C : \mathbb{F}^{\ell+n} \to \mathbb{F}$ is a ROM circuit over $\mathbb{F}$ with respect to $M$, such that $c^M(x, w) = 0$ and $\mathcal{C} = \{C_1, \ldots, C_k\}$ is a $k$-partition of $\mathcal{C}$.

In Figure 3, we provide a formal description of our reduction $c2r = (c2r_1, c2r_x, c2r_w)$ from $k$-CSAT to $k$-RCSAT that removes shared wires between subcircuits. The following lemma follows from the construction of the reduction.

**Lemma 4.6.** The function $c2r = (c2r_1, c2r_x, c2r_w)$ defined in Fig. 3 is a reduction from $k$-CSAT to $k$-RCSAT. That is, $(i, x, w) \in k$-CSAT if and only if $(i_M, x_M, w_M) \in k$-RCSAT, where $i_M = c2r_1(i)$, $x_M = c2r_x(x)$, and furthermore, there are no shared wires between the subcircuits in $i_M$.

### 4.3 Reducing partitioned ROM circuits to committable circuits

We now show how to reduce the problem of checking satisfiability of an instance of $k$-RCSAT to that of checking simultaneous satisfaction of $k$ instances of CSAT. For this we use the notion of a memory trace [BCGT13; BCGTV13; BCTV14b; ZGKPP18]. Informally, a memory trace is a list of entries recording memory reads performed by a ROM circuit.

**Definition 4.7.** A memory trace entry is a tuple $e = (t, i, v)$, where $t \in [s]$ is the index of the subcircuit that performed this operation, $i$ is the memory address accessed, and $v$ is the value read from $M[i]$.

A memory trace is a list of memory trace entries, one for each read gate in a ROM circuit.

We will consider memory traces whose entries are sorted by subcircuit index (denoted $T$), and by memory address (denoted $A$).

**Notation.** We now introduce some notation for partitioning memory traces arising from a $k$-partitioned ROM CSAT instance $(\mathbb{F}, \ell, n, s, R, \mathcal{R} = \{R_1, \ldots, R_k\})$. Denote by $s_i$ the number of read gates in the $i$-th subcircuit $R_i$.
Figure 3: Reduction from k-CSAT to k-RCSAT.

by $k_i = \sum_{j=1}^{i-1} s_j$ the cumulative number of read gates in circuits $R_1, \ldots, R_{i-1}$, and by $S_i = \{k_i + 1, \ldots, k_{i+1}\}$ the set of (indices of) all read gates in $R_i$.

Let $T = (T_1, \ldots, T_s)$ and $A = (A_1, \ldots, A_s)$ denote the subcircuit-sorted and the address-sorted memory traces respectively. Then the $i$-th subcircuit-sorted subtrace $T_i$ is defined as $T_i := (T_{k_i+1}, \ldots, T_{k_{i+1}})$. Note that, by construction, the subcircuit index of each entry in $T_i$ equals $i$, i.e., $T_i$ contains only those trace entries that correspond to the memory reads made by $R_i$. Denote by $t_i$ the last entry in $T_i$. Similarly, the $i$-th address-sorted subtrace is defined as $A_i := (A_{k_i+1}, \ldots, A_{k_{i+1}})$, and $a_i$ denotes the last entry in $A_i$.

**Construction intuition.** We begin by noting that valid memory traces for ROM circuits produced by the reduction in Section 4.2 should satisfy the following properties:

1. **Values are consistent:** If consecutive entries $(i, i, v)$ and $(i', i', v')$ in the trace $A$ have the same address (i.e., $i = i'$), they must have the same values ($v = v'$).
2. **Every subcircuit reads:** Consecutive entries $(i, i, v)$ and $(i', i', v')$ in $T$ satisfy the constraint that either $t = t'$ or $t = t' + 1$. If $t = t'$, then it must be that $i \leq i'$.
3. **Every address is read:** Consecutive entries $(i, i, v)$ and $(i', i', v')$ in $A$ must satisfy the constraint that either $i = i'$ or $i = i' + 1$. If $i = i'$, then it must be that $t \leq t'$.
4. **Traces are consistent:** $T$ and $A$ are permutations of each other.

We now describe the intuition behind our reduction $f^k = (f^k_1, f^k_{w_1}, f^k_{\pi, w_2})$ from k-RCSAT to k-CSAT, referring to Figure 4 the formal pseudocode of the reduction.

The function $f^k_{s}$ constructs a set of $k$ ‘commit-carrying’ circuits $R'_i$ by relying on the following observations. Checks 1 to 3 are entirely local as they inspect only consecutive entries in $T_i$ and $A'_i$, and can thus be enforced by each subcircuit $R_i$ independently. Check 4, on the other hand, is a global property that requires coordination.
between the subcircuits. To enable this coordination, we store the running trace evaluations (as well as the last entries of the sorted traces) in a Merkle tree, and give to each subcircuit the root of this tree as public input. In more detail, each circuit $R'_i$ is constructed as follows.

- $R'_i$ takes as public input $x' = (1, rt, \alpha, \beta)$ where $\alpha, \beta \in \mathbb{F}$ are field elements, and $rt$ is a Merkle tree root. (We will specify how these are obtained below.)
- $R'_i$ takes as input a witness that can be split two parts, $w_{i,1}$ and $w_{i,2}$, where the first contains the subtraces $T_i$ and $A_i$, while the second contains the witness wires of $R_i$ and two Merkle paths.
- $R'_i$ enforces (a) that the input Merkle paths are valid paths for the $i$-th and the $i$-th leaves in a Merkle tree with root $rt$; (b) that the $i$-th leaf of the Merkle tree contains the last entries $t_i$ and $a_i$ of $T_i$ and $A_i$ respectively; (c) the correctness of running trace evaluations $T_i(\alpha, \beta)$ and $A_i(\alpha, \beta)$, where $T_i(X, Y) := \prod_{j=1}^{m} (X - (t_j^X + Y \cdot i_j^X + Y^2 \cdot v_j^X))$ and $A_i(X, Y) := \prod_{j=1}^{m} (X - (t_j^X + Y \cdot i_j^X + Y^2 \cdot v_j^X))$; and (d) that the 0-th leaf holds the public input $x$ of $R$.

The reduction satisfies the following lemma, which roughly says that when the inputs $\alpha, \beta$ are chosen by hashing commitments to the memory traces, the output circuits are satisfiable if and only if the input ROM circuit is satisfiable.

**Lemma 4.8.** Let $CM = \text{(Setup, Commit)}$ be a commitment scheme. Then $f^k = (f^k_1, f^k_{w_1}, f^k_{x_1, w_2})$ defined in Fig. 4 on input $(i, x, (\mathbb{w}, M), \alpha, \beta)$, outputs $([i]_i^k, x', [w_{i,1}]_i^k, [w_{i,2}]_i^k)$ such that the following properties hold:

- **Completeness:** For all $\alpha, \beta \in \mathbb{F}$, if $(i, x, (\mathbb{w}, M)) \in k$-RCSAT, then $(i, x', (w_{i,1}, w_{i,2})) \in CSAT$ for all $i \in [k]$ and $x$ is the 0-th leaf of $rt$.
- **Knowledge soundness:** There exists an efficient extractor $\mathcal{E}$ such that for every $A$ the following probability is negligible:

\[
\Pr \left[ \begin{array}{c}
(i, x, (\mathbb{w}, M)) \notin k$-RCSAT \\
\forall i \in [k] : (i, x', w_{i,1}, w_{i,2}) \in CSAT \\
x \text{ is the 0-th leaf of } rt
\end{array} \right] \geq \begin{array}{c}
\rho \leftarrow \mathcal{U}(\lambda) \\
[i]_i^r = f_1(i, \mathcal{C}) \\
ck \leftarrow CM$-Setup$(1^k, \mathcal{M}i) \\
aux_{\mathcal{V}} \leftarrow D(i) \\
([w_{i,1}]_i^m, rt) \leftarrow \mathcal{V}^p(i, aux_{\mathcal{V}}) \\
cm \leftarrow CM$-Commit$(ck, [w_{i,1}]_i^m) \\
(\alpha, \beta) \leftarrow \rho(cm) \\
x' = (1, rt, \alpha, \beta) \\
(\mathbb{x}, (\mathbb{w}, M)) \leftarrow \mathcal{E}_i^p(\mathbb{i}, aux_{\mathcal{V}})
\end{array}
\]

**Proof.** Completeness. Let $\alpha, \beta$ be arbitrary elements of $\mathbb{F}$ and parse $i$ as $(\mathbb{F}, \ell, n, s, R, \mathcal{R})$ and $x'$ as $(1, rt, \alpha, \beta)$. By construction, $x$ is the 0-th leaf of the Merkle tree with root $rt$.

If $(i, x, (\mathbb{w}, M)) \in k$-RCSAT, then $R^M(x, \mathbb{w}) = 0$, which in turn implies that we can construct appropriate memory sub-traces $[w_{i,1}]_i^k = (T_i, A_i)_i^k$ and witnesses $[w_{i,2}]_i^k$ such that $(i, x', (w_{i,1}, w_{i,2})) \in CSAT$ for all $i \in [k]$.

**Knowledge soundness.** We construct an extractor $\mathcal{E}$ such that if $(i, x', (w_{i,1}, w_{i,2})) \in CSAT$ for all $i \in [k]$, the probability that $\mathcal{E}$ fails to produce $(\mathbb{x}, (\mathbb{w}, M))$ such that $(i, x, (\mathbb{w}, M)) \in k$-RCSAT and $x$ is the 0-th leaf of $rt$ is negligible. $\mathcal{E}$ works as follows:
**Figure 4:** Reduction from $k$-RCSAT to $k$ instances of CSAT.
We prove the following claim about the extractor $\mathcal{E}$.

**Claim 4.10.** Given that $\left(\vec{i}, \vec{z}', (\vec{w}_{i,1}, \vec{w}_{i,2})\right) \in \text{CSAT}$ for all $i \in [k]$, if the extractor $\mathcal{E}$ does not abort, it outputs a valid $(\vec{z}, (\vec{w}, M))$ such that $(\vec{i}, \vec{z}, (\vec{w}, M)) \in k\text{-RCSAT}$.

**Proof.** Since the public input of $R$ is used only in the first subcircuit $R_1$, the fact that $\left(\vec{i}_1, \vec{z}', (\vec{w}_{i,1}, \vec{w}_{i,2})\right) \in \text{CSAT}$ and the extracted $\vec{z}$ is exactly the 0-th leaf of $rt$ implies that the public input of $R$ is consistent with $\vec{z}$, by construction of $C_1$.

Since $\mathcal{E}$ does not abort, this implies that all the values read from the Merkle tree are consistent amongst all $k$ subcircuits and $T$ and $A$ are permutations of each other.

By construction of each index $\left[1\right]_{i=1}^{k}$, and because for each $i$, $\left(\vec{i}_i, \vec{z}', (\vec{w}_{i,1}, \vec{w}_{i,2})\right) \in \text{CSAT}$, we have that the witness $\vec{w}$ and memory $M$ extracted by $\mathcal{E}$ are such that $R^M(\vec{z}, \vec{w}) = 1$. 

It is now sufficient to prove that the extractor aborts with negligible probability, which we do in the following claim.

**Claim 4.11.** Given that $\left(\vec{i}_i, \vec{z}', (\vec{w}_{i,1}, \vec{w}_{i,2})\right) \in \text{CSAT}$ for all $i \in [k]$, the extractor $\mathcal{E}$ aborts with negligible probability.

**Proof.** The extractor only aborts if the values read from the Merkle tree are not consistent amongst all $k$ subcircuits, or if $T$ and $A$ are not permutations of each other.

The probability that the adversary can open the same leaf to two different values is negligible, and so it must be the case that, except with negligible probability, the values read from the Merkle tree are consistent amongst all $k$ subcircuits.

Now, $\left(\vec{i}_i, \vec{z}', (\vec{w}_{i,1}, \vec{w}_{i,2})\right) \in \text{CSAT}$ for all $i \in [k]$, implies that $T$ is a valid time sorted trace, $A$ is a valid address sorted trace, and $T(\vec{z}, \vec{w}) = A(\vec{z}, \vec{w})$, where $T(\vec{z}, \vec{w}) := \prod_{j=1}^{n} (X - (t_j^T + Y \cdot i_j^T + Y^2 \cdot v_j^T))$ and $A(\vec{z}, \vec{w}) := \prod_{j=1}^{n} (X - (t_j^A + Y \cdot i_j^A + Y^2 \cdot v_j^A))$. Lemma B.1 now implies that, except with negligible probability, $T(X, Y) = A(X, Y)$, thus implying that $T$ and $A$ are permutations of each other. Thus overall the probability of abort is negligible.
5 Aggregation scheme for Mirage

We now describe our aggregation scheme for the Mirage commit-carrying zkSNARK (Section 5.1.1). We describe the aggregation scheme and prove it secure in Section 5.3. A key building block of our construction is a method to prove the correctness of multiple pairing products simultaneously. We describe this building block in Section 5.2.

Throughout this section we assume that the public inputs to each subcircuit are of the form $(1, rt, \alpha, \beta)$ as this is what the reduction in Section 4 mandates. We also assume $\text{Agg} \cdot C$ does not take any randomness as input as $\text{Agg}$ does not have to be hiding.

5.1 Background

We begin by recalling some results from prior work [BMMTV21; GMN22] that we will use in our construction.

Commitment schemes. The commitment scheme $\text{CM}_D$ [GMN22] is defined as follows.

<table>
<thead>
<tr>
<th>$\text{CM}_D.\text{Setup}(G^n_1, G^n_2)$:</th>
<th>$\text{CM}_D.\text{Commit}(ck_D, m = (A, B))$:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Sample $a, b \leftarrow F$.</td>
<td>1. Parse $ck_D$ as $(v_1, v_2, w_1, w_2)$.</td>
</tr>
<tr>
<td>2. Set $v_1 := [a H]<em>{i=0}^{n-1}$, and $v_2 := [b H]</em>{i=0}^{n-1}$.</td>
<td>2. Parse $v_1$ as $[a H]<em>{i=0}^{n-1}$, and $v_2$ as $[b H]</em>{i=0}^{n-1}$.</td>
</tr>
<tr>
<td>3. Set $w_1 := [a G]<em>{i=n}^{2n-1}$, and $w_2 := [b G]</em>{i=n}^{2n-1}$.</td>
<td>3. Parse $w_1$ as $[a G]<em>{i=n}^{2n-1}$, and $w_2$ as $[b G]</em>{i=n}^{2n-1}$.</td>
</tr>
<tr>
<td>4. Output $ck_D := (v_1, v_2, w_1, w_2)$.</td>
<td>4. Parse $A$ as $[A]<em>{i=0}^{n-1}$ and $B$ as $[B]</em>{i=0}^{n-1}$.</td>
</tr>
<tr>
<td>5. Set $T := \prod_{i=0}^{n-1} e(A_i, a H) \cdot \prod_{i=0}^{n-1} e(a^{-i} G, B_i)$.</td>
<td>5. Set $T := \prod_{i=0}^{n-1} e(A_i, b H) \cdot \prod_{i=0}^{n-1} e(b^{-i} G, B_i)$.</td>
</tr>
<tr>
<td>6. Set $U := \prod_{i=0}^{n-1} e(A_i, b H) \cdot \prod_{i=0}^{n-1} e(b^{-i} G, B_i)$.</td>
<td>6. Set $U := \prod_{i=0}^{n-1} e(A_i, b H) \cdot \prod_{i=0}^{n-1} e(b^{-i} G, B_i)$.</td>
</tr>
<tr>
<td>7. Output $C := (T, U)$.</td>
<td>7. Output $C := (T, U)$.</td>
</tr>
</tbody>
</table>

We will also use $\text{CM}_1$ and $\text{CM}_2$, two special cases of the above commitment scheme. In the former, the $G_2$ component of the message is ignored, while in the latter, the $G_1$ component is ignored. That is, $\text{CM}_1$ is a commitment scheme with message space $G^n_1$, while $\text{CM}_2$ is a commitment scheme with message space $G^n_2$. In fact, $\text{CM}_1$ is the same as the commitment scheme $\text{CM}_5$ from [GMN22].

We now define two relations associated with the above commitment schemes. These are slightly modified versions of the relation $\mathcal{R}_{\text{TIPP}}$ defined in [BMMTV21]:

**Definition 5.1** (separate pairing product relation). The indexed relation $\mathcal{R}_{\text{SPP}}$ is the set of triples

$$(i, x, w) = (ck_D, \text{cm}^1_A, \text{cm}^2_B, Z, r, (A', B))$$

where $ck_D$ is a commitment key for the commitment schemes $\text{CM}_1$ and $\text{CM}_2$ (Section 3.1), $\text{cm}^1_A$ and $\text{cm}^2_B$ are commitments under $\text{CM}_1$ and $\text{CM}_2$, $Z$ is the claimed pairing product of $A'$ and $B$, $r$ is a field element, and $A' \in G^n_1$ and $B \in G^n_2$ are vectors of group elements such that

$$\text{cm}^1_A = \text{CM}_1.\text{Commit}(ck_D, r^{-1} \circ A')$$

$$\text{cm}^2_B = \text{CM}_2.\text{Commit}(ck_D, B)$$

$$Z = A' \cdot B$$

**Definition 5.2** (combined pairing product relation). The indexed relation $\mathcal{R}_{\text{CPP}}$ is the set of triples

$$(i, x, w) = (ck_D, \text{cm}^D, Z, r, (A', B))$$

20
Theorem H.1 from [CFQ19] can easily be extended to prove knowledge soundness for ARG
Mirage. Obtain partial QAP witness $D$. For each $i$, $\ell \in [\ell_i, m_i, n_i, C_i]$ and construct from it the QAP index $(\mathbb{F}, \ell_i(X), a_i(X), b_i(X), c_i(X))$.
2. Sample $\alpha, \beta, [\gamma]_{i=1}, \delta_i \leftarrow \mathbb{F}$ and $[\delta_i]_{n=1}, [\eta]_{n=1} \leftarrow \mathbb{F}$.
3. For each $i \in [n]$ and $j \in [\ell_i + m_i + n_i]$, define $p_{ij}(X) := \beta a_{ij}(X) + \alpha b_{ij}(X) + c_{ij}(X)$.
4. For each $i \in [n]$, construct the verifying, committing, and proving keys, and output these:
   $$\begin{align*}
   ivk_i & := \left( e([\alpha]_1, [\beta]_{i=2}, [\gamma]_2, [\delta]_{i=2}, \left\{ (p_{ij}(s_i)/\gamma)_{i=1}^{\ell_i-1} \right\}_{j=0}^{\ell_i-m_i-n_i}, (p_{ij}(s_i)/\delta)_{i=1}^{\ell_i-m_i-n_i}, \left\{ (s_i/t_i(s_i)/\delta)_{i=1}^{\ell_i-m_i-n_i} \right\}_{j=0}^{\ell_i-m_i-n_i} \right) ;
   ck_i & := \left( [\delta]_1, \{ (p_{ij}(s_i)/\gamma)_{i=1}^{\ell_i-1} \right\}_{j=0}^{\ell_i-m_i-n_i},
   \right); 
   ic_k_i := ipk_i \end{align*}.$$
5. Output $(\{ipk_i\}_{i=1}^n, \{ic_k_i\}_{i=1}^n, \{ivk_i\}_{i=1}^n)$.

5.1.1 Mirage: commit-carrying Groth16

Below we describe the commit-carrying zkSNARK Mirage [CFQ19; KPPS20]. We note that while [KPPS20] does not have a formal statement regarding the binding property of Mirage, the ccGroth16 construction in [CFQ19] is morally the same as Mirage, except that the latter handles public inputs as well. As a result, Theorem H.1 from [CFQ19] can easily be extended to prove knowledge soundness for Mirage. For simplicity, in our use case, we use a construction of ARG where $ipk_i = ic_k_i$ for all $i \in [n]$.

 Mirage Setup($\lambda, \xi^n_{i=1}) \rightarrow (\{ipk_i\}_{i=1}^n, \{ic_k_i\}_{i=1}^n, \{ivk_i\}_{i=1}^n)$:
1. For each $i \in [1, \ldots, n]$, parse $i \ell_i$ as $(\mathbb{F}, \ell_i, m_i, n_i, C_i)$ and construct from it the QAP index $(\mathbb{F}, \ell_i(X), a_i(X), b_i(X), c_i(X))$.
2. Sample $\alpha, \beta, [\gamma]_{i=1}, \delta_i \leftarrow \mathbb{F}$ and $[\delta_i]_{n=1}, [\eta]_{n=1} \leftarrow \mathbb{F}$.
3. For each $i \in [n]$ and $j \in [\ell_i + m_i + n_i]$, define $p_{ij}(X) := \beta a_{ij}(X) + \alpha b_{ij}(X) + c_{ij}(X)$.
4. For each $i \in [n]$, construct the verifying, committing, and proving keys, and output these:
   $$\begin{align*}
   ivk_i & := \left( e([\alpha]_1, [\beta]_{i=2}, [\gamma]_2, [\delta]_{i=2}, \left\{ (p_{ij}(s_i)/\gamma)_{i=1}^{\ell_i-1} \right\}_{j=0}^{\ell_i-m_i-n_i}, (p_{ij}(s_i)/\delta)_{i=1}^{\ell_i-m_i-n_i}, \left\{ (s_i/t_i(s_i)/\delta)_{i=1}^{\ell_i-m_i-n_i} \right\}_{j=0}^{\ell_i-m_i-n_i} \right) ;
   ck_i & := \left( [\delta]_1, \{ (p_{ij}(s_i)/\gamma)_{i=1}^{\ell_i-1} \right\}_{j=0}^{\ell_i-m_i-n_i},
   \right); 
   ic_k_i := ipk_i \end{align*}.$$
5. Output $(\{ipk_i\}_{i=1}^n, \{ic_k_i\}_{i=1}^n, \{ivk_i\}_{i=1}^n)$.

 Mirage Commit($ic_k, w_{1, \ell}^c, c_k) \rightarrow D$:
1. Obtain $ck$ from $ic_k$ and parse $ck$ as
   $$\begin{align*}
   \left( [\delta]_1, \left\{ (p_{ij}(s_i)/\gamma)_{i=1}^{\ell_i-1} \right\}_{j=0}^{\ell_i-m_i-n_i}, \left\{ (s_i/t_i(s_i)/\delta)_{i=1}^{\ell_i-m_i-n_i} \right\}_{j=0}^{\ell_i-m_i-n_i} \right)
   \end{align*}.$$
2. Obtain partial QAP witness $w' = [w_i]_{i=0}^{\ell_i-1}$ from $w_{1, \ell}^c$.
3. Define $V_1(X) := \sum_{i=0}^{\ell_i-1} w_i p_{ij}(X)$.
4. Output $D := [V_1(s)/\gamma]_1 + [\kappa \delta]_1$.

 Mirage Verify($ivk, z_c, D, \pi) \rightarrow \{0, 1\}$:
1. Obtain the QAP instance $z_c$ from $z_c$.
2. Parse proof $\pi$ as $(A, B, C)$.
3. Check that $e(A, B) =
   e([\alpha]_1, [\beta]_{i=2} e\left( \sum_{i=0}^{\ell_i-1} \frac{z_c(p_{ij}(s))}{\gamma} \right)_{i=1}^{\ell_i-1}, [\gamma]_{i=1}^{\ell_i-1} e(C, [\delta]_2) e(D, [\eta]_2)$.

 Mirage Prove($ipk, z_c, (w_{1, \ell}^c, w_{1, m}^c, c_k, \eta) \rightarrow \pi$:
1. Parse $ipk$ as $\left( [\delta]_1, \left\{ (p_{ij}(s_i)/\gamma)_{i=1}^{\ell_i-1} \right\}_{j=0}^{\ell_i-m_i-n_i}, \left\{ (s_i/t_i(s_i)/\delta)_{i=1}^{\ell_i-m_i-n_i} \right\}_{j=0}^{\ell_i-m_i-n_i} \right)$.
2. Obtain the QAP instance $z_c$ from $z_c$ and the QAP witness $(w', w)$ from $(w_{1, \ell}^c, w_{1, m}^c)$.
3. Parse $\pi$ as the public input variables $[w_i]_{i=0}^{\ell_i-1}, w'$ as $[w_i]_{i=0}^{\ell_i-1}$ and $w$ as $[w_i]_{i=0}^{\ell_i-1}$.
4. Sample randomizers $\kappa_A, \kappa_B \leftarrow \mathbb{F}$, and set
   $$\begin{align*}
   a(A) & := \sum_{i=0}^{\ell_m-n-1} w_i a_i(X),
   p_A(X) := \alpha + p_A(X) + \kappa_A \delta.
   \end{align*}$$
   $$\begin{align*}
   b(B) & := \sum_{i=0}^{\ell_{m-n}} w_i b_i(X),
   p_B(X) := \beta + p_B(X) + \kappa_B \delta.
   \end{align*}$$
   $$\begin{align*}
   c(C) & := \sum_{i=0}^{\ell_{m-n-1}} w_i c_i(X),
   p_C(X) := \frac{h(x)(s)}{\delta} + [V_2(s)/\delta]_1 + \kappa_B \delta + \kappa_A \left[ p_B(s) \right]_1 - [\kappa_A \kappa_B \delta]_1 - [\kappa_B \eta]_1.
   \end{align*}$$
5. Compute $A := \frac{p_A(s)}{\delta}, B := \left[ p_B(s) \right]_2, C := \frac{h(x)(s)}{\delta} + [V_2(s)/\delta]_1 + \kappa_B A + \kappa_A \left[ p_B(s) \right]_1 - [\kappa_A \kappa_B \delta]_1 - [\kappa_B \eta]_1$.
6. Output $\pi := (A, B, C)$.
5.2 Proving multiple pairing products simultaneously

We now describe our protocol for showing the correctness of multiple pairing products, where the inputs to each pairing product are vectors of group elements committed under the commitment schemes described in Section 5.1. Concretely, we describe a protocol for the following relation.

**Definition 5.3 (multiple pairing product relation).** The indexed relation $\mathcal{R}_{MPP}$ is the set of triples

$$
\begin{pmatrix}
  i = \langle cm_D^{AB} \rangle \\
  x = \langle cm_1^{AB}, cm_2^{Z_{SY}} \rangle \\
  w = \langle A', B \rangle,
\end{pmatrix}
\begin{pmatrix}
  \text{ck} \\
  \langle Z_{AB} \rangle, \langle Z_{C_\delta} \rangle, \langle Z_{D_\eta} \rangle \\
  \langle \delta \rangle,
\end{pmatrix}
$$

where $\text{ck}$ is a commitment key for the commitment schemes $CM_D$, $CM_1$, and $CM_2$ (Section 3.1), and

- $(ck, \langle cm_D^{AB}, Z_{AB}, r \rangle, \langle A', B \rangle) \in \mathcal{R}_{CPP}$.
- $(ck, \langle cm_1^{AB}, cm_2^{Z_{SY}}, r \rangle, \langle \delta \rangle) \in \mathcal{R}_{SPP}$.
- $(ck, \langle cm_1^{C_\delta}, cm_2^{Z_{C_\delta}}, r \rangle, \langle C', \delta \rangle) \in \mathcal{R}_{SPP}$.
- $(ck, \langle cm_1^{D_\eta}, cm_2^{Z_{D_\eta}}, r \rangle, \langle D', \eta \rangle) \in \mathcal{R}_{SPP}$.

**Theorem 5.4.** Given a SNARK TIPP for the $\mathcal{R}_{CPP}$ NP relation, the construction MPP in Fig. 5 is a SNARK for the $\mathcal{R}_{MPP}$ relation. MPP achieves the following efficiency properties:

- The prover’s running time is the cost of running TIPP.Prove once.
- The proof size $|\pi|$ is $|\pi_{TIPP}| + 12|G_T|$.

**Proof.** Completeness. Completeness follows from the completeness of TIPP.

**Knowledge soundness.** We reduce to the knowledge soundness of the underlying TIPP protocol by using any successful adversary $A$ against MPP to construct an adversary $B$ against TIPP. We then use the extractor for the latter to construct an extractor $E_A$ for $A$.

$$
B'(pk_{TIPP}, \text{aux}) \rightarrow (x_{TIPP}, \pi_{TIPP})$

1. Set $pk_{MPP} = pk_{TIPP}$.
2. Obtain $(x, \pi) \leftarrow A'(pk_{MPP}, \text{aux})$.
3. Construct $x_{TIPP} := (\langle cm_1^{D_R}, Z_{L_R} \rangle, r)$ as in MPP.Prove.
4. Output $(x_{TIPP}, \pi_{TIPP})$.

Now we construct $E_A$ as follows. The high level idea is to have $E_A$ rewind $A$ with the same $x$, but obtaining TIPP proofs for $x$ with respect to 4 different random challenges $(s_1, t_1), (s_2, t_2), (s_3, t_3)$ and $(s_4, t_4)$. We then use the TIPP extractor $E_B$ to extract witnesses for each of these TIPP instances, and interpolate to obtain a valid witness $w$ for $\mathcal{R}_{MPP}$.

$$
E_A'(pk_{MPP}, \text{aux}) \rightarrow w$

1. Set $pk_{TIPP} = pk_{MPP}$.
2. Compute $(x, \pi = (Z_{CP}, \pi_{TIPP})) \leftarrow A'(pk_{MPP}, \text{aux})$.
3. For $i \in \{1, 2, 3, 4\}$:
   - (a) If $i \neq 1$, sample $(s_i, t_i) \leftarrow \mathbb{F}_2$ and program $p$ to return $(s_i, t_i)$ at $(vk_{TIPP}, x, Z_{CP})$.
   - (b) Extract TIPP witness $(L_i, R_i) \leftarrow E_B'(pk_{TIPP}, \text{aux})$. 

22
MPP.Setup($\lambda, i$) \rightarrow (pk_{\text{MPP}}, vk_{\text{MPP}}):
1. Sample $(pk_{\text{TIPP}}, vk_{\text{TIPP}}) \leftarrow \text{TIPP.Setup}(\lambda, i)$.
2. Set $(pk_{\text{MPP}} := (pk_{\text{TIPP}}, vk_{\text{TIPP}}), vk_{\text{MPP}} := vk_{\text{TIPP}})$.

MPP.Prove$^0(pk_{\text{MPP}}, x, w) \rightarrow \pi$:
1. Parse $pk_{\text{MPP}}$ as $(pk_{\text{TIPP}}, vk_{\text{TIPP}})$.
2. Parse $x$ as $(cm_{AB}^0, \ldots, cm_0^3, \ldots, cm_{\gamma}^3, \ldots, cm_{\eta}^3, \ldots, Z_{ip}, r)$ and $w$ as $((A', B), (\gamma'), (C', \delta), (D', \eta))$.
3. Parse the claimed inner products $Z_{ip}$ as $(Z_{AB}, Z_{SB}, Z_{SG}, Z_{DN})$.
4. Compute the cross-products $Z_{IP} := \{Z_{XY} = X \cdot Y : X \in \{A, S, C, D\}, Y \in \{B\gamma\delta\eta\}\} \setminus Z_{ip}$.
5. Compute random linear combination challenges $(s, t) := \rho(vk_{\text{TIPP}}, x, \bar{Z}_{CP})$.
6. Compute linear combinations of left and right inputs: $L = A' + s \cdot \gamma' \cdot \delta' \cdot \eta', C' + s \cdot \gamma' \cdot \delta' \cdot \eta', D' \cdot \delta' + s \cdot \gamma' \cdot \delta' \cdot \eta'$.
7. Compute inner product of combined left and right inputs: $Z_{LR} = L \cdot R = \prod_{X \in \{A, S, C, D\}, Y \in \{B\gamma\delta\eta\}} Z_{XY}^{(X)(Y)}$ where $s(A) = 1, s(S) = s, s(C) = s^2, s(D) = s^3, t(B) = 1, t(\gamma) = t, t(\delta) = t^2, t(\eta) = t^3$.

9. Assemble the $\mathcal{R}_{\text{TIPP}}$ instance and witness: $z' := ((cm_{LR}^0, Z_{LR}), r)$ and $w' := (L, R)$.
10. Run the TIPP prover to obtain $\pi_{\text{TIPP}} := \text{TIPP.Prove}(pk_{\text{TIPP}}, z', w')$.
11. Output $\pi := (Z_{CP}, \pi_{\text{TIPP}})$.

MPP.Verify$^0(pk_{\text{MPP}}, x, \pi) \rightarrow \{0, 1\}$:
1. Parse $vk_{\text{MPP}}$ as $vk_{\text{TIPP}}$.
2. Parse $x$ as $(cm_{AB}^0, \ldots, cm_0^3, \ldots, cm_{\gamma}^3, \ldots, cm_{\eta}^3, \ldots, Z_{ip}, r)$.
3. Parse $\pi$ as $(Z_{CP}, \pi_{\text{TIPP}})$.
4. Parse $Z_{CP}$ as $\{Z_{XY} = X \cdot Y : X \in \{A, S, C, D\}, Y \in \{B\gamma\delta\eta\}\}$ \setminus $Z_{ip}$.
5. Compute random linear combination challenges $(s, t) := \rho(vk_{\text{TIPP}}, x, \bar{Z}_{CP})$.
6. Reconstruct commitment: $\text{cm}_{LR}^0 = \text{cm}_{AB}^0 \cdot \text{cm}_{\gamma}^3 \cdot \text{cm}_{\delta}^2 \cdot \text{cm}_{\eta}^3$.
7. Reconstruct inner product of combined left and right inputs: $Z_{LR} = \prod_{X \in \{A, S, C, D\}, Y \in \{B\gamma\delta\eta\}} Z_{XY}^{(X)(Y)}$.

8. Assemble the $\mathcal{R}_{\text{TIPP}}$ instance: $z' := ((\text{cm}_{LR}^0, Z_{LR}), r)$.
9. Output $\text{TIPP.Verify}(vk_{\text{TIPP}}, z', \pi_{\text{TIPP}})$.

Figure 5: The MPP Protocol
4. Interpolate the extracted TIPP witnesses to obtain \( \mathsf{w} := (A, B, C, D) \).

The correctness of the extractor \( \mathcal{E}_A^D \) follows from Lemma B.1.

### Efficiency

The efficiency claims follow from inspection:

- \( \pi_{\text{MPP}} \) is of the form \((Z_{CP}, \pi_{\text{TIPP}})\), where \( Z_{CP} \) consists of 12 \( \mathbb{G}_T \) elements.

#### 5.3 Construction

We present the full construction of our aggregation scheme for Mirage in Fig. 6.

##### 5.3.1 Completeness and knowledge soundness proofs

**Completeness.** Completeness follows from the completeness of MPP and \( \text{CM}_D.\text{Commit} \).

**Knowledge Soundness.** Knowledge soundness reduces to the knowledge soundness of MPP and the binding property of \( \text{CM}_D.\text{Commit}. \) That is, given a successful adversary \( \mathcal{A} \) against Agg, we will construct an adversary \( \mathcal{B} \) against MPP. We will then use the knowledge soundness of the latter to construct an extractor \( \mathcal{E}_A \) for \( \mathcal{A} \). We begin by constructing \( \mathcal{B} \) for MPP as follows:

\[
B^D(\mathsf{pk}_{\text{MPP}}, \mathsf{aux}_B) \rightarrow (\pi_{\text{MPP}}, \pi_{\text{TIPP}}):
\]

1. Parse \( \mathsf{aux}_B \) as \((\mathsf{aux}_A, \mathsf{pk}_{\text{MPP}}))
2. Parse \( \mathsf{pk}_{\text{MPP}} \) as \((\mathsf{pk}_{\text{MPP}}, \mathsf{vk}_{\text{MPP}}, S, S', S, \gamma, \delta, \eta)\).
3. Parse verification key \( \mathsf{vk}_{\text{MPP}} = (\mathsf{vk}_{\text{MPP}}, \mathsf{cm}_1, \mathsf{cm}_1', \mathsf{cm}_2, \mathsf{cm}_2', \mathsf{cm}_3, \mathsf{cm}_3')\).
4. Obtain \((\pi, \mathsf{cm}_{\text{MPP}}), \pi) \leftarrow \mathcal{A}(\mathsf{pk}_{\text{MPP}}, \mathsf{aux})\).
5. Parse public input \( \mathsf{x} \) as \((1, \mathsf{rt}, \alpha, \beta)\), commitment \( \mathsf{cm}_{\text{MPP}} \) as \( \mathsf{cm}_D \), and proof \( \pi \) as \((\mathsf{cm}_D, \mathsf{cm}_C, \mathsf{Z}_{AB}, \mathsf{Z}_{SY}, \mathsf{Z}_{CD}, \mathsf{Z}_{D\eta})\).
6. Construct \( \pi_{\text{MPP}} \) and \( \pi_{\text{TIPP}} \) as in \( \mathcal{A}(\mathcal{P}, \mathcal{H}) \), and output these.

Agg.\( \mathcal{V} \) accepting implies that the batched SNARK verification equation holds and MPP.\( \text{Verify} \) accepts. The latter implies that \( \mathcal{E}_B \) outputs a valid witness \( \mathsf{w}_{\text{MPP}} = (A', B, S', (C'), (\delta'), (D', \eta)) \) for \( \pi_{\text{MPP}} \), except with negligible probability.

After rescaling \( A = r^{-1} \circ A', S = r^{-1} \circ S', C = r^{-1} \circ C' \) and \( [\mathsf{cm}]_{i=1}^n = r^{-1} \circ D' \), the batched SNARK verification equation \( \mathsf{Z}_{AB} = (e(\mathsf{[\alpha]}, \mathsf{[\beta]}), \mathsf{Z}_{SY}, \mathsf{Z}_{CD}, \mathsf{Z}_{D\eta}) \) implies that, except with negligible probability, for all \( i \in [n] \), Mirage.\( \text{Verify} (\mathsf{iv}_i, (1, \mathsf{rt}, \alpha, \mathsf{cm}_i, (A_i, B_i, C_i)) = 1 \) (by Lemma B.1).

This allows us to argue the success of the extractor \( \mathcal{E}_A \) constructed below:

\[
\mathcal{E}_A^D(\mathsf{pk}_{\text{MPP}}, \mathsf{aux}_A) \rightarrow \mathsf{w}_{\text{MPP}}:
\]

1. Set \( \mathsf{aux}_B := (\mathsf{aux}_A, \mathsf{pk}_{\text{MPP}}))
2. Obtain \( \mathsf{w}_{\text{MPP}} \leftarrow \mathcal{E}_B^D(\mathsf{pk}_{\text{MPP}}, \mathsf{aux}_B)\).
3. Parse \( \mathsf{w}_{\text{MPP}} \) as \((A', B', S', (C'), (\delta'), (D', \eta))\).
4. Rescale \( A = r^{-1} \circ A', S = r^{-1} \circ S', C = r^{-1} \circ C' \) and \( [\mathsf{cm}_i]_{i=1}^n = r^{-1} \circ D' \), c
5. Output \( \mathsf{w}_1 := D, \mathsf{w}_2 := (A_i, B_i, C_i)_{i=1}^n \).

#### Efficiency

The efficiency claims follow from inspection:

- \( \text{Agg.}\mathcal{P} \) runs MPP.\( \text{Prove} \) once, \( \text{CM}_D.\text{Commit} \) once, and \( \text{CM}_1.\text{Commit} \) once.
- \( \pi_{\text{Agg}} \) is of the form \((\mathsf{cm}_D, \mathsf{cm}_C, \mathsf{Z}_{IP}, \pi_{\text{MPP}})\), where \( \mathsf{Z}_{IP} \) consists of 4 \( \mathbb{G}_T \) elements and \( \mathsf{cm}_D \) and \( \mathsf{cm}_C \) are 2 \( \mathbb{G}_T \) elements each.
\textbf{Figure 6:} Our aggregation scheme for Mirage.
6 Divide-and-aggregate zkSNARKs

The following theorem describes our main result: a divide-and-aggregate (DNA) zkSNARK for CSAT that supports both the distributed and low-memory prover workflows.

**Theorem 6.1.** Consider the following ingredients:
- an inner commit-carrying zkSNARK ARG for CSAT (Section 3.2), and
- an aggregation scheme Agg for ARG (Section 3.3).

Then the construction in Section 6.1 is a zkSNARK for CSAT that has a horizontally-scalable distributed prover (Section 6.2).

**Remark 6.2.** Our construction can be instantiated with a variety of different ingredients, including aggregation schemes based on proof-carrying data [CT10; BCTV14a; BCMS20; BCLMS21]. However, it fails to capture divide-and-aggregate zkSNARKs which leverage fine-grained aggregation *within* the execution of the inner zkSNARK. An example of such a scheme is the aPlonk SNARK [ABST23], which aggregates separately the polynomial commitments generated in each round of the Plonk SNARK. We leave the task of extending our construction to capture such zkSNARKs to future work.

6.1 Construction

**Generator** DNA, $G$. On input the security parameter $\lambda$ and the CSAT index $i = (F, k, \ell, n, C, \mathcal{C})$, compute the index-specific proving and verifying keys ($ipk, ivk$) as follows.
1. Reduce $i$ to multiple CSAT indices: $[i]_i^n := f_i(i)$.
2. Sample keys for CSAT indices: $(|ipk|_i^n, |ivk|_i^n, [ivk]_i^n) \leftarrow ARG.G(\lambda, [i]_i^n)$.
3. Construct Agg index: $i_{Agg} := [ivk]_i^n$.
4. Sample keys for Agg: $(pk_{Agg}, ck_{Agg}, vk_{Agg}) \leftarrow ARG.G(\lambda, i_{Agg})$.
5. Construct verifying key: $ivk := vk_{Agg}$ and proving key $ipk := (i, ivk, [ipk]_i^n, pk_{Agg})$.
6. Output $(ipk, ivk)$.

**Prover** DNA, $P$. Given oracle access to a random oracle $\rho$, and on input the proving key $ipk$, the CSAT instance $\varpi$, and the CSAT witness $w$, compute the proof $\pi$ as follows.
1. Parse the proving key as $ipk = (i, ivk, [ipk]_i^n, pk_{Agg})$.
2. Compute the memory subtraces: $[w]_i^n = f_{w1}(i, \varpi, w)$.
3. For each $i$ in $\{1, \ldots, n\}$, obtain $ick_i$ from $ipk$, and commit to $w_i': cm_i \leftarrow ARG.C(ick_i, w'_i)$.
4. Obtain $ck_{Agg}$ from $pk_{Agg}$ and commit to $[cm]_i^n: cm_{Agg} \leftarrow Arg.C^{\rho}(ck_{Agg}, [cm]_i^n)$.
5. Compute the challenges $(\alpha, \beta) = \rho(ipk, cm_{Agg})$.
6. Compute instances and witnesses for subcircuits: $(\varpi_{Agg}, [w_i]_i^k) = f_{x,w2}(i, \varpi, w, \alpha, \beta)$.
7. For each $i$ in $\{1, \ldots, n\}$, compute the cc-zkSNARK proofs: $\pi_i \leftarrow ARG.P(ipk, i, \varpi_{Agg}, (w'_i, w_i))$.
8. Parse $\varpi_{Agg}$ as $(1, rt, \alpha, \beta)$.
9. Compute a membership proof $\pi_{rt}$ asserting that the first leaf of the Merkle tree with root $rt$ is $\varpi$.
10. Assemble $\mathcal{H}_{Agg}$ witness: $w_{Agg} := ([cm]_i^n, [\pi]_i^n)$.
11. Compute aggregated proof: $\pi_{Agg} \leftarrow Arg.P^{\rho}(pk_{Agg}, \varpi_{Agg}, w_{Agg}, cm_{Agg})$.
12. Output $\pi := (\varpi_{Agg}, cm_{Agg}, \pi_{Agg}, \pi_{rt})$.

**Verifier** DNA, $V$. Given oracle access to a random oracle $\rho$, and on input the verification key $ivk$, the CSAT instance $\varpi$, and the proof $\pi$, compute the verdict $b$ as follows.
1. Parse the verification key as $ivk = vk_{Agg}$.
2. Parse $\pi$ as $(\varpi_{Agg}, cm_{Agg}, \pi_{Agg}, \pi_{rt})$. 26
3. Parse \( x_{\text{Agg}} \) as \((1, rt, \alpha, \beta)\).
4. Check that \( \pi_{rt} \) is a valid proof that \( x \) is the first leaf of the Merkle tree with root \( rt \).
5. Check that \((\alpha, \beta) = \rho(vk_{\text{Agg}}, cm_{\text{Agg}})\).
6. Check that \( \text{Agg.} \psi(vk_{\text{Agg}}, x_{\text{Agg}}, cm_{\text{Agg}}, \pi_{\text{Agg}}) = 1 \).

6.2 Distributed prover workflow

We now describe at a high level the workflow for distributed proving of the foregoing zkSNARK. Proving responsibilities are split between a central coordinator and \( n \) workers.

1. **Setup:** The coordinator generates and partitions the memory traces, and distributes the subtraces and witness shares to the respective workers.

2. **Commit:** In parallel, each worker invokes \( \text{ARG}.C \) to commit to its subtrace, and sends the resulting commitment back to the coordinator.

3. **Challenge:** The coordinator aggregates the received commitments via \( \text{Agg}.C \), and uses the resulting commitment to derive a challenge \( r \) which it sends to each worker.

4. **Prove:** In parallel, each worker invokes \( \text{ARG}.P \) to compute its proof \( \pi_i \) for its subcircuit, and sends the resulting proof back to the coordinator.

5. **Aggregate:** The coordinator aggregates the received proofs via \( \text{Agg}.P \) to compute the final proof \( \pi \).

Note that “Commit” and “Prove” steps above are highlighted in **brown** in Section 6.1.

**Efficiency analysis.**

- Per-worker running time is the cost of running \( \text{ARG}.C \) and \( \text{ARG}.P \) on input \( C_i \).
- Per-worker communication cost is the size of a single proof and a single commitment of \( \text{ARG} \), \(|\pi_{\text{ARG}}| + |CM_{\text{ARG}}|\).
- The primary node’s active compute time is dominated by the cost of running \( \text{Agg}.C \) and \( \text{Agg}.P \) on inputs of size \( n \).
- Proof size is \(|\pi| = |\pi_{\text{Agg}}| + |cm_{\text{Agg}}| + |\pi_{rt}| + 3|\pi|\).

6.3 Optimizations

When the underlying ccSNARK \( \text{ARG} \) relies on circuit-specific setup, it is not clear how to obtain an SRS that grows only with the number of unique subcircuits because the construction in Section 6.1 creates subcircuits which have hardcoded in them the subcircuit number as well as the indices of the accessed shared wires. Since the latter are unique to each subcircuit, the SRS size would grow with the number of subcircuits.

To address this, we can modify the construction in Section 6.1 via the following optimization: instead of hardcoding the aforementioned values, we will instead provide these as **public input** to each subcircuit; since these values are known at setup time, the computations required for this public input handling can also be done during setup. This would ensure that we only need to perform one setup per unique subcircuit independent of how it is wired with respect to other subcircuits or how it is numbered. We provide details next. Assuming each subcircuit takes at most \( n \) public inputs, they can be preprocessed in \( \text{Agg}.G \) as follows:

1. For all \( i \in [n] \) parse ivk, as \( (e([\alpha]_1,[\beta]_2),[\gamma]_2,[\delta]_2,[\eta]_2,\{[p_{i,j}(s)]_{j=1}^n\}) \).
2. For each \( i \in [n] \) let the \( i \)-th subcircuits public inputs be \((1, rt, \alpha, \beta, v_{i,5}, \ldots, v_{i,u})\).
3. For each \( j \in \{5,6,\ldots,u\} \), multiply the \( G_1 \) elements of \((ivk)_{i=1}^n\) corresponding to public inputs with the appropriate coefficients to obtain \( S_j := (v_{1,j}, [p_{1,j}(s)]_{1},\ldots,v_{n,j}, [p_{n,j}(s)]_{1}) \).
4. Replace \( S_1 \) with \( \check{S} := S_1 + \sum_{j=5}^u S_j \).
5. Replace \( cm_{S_1}^1 \) with a commitment to \( \tilde{S} \): \( cm_{S_1}^1 \leftarrow CM_1 . \text{Commit}(ck_D, \tilde{S}) \).

6. Set \( vk_{Agg} := (vk_{MPP}, cm_{S_1}^1, cm_{S_2}^1, cm_{S_1}^2, cm_{S_1}^3, cm_{S_1}^4) \).

7. Set \( pk_{Agg} := (pk_{MPP}, ck_{Agg}, vk_{Agg}, \tilde{S}, S_2, S_3, S_4, \gamma, \delta, \eta) \).

Now in \( \text{Agg.} \mathcal{P} \) and \( \text{Agg.} \mathcal{V} \) we can:

1. Compute public input vector \( S = \tilde{S} + rt \cdot S_2 + \alpha \cdot S_3 + \beta \cdot S_4 \).
2. Compute public input commitment \( cm_{S_1}^1 := (cm_{S_1}^1)^{rt} \cdot (cm_{S_3}^1)^{\alpha} \cdot (cm_{S_4}^1)^{\beta} \).
3. We can then proceed by using \( S \) and \( cm_{S_1}^1 \) as usual.
7 Implementation

We implemented HEKATON in 5400 lines of Rust code using the arkworks framework [Ark22]. Our library implements: (1) a generic API to describe partitionable circuits with shared wires, (2) an implementation of the Mirage cc-zkSNARK, (3) the aggregation scheme for Mirage described in Section 5, (4) an OpenMPI implementation of coordinator and worker nodes for distributed proving described below, and (5) implementations of the applications described in Section 1 atop HEKATON. Our code will be open-sourced shortly.

**Design choices for distributed proving.** Our library is cluster-agnostic and highly configurable. Due to space constraints, we defer further discussion of design decisions to Appendix A.

We use the OpenMPI message passing framework for communication, and SLURM for job orchestration. We implement our system using two rounds of MPI scatter-gather: one for the *Commit phase* and one for the *Proving phase*. Our system is highly configurable: each MPI node has multiple threads to run tasks, and the number of nodes, threads per node, and memory per thread is configurable at runtime.
8 Evaluation

In this section we evaluate the performance of our built system HEKATON via a set of scalability experiments (Section 8.2) and via two applications: verifiable key directories (Section 8.3) and verifiable RAM computation (Section 8.4). Overall, our experiments suggest that HEKATON indeed achieves its goal of horizontal scalability, and furthermore, improves upon existing solutions for the aforementioned applications.

8.1 Experimental setup

**Hardware.** Our experiments are run on a large multi-tenant HPC cluster. The cluster consists of nodes with AMD EPYC 7763 processors with 128 cores and 512GB of RAM. The cluster costs $0.00232 per core-hour. (For comparison, on-demand pricing for the equivalent AWS instance (m6a.32xlarge) is $0.0432 per core-hour).

**Circuit choice.** For most zkSNARKs, prover performance is independent of the circuit structure. This is true for Mirage, and so should also hold (modulo shared wires) for HEKATON. We want a circuit that, when partitioned into subcircuits, exercises the full features of HEKATON, yielding heterogeneous subcircuits and shared wires. We generate 5 distinct subcircuits, each with a tunable number of shared wires between them and an adjustable number of constraints. This adjustability is necessary both for benchmarking, and also for prototyping and performance tuning early iterations of HEKATON. We define the larger circuit as a power-of-two number of subcircuits.

**Shared wire costs are low.** We find that shared wires cost 13 constraints each. At the scale we aim for, this adds a negligible overhead to the cost of subcircuit proving. Indeed, for any reasonable computation, the additional cost of handling shared wires is negligible compared the cost of the core computation. For our end-to-end experiments, we arbitrarily fix the proportion of constraints that come from shared wires to 10%.

**Parameters for distributed proving.** As described in Section 7, HEKATON’s distributed prover is highly configurable. To determine an appropriate configuration for our cluster, we ran preliminary experiments which confirmed our intuition to allocate each subcircuit its own core (this avoids diminishing returns of multithreading subcircuit proof generation). Further exploration indicated that, accounting for miscellaneous memory overhead, we could not reliably exceed 3.5GB of memory for the prover. This translates to proving a 1.3 million constraint subcircuit. For our distributed experiments, we fix this as the subcircuit size and vary total circuit size by adding subcircuits.

We found that a configuration where each MPI node runs 32 provers, each equipped with one core and 4GB of memory, was the optimal setting for our workload.

8.2 Scaling experiments

Figure 7 plots latency and throughput for HEKATON, for a varying number of cores, as the size of the circuit increases. We stress that, in a departure from previous work [WZCPS18; LXZSZ24], the numbers are end-to-end, counting the full cost of pre-positioning data, serialization and deserialization, data transmission, and computation.

**Prover latency and throughput.** Both latency and throughput scale well. As shown in the latency graph, Figure 7a, workers appear to initially starve, having insufficient work to achieve full utilization. However, after initial starvation, latency stabilizes. The throughput in Figure 7b follows the same pattern.

For an ideal system, we would expect end-to-end latency to scale linearly; doubling the circuit size should double the runtime Indeed this is the largely case: of all the plots in Figure 7a, the worst fitting simple linear
Figure 7: HEKATON proving performance for a fixed set of cores as total circuit size increases.

regression line is a 98.58% fit.4

Communication costs. We report the communication costs per subcircuit in Table 2. The largest overhead is less than 1MB per subcircuit. As expected, commit and proving phase responses sizes do not vary with any circuit parameters. Commit phase request sizes do not vary with number of circuits.

<table>
<thead>
<tr>
<th>Number of subcircuits</th>
<th>Commit phase request response</th>
<th>Proving phase request response</th>
</tr>
</thead>
<tbody>
<tr>
<td>$2^8$</td>
<td>↑</td>
<td>577B ↑</td>
</tr>
<tr>
<td>$2^{11}$</td>
<td>923kB ↑</td>
<td>673B 496B ↑</td>
</tr>
<tr>
<td>$2^{14}$</td>
<td>↓</td>
<td>769B ↓</td>
</tr>
</tbody>
</table>

Table 2: Maximum communication costs per subcircuit for each round of HEKATON.

Proof size and verification time. Proofs size and verifier time grow logarithmically. Concretely, in our largest test at $2^{18}$ subcircuits, proofs are 32kB and take 83ms to verify. We did not attempt to optimize the latter cost, but expect that it can be reduced greatly.

Aggregation times. We report aggregation times in Fig. 8. As expected, aggregation time grows linearly with number of subcircuits, topping out at $\approx 100100$ s for $2^{16}$ subcircuits. While this is non-trivial, we expect that the aggregation itself can be distributed across multiple nodes. We leave this task to future work.

Overhead relative to existing provers. Recall that HEKATON is modular, using an existing monolithic inner cc-zkSNARK ARG to prove subcircuits. We now investigate the overhead added by HEKATON relative to ARG (which in the experiments is Mirage).

Overhead comes from two sources, First, for every subcircuit, HEKATON adds extra constraints for shared wires that are not needed in a single monolithic prover. Second, HEKATON imposes overhead for the additional machinery for the commit-carrying aggregation scheme and extra communication.

4Computed as $r^2$, or the coefficient of determination of the model.
To measure total overhead, we compare normalized throughput in constraints per core-second (cpcs) of a monolithic Mirage prover versus HEKATON. To set the baseline maximum normalized throughput, we run the monolithic prover using just 1 thread. This operates at 38kcpcs. For comparison, HEKATON achieves 34kcpcs (90% baseline) at 4 threads.

We now look at how this overhead scales when we add more compute cores. For context, the normalized throughput for a perfectly parallelizable problem would be identical regardless of the number of cores. When we scale up our baseline monolithic Groth16 prover by $32\times$, the normalized throughput drops dramatically to 1.2kcpcs (5% of the baseline). Meanwhile, when we scale up HEKATON by $32\times$, the normalized throughput drops to 33kcpcs (87% baseline). When we scale again by $32\times$, it drops to 12kcpcs (64% baseline). This is promising: HEKATON is able to utilize its additional resources much more effectively than the monolithic prover.

**Comparison to Pianist.** As noted in Section 1.2, the most closely related prior work to ours is Pianist [LXZSZ24], and so we tried to benchmark their implementation on our cluster. Unfortunately, despite significant effort, we were unable to get their implementation to run stably on our cluster for circuit sizes larger than $2^{21}$ constraints, or with more than 2 worker nodes. The numbers that we did obtain, however, are in line with the numbers reported in Pianist.

For example, the running time of Pianist when trying to prove $2^{21}$ constraints with 128 cores, each with 4GB of memory, is the same in our setup as in theirs, at around 9.3s.\(^5\) This allows us to perform a rough extrapolated comparison of the performance of the two systems, which we do next.

Pianist reports a runtime of 1 second for a circuit of size $2^{22}$ and, for their largest measured circuit, sub-10s runtimes for $2^{25}$ constraints. Both of these are on 2048 cores, the most powerful hardware configuration they benchmark for general circuits. In contrast, the smallest circuit size we benchmark ($2^{27}$) takes 11.6s to prove on 2048 cores in the same configuration. Extrapolating Pianist’s numbers to this many constraints indicates that HEKATON is able to achieve $\approx 3 \times$ lower latency.

This gap is despite a couple of factors in Pianist’s favor. First, Pianist uses the BN254 curve, which is faster than BLS12-381 used by HEKATON, but has worse security, estimated to be 100 bits [BD19]. Updating Pianist to use the latter curve would likely lead to a further slowdown in their reported numbers.

Second, Pianist’s circuit sizes are for Plonk constraints [GWC19]. Liu et al. [LXZSZ24] themselves report that equivalent R1CS circuits would be $\approx 7$ times smaller, meaning that the effective number of constraints is much smaller.

---

\(^5\)See [LXZSZ24, Figure 2]. Numbers are approximate as this is a log-log plot.
constraints proved by Pianist in a given time is $\approx 7$ times smaller than that proved by HEKATON. (However, keep in mind that optimized circuits and custom Plonk gates might reduce this overhead.)

Overall, these factors combine to indicate that HEKATON is able to achieve better performance in practice.

8.3 Application: verifiable key directories

A Verifiable Key Directory (VKD) [MBBFF15; CDGM19] is a primitive that allows clients to monitor a key-value mapping maintained by an untrusted server that a service is distributing on their behalf. VKDs are commonly used to maintain public key registries for end-to-end encrypted messaging services [EA23; LL23] in a way that prevents the server from tampering with the registry by changing users’ public keys undetectably.

In a VKD, the server maintains a mapping from a client-specific identifier $k$ to a tuple $(i, v)$, where $i$ is a version number recording the number of times this entry has been updated, and $v$ is the current value. A client can request the server to either insert a new key-value pair $(k, (0, v))$, or to update an existing key $k$ to a new value $v'$ while incrementing the version number. This mapping is committed inside a Merkle tree, and both operations come with “proofs” that the server has indeed performed the requested operation correctly. Periodically at each epoch $t$, the server signs and publishes a digest $d_t$ corresponding to the current state of the tree.

To verify the integrity of its stored value, a client can request from the server a proof attesting that the value $v$ is indeed the one committed in the latest published digest $d_t$. Existing VKDs use this mechanism to allow clients to audit the integrity of their mappings over time: they can request proofs for the value of their key at different epochs, and verify that the server has not tampered with their key-value mapping. Unfortunately, this mechanism is not scalable: if the client goes offline between epochs $t$ and $t'$ that are far apart, they would need to download and verify all intervening digests and proofs to ensure integrity.

Baseline approach. Recent work [TFZBT22] proposes to improve the efficiency of this process via invariant proofs. Roughly, such a proof would assert that if at any point between $t$ and $t'$ the server updated the value of $k$, then the server must have also incremented the version number. Now, if version numbers $i_t$ and $i_{t'}$ are equal, then the invariant proof guarantees that the value $v$ was unchanged between epochs $t$ and $t'$.

Tyagi et al. [TFZBT22] construct constant-sized invariant proofs by replacing Merkle trees with RSA-based authenticated dictionaries, and using SNARKs to prove the preservation of the invariant. Unfortunately, their reliance on RSA accumulators makes them incompatible with existing systems that already use Merkle tree based registries.

Our approach. We leverage HEKATON to provide invariant proofs for existing registries by proving that all operations performed by the server on the key-value mapping preserve the versioning invariant and correctly update the Merkle tree.

In more detail, we consider a VKD with a SHA-256-based Merkle tree of depth 128. Our circuit verifies a batch of update operations on the VKD, each of which requires incrementing the user’s version number and checking two Merkle paths. Because the task of checking even a single Merkle path is too onerous for a single subcircuit, we partition each path check into four subcircuits. Each subcircuit passes its intermediate computation to the next, until the root is computed and checked against the expected root.

Experimental comparison. We measure total proving latency as we increase the size of the batch of VKD updates. We use 4096 cores, with the same core allocation as in the previous section. Each subcircuit is 2.4M constraints. Prover performance is illustrated in Fig. 9; observe that prover throughput increases as we increase the batch size.
Figure 9: End-to-end runtime for proving batches of updates in a Merkle-tree-based VKD; the x-axis indicates the batch size.

8.4 Application: verifiable RAM computation

Numerous applications are most easily and efficiently expressed as RAM computations. As a result, much industrial and academic effort has been invested in constructing efficient “zkVMs” that can verify the correctness of RAM computations [BCGTV13; BCTV14b; ZGKPP18; Sta21; AST24].6 In this section, we show that HEKATON can scale to prove such RAM computations very efficiently, focusing on the TinyRAM instruction set.

Circuit construction. Our circuit consists of repeated CPU “cycle” circuits. Each of the latter contains two modules: an ALU that executes TinyRAM instructions, and a memory checker that verifies memory accesses. We instantiate the ALU module with dummy constraints that do not actually execute TinyRAM instructions, but whose cost equals that of a real ALU. The resulting dummy ALU module costs roughly 1,114 constraints (we obtained this number from the breakdown in [BCTV14b, Figure 7]).

We instantiate the memory checking module in two ways. The first constitutes our baseline approach that uses online memory checking via Merkle trees. Indeed, as discussed in Section 2.1.1, prior to our work, this was the only way to verify memory accesses in a distributed setting. We use a Merkle tree of depth 32 with the Poseidon hash function (concretely, two Merkle path verifications).

The second instantiation is obtained by adapting our permutation-based memory checker to the read-write setting in a standard way [BCTV14b]. This adaptation does not fundamentally change any asymptotic costs, and only increases the cost of the local checks by a constant factor.

Again, we use 4096 cores, with the same core allocation as in the previous section. Each Merkle-backed subcircuit is 2.4M constraints, and each shared wire-backed subcircuit is 1.6M constraints.

Experimental comparison. The performance of the two approaches is illustrated in Fig. 10. It can be clearly seen that the permutation-based memory checker offers much higher throughput (in terms of cycles per second) than the Merkle tree-based memory checker.

---

6The name zkVM is a misnomer, as many of these works do not provide zero knowledge, but rather only succinct proofs.
Figure 10: Comparison of end-to-end runtime for proving a TinyRAM computation using a Merkle tree-based memory checker versus a permutation-based memory checker.
Acknowledgements

We thank Aakash Dutt for contributions to an early stage of our implementation.
A Design choices and cluster architecture

There are many choices for using a cluster for the above fan-out/fan-in approach. Below we consider the design choices.

Breaking down the job. Firstly, we must define precisely what a worker does. In the SLURM clustering architecture, a job (a complete end-to-end computation) is broken up into multiple tasks. Each task can be thought of as an individual process, with access to some number of threads, in communication with the other tasks, which possibly reside on the same node (physical machine) or other nodes. For HEKATON, a worker task could reasonably be a single-threaded or multi-threaded process which proves one subcircuit or many subcircuits. After experimentation, we found it is most efficient to have a worker task be a multithreaded process, which batch executes a many (specifically 32) single-threaded provers. For a fixed number of subcircuits, it is more efficient to have many single-threaded provers than few multi-threaded provers because of the sublinear speedup conferred by adding threads to proving.

Scheduling. Beyond the structure of tasks, there is also the choice of scheduling of tasks. Two options here are a work-stealing queue, wherein workers ingest subcircuit proof requests as threads open up, or fixed allocation, wherein every worker is given its set of proof requests in advance, and it must work through all of them. There are meaningful tradeoffs in both cases. A work-stealing queue is an adaptive strategy, and thus optimal in a theoretic sense because it avoids over- or under-utilization due to environmental conditions of the cluster. On the other hand, this method requires a large amount of coordination in the form of a literal queue, which must be available and highly responsive to every worker performing the job. The fixed allocation method is suboptimal from a utilization perspective, since you may have workers which complete all their requests while other workers still have many remaining. On the other hand, far less coordination is required, as there is no queue to maintain. For simplicity in our implementation, we choose the fixed array allocation. But we leave the queue method as an interesting avenue for future optimizations.

Our final note on scheduling: rather than bringing up N workers for committing, tearing them down, and bringing up N workers for proving, we simply bring them up once and use them for both committing and proving. Again, this is not perfectly optimal in terms of utilization, since some workers may be idle while stragglers finish their work. But we avoid an extra set of setup and teardown costs for our job. In addition, it simplifies our deployment, as our entire worker job is encapsulated by a single binary, which we execute via mpirun on our cluster.

Communication channels. Communication in a cluster can operate one of two ways: via MPI channels for small, low-latency transfers; and via the cluster filesystem (BeeGFS) for large, high-bandwidth transfers. The largest single piece of data in our experiments is the bundle of Groth16 proving keys, which reaches up to 5GB. Since this data must be precomputed in order for the protocol to run, it must exist in nonvolatile storage anyway. From experimentation, we found that the optimal data loading procedure is to have the coordinator load the proving keys from the filesystem, and use MPI for all further communications, including sending workers the proving key bundle.

B Additional definitions and lemmas

We now recall formal definitions and lemmas that we will use in our proofs and algorithm descriptions in subsequent appendices.
B.1 Zero-finding lemma

We recall a simplification of a useful lemma from \[BCMS20; BCLMS21\] that bounds the probability that applying a random oracle to a commitment to a polynomial yields a root of the polynomial.

**Lemma B.1.** Let $CM = (\text{Setup}, \text{Commit})$ be a commitment scheme for a message space universe $\{M_i\}_{i \in \mathbb{N}}$. Let $\mathbb{F}$ be a finite field, $M \in \mathbb{N}$ a number of variables, and $D \in \mathbb{N}$ a total degree bound. Then, for large enough message size $i \in \mathbb{N}$, every family of (possibly inefficient) functions $\{f : M_i \rightarrow \mathbb{F}^{\leq D}[X_1, \ldots, X_M]\}$ mapping messages to polynomials of degree at most $D$ over $\mathbb{F}$, and every $t$-query oracle algorithm $A$ that runs in expected polynomial time, the following holds:

$$
\Pr\left[ \begin{array}{c}
p \neq 0 \\
p(z) = 0
\end{array} \right] \leq \operatorname*{negl}(\lambda).
$$

If $CM$ is perfectly binding, then the above holds also for computationally-unbounded $t$-query adversaries $A$.  

38
C Proof of Theorem 6.1

We now prove Theorem 6.1 which asserts that the protocol DNA in Section 6.1 satisfies the completeness and knowledge soundness properties. We assume that a partitioning of the circuit is provided during setup and proving.

C.1 Completeness

Completeness follows from the completeness of $f$, ARG and Agg. To see this, we first recall the following definitions.

The completeness of $f = (f_1, f_{w_1}, f_{w_2})$ defined in Section 4 gives that for all $\alpha, \beta \in \mathbb{F}$ and $(i, x, w) \in k$-CSAT, on input $(i, x, w, \alpha, \beta)$, $f$ outputs $([i]_i^{n}, x', [w_{i,1}]_i^{n}, [w_{i,2}]_i^{n})$, such that $(i, x', (w_{i,1}, w_{i,2})) \in$ CSAT for all $i \in [n]$ and $x$ is the 0-th leaf of $rt$.

The completeness of Agg gives that for every index $i_{Agg}$ and every adversary $B$, the following probability is 1.

$$\Pr \left[ \begin{array}{c}
(i_{Agg}, x_{Agg}, w_{Agg}) \notin \mathcal{R}_{Agg} \\
Agg.\mathcal{V}^\rho(v_{Agg}, x_{Agg}, c_{Agg}, \pi_{Agg}) = 1
\end{array} \right] = \mathbb{P}(\rho \leftarrow U(\lambda))$$

The completeness of ARG gives that, for every set of indices $[i]_i^{n}$ and every adversary $C$, the following probability is 1.

$$\Pr \left[ \begin{array}{c}
\forall i \in [n] : \\
\downarrow
\end{array} \begin{array}{c}
(i_i, x_i, w_i) \in \mathcal{R} \\
Agg.\mathcal{V}^\rho(v_{ik}, x_i, c_{ik}, \pi_i) = 1
\end{array} \right] = \mathbb{P}(\rho \leftarrow U(\lambda))$$

Construction of adversaries. Given an index $i$ and an adversary $A$ for DNA, we can define $i_{Agg}$ and $B$ for Argg and $[i]_i^{n}$ and $C$ for ARG as follows:

1. Obtain the partitioned indices $[i]_i^{n} \leftarrow f_i(i)$.
2. Sample ARG public parameters: $([ipk]_i^{n}, [ick]_i^{n}, [ivk]_i^{n}) \leftarrow ARG.G^\rho(1^\lambda, [i]_i^{n})$.
3. Set $i_{Agg} := [ivk]_i^{n}$.

If $A(ipk)$ outputs $(x, w)$, do the following:

1. Partition $w$ to obtain the subcircuit witnesses to be committed to: $[w_{i,1}]_i^{n} \leftarrow f_{w_1}(i, x, w)$.
2. For each $i \in \{1, \ldots, n\}$, commit to $w_{i,1}$: $c_{i} \leftarrow ARG.C(ick_i, w_{i,1})$.
3. Commit to $[cm]_i^{n} : cm_{Agg} \leftarrow Agg.C^\rho(c_{Agg}, [cm]_i^{n})$.
4. Compute the challenge $(\alpha, \beta) := \rho(v_{Agg}, c_{Agg})$.
5. Obtain the witnesses that depend on the committed witnesses: $(x_{Agg}, w_{i,2}]_i^{n} \leftarrow f_{x, w_2}(i, x, w, \alpha, \beta)$.
6. For each $i \in \{1, \ldots, n\}$, compute the cc-SNARK Proofs: $\pi_i \leftarrow ARG.P(ipk, x_{Agg}, (w_{i,2}^i, w_i))$. 

39
7. Set $w_{Agg} := ([cm]_{i=1}^n, [\pi]_{i=1}^n)$. If $A$ outputs $\mathcal{R}$ and $C$ outputs $\mathcal{Z}_{Agg, w}$. 

**Completeness analysis.** We now show that for every index $i$ and every adversary $A$, the following probability is 1.

$$Pr\left[\begin{array}{l}
(i, x, w) \not\in \mathcal{R} \\
\text{DNA.} \mathcal{V}^p(\text{ivk}, x, cm, \pi) = 1
\end{array}\right] = 1.$$ 

If $(i, x, w) \not\in \mathcal{R}$, the condition trivially holds. If $(i, x, w) \in \mathcal{R}$, then the completeness of $f$ implies that $\forall i \in [n]$, $(i, x, Arg, (w_{i,1}, w_{i,2})) \in \mathcal{R}_{ARG}$ and $\pi_r$ is a valid proof that $x$ is the first leaf of the Merkle tree with rooth $x$.

The completeness of $ARG$ thus implies that $\forall i \in [n]$, $ARG. \mathcal{V}^p(\text{ivk}_i, x_i, cm_i, \pi_i) = 1$, implying that $(i_{Agg}, x_{Agg}, w_{Agg}) \in \mathcal{R}_{Agg}$.

Now the completeness of $Arg$ gives that $Arg. \mathcal{V}^p(\text{ivk}_{Arg}, x_{Arg}, cm_{Arg}, \pi_{Arg}) = 1$. This, along with the fact that $\pi_r$ is a valid proof that $x$ is the first leaf of the Merkle tree with rooth $x$ implies that $DNA. \mathcal{V}^p(\text{ivk}, x, cm, \pi) = 1$.

**C.2 Knowledge soundness**

Let $A$ be an adversary that succeeds against $E_A$ constructed above with probability $\varepsilon(\lambda)$, that is, for some $i$:

$$Pr\left[\begin{array}{l}
(i, x, w) \notin \mathcal{R} \\
\text{DNA.} \mathcal{V}^p(\text{ivk}, x, cm, \pi) = 1
\end{array}\right] \geq \varepsilon(\lambda).$$

Given this $i$, do the following:

1. Reduce $i$ to multiple CSAT indices: $[i]_{i=1}^n = f_i(i)$.
2. Sample keys for the CSAT indices: $(\text{ipk}_i)_{i=1}^n, ([\text{ick}]_{i=1}^n, [\text{ivk}]_{i=1}^n) \leftarrow ARG. G(\lambda, [i]_{i=1}^n)$.
3. Construct $Agg$ index: $i_{Agg} := [\text{ivk}]_{i=1}^n$.

For this value of $i$, Lemma 4.1 says that there exists an efficient extractor $E'_W$ such that for every efficient adversary $W$ and auxiliary distribution $D_W$, the following probability is negligible:

$$Pr\left[\begin{array}{l}
(i, x, w) \notin k-CSAT \\
\forall i \in [k] : (i, x', (w_{i,1}, w_{i,2})) \in \text{CSAT} \\
x' \text{ is the 0-th leaf of rt}
\end{array}\right] = 1.$$ 

For the aforementioned values $[i]_{i=1}^n$, multi-instance knowledge soundness of $ARG$ says that for every adversary $C$ and auxiliary input distribution $D_C$, there exists an efficient extractor $E'_C$ such that the following probability is negligible.\footnote{Recall that for simplicity, we use a construction of ARG where ipk$_i = \text{ick}_i$ for all $i \in [n]$.}

\[\]
As well as the extractor

Constructing side must be non-negligible, leading to a contradiction.

Finally, for this $i_{\text{Agg}}$, knowledge soundness of Agg gives that for every adversary $B$ and auxiliary input distribution $D_B$, there exists an efficient extractor $E_B$ such that the following probability is negligible.

\[
\Pr \left[ \exists i \in [n] : \left( (i, z_i, w_i) \not\in \mathcal{R} \right) \lor \left( C^p(\text{ipk}, w_i, r_i) \neq cm \right) \land \forall \rho(\text{ivk}, z_i, cm, \pi_i) = 1 \right] = \Pr \left[ \left( [\text{ipk}]^n_{i=1}, [\text{ick}]^n_{i=1}, [\text{ivk}]^n_{i=1} \right) \leftarrow \mathcal{U}(\lambda) \right] \\
\Pr \left[ \left( (i_{\text{Agg}}, z_{\text{Agg}}, w_{\text{Agg}}) \not\in \mathcal{R}_{\text{Agg}} \right) \lor \left( \text{Agg}, C^p(\text{ck}_{\text{Agg}}, w_{\text{Agg}}) \neq cm_{\text{Agg}} \right) \land \forall \rho(\text{vk}_{\text{Agg}}, z_{\text{Agg}}, cm_{\text{Agg}}, \pi_{\text{Agg}}) = 1 \right] = \Pr \left[ \left( \text{pk}_{\text{Agg}}, \text{ck}_{\text{Agg}}, \text{vk}_{\text{Agg}} \right) \leftarrow \text{Agg}, \mathcal{G}^p(1^{\lambda}, \text{Agg}) \right] \\
\Pr \left[ \left( \text{vk}_{\text{Agg}}, z_{\text{Agg}}, cm_{\text{Agg}}, \pi_{\text{Agg}} \right) \leftarrow \text{B}^p(\text{pk}_{\text{Agg}}, \text{aux}_{\text{B}}) \right] = \Pr \left[ \left( \text{vk}_{\text{Agg}}, z_{\text{Agg}}, cm_{\text{Agg}}, \pi_{\text{Agg}} \right) \leftarrow \text{E}_B(\text{pk}_{\text{Agg}}, \text{aux}_{\text{B}}) \right]
\]

Construction of adversaries and auxiliary input distribution. We describe how to use $A$ to build an adversary $B$ and auxiliary distribution $D_A$ against Agg, an adversary $C$ and auxiliary distribution $D_B$ against ARG, and an adversary $W$ and auxiliary distribution $D_C$ that breaks Lemma 4.1, such that

\[
\Pr \left[ B \text{ breaks knowledge soundness of Agg for } i_{\text{Agg}} \right] + \Pr \left[ C \text{ breaks multi-instance knowledge soundness of ARG for } [i]^n_{i=1} \right] + \Pr \left[ W \text{ breaks Lemma 4.1 for } i \right] \geq \epsilon(\lambda)
\]

This is sufficient, because if $\epsilon(\lambda)$ is non-negligible, then at least one of the three probabilities on the left-hand side must be non-negligible, leading to a contradiction.

Constructing $B$. We construct adversary $B$ from the adversary $A$ as follows:

1. Parse aux$_B$ as (aux$_A$, ipk).
2. Obtain $(z, \pi) \leftarrow A^p(\text{ipk}, \text{aux}_A)$.
3. Parse $\pi$ as $(z_{\text{Agg}}, cm_{\text{Agg}}, \pi_{\text{Agg}})$.  
4. Output $(z_{\text{Agg}}, cm_{\text{Agg}}, \pi_{\text{Agg}})$.

Constructing $C$. Denote by $E_B$ the extractor corresponding to $B$ for Agg. Then we construct $C$ from $B$ as well as the extractor $E_B$ as follows:

1. Parse aux$_C$ as (aux$_A$, ipk).
2. Obtain $\text{pk}_{\text{Agg}}$ from ipk.
3. Obtain $(z_{\text{Agg}}, cm_{\text{Agg}}, \pi_{\text{Agg}}) \leftarrow B^p(\text{pk}_{\text{Agg}}, \text{aux}_{\text{C}})$.
4. Extract $w_{\text{Agg}} \leftarrow E_B(\text{pk}_{\text{Agg}}, \text{aux}_{\text{C}})$.
5. Parse $w_{\text{Agg}}$ as $([cm]^{n}_{i=1}, [\pi_i]^{n}_{i=1})$.
6. Output $([z_{\text{Agg}}, cm_i, \pi_i])^{n}_{i=1}$.

Constructing $W$. Denote by $E_C$ the extractor corresponding to $C$ for ARG. Then we construct $W$ from the adversary $C$ as well as the extractor $E_C$ as follows:
\( \mathcal{W}(1^k, i, \text{aux}_w) : \\
1. \text{Parse aux}_w \text{ as } \text{aux}_A, \text{ipk}. \\
2. \text{Obtain } [\text{ipk}]_{n=1} \text{ from ipk.} \\
3. \text{Extract } [([w_i])_{i=1}^n] \leftarrow \mathcal{E}_C([\text{ipk}]_{i=1}^n, \text{aux}_w). \\
4. \text{Use } x, \alpha, \beta, T \text{ and } A \text{ from } ([w_i])_{i=1}^n \text{ to construct } rt \text{ as} \\
   \text{in Fig. 4.} \\
5. \text{Output } ([([w_i])_{i=1}^n, rt).
\)

**Constructing an extractor \( \mathcal{E}_A \) for \( A \).** We now use the extractor for \( \mathcal{W} \) to construct an extractor for \( A \).

\( \mathcal{E}_A(\text{ipk}, \text{aux}_A) \rightarrow w: \\
1. \text{Obtain } i \text{ from ipk.} \\
2. \text{Set } \text{aux}_w := (\text{aux}_A, \text{ipk}). \\
3. \text{Run } (x, w) \leftarrow \mathcal{E}_W(i, \text{aux}). \\
4. \text{Output } w.
\)

**Success probability of \( \mathcal{E}_A \).** For \( i \) in Eq. (1) we have:

\[
\Pr \left[ \begin{array}{c}
(i, x, w) \notin R \\
\text{DNA.}V^p(ivk, x, cm, \pi) = 1 \\
\text{Agg.}V^p(\text{vk}_{Agg}, x_{Agg}, cm_{Agg}, \pi_{Agg}) = 1
\end{array} \right] \left[ \begin{array}{c}
\rho \leftarrow U(\lambda), \\
(\text{ipk}, ick, ivk) \leftarrow \text{DNA.}\mathcal{G}^p(1^k, i), \\
\text{aux} \leftarrow D(\text{ipk}), \\
(x, cm, \pi) \leftarrow \mathcal{A}(\text{ipk}, aux), \\
w \leftarrow \mathcal{E}(\text{ipk}, aux)
\end{array} \right] \geq \varepsilon(\lambda).
\]

Parsing \( \pi \) as \( \pi_{Agg}, cm_{Agg}, \pi_{Agg}, \pi_{ivk} \) and \( ivk \) as \( \text{vk}_{Agg}, \), by construction of DNA, \( \text{DNA.}V^p(ivk, x, cm, \pi) = 1 \) implies that \( \text{Agg.}V^p(\text{vk}_{Agg}, x_{Agg}, cm_{Agg}, \pi_{Agg}) = 1 \). Thus:

\[
\Pr \left[ \begin{array}{c}
(i, x, w) \notin R \\
\text{Agg.}V^p(\text{vk}_{Agg}, x_{Agg}, cm_{Agg}, \pi_{Agg}) = 1
\end{array} \right] \left[ \begin{array}{c}
\rho \leftarrow U(\lambda), \\
(\text{ipk}, ick, ivk) \leftarrow \text{DNA.}\mathcal{G}^p(1^k, i), \\
\text{aux} \leftarrow D(\text{ipk}), \\
(x, cm, \pi) \leftarrow \mathcal{A}(\text{ipk}, aux), \\
w \leftarrow \mathcal{E}(\text{ipk}, aux)
\end{array} \right] \geq \varepsilon(\lambda).
\]

**'Factoring out' the probability of \( \mathcal{E}_W \) failing.** \( (i, x, w) \notin R \) either when

1. \( \forall i \in [n] : (i, x', (w_{i,1}, w_{i,2})) \in \text{CSAT but } \mathcal{E}_W \text{ fails to extract a valid } w \). The probability of this happening is precisely:

\[
\Pr \left[ \begin{array}{c}
\mathcal{W} \text{ breaks } \\
\text{Lemma 4.1 for } i
\end{array} \right]
\]

2. \( \exists i \in [n] : (i, x', (w_{i,1}, w_{i,2})) \notin \text{CSAT.} \)

**'Factoring out' the probability of \( \mathcal{E}_C \) failing.** \( \exists i \in [n] : (i, x', (w_{i,1}, w_{i,2})) \notin \text{CSAT either when} \\
1. \forall i \in [n] : \text{ARG.}V^p(ivk_i, x_i, cm_i, \pi_i) = 1 \text{ but } \mathcal{E}_C \text{ fails to extract valid } [w_i]_{i=1}^n. \text{ The probability of this happening is precisely:} \\
\]

\[
\Pr \left[ \begin{array}{c}
\mathcal{C} \text{ breaks multi-instance knowledge soundness} \\
\text{of ARG for } [i]_{i=1}^n
\end{array} \right]
\]

2. \( \exists i \in [n] : \text{ARG.}V^p(ivk_i, x_i, cm_i, \pi_i) \neq 1. \)
‘Factoring out’ the probability of $E$ failing. The probability that $\text{Agg}.\mathcal{V}(\nu\kappa_{\text{Agg}}, \nu\pi_{\text{Agg}}, \nu\tau_{\text{Agg}}, \nu\rho_{\text{Agg}}) = 1$ and $\exists i \in [n]: \text{ARG}.\mathcal{V}(\nu\kappa_i, \nu\pi_i, \nu\tau_i, \nu\rho_i) \neq 1$ is exactly:

$$\Pr \left[ \begin{array}{c} \text{B breaks knowledge soundness of Agg for } i_{\text{Agg}} \\ \end{array} \right]$$

Putting the above together, we obtain our desired contradiction, i.e., that

$$\Pr \left[ \begin{array}{c} \text{B breaks knowledge soundness of Agg for } i_{\text{Agg}} \\ \end{array} \right] + \Pr \left[ \begin{array}{c} \text{C breaks multi-instance knowledge soundness of ARG for } [i]_1^n \\ \end{array} \right] + \Pr \left[ \text{W breaks Lemma 4.1 for } i \right] \geq \epsilon(\lambda),$$

### C.3 Zero-knowledge

The construction in Section 6.1 can be turned zero-knowledge by using a commit-carrying SNARK ARG with multi-instance zero-knowledge and the hiding version of Merkle commitments.

Zero-knowledge then follows from the multi-instance zero-knowledge property of DNA and the zero-knowledge property of hiding Merkle proofs. The former implies the existence of $S^{\text{ARG}} = (S_1^{\text{ARG}}, S_2^{\text{ARG}})$ for ARG.

The Simulator $S = (S_1, S_2)$ for the distributed zkSNARK DNA is presented in Figure 11.

#### Figure 11: Simulator for the Distributed zkSNARK

- **$S_1^0(i^k, i) \rightarrow (ipk, \tau)$:**
  1. Obtain the partitioned indices: $[i^k]_i^{n} := f_i(i)$.
  2. Run $S_1^{\text{ARG}}$ to obtain $(ipk|_{i=1}^{n}, \tau) \leftarrow S_1^{\text{ARG}}(1^k, [i^k]_i^{n})$.
  3. Obtain $[ivk]_i^{n}$ from $[ipk]_i^{n}$ and set $i_{\text{Agg}} = [ivk]_i^{n}$.
  4. Sample public parameters for Agg: $(pk_{\text{Agg}}, \nu\tau_{\text{Agg}}, \nu\kappa_{\text{Agg}}) \leftarrow \text{Agg}\cdot G(1^k, i_{\text{Agg}})$.
  5. Output $(ipk := (i, ivk, [ipk]_i^{n}, pk_{\text{Agg}}), \tau)$.

- **$S_2^0(ipk, \nu, \tau) \rightarrow (\nu, \mu)$:**
  1. Parse proving key as ipk $= (i, ivk, [ipk]_i^{n}, pk_{\text{Agg}})$.
  2. Sample random $(\alpha, \beta) \leftarrow \mathbb{F}_2^2$.
  3. Compute a random Merkle tree root and proof $\pi_r$ that $\nu$ is its first leaf.
  4. Set $\pi_{\text{Agg}} := (1, rt, \alpha, \beta)$.
  5. Run $S_2^{\text{ARG}}$ to obtain $([cm]_i^{[\pi_i]^{n}}, \mu) \leftarrow S_2^{\text{ARG}}(1^{\nu}, [\pi_{\text{Agg}}]^{n})$.
  6. Commit to $[cm]_i^{[\pi_i]^{n}}$: $\nu\tau_{\text{Agg}} \leftarrow \text{Agg}\cdot C(\nu\kappa_{\text{Agg}}, [cm]_i^{[\pi_i]^{n}})$.
  7. Program $\mu(ipk, cm_{\text{Agg}}) := (\alpha, \beta)$.
  8. Assemble the $\pi_{\text{Agg}}$ instance and witness:

$$\pi_{\text{Agg}} \cdot W_{\text{Agg}} = ((1, rt, \alpha, \beta), ([cm]_i^{[\pi_i]^{n}}, \nu_{\text{Agg}}, \nu_{\text{Agg}}))$$

9. Compute aggregated proof: $\nu_{\text{Agg}} \leftarrow \text{Agg}\cdot D(\nu\kappa_{\text{Agg}}, \nu_{\text{Agg}}, \nu_{\text{Agg}}, cm_{\text{Agg}})$.
  10. Output $(\nu := (\pi_{\text{Agg}}, cm_{\text{Agg}}, \nu_{\text{Agg}}, \nu_{\text{Agg}}), \mu)$.

**Proof sketch.** The zero-knowledge of DNA follows from the multi-instance zero-knowledge property of ARG and the zero-knowledge property of hiding Merkle proofs. At a high level, the proof $\nu$ generated by DNA consists of $\pi_{\text{Agg}} = (1, rt, \alpha, \beta), \pi_{\text{Agg}}$ and $cm_{\text{Agg}}, \pi_{\text{Agg}}$. $(\alpha, \beta)$ are randomly sampled and $(rt, \pi_{\text{Agg}})$ hide all information about their leaves because of the hiding property of Merkle trees. Lastly, $cm_{\text{Agg}}$ and $\pi_{\text{Agg}}$ are
also indistinguishable from the honest case, as they are deterministic functions of \([cm_i]_{i=1}^n\) and \([\pi_i]_{i=1}^n\) and the latter hide all information about the underlying witness since ARG has multi-instance zero-knowledge.

Note that the simulator can fail if the simulator \(S_{\text{ARG}}\) for ARG also programs the random oracle at the point \((ivk, cm_{\text{Agg}})\) to a value different from \((\alpha, \beta)\). This can happen with non-negligible probability only if \(S_{\text{ARG}}\) is ‘adversarial’ and purposely programs that point. However, for all natural constructions of ARG (including Mirage, the cc-zkSNARK we use in HEKATON), the simulator is benign. In fact, the simulator for Mirage does not program the random oracle at all.

C.4 Efficiency

The efficiency claims follow from inspection:
• Each worker node runs \(\text{ARG.C}\) and \(\text{ARG.P}\) once.
• Each worker node’s communication cost is the size of a single commitment of ARG and a single proof of ARG.
• The primary node runs \(\text{Agg.C}\) and \(\text{Agg.P}\) once on inputs of size \(n\).
• Proof size is the size of an Agg proof and one Merkle proof for the tree with root \(rt\).
References


