zkSaaS: Zero-Knowledge SNARKs as a Service

Sanjam Garg, *University of California, Berkeley, and NTT Research;* Aarushi Goel, *NTT Research;* Abhishek Jain, *Johns Hopkins University;* Guru-Vamsi Policharla and Sruthi Sekar, *University of California, Berkeley*

https://www.usenix.org/conference/usenixsecurity23/presentation/garg

This paper is included in the Proceedings of the 32nd USENIX Security Symposium.

August 9–11, 2023 • Anaheim, CA, USA

978-1-939133-37-3

Open access to the Proceedings of the 32nd USENIX Security Symposium is sponsored by USENIX.

zkSaaS: Zero-Knowledge SNARKs as a Service

Sanjam Garg
UC Berkeley and NTT Research

Aarushi Goel
NTT Research

Abhishek Jain
Johns Hopkins University

Guru-Vamsi Policharla
UC Berkeley

Sruthi Sekar
UC Berkeley

Abstract

A decade of active research has led to practical constructions of zero-knowledge succinct non-interactive arguments of knowledge (zk-SNARKs) that are now being used in a wide variety of applications. Despite this astonishing progress, overheads in proof generation time remain significant.

In this work, we envision a world where consumers with low computational resources can outsource the task of proof generation to a group of untrusted servers in a privacy-preserving manner. The main requirement is that these servers should be able to collectively generate proofs at a *faster* speed (than the consumer). Towards this goal, we introduce a framework called zk-SNARKs-as-a-service (zkSaaS) for faster computation of zk-SNARKs. Our framework allows for distributing proof computation across multiple servers such that each server is expected to run for a shorter duration than a single prover. Moreover, the privacy of the prover's witness is ensured against any minority of colluding servers.

We design custom protocols in this framework that can be used to obtain faster runtimes for widely used zk-SNARKs, such as Groth16 [EUROCRYPT 2016], Marlin [EUROCRYPT 2020] and Plonk [EPRINT 2019]. We implement proof of concept zkSaaS for the Groth16 and Plonk provers. In comparison to generating these proofs on commodity hardware, we can not only generate proofs for a larger number of constraints (without memory exhaustion), but can also get $\approx 22\times$ speed-up when run with 128 parties for 2^{25} constraints with Groth16 and 2^{21} gates with Plonk.

1 Introduction

zk-SNARKs are zero-knowledge succinct non-interactive arguments of knowledge [10], that allow a prover to non-interactively convince a verifier of the knowledge of a witness attesting to the validity of an NP relation, without revealing any information about the witness. zk-SNARKs have been a topic of extensive research in recent years [18, 19, 78, 68, 20, 57, 56, 21, 48, 27, 37, 81]. Their flexibility and expressiveness make them applicable to a wide variety of scenarios such as private transactions [7, 63], roll-ups [84], private smart contracts [52, 22], access control in compliance with KYC regulations [65, 66], social networks with private reputation monitoring [25], proving existence of bugs in zero-knowledge [47], static program analysis [35], zero-knowledge middleboxes for enforcing network policies on encrypted traffic [50], verifiable inference of machine learning [80, 59, 58, 76] and verifiable database queries [82].

Despite recent advances [11, 51, 14, 13, 48, 27, 37, 29], generation of zk-SNARKs remains thousands of times [75, 61] slower than checking the relation directly for typical applications, with large memory usage — effectively gate keeping users without access to large machines. A natural way out for such users is to outsource proof generation to more powerful servers. While one could use a cloud server such as AWS, GCP or Azure to generate proofs, this approach requires sharing the witness in the clear with the cloud server. As such, this solution offers *no privacy* against insider threats such as rogue administrators [31] who may compromise data privacy for financial gains. Even if the cloud service provider were trusted, the witness might consist of sensitive data such as patient medical records that legally cannot be placed off-premises due to data protection laws.

To address the privacy problem, recently, Ozdemir et al. [61] introduced the idea of *collaborative zk-SNARKs* for distributed generation of zk-SNARKs. Collaborative zk-SNARKs are essentially secure multiparty computation (MPC) protocols that allow a group of parties holding shares of the witness to collectively generate a *single* succinct proof. The key security guarantee is that the witness remains hidden as long as only a subset of the parties collude. Ozdemir et al. design collaborative zk-SNARK analogs of Groth16 [48], Marlin [27] and

Plonk [37]. In their protocols, all parties run in parallel and each of them performs as much work as the (single) prover of the underlying zk-SNARK. This results in approximately the same runtime as the single prover.[1]

We posit that requiring each of the parties running in parallel to do as much work as the zk-SNARK prover is an overkill. Indeed, a previous work of Wu et al. [77] leveraged parallelism to distribute proof computation across different machines in a compute cluster to achieve *faster* proof generation times. Their approach, however, requires leaking the witness to the cluster, resulting in a loss of privacy.

In this work, we explore the possibility of combining the best features of collaborative zk-SNARKs [61] and the work of Wu et al. [77]. We ask:

Is it possible to outsource zk-SNARK proof generation to a group of parties in a privacy-preserving manner for faster proof generation?

Our Contributions. Our contributions are as follows:

1. We present a general framework for *zk-SNARKs-as-a-service* (zkSaaS), where a client delegates proof computation to a group of untrusted servers in a privacy preserving manner. Each of these servers is expected to run for a shorter duration than a single local prover.

2. We instantiate this framework with custom protocols to obtain faster runtimes than local provers for widely used zk-SNARKs, such as Groth16 [48], Marlin [27] and Plonk [37].

3. Finally, we implement prototypes of zkSaaS for the Groth16 [48] and Plonk [37] proof systems. Concretely, we show that when creating a proof for 2^{25} constraints in the case of Groth16 (and 2^{21} constraints with Plonk), the zkSaaS protocol with 128 servers is $\approx 22\times$ faster than a local prover. We also show that deploying more servers helps us get a further speed-up. For instance, when creating a Groth16 proof for 2^{19} constraints, we see an improvement from $\approx 1.9\times$ to $\approx 22\times$ when the number of servers is increased from 8 to 128. This is in contrast to collabortive zk-SNARKs [61] which do not obtain any speedup.

4. We also estimate the financial cost of using zkSaaS to compute a Groth16 proof for an instance of size 2^{19} to be under \$0.23 with 128 parties using a 64 Mbps link between servers.

[1] This is interesting since they manage to avoid additional security parameter overhead that is usually incurred when *securely* computing a function.

1.1 Overview of Our Approach

Similar to [61], our initial idea is to identify common building blocks within widely used zk-SNARKs and design custom secure multiparty computation (MPC) protocols to compute them efficiently. We then stitch them together to obtain zkSaaS, an efficient MPC protocol for the corresponding zkSNARK prover.

An Important Observation. One of the key observations made in [61] is that it is possible to directly secret share points on the elliptic curve and fields and apply MPC techniques on these shares, which avoids the large overheads incurred when using generic MPC techniques with very few rounds of communication.[2] We go one step ahead and observe that these building blocks can be rewritten in a way that allows us to leverage significant SIMD structure that appears within their computation.

To leverage this SIMD structure, we make use of a tool called *packed secret sharing* (PSS) [36].[3] This is a more efficient sibling of Shamir's polynomial-based secret sharing scheme [69], that allows secret-sharing a *vector of values* amongst a set of parties. In particular, at the cost of a slight reduction in the corruption threshold, using PSS we can "hide" $\ell = \mathcal{O}(n)$[4] secrets (where n is the total number of servers) in a polynomial and each of the n servers receives an evaluation at a single point in this polynomial. Such sharings allow parties to efficiently perform SIMD computations on secret shared data, while reducing the workload on each of them. We use this in the design of each of our sub-protocols.

Multi-Scalar Multiplications (MSM). One of the main building blocks in the zkSNARKs we consider is MSM, which are operations of the form $\Pi_{i\in[m]} g_i^{a_i}$, where g_i's are points on an elliptic curve. This is by far the most expensive component.[5] We design a bespoke MPC protocol where the total work (as well as the asymptotic space requirement) of each server is a factor of ℓ less than that of a single prover.

Next we discuss the remaining components i.e., Product Check, Fast Fourier Transform (FFT) and polynomial computations, which exclusively involve field operations that are much cheaper than elliptic curve operations.

Product-Check. Product-Check requiring computations of the form $\prod_{j\in[i]} x_j$, for all $i \in [m]$, which are referred to as *partial products*. Similar to MSM, we design a special-purpose MPC protocol for partial products, that

[2] This is mainly due to generic MPC techniques making non-black-box use of the elliptic curve.

[3] This is also the main building block used in the design of all of general-purpose MPC protocols that support some division of work (See Section 1.3 for more details.)

[4] In the implementation we set this to be $n/4$.

[5] In Figure 1 we show the fraction of time spent computing MSMs for Groth16.

allows us to divide the work of each server by a factor of ℓ less than that of a single prover.

Fast Fourier Transform. The standard description of the FFT algorithm on a polynomial with m coefficients can be divided into $\log m$ steps, with $O(m)$ field multiplications at each step. For the first $\log m/\ell$ steps, we are able to divide the work of each server by a factor of ℓ. For the remaining $\log \ell$ steps, however, we require one of the parties to do $O(m)$ field operations and have $O(m)$ memory, while the work and memory requirement of the remaining parties gets divided by ℓ.

Polynomial Multiplication and Division. Finally, we show how to combine standard packed secret sharing based subprotocols for addition and multiplication along with our custom MPC for FFT to enable secure distributed polynomial computations.

Composing different subprotocols based on packed secret sharing is not straightforward, and requires care. We show how to combine the above subprotocols to obtain zkSaaS for faster generation of zk-SNARKs such as Groth16, Marlin and Plonk.

Communication over a Star Topology Network. Our distributed sub-protocols for all of the functions described above, do not require servers to communicate with all other servers. Aside from receiving shares of the extended witness from the client, we require the servers to only communicate with the one large server, throughout the rest of the computation. As a result, we only need communication channels between the client and each server and between the large server and every other server.

Instantiating zkSaaS. As discussed above, the distributed FFT protocol requires one party which has memory proportional to the size of the relation but the computational resources demanded from all other parties is reduced by a factor of ℓ. Therefore, a zkSaaS deployment requires one large server. While not ideal, we argue that this is still reasonable for two reasons — (1) even if the private view of this large server is leaked, it does not compromise a client's secrets unlike when a client simply rents a large server to generate the proof. (2) it is very easy and quite cheap to rent a large server from a cloud service provider.

Finally, we remark that proof generation in the zk-SNARKs that we consider proceeds in two steps: first, the prover uses its (short) witness to evaluate the relation circuit and obtain a corresponding *extended witness*, which is then used to generate the proof (See Section 3 for a detailed discussion). Similar to collaborative zk-SNARKs [61], in this work, we focus on designing secure distributed protocols for proof generation and assume that the client computes and shares the extended witness with the servers. We view faster generation of the extended witness as an important orthogonal question but such protocols would need to essentially be redesigned for every application.

Security. We now discuss the key aspects of our security model and the guarantees provided by zkSaaS.

We assume that a majority of the servers are honest. Specifically, let n be the total number of servers and ℓ be the total number of secrets that we can pack in a single packed secret sharing. We require that at most $t < \frac{n}{2} - \ell$ of the servers can be corrupted. Further, we assume that the corruptions are *semi-honest*.

Our zkSaaS framework retains the soundness property of the underlying zk-SNARK and provides the following completeness and zero-knowledge guarantees:

— *Completeness*: For any true statement, an execution of zkSaaS involving an honest client and honest servers outputs an accepting proof.[6]

— *t-Zero Knowledge:* In any execution of zkSaaS, the view of the t corrupt servers can be efficiently simulated without the client's witness. This, in particular, implies that the corrupt servers learn nothing about the client's witness.

We conclude with a few remarks on security against *malicious* servers. We first note that proofs output by zkSaaS remain sound even when *all* servers are malicious. Next, we conjecture that our protocols can be augmented to achieve t-zero knowledge against malicious servers by using highly efficient compilers from the recent MPC literature [40, 39, 6, 44, 45]. In essence, these compilers show that semi-honest MPC protocols that are "secure against malicious corruptions up to linear attacks" can be compiled into maliciously secure protocols with a small constant (typically, at most two) overhead. We conjecture that our semi-honest zkSaaS protocols already satisfy the properties required for these efficient compilers; a formal treatment of the same, however, is outside the scope of this work.

1.2 Example Applications of zkSaaS

We now discuss some real-world applications where we envision our zkSaas-framework to be useful.

Private Transactions and Smart Contracts. A simple spend transaction on a private chain such as ZCash[7] already involves $\approx 130,000$ R1CS constraints[7] which takes roughly 10 seconds on a high-end laptop in single threaded mode. This would take even longer on weaker

[6] The completeness property, in fact, holds even if the servers are semi-honest (since such servers follow protocol instructions.)

[7] https://github.com/zcash/librustzcash

devices such as smart phones making the process quite tedious for users.

Private smart contracts are immutable programs running on blockchains, which provide confidentiality of the computation carried out on blockchains [52, 22]. Although these chains can be designed more carefully to reduce the overhead for simple transactions, they aim to support general computation on the smart contracts which can blow up to a very large number of constraints as there may be very complicated logic that gets executed involving cryptographic functions such as signature verification. With zkSaaS, users can potentially pay a tiny transaction fee in exchange for a seamless experience akin to current centralized payment methods.

Statements involving Ethereum Wallets. Ethereum uses EdDSA signatures for authentication of transactions over the ed25519 elliptic curve which is not proof friendly. As a result, one needs to emulate non-native 256-bit field arithmetic which is quite expensive. The verification circuit of an EdDSA signature costs over 2.5 million R1CS constraints[8] and the proving key itself is 1.6 GB in size. Common tasks such as proof of membership viz. "I own an Ethereum wallet out of these set of 1024 wallets" become impractical on mobile devices as the statements are simply too large and lead to memory exhaustion.

Combating Disinformation. It was shown that zero-knowledge proofs can be used to prove that images appearing in news articles underwent an approved set of transformations from the time of creation [60]. This is particularly helpful in allowing reporters to hide sensitive content while at the same time proving authenticity of the image. While fast generation of zk-SNARKs is possible for images, doing the same for compute-heavy video files is currently far from practical and our zkSaaS-framework could aid in carrying out such a computation.

Verifiable Private ML Inference. A user can commit to a machine learning model and provide a proof of inference on this machine learning model, which can be used to verify accuracy of a machine learning model or to ensure that a certain entity who claims to use AI for a task is actually producing predictions using a machine learning model. Since circuits for inference can be quite large as the models grow in size, they quickly become impractical to prove even on consumer grade laptops. Again, the data used to train this model could be patient health records for example which cannot be placed on servers that do not comply with HIPAA[9]. Hence, a solution is to use zkSaaS to compute proofs of inferences where no server sees sensitive information.

[8] https://github.com/Electron-Labs/ed25519-circom

[9] https://www.hipaajournal.com/

1.3 Related Work

Some prior works [53, 30] have considered building MPC-as-a-service, which involves deploying MPC in a volunteer-operated network (like blockchains). However, such protocols are built for generic functionalities and do not offer efficient solutions for our specific goal. In a different line of work (unrelated to our goal), MPC has been used to securely sample the common parameters used in zk-SNARKs [8, 23, 55].

A different line of work has considered the problem of speeding-up the zk-SNARK prover time, but they either do not hide the witness [77, 67], or lead to [24] linear verification time with a security guarantee that is weaker than both our framework and the collaborative zk-SNARK framework [61]. Some prior works have also studied other distributed models of proof systems, including ones where the statement is shared amongst multiple verifiers [1, 16, 17], or where there are two (or more) non-colluding provers [4, 12].

Our goal in some sense is very similar to the design of MPC protocols, where the total computation and communication is independent of the number of parties, which has been the focus of a significant line of research [34, 33, 39, 43, 6, 44, 46]. However for arithmetic circuits, most of these require round complexity linear in the multiplicative depth of the circuit, which is not ideal in our setting since the prover algorithms in zk-SNARKs typically don't have a constant multiplicative depth. Moreover, representing cryptographic operations such as group exponentiations as an arithmetic circuit and computing them inside an MPC is extremely inefficient. Therefore, naïvely using these protocols in computing a zk-SNARK will result in inefficient solutions.

More recently, in a concurrent and independent work, Chiesa et al. [28] also considered the problem of private delegation of zk-SNARKs for faster proof generation. However, their model is quite different from ours. In particular, they assume that the client remains online throughout the computation and actively participates in the zk-SNARK computation along with the servers. For this, they design an MPC protocol (that is run between the servers and the clients) for zk-SNARK computation leveraging the fact that one of the parties (i.e., the client) is always honest and the witness need not be hidden from them. Their goal is to essentially "reduce" the work done by the client. In contrast, in our setting, after sharing the extended witness, the client does not need to do any work and can delegate the *entire* zk-SNARK computation to the servers.

1.4 Future Directions

A promising direction for future work would be to eliminate the need of a single large server in zkSaaS. In our current solution, this large server is only needed for our distributed protocol for FTT. Potential approaches for avoiding this could be – (1) Designing a more efficient subprotocol for for distributed computation of FFT or (2) designing zkSaaS for zk-SNARKs that do not use FFT operations e.g. Orion [79], Brakedown [42], Hyperplonk [26]. Another interesting problem would be to enable faster generation of the extended witness in a similar framework. Finally, as discussed in Section 1.1, it would be interesting to formally demonstrate how the protocols developed in this work can be augmented to achieve security against malicious servers.

Paper Organization. We start by establishing some notations in Section 2. We then formally define zkSaaS-in Section 3, and give an overview of the popular zk-SNARKs of interest (Groth, Marlin and Plonk) in Section 4. A detailed technical exposition of each of our distributed sub-protocols (FFT, MSM and sum of partial products) appears in Section 5, and we show how to build a zkSaaS for a specific class of zk-SNARKs in Section 6. Finally, we discuss the concrete efficiency of our scheme in Section 7.

2 Preliminaries

For any $n \in \mathbb{N}$, we use $[n]$ to denote the set $\{1,\cdots,n\}$ and for $i,j \in \mathbb{N}$ with $i < j$, we use $[i,j]$ to denote the set $\{i, i+1, \ldots, j\}$. We denote a vector of ℓ elements from a field $\mathbb{F}$, $(x_1,\cdots,x_\ell)$, by $\mathbf{x}$. For any two vectors $\mathbf{x}$ and $\mathbf{y}$, the component-wise multiplication is denoted by $\mathbf{x} \odot \mathbf{y} := (x_1 \cdot y_1, \cdots, x_\ell \cdot y_\ell)$. We always use capital letters (e.g. X) to denote elements from a group $\mathbb{G}$, and correspondingly $\mathbf{X}$ denotes a vector of ℓ group elements $(X_1,\cdots,X_\ell)$. We use a multiplicative notation for our group operations throughout the paper.

Linear Secret Sharing Schemes. In this work we make use of polynomial based, regular threshold secret sharing scheme as well as a packed secret sharing scheme. For regular threshold secret sharing, we use $[x]$, $\langle x \rangle$ to denote shares of a value x, w.r.t. to a degree t and $n-1$ polynomial respectively. For packed secret sharing, we use $[\![\mathbf{x}]\!]$, $\langle\!\langle \mathbf{x} \rangle\!\rangle$ to denote shares of a vector $\mathbf{x}$ w.r.t. to a degree D and $n-1$ polynomial respectively, where we assume that the length of $\mathbf{x}$ is $\ell \in \mathcal{O}(n)$ and $D = t + \ell$. We use $[x]_i, \langle x \rangle_i, [\![\mathbf{x}]\!]_i, \langle\!\langle \mathbf{x} \rangle\!\rangle_i$ to denotes shares held by a party P_i and $[x]_S, \langle x \rangle_S, [\![\mathbf{x}]\!]_S, \langle\!\langle \mathbf{x} \rangle\!\rangle_S$ to denote the shares held by a subset S of the parties. Finally, we use functions $[x] \leftarrow \mathsf{share}(\mathbb{F}, x, t)$ and $[\![\mathbf{x}]\!] \leftarrow \mathsf{pshare}(\mathbb{F}, \mathbf{x}, D)$ to compute shares and $\mathsf{open}(\mathbb{F}, [\![\mathbf{x}]\!], D)$ and $\mathsf{open}(\mathbb{F}, [x], t)$ to reconstruct shares.

3 zkSaaS Framework

As discussed earlier, zkSaaS is a collaborative zk-SNARK [61] framework, where a client delegates the task of computing a zk-SNARK to n-servers. To realize our goal of enabling fast computation of zk-SNARKs, the resultant zkSaaS protocol must ensure that – (1) the work done by the clients is minimized and (2) the work required to compute the zk-SNARK gets divided across all servers.

RICS Format. Let us briefly recall the structure of existing zk-SNARKs. Different zk-SNARKs work with different representations of the relation R - e.g., quadratic arithmetic programs [62, 67], low-depth circuits [15, 32, 41, 71, 72, 73, 74, 78, 83], binary arithmetic circuits [37], etc. The most popular representation amongst state-of-the-art proof systems [9, 27, 48, 49] is known as the rank-1 constraint systems (R1CS) that generalizes arithmetic circuits.

Proof systems working with this representation proceed in two steps: (1) First, extend the given (short) statement-witness pair (ϕ, w) into a satisfying assignment $\mathbf{z}$ for the RICS relation. The length of this satisfying assignment is proportional to the size of the relation. (2) second, give an argument of knowledge for this satisfying assignment for the RICS relation. Step 1 is inexpensive and only requires non-cryptographic field operations, while Step 2 requires more expensive cryptographic operations and is typically the bottleneck.

Our Framework. zkSaaS is essentially a secure multiparty computation (MPC) protocol for computing zk-SNARKs, between a client and n-servers.

Client. Since Step 1 of the proof generation is inexpensive, we assume that client performs this step and "securely" shares the resulting satisfying assignment $\mathbf{z}$ with the servers at the start of the protocol. To compute this satisfying assignment, the client essentially needs to represent the original relation R as an arithmetic circuit C_R and compute all the values induced on the intermediate wires in circuit C_R when evaluated on inputs (ϕ, w). These intermediate values form the satisfying assignment $\mathbf{z}$ for the corresponding RICS relation. Looking ahead, in our construction, the client (pack) secret shares [36] the vector $\mathbf{z}$ with the servers.

Computing and sharing this satisfying assignment requires $\mathcal{O}(|R| \log N) \approx \mathcal{O}(|R|)$ field operations and very little space. Indeed, observe that the client can do this computation in a streaming fashion, where it computes a

Definition 1 (zkSaaS)**.** *Let λ be the security parameter, n be the number of parties, $R \in \mathcal{R}_\lambda$ be an NP-relation and $\Sigma_R = (\mathsf{Setup}, \mathsf{Prove}, \mathsf{Ver}, \mathsf{Sim})$ be a zk-SNARK for R such that: the prover computation time is $T_{\mathsf{prover}} = T_{\mathsf{field}} + T_{\mathsf{crypto}}$, where T_{field} and T_{crypto} are the times taken by the prover for the field operations and cryptographic operations, respectively; the prover space complexity is S_{prover}. Let $(\mathsf{Preprocessing}, \Pi_{\mathsf{online}})$ be a tuple defined as follows:*

— $\mathsf{Preprocessing}(\mathsf{crs}, 1^n) \to \mathsf{pre}_1, \ldots, \mathsf{pre}_n$: *This is a PPT algorithm that takes the* crs *output by* Setup *and the number of servers n as input and outputs correlated randomness pre_i for each server P_i (for $i \in [n]$).*

— $\Pi_{\mathsf{online}}(\mathsf{crs}, \phi, w, \mathsf{pre}_1, \ldots, \mathsf{pre}_n) \to \pi$: *This is an MPC protocol between a client $\mathcal{C}$ and n servers $\mathsf{P}_1, \ldots, \mathsf{P}_n$. The client has the statement ϕ and private input w. It sends messages to each of the n servers in a single round. Given these messages, ϕ and their respective correlated randomness $\mathsf{pre}_1, \ldots, \mathsf{pre}_n$, the servers then engage in an interactive protocol amongst each other to compute a proof π.*

We say that $\Pi = (\mathsf{Preprocessing}, \Pi_{\mathsf{online}})$ is a zkSaaS *for Σ_R, if the following properties are satisfied.*

1. ***Completeness:*** *For all $(\phi, w) \in R$, the following holds:*

$$\Pr\left[\begin{array}{c} \mathsf{crs} \leftarrow \mathsf{Setup}(1^\lambda) \\ \mathsf{pre}_1, \ldots, \mathsf{pre}_n \leftarrow \mathsf{Preprocessing}(\mathsf{crs}, 1^n) \\ \pi \leftarrow \Pi_{\mathsf{online}}(\mathsf{crs}, \phi, w, \mathsf{pre}_1, \ldots, \mathsf{pre}_n) \end{array} \middle| \mathsf{Ver}(\mathsf{crs}, \phi, \pi) = 0 \right] \leq \mathsf{negl}(\lambda)$$

2. ***t-zero-knowledge:*** *Let $\mathsf{crs} \leftarrow \mathsf{Setup}(1^\lambda)$, $\mathsf{pre}_1, \ldots, \mathsf{pre}_n \leftarrow \mathsf{Preprocessing}(\mathsf{crs}, 1^n)$. For all semi-honest PPT adversaries $\mathcal{A}$ controlling at most a t-sized subset $\mathsf{Corr} \subset [n]$ of the servers, there exists an efficient Simulator $\mathsf{Sim}_{\mathsf{zkSaaS}}$, such that the following holds for all ϕ, w (where $b \leftarrow R(\phi, w) \in \{0,1\}$):*

$$\left\{\mathsf{view}^{\mathcal{A}}_{\Pi_{\mathsf{online}}}[\phi, w]\right\} \approx_c \left\{\mathsf{Sim}_{\mathsf{zkSaaS}}(\mathsf{crs}, \phi, b, \{\mathsf{pre}_i\}_{i \in \mathsf{Corr}})\right\}$$

   here $\mathsf{view}^{\mathcal{A}}_{\Pi_{\mathsf{online}}}[\phi, w]$ denotes the view of $\mathcal{A}$ in an execution of $\Pi_{\mathsf{online}}(\mathsf{crs}, \phi, w, \mathsf{pre}_1, \ldots, \mathsf{pre}_n)$, and we use $\approx_c$ to denote computational indistinguishability between the two distributions.

3. ***Efficiency:*** *$\Pi = (\mathsf{Preprocessing}, \Pi_{\mathsf{online}})$ satisfy the following efficiency requirements:*

   Preprocessing: The computation complexity of the Preprocessing *algorithm is $o(T_{\mathsf{field}})$. For each $i \in [n]$, size of the correlated randomness $|\mathsf{pre}_i| \in \mathcal{O}(S_{\mathsf{prover}}/n)$.*

   Client: Computation complexity of the client $o(T_{\mathsf{field}})$ field operations.

   Special Server P_1: The first server has a computation complexity of $o(T_{\mathsf{field}})$ field operations and $\mathcal{O}(T_{\mathsf{crypto}}/n)$ cryptographic operations. It's space and communication complexities are $\mathcal{O}(S_{\mathsf{prover}})$ and $o(T_{\mathsf{field}})$, resp.

   Other Servers $\mathsf{P}_2, \ldots, \mathsf{P}_n$: All other servers have computation complexity of $o(T_{\mathsf{field}}/n)$ field and $\mathcal{O}(T_{\mathsf{crypto}}/n)$ cryptgraphic operations. Their space and communication complexities are $\mathcal{O}(S_{\mathsf{prover}}/n)$ and $o(T_{\mathsf{field}}/n)$, resp.

fraction of the circuit at a time and secret shares the resulting wire values before proceeding with evaluating the next part of the circuit, thereby minimizing the required space. However, if the client is unwilling, this computation can also be done via an MPC protocol, where the client only needs to share the original statement-witness pair (ϕ, w) with the servers, who then run a generic MPC to compute C_R and obtain shares of $\mathbf{z}$.

In this work, we focus on designing an efficient protocol for Step 2, which is the main bottleneck in the computation of existing zk-SNARKs.

Servers. Given shares of the extended witness, Step 2 is executed via an MPC protocol between the servers. We want the *total* computation and space complexity of this protocol to be asymptotically identical to that of computing the zk-SNARK by a monolithic entity.

Let the space complexity of the underlying zk-SNARK be S_{prover} and the computation complexity be T_{field} field operations and T_{crypto} cryptographic operations. Then, our zkSaaS-framework guarantees the following efficiencies: first, in terms of space complexity, one of the servers requires a $\mathcal{O}(S_{\mathsf{prover}})$ space, while all others require a $\mathcal{O}(S_{\mathsf{prover}}/n)$ space; second, in terms

of computation, the cryptographic operations are almost equally divided amongst all the n servers, i.e., all the servers perform $\mathcal{O}(T_{\mathsf{crypto}}/n)$ cryptographic operations, and the remaining field operations are divided amongst the servers such that the server with more memory performs $o(T_{\mathsf{field}})$[10] field operations, while the remaining $(n-1)$ low-memory servers each perform $\mathcal{O}(T_{\mathsf{field}}/n)$ field operations.

Pre-Processing. Finally, we assume that the servers get access to some correlated-randomness at the start of this MPC protocol. This *relation-independent* correlated-randomness can be generated as part of a pre-processing step with $o(T_{\mathsf{field}})$ computation complexity (requiring only non-cryptographic operations). This can be either be generated by the client or by the servers themselves using a generic MPC protocol. Since the computation in this pre-processing phase is independent of the relation, it can be pre-computed by the servers during downtime.

Communication and Round Complexity. Finally, we remark that in order to minimize the overhead from communication, we want to limit the total communication to $\mathcal{O}(T_{\mathsf{field}})$ and restrict the number of rounds of interaction between the servers to a constant value. zkSaaS-framework is formalized in Definition 1.

4 Overview of Groth, Marlin and Plonk

In this section, we review the design of the prover algorithms in three widely used zk-SNARK construction: Groth16 [48], Marlin [27] and Plonk [37]. We start by discussing the key components (that are often the main bottleneck) used in the generation of Groth16-, Marlin- and Plonk-proofs.

Fast Fourier Transform, Polynomial Multiplication and Division. One of the key components used to optimize prover computation is the Fast Fourier Transform (FFT), which helps evaluate a given polynomial at m points in $\mathcal{O}(m \log m)$ time. The prover computations also typically require additional FFT-based computations: (1) *Inverse FFT (iFFT)* is used to convert m evaluations of a polynomial to the coefficient representation of the polynomial in time $\mathcal{O}(m \log m)$; (2) *Polynomial Multiplication*, which given the coefficient representation, can just be computed using one call each to FFT and iFFT along with m field multiplications and takes $\mathcal{O}(m \log m)$ time; (3) *Polynomial Division*, which can actually be run by making two calls to polynomial multiplication –takes $\mathcal{O}(m \log m)$ time, where m is the degree of the dividend.

Multi-scalar Multiplications (MSMs). Multi-scalar multiplications are of the form $\prod_{j \in [m]} (X_j)^{y_j}$, where $y_1, \ldots, y_m \in \mathbb{F}$, and $X_1, \ldots, X_m \in \mathbb{G}$.

Polynomial Commitments. In interactive oracle proofs (IOP)-based zkSNARKs, like Marlin and Plonk, in each round, the prover sends polynomial oracles to the verifier, which are essentially encodings of the witness, that the verifier can query. To convert these polynomial-IOPs to SNARKs, the prover commits to these polynomial oracles using the KZG polynomial commitment scheme [54]. These commitments allow a prover to commit to a univariate polynomial $p(X) \in \mathbb{F}[X]$ and get a com, such that the prover can later open to an evaluation of $p(X)$ at any point z, while giving a proof π of correct evaluation. In KZG commitment, to commit to a polynomial of degree d, MSM function is evaluated on d field and group elements, and to generate an opening proof the prover needs to perform one polynomial division followed by an MSM operation.

Sumcheck Protocol. Marlin relies heavily on what is known as a sumcheck protocol for univariate polynomials. For a polynomial oracle $p(X) \in \mathbb{F}[X]$ sent by the prover, this involves giving a proof of the fact that evaluations of $p(X)$ on the set $S := \{1, \omega, \ldots, \omega^{m-1}\}$, sums to some value $\sigma \in \mathbb{F}$, where ω is the m-th primitive root of unity in the field $\mathbb{F}$. It was shown in [9] that $\sum_{x \in S} p(x) = \sigma$, if and only if $p(X)$ can be written as $q(X) \cdot X + \sigma/|S|$, for some $q(X) \in \mathbb{F}[X]$. Thus, the prover in the polynomial-IOP first evaluates the polynomial $q(X)$ by dividing the polynomial $p(X) - \sigma/|S|$ by X and sends $q(X)$ as an oracle. Hence, for running each sumcheck, the prover in Marlin invokes polynomial division before committing to this polynomial using the polynomial commitment.

Partial Products. In Plonk, the prover needs to compute (for reference, see round 2 of [37, Section 8.3]) partial products of the form $\prod_{i \in [j]} p(\omega^{i-1})$ for all $j \in |S|$, where $p(X) \in \mathbb{F}[X]$ is some polynomial (which in turn is some encoding of the witness), and $S := \{1, \omega, \ldots, \omega^{m-1}\}$, where ω is the m-th primitive root of unity in the field $\mathbb{F}$. The prover uses this in computing the polynomial $z(X)$ obtained by additional polynomial multiplication and addition operations and sends it as an oracle in the polynomial-IOP protocol. In the final Plonk protocol z is committed using the polynomial commitment.

4.1 Groth16, Marlin, and Plonk Provers

We now briefly describe how each of the three zk-SNARKs that we consider can be computed via some combination of the above functions. We defer the details on how Groth and Plonk can be computed using our zkSaaS framework to the full version [38].

Groth [48]. Groth16 is the smallest, non-interactive

[10] We note that here we are hiding a logarithmic factor in the n

zk-SNARK, where the prover only sends 3 group elements to the verifier. This construction makes use of a structured CRS consisting of group elements proportional to the number of constraints. To generate the proof, the prover needs to compute a polyomial multiplication, polynomial division and a constant number of MSMs with the group elements in the CRS.

Marlin [27]. Marlin is a six round protocol, where overall, the prover generates the KZG polynomial commitments of 21 polynomials, and requires the following operations, each called for a small constant number of times: three sequential calls to the sumcheck protocol with each call additionally needing a call to polynomial division, all involving polynomials with a degree bound of the size of the relation, and polynomial additions. As a final step, this interactive protocol is converted to a SNARK using the Fiat-Shamir transformation.

Plonk [37]. Plonk on the other hand is a five round protocol, where overall the prover generates the KZG polynomial commitments of 9 polynomials, and requires the following operations, each called a small constant number of times, to generate these polynomials: polynomial multiplication and division involving polynomials with a degree bound of the size of the relation, partial products, and polynomial additions. This interactive protocol is also converted to a SNARK in the RO model.

5 Distributed Sub-Protocols for the zkSaaS Framework

For each of the sub-functions (FFT, MSM, Sum of Partial Products) discussed in section 4, we build custom MPC protocols. Looking ahead, we show how to compose these protocols, to design a zkSaaS for a specific subclass of zk-SNARKs. The full description of these protocols is deferred to the full version of our paper [38].

5.1 Distributed Fast Fourier Transform

The fast Fourier transform (FFT) algorithm is a recursive divide and conquer algorithm that helps evaluate a polynomial at multiple points efficiently. In particular, to evaluate a polynomial $p(x) \in \mathbb{F}[X]$ of degree $m-1$ at the points $S = \{\omega^i : i \in [m]\}$, where ω is the m-th primitive root of unity in the field $\mathbb{F}$, FFT does the following: At level $i = \log m$, each of the m polynomials will be evaluated at a single point, which is the identity element in $1 \in \mathbb{F}$. Subsequently, given the evaluations at level i (for any $i \in [\log(m), 1]$), the FFT algorithm gives us evaluations at the level $i-1$ in the following way: For each $j \in [2^{i-1}]$, and each $k \in [m/2^i]$, $x_j^{i-1,k} := x_{2j-1}^{i,k} + \omega^{k2^{i-1}} \cdot x_{2j}^{i,k}$ and $x_j^{i-1,(m/2^i)+k} := x_{2j-1}^{i,k} + \omega^{((m/2^i)+k)2^{i-1}} \cdot x_{2j}^{i,k}$. At the end ($i = 0$), the algorithm outputs all the m evaluations of p at S. We represent the recursion by the following function defined for each $i = \log m, \ldots, 1$:

$$F_{\mathsf{FFT}}^i(\{x_j^{i,k}\}_{j\in[2^i],k\in[m/2^i]}) := \{x_j^{i-1,k}\}_{j\in[2^{i-1}],k\in[m/2^{i-1}]}, \tag{1}$$

Here, the input to $F_{\mathsf{FFT}}^{\log m}$ are the values $x_j = x_j^{\log m,1}$ for each $j \in [m]$, representing the evaluations[11] of each of the m polynomials of level $i = \log m$ at the single point 1. Note that F_{FFT}^1 outputs the required evaluations $\{p(1), p(\omega), \ldots, p(\omega^{m-1})\}$. Using this notation, the FFT algorithm can be written as $F_{\mathsf{FFT}}(x_1, \ldots, x_m) = F_{\mathsf{FFT}}^1(F_{\mathsf{FFT}}^2(\ldots F_{\mathsf{FFT}}^{\log m}(x_1, \ldots, x_m)))$.

An Alternate View of the FFT algorithm. Our goal is to compute F_{FFT} via a secure MPC protocol. Notice that FFT is a logarithmic step algorithm, while we want a constant round MPC protocol for computing it. Towards designing such a protocol, we begin by making some key observations about FFT, and abstracting the main idea behind our final MPC protocol. For ease of exposition, consider an example where the input size is $m = 32$. We want to convert the linear function evaluation on each pair of values (at each recursive level i), into a linear function evaluation on a pair of vectors (of say $\ell = 4$ values) and be able to recurse on these vectors. Looking ahead, this will help us to pack share these vectors together and locally compute on them.

1. *FFT Step I:* Since $\log m = \log 32 = 5$, the inputs to the FFT algorithm will be $x_1^{5,1}, \ldots, x_{32}^{5,1}$. The next level $i = 4$ of the recursion is computed as: $x_j^{4,1} = x_{2j-1}^{5,1} + \omega^{16} \cdot x_{2j}^{5,1}$ and $x_j^{4,2} = x_{2j-1}^{5,1} + \omega^{32} \cdot x_{2j}^{5,1}$, for each $j \in [16]$. We now look at the same step as being computed on a vector instead of individual values. Suppose we group $\ell = 4$ elements to get the following 8 vectors at level $i = 5$: $\mathbf{x}_1^{5,1} = (x_1^{5,1}, x_3^{5,1}, x_5^{5,1}, x_7^{5,1})$, $\mathbf{x}_2^{5,1} = (x_2^{5,1}, x_4^{5,1}, x_6^{5,1}, x_8^{5,1})$, $\ldots$, $\mathbf{x}_8^{5,1} = (x_{26}^{5,2}, x_{28}^{5,2}, x_{30}^{5,2}, x_{32}^{5,2})$.

   Observation I. The key observation here is that, each pair of vectors $\mathbf{x}_{2j}^{5,1}$ and $\mathbf{x}_{2j-1}^{5,1}$ are used to compute two vector evaluations $\mathbf{x}_j^{4,1}$ (which uses ω^{16}) and $\mathbf{x}_j^{4,2}$ at the level $i = 4$ (which uses ω^{32}), for each $j \in [4]$. In other words:

$$\forall j \in [4], \mathbf{x}_j^{4,1} = \mathbf{x}_{2j}^{5,1} + \omega^{16} \cdot \mathbf{x}_{2j-1}^{5,1} \;\&\; \mathbf{x}_j^{4,2} = \mathbf{x}_{2j}^{5,1} + \omega^{32} \cdot \mathbf{x}_{2j-1}^{5,1}.$$

2. *FFT Step II:* We want to continue to compute on pairs of vectors linearly to obtain the next recursive step $i = 3$. However, note that now $\mathbf{x}_{2j-1}^{4,1}$ and

[11] For $p(X) = \sum_{i\in[m]} c_{i-1}x^{i-1}$, the x_i's are just a rearrangement of the c_i's, obtained by recursively reordering the c_i's as: put the even indexed terms at each level on the left and the odd indexed terms on the right. Continue the recursion for $\log m$ steps.

$\mathbf{x}_{2j}^{4,1}$ do not contain the values that are linearly combined in the next recursive step of FFT. For instance we have vectors $\mathbf{x}_1^{4,1} = (x_1^{4,1}, x_2^{4,1}, x_3^{4,1}, x_4^{4,1})$ and $\mathbf{x}_2^{4,1} = (x_5^{4,1}, x_6^{4,1}, x_7^{4,1}, x_8^{4,1})$, while the values that we want to combine are $x_1^{4,1}$ with $x_2^{4,1}$, and $x_3^{4,1}$ with $x_4^{4,1}$.

Observation II. The key observation that we make to resolve this issue is that, if at level $i = 5$, we had started with vectors: $\mathbf{x}_1^{5,1} = (x_1^{5,1}, x_5^{5,1}, x_9^{5,1}, x_{13}^{5,1})$, $\mathbf{x}_2^{5,1} = (x_2^{5,1}, x_6^{5,1}, x_{10}^{5,1}, x_{14}^{5,1})$, etc., then Step 1 would have led to the vectors (at $i = 4$): $\mathbf{x}_1^{4,1} = (x_1^{4,1}, x_3^{4,1}, x_5^{4,1}, x_7^{4,1})$, $\mathbf{x}_2^{4,1} = (x_2^{4,1}, x_4^{4,1}, x_6^{4,1}, x_8^{4,1})$, etc. These vectors $\mathbf{x}_1^{4,1}$ and $\mathbf{x}_2^{4,1}$ can now be combined using a linear combination, the same way as we did in step 1, to get $\mathbf{x}_1^{3,1}$ and $\mathbf{x}_1^{3,2}$ (and similarly others) of level $i = 3$. But, this reordering will only help us for this level and we run into the same issue when computing the next level $i = 2$.

Our Main Idea: The key point from observations I and II above is that in order to continue doing the FFT computations for each recursive level as a linear combination of vectors, instead of individual values, we must take the initial vectors at level $i = 5$ to be such that the values $x_j^{5,k}$'s in the same vector have the j's as far away as possible– this ensures that we push our problem down to as far a recursive layer as possible. However, even with the best ordering that we start with, we would reach a recursive level beyond which we cannot hope to compute on the vectors through a linear combination. Our MPC protocol combines the above key idea of packing the inputs into vectors such that local computations can be performed on them for as long as possible, along with additional techniques to overcome the challenge in computing the remaining recursive layers.

MPC for FFT. For m inputs, we start with a packed sharing of ℓ-sized vectors at level $i = \log m$: $\mathbf{x}_j^{\log m,1} = (x_j^{\log m,1}, x_{\frac{m}{\ell}+j}^{\log m,1}, \ldots, x_{\frac{m(\ell-1)}{\ell}+j}^{\log m,1})$, for each $j \in [m/\ell]$. As discussed in our main idea, this allows us to locally compute on the shares at each recursive layer (as in Steps 1 and 2) until level $i = \log \ell + 1$, beyond which we cannot do a local computation.

Beyond $i = \log \ell$: One approach that comes to mind is to rearrange the elements packed together (using some interaction) in such a way that a similar local computation can be done. However, one can observe that doing such a rearrangement actually leads to another problem– each pair of vectors are combined now using a vector of the ω^i's (instead of a single one), which leads to an "interactive" multiplication protocol at each level. This approach does give a feasible solution, but requires the parties to communicate at each of the remaining $\log \ell$ levels. Furthermore, the total communication in each round is $\mathcal{O}(m)$ which will become a bottleneck when dealing with a large constraint size.

We minimize the number of communication rounds and give a more efficient solution than the above by making use of our all powerful server and just two rounds of communication. On a high level, this uses the fact that FFT (and each of its recursive layer) is a linear function, i.e.: $F_{\mathsf{FFT}}^i((x_1 + r_1), \ldots, (x_m + r_m)) = F_{\mathsf{FFT}}^i(x_1, \ldots, x_m) + F_{\mathsf{FFT}}^i(r_1, \ldots, r_m)$. Suppose that the parties get packed shares of random values $(r_1, \ldots, r_m)$ and packed shares of $(s_1, \ldots, s_m)$ generated as: $(s_1, \ldots, s_m) = F_{\mathsf{FFT}}^1(F_{\mathsf{FFT}}^2(\ldots F_{\mathsf{FFT}}^{\log \ell}(r_1, \ldots, r_m)))$.

Then, the packed shares of the level $i = \log \ell$ are masked using the packed shares of $(r_1, \ldots, r_m)$ locally, and sent to the powerful party P_1. Now, P_1 reconstructs, computes the remaining recursive levels of FFT until $i = 1$, and sends the packed shares of the output to all parties. By virtue of linearity, the parties can obtain packed shares of the FFT output by locally subtracting the packed shares of $(s_1, \ldots, s_m)$. This securely reduces the communication rounds to two[12].

Complexity of our distributed FFT. Our protocol runs in two rounds, where in the first round each party communicates $\mathcal{O}(m/\ell)$ field elements and in the second round, party P_1 communicates $\mathcal{O}(m/\ell)$ field elements to each of the remaining parties. P_1 does $\mathcal{O}((\log \ell + \log n)m + m(\log m - \log \ell)/\ell)$ field operations and has a space complexity of $\mathcal{O}(m)$. The remaining parties perform $\mathcal{O}(m(\log m - \log \ell)/\ell)$ field operations and require $\mathcal{O}(m/\ell)$ space.

5.2 Distributed Partial Products

In this section, we discuss our ideas for securely computing functions of the form $F_{\mathsf{part}}(x_1, \ldots, x_m) = (\prod_{j \in [i]} x_j)_{i \in [m]}$, in a distributed way. When computing on a single machine, this function requires computing the $x_{[1,i]} := x_1 \cdots x_i$ values for each $i \in [m]$ in a sequential order. Simply implementing this approach inside an MPC protocol will require $\mathcal{O}(m)$ rounds. Moreover, since each step only requires multiplying two values at a time (i.e., $x_{[1,i-1]}$ and x_i), it is unclear how to leverage packed sharing to get a division of work amongst the parties.

Our goal is to design a computation mechanism that is more amenable to parallelism and where we can meaningfully use an approach based on packed secret sharing.

The key idea to achieve this comes from rewriting F_{part} in the following way: $F_{\mathsf{part}}(x_i, \ldots, x_j) = (x_i, x_{[i,i+1]}, \ldots, x_{[i,j]}) = (F_{\mathsf{part}}(x_{\frac{(i-1)m}{\ell}+1}, \ldots, x_{\frac{im}{\ell}}) \cdot$

[12] A curious reader might wonder why the linearity doesn't help us use the powerful server to compute all the recursive levels. For input size m (which is as large as the constraint size), this solution leads to a $\mathcal{O}(m \log m)$ compute time for both the pre-processing step and the server time, as opposed to our demand of $\mathcal{O}(m)$ compute time for both.

$x_{[1,\frac{(i-1)m}{\ell}]})_{i\in[\ell]}$. Observe that the $F_{\mathsf{part}}(x_{\frac{(i-1)m}{\ell}+1},\ldots,x_{\frac{im}{\ell}})$'s depend on disjoint subsets of the input x_i's. Thus, they can all be computed in parallel. In fact, since each of these $F_{\mathsf{part}}(x_{\frac{(i-1)m}{\ell}+1},\ldots,x_{\frac{im}{\ell}})$ are computed identically, albeit on a different set of inputs, this is exactly the kind of SIMD computation for which packed secret sharing is most helpful.

MPC for F_{part}. We start with a packed secret sharing of vectors $\mathbf{x}_1,\ldots,\mathbf{x}_{m/\ell}$, where $\mathbf{x}_j$ is an ℓ-sized vector consisting of the j^{th} inputs i.e., $\mathbf{x}_j = (x_j, x_{\frac{m}{\ell}+j},\ldots,x_{\frac{m(\ell-1)}{\ell}+j})$, for each $j\in[m/\ell]$. We now compute $F_{\mathsf{part}}([\![\mathbf{x}_1]\!],\ldots,[\![\mathbf{x}_{m/\ell}]\!])$ to obtain packed shares of $(\mathbf{y}_j = (x_{[\frac{(i-1)m}{\ell}+1,\frac{(i-1)m}{\ell}+j]})_{i\in[\ell]})_{j\in[m/\ell]}$ using known techniques with $\mathcal{O}(m)$ total computation and communication.

A careful reader might have observed that while the above idea allows us to compute $\{\mathbf{y}_j\}_{j\in[m/\ell]}$ simultaneously, doing this naïvely will still require $\mathcal{O}(m/\ell)$ rounds. To avoid this, we observe that a slightly modified version of Bar-Ilan and Beaver's [5] constant-round MPC for unbounded multiplication can be used to compute this in a constant number of rounds.[13] We defer more details about this protocol and the modification to the technical sections.

Finally, to compute $F_{\mathsf{part}}(x_1,\ldots,x_m)$, given the packed secret sharings of $\{\mathbf{y}_j\}_{j\in[m/\ell]}$ from the previous step, we note that while computing these packed shares of $\{\mathbf{y}_j\}_{j\in[m/\ell]}$, the parties also inevitably end up computing a packed secret sharing of the vector $\mathbf{z} = (x_{[1,m/\ell]}, x_{[\frac{m}{\ell}+1,\frac{2m}{\ell}]},\ldots,x_{[\frac{m(\ell-1)}{\ell},m]})$.
Given $\{[\![\mathbf{y}_j]\!]\}_{j\in[m/\ell]}$ and $[\![\mathbf{z}]\!]$, our final step computes the shares of desired output:
(1) Convert a packed sharing of $\mathbf{z}$ into regular threshold shares of the individual elements in $\mathbf{z}$, i.e., $[x_{[1,m/\ell]}],\ldots,[x_{[\frac{m(\ell-1)}{\ell},m]}]$.
(2) Use the above modified version of Bar-Ilan and Beaver's protocol on these shares to compute shares $[x_{[1,m/\ell]}],[x_{[1,2m/\ell]}]\ldots,[x_{[1,m]}]$.
(3) Finally, for each $j\in[m/\ell]$, compute an inner product between $[\![\mathbf{y}_j]\!]$ and packed shares of vector $(1, x_{[1,m/\ell]}, x_{[1,2m/\ell]},\ldots,x_{[1,\frac{m(\ell-1)}{\ell}]})$.

Complexity of our distributed Partial Products Protocol. Our protocol runs in constant rounds, where each of the small servers communicate $\mathcal{O}(m/\ell)$ field elements. They perform $\mathcal{O}(m/\ell)$ field operations and require a space complexity of $\mathcal{O}(m/\ell)$. While the big server, P_1 has a space complexity of $\mathcal{O}(m)$ and performs $\mathcal{O}(m)$ field operations and communicates $\mathcal{O}(m)$ field elements. This is because of the sub-protocol that we adapt from Bar-Ilan and Beaver's [5] constant-round MPC for unbounded multiplication. Since our distributed FFT protocol already assumes that one of the servers has more memory and computational resources, this is the version we use in our implementation of Plonk. However, in the full version [38], we present an alternate protocol for distributed computation of partial products, where the total work gets equally divided amongst all parties. In particular, *each server* in that protocol requires $\mathcal{O}(m/\ell)$ field operations, a space complexity of $\mathcal{O}(m/\ell)$, and each server communicates $\mathcal{O}(m/\ell+n)$ field elements.

[13] We note that this protocol crucially relies on the fact that none of the values being multiplied are zero. Which is indeed the case (w.h.p.) for our use-case in Plonk.

5.3 Distributed Multi-Scalar Multiplications

Polynomial-based secret sharing schemes typically only support arithmetic operations over a finite field. Several zk-SNARKs perform many elliptic curve group operations, such as multiplying group elements or group exponentiations. Representing these group operations as an arithmetic circuit over a finite field and computing it inside an MPC protocol is not feasible.

Prior works [70, 61] have explored generalizations of polynomial-based secret sharing schemes for group operations. Let $\mathbb{G}$ be a group of order p, with generator g, such that each element $A\in\mathbb{G}$ can be represented as g^a, where $a\in\mathbb{Z}_p$. The main idea in these works is to first compute secret shares (say $s_1,\ldots,s_n$) of the field element a, and then compute the shares of A as $g^{s_1},\ldots,g^{s_n}$. This allows us to perform arithmetic field operations in the exponent which can be used for group exponentiation and for multiplying group elements.
(1) *Addition in the exponent.* Given packed secret shares of another vector $\mathbf{B} = (B_1,\ldots,B_\ell)\in\mathbb{G}^\ell$, each party P_i can locally multiply their shares $[\![\mathbf{A}]\!]_i\cdot[\![\mathbf{B}]\!]_i$, to get a valid packed secret sharing of $\mathbf{C} = (A_1\cdot B_1,\ldots,A_\ell\cdot B_\ell)$.
(2) *Multiplication in the exponent.* Given packed secret shares of another vector of field elements $\mathbf{b} = (b_1,\ldots,b_\ell)\in\mathbb{Z}_p^\ell$, each party P_i can locally compute $[\![\mathbf{A}]\!]_i^{[\![\mathbf{b}]\!]_i}$ to get a packed secret sharing of $\mathbf{C} = (A_1^{b_1},\ldots,A_\ell^{b_\ell})$. However, in this case, since the shares of $\mathbf{a}$ and $\mathbf{b}$ get multiplied in the exponent, the degree of the resulting sharing will be twice that of the original sharings. To reduce the degree, we can use the standard ideas for degree reduction, albeit in the exponent.

MPC for MSM. Given the above observations, our idea for computing multi-scalar multiplications of the form $F_{\mathsf{MSM}}(A_1,b_1\ldots,A_m,b_m) = \prod_{i\in[m]}A_i^{b_i}$ is to first observe that this decomposes as is quite intuitive. Observe, this computation can be decomposed as:

$\prod_{i\in[\ell]}\left(F_{\mathsf{MSM}}(A_i, b_i, A_{\ell+i}, b_{\ell+i}, \ldots, A_{(\frac{m}{\ell}-1)\ell+i}, b_{(\frac{m}{\ell}-1)\ell+i})\right)$. This is essentially equivalent to computing ℓ instances of F_{MSM} in parallel and then multiplying the ℓ outputs. We compute this using PSS as follows: (1) Assuming that the parties have packed secret shares of vectors $\mathbf{A}_j = (A_{(j-1)\ell+i})_{i\in[\ell]}$ and $\mathbf{b}_j = (b_{(j-1)\ell+i})_{i\in[\ell]}$ for each $j \in [m/\ell]$, the parties compute F_{MSM} function on these packed shares to get packed shares of a vector $\mathbf{C}$. (2) Convert $[\![\mathbf{C}]\!]$ to regular threshold shares $[C_1], \ldots, [C_\ell]$ of the individual elements in $\mathbf{C}$. (3) Finally, the parties locally multiply these shares to get a sharing of the desired output.

Complexity of distributed MSM. Our protocol runs in constant rounds, where each party communicates $\mathcal{O}(1)$ group elements. All parties perform $\mathcal{O}(m/\ell)$ group exponentiations and have a space complexity of $\mathcal{O}(m/\ell)$.

6 zkSaaS for Admissible zk-SNARKs

In this section, we formally define a notion of admissible zk-SNARKs and show that our techniques from Section 5 can be used to obtain a zkSaaS for them.

Admissible zk-SNARKs. We start by formalizing a class of zk-SNARKs that are amenable to our zkSaaS framework, and refer to them as admissible zk-SNARKs. Informally speaking, we say that a zk-SNARK with computation complexity $T_{\mathsf{field}} + T_{\mathsf{crypto}}$ is admissible if the prover algorithm is composed of some combination of a subset or all of the following six types of operations on the satisfying assingment $\mathbf{z}$ for the RICS relation R — (1) multi-scalar multiplications (MSMs), (2) Fast Fourier Transforms (FFT), (3) sum of partial-products, (4) multiplication/Hadamard product, (5) additions and (6) permutations.

To formally capture this, our initial idea is to say that the prover algorithm in admissible zk-SNARKs can be represented as a polynomial-sized circuit consisting of special gates with "multi-ary" inputs and outputs, where each of these special gates correspond to one of the above six operations. However, this is alone is not sufficient. To capture the efficiency requirements of a zkSaaS (as discussed in Section 3), we need to further restrict the number of times a particular gate with a certain ari-ty can appear in this circuit.

Indeed, consider for instance a circuit, where two-input multiplication gates appear $\mathcal{O}(T_{\mathsf{field}})$ times in the circuit. Since the only distributed protocol that we can use for evaluating such gates is π_{mult} (c.f. Figure **??**), which requires a total communication and computation of $\mathcal{O}(n)$, the total communication and computation incurred in evaluating $\mathcal{O}(T_{\mathsf{field}})$ such gates would be $\mathcal{O}(n.T_{\mathsf{field}})$. This clearly violates the efficiency requirement of zkSaaS. Therefore, we must limit the number of such gates with low-ary inputs that appear in this circuit to ensure that the cost of computing them does not surpass the asymptotic bound that we have on the total computation complexity of zkSaaS. More concretely, in order to use our packed secret sharing based sub-protocols, we must limit the number of gates with $o(n)$ inputs. Hence, we define the notion of admissibility w.r.t. the number of parties n. This is formalized in Definition 2.

We now present our main composition theorem and show that the three zk-SNARKs that we discussed in Section 4.1 are admissible.

Theorem 1. *Let λ be the security parameter, $R \in \mathcal{R}_\lambda$ be an NP-relation and $\Sigma = (\mathsf{Setup}, \mathsf{Prove}, \mathsf{Ver}, \mathsf{Sim})$ be an n-admissible zk-SNARK for relation R. Then, there exists a secure n-server* zkSaaS *Π for Σ, which securely computes* Prove *in the $f_{\mathsf{double-prand}}, f_{\mathsf{prand}}, f_{\mathsf{pack-mult}}, f_{\mathsf{rand}}, f_{\mathsf{psstoss}}, f_{\mathsf{mult}}, f_{\mathsf{sstopss}}$, $f_{\mathsf{FFT}}, f_{\mathsf{MSM}}, f_{\mathsf{permute}}, f_{\mathsf{part-product}}, f_{\mathsf{poly-mult}}, f_{\mathsf{poly-divide}}$-hybrid model.*

We give the formal proof of this theorem along with a formal description of all the ideal functionalities in the full version [38] of our paper.

Instantiation. We note that, as discussed in Section 4.1, the provers of Groth16 [48], Marlin [27] and Plonk [37] only call the functionalities listed in defintion 2. Furthermore, the number of gates with $o(n)$-inputs in each of these is at most a constant number. Hence, using Theorem 1 we can directly get a zkSaaS for Groth16, Marlin and Plonk. Note here that Marlin and Plonk are described as interactive protocols, but as mentioned in Section 4.1 they can be converted to non-interactive protocols in the random oracle model, by using the Fiat-Shamir transformation. Specifically, this would require the prover to make random oracle queries on parts of the transcript– in our zkSaaS this translates to each party reconstructing shares of the transcript, to make these random oracle queries locally. The protocol clearly remains zero-knowledge.

7 Implementation and Evaluation

To evaluate the concrete performance of our techniques, we implemented a proof-of-concept zkSaaS framework supporting Groth16 [48] and Plonk [37] in Rust. We use the `arkworks` [3] library for finite field, pairing-friendly curves and, FFT implementations and the `mpc-net` crate from the collaborative-snarks implementation [2] to facilitate communication between parties. Our code is available on Github.[14] All of our experiments are

[14] https://github.com/guruvamsi-policharla/zksaas

Definition 2 (n-Admissible zk-SNARKs)**.** *Let λ be the security parameter, $R \in \mathcal{R}_\lambda$ be an NP-relation and $\Sigma = (\mathsf{Setup}, \mathsf{Prove}, \mathsf{Ver}, \mathsf{Sim})$ be a zk-SNARK for R. We say that Σ is n admissible if $n < T_{\mathsf{field}}$, $n < T_{\mathsf{crypto}}$ and the* Prove *algorithm can be represented as a circuit* C *comprising of gates implementing the following functionalities:*

Multi-Scalar Multiplication: $F_{\mathsf{MSM}}(y_1,\ldots,y_m,X_1,\ldots,X_m) = \prod_{j\in[m]}(X_j)^{y_j}$.
Fast Fourier Transform: $F_{\mathsf{FFT}}(x_1,\ldots,x_m) = F^1_{\mathsf{FFT}}(F^2_{\mathsf{FFT}}(\ldots F^{\log m}_{\mathsf{FFT}}(x_1,\ldots,x_m)))$*, where each* F^i_{FFT} *is the recursive function described in equation 1.*
Sum of Partial Products: $F_{\mathsf{part-prod}}(x_1,\ldots,x_m) = \sum_{j\in[m]}\prod_{i\in[j]} x_i$.
Multiplication/Hadamard Product: $F_{\mathsf{prod}}(x_1,\ldots,x_m,y_1,\ldots,y_m) = (x_1\cdot y_1),\ldots,(x_m\cdot y_m)$.
Addition: $F_{\mathsf{sum}}(x_1,\ldots,x_m,y_1,\ldots,y_m) = (x_1+y_1),\ldots,(x_m+y_m)$.
Permutation: $F_{\mathsf{perm}}(x_1,\ldots,x_m) := (x_{\mathsf{perm}(1)},\ldots,x_{\mathsf{perm}(m)})$*, where* perm *is a permutation function on* $[m]$.

Furthermore, the total number of MSM gates with $m \in o(n)$ inputs is limited to $\mathcal{O}(T_{\mathsf{crypto}}/n)$ and all other types of gates $m \in o(n)$ inputs are limited to $\mathcal{O}(T_{\mathsf{crypto}}/n)$.

run on the Google Cloud Platform (GCP) using two types of machines – all servers with low memory requirements are custom `N1` instances with 1 vCPU and 2GB of memory, while the powerful server is a custom `N1` instance with 96 vCPUs and 128 GB of RAM.

We compare the performance of our zkSaaS protocol against a prover running locally on an `N1` instance with 1 vCPUs and 4 GB of RAM, emulating a mid tier consumer laptop and hence refer to this as the consumer machine. Our VM configuration choices aim to reflect realistic deployment scenarios for zkSaaS, where one powerful instance is hired to aid many weak — volunteer-run nodes, often on outdated and older hardware. Throughout the analysis when the zkSaaS protocol is run with n parties, the corruption threshold is set to be $t = n/4 - 1$ and the number of secrets packed together to be $\ell = n/4$. All numbers reported are the average of five trials.

We view our implementation as a proof-of-concept to estimate running times and network delays and do not implement multi-threading on the powerful server. In a production-level implementation, we expect the powerful server to use multi-threading in the FFT protocol which includes packing/opening shares and communicating with parties. The data we present takes this into account by dividing the time taken during computation by the number of threads on the server and the time spent during communication by $\min(n, \#\text{ of threads})$. Finally, we implement a variant of our distributed partial products protocol which avoids *all-to-all* communication but the king party does linear work. This does not affect our speedup as we assume the king is multi-threaded and simplifies the communication to a *star* like structure.

Pre-processing. Our goal is to analyze the *online* work carried out by the servers. In both the single prover baseline and the zkSaaS protocol, we do not benchmark the time taken to prepare the witness, since we assume this is given to the zkSaaS servers by the clients. We do not evaluate pre-processing as it can be carried out during periods of low demand when spare compute and bandwidth are available. Prior work [61] also omits this analysis.

Evaluation. We now evaluate the performance of our zkSaaS framework against a prover running locally on a consumer machine for Groth16 and Plonk. In particular, we aim to answer three main questions: (1) How does the performance of zkSaaS compare to a Groth16/Plonk prover running on a single consumer machine? We are interested in two main metrics — (a) the largest number of constraints that can be supported without memory exhaustion and (b) the time required to generate the proof. (2) Does the performance of zkSaaS improve as we increase the number of servers? (3) How does the performance of zkSaaS vary with network bandwidth?

Varying Constraints. For our first experiment, we run benchmarks on a high speed network (4Gbps). We compare the running time of zkSaaS with 128 parties against a local execution of Groth16/Plonk prover on a single consumer machine, by varying the number of constraints. We summarize our results in Figure 1.

The performance of our zkSaaS for larger constraints is approximately 22× better better than the consumer machine. Here, we incur a loss from the theoretically expected savings of 32× due to a few factors:

— Pippenger's algorithm [64] provides a way to compute a multi-scalar multiplication $\Pi_{i=1}^{N} g_i^{\alpha_i}$ using $O(N/\log N)$ group operations as opposed to $O(N)$ in the naive strategy. However, Pippenger's algorithm is not conducive to our packed secret sharing MPC techniques as it lacks sufficient SIMD structure. With an optimal division of work, each weak server would carry out $O(N/(\ell \cdot \log N))$ group operations. While we do not attain an optimal division of

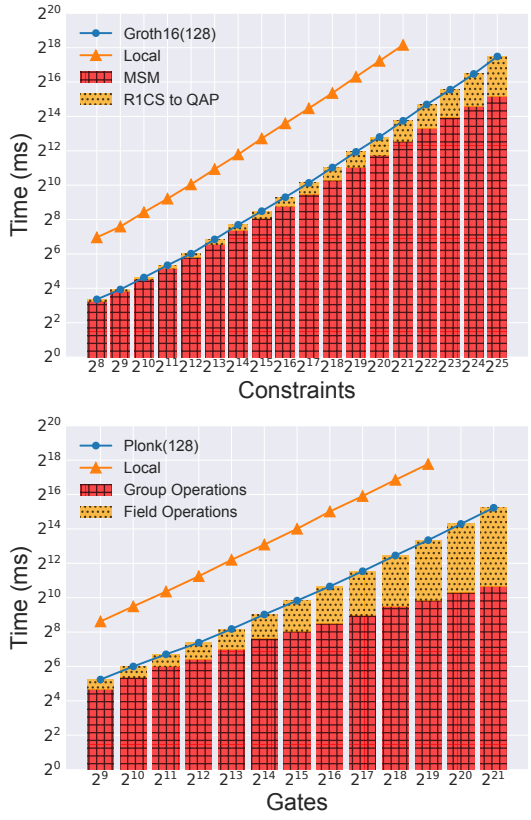

Figure 1: *Comparison between proof generation time for a local prover (Groth16 and Plonk resp.) run on the consumer machine against that of the* zkSaaS *protocol with 128 servers on a 4 Gigabit link. The bar graph indicates the* fraction *of time spent computing Field Operations (T_{field}) vs Group Operations (T_{crypto}). Missing data points on the local curve indicates memory exhaustion.*

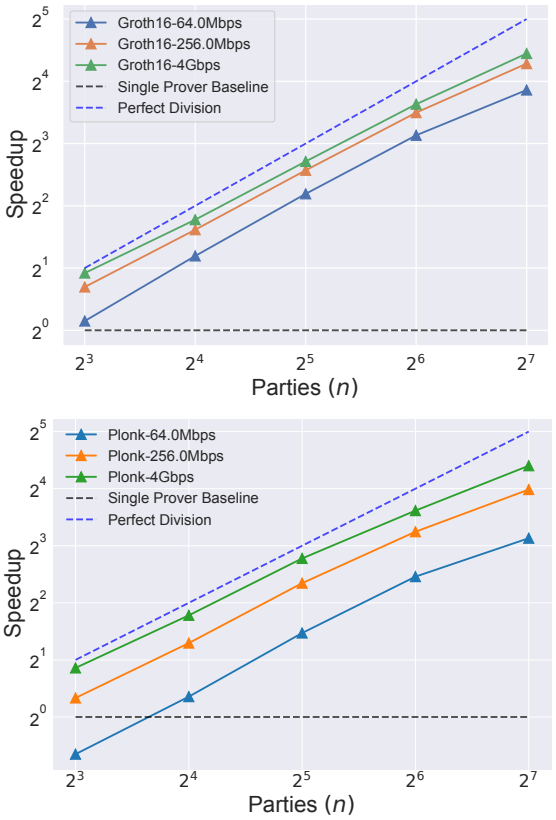

Figure 2: *Proving time versus number of parties, normalized by a single-prover time for 2^{19} constraints (gates) denoted by the dotted black line.*

work, we come very close with a per server complexity of $O(N/\ell \cdot \log N/\ell)$. In fact, since the constants inside the big O notation are the same, we can theoretically predict the percentage *loss* in performance under infinite bandwidth conditions by simply dividing the two asymptotics. As can be seen in Figure 2, a 4Gbps connection closely emulates infinite bandwidth and our implementation indeed comes very close to the theoretical prediction.

— In our distributed FFT/partial product protocol, during degree reduction, the King unpacks and repacks shares adding additional overhead. Our FFT implementation is also not as carefully optimized as the `arkworks` implementation which is used by the consumer machine.

Also, observe in Figure 1, that the fraction of time spent computing the field operations (refered to as R1CS to QAP mapping in the case of Groth16) increases with the number of constraints. This is because the FFT operation is asymptotically more expensive $(O(m \log m))$ than the MSM $(O(m/\log m))$.

Varying Parties and Network Speeds. For our next experiment, we show how zkSaaS scales as the number of parties increase, at varying network speeds. The dotted black line denotes the time taken by a local prover. We simulate slower connections by scaling up the time spent on communication by the network slowdown factor, comparing this to an implementation of Groth16/Plonk on a single consumer machine and present our findings in Figure 2. Even at lower network speeds (64 Mbps), we observe that the performance degradation is $\approx 2\times$.

Discussion on Financial Costs. We now estimate the costs of providing zk-SNARKs as a service. The powerful VM costs has a spot pricing of \$0.79/hr[15] and cross continent egress traffic pricing is \$0.08/GB[16]. Being very conservative, our estimates show that with 128 parties, generating a Groth16 proof for an R1CS instance of size

[15] https://cloud.google.com/compute/vm-instance-pricing

[16] https://cloud.google.com/compute/network-pricing

2^{19} takes under 1 minute on a 4-Gbps link and under 5 minutes on a 64 Mbps link, with the total outgoing communication from the server < 1.85GB. Hence, creating this proof would cost < 18 cents with a 4 Gbps link and < 23 cents on a 64 Mbps link.

Acknowledgements. Sanjam Garg, Guru-Vamsi Policharla and Sruthi Sekar are supported in part by DARPA under Agreement No. HR00112020026, AFOSR Award FA9550-19-1-0200, NSF CNS Award 1936826, and research grants by the Sloan Foundation, and Visa Inc. Guru-Vamsi Policharla is also supported by the UC Berkeley Center for Long-Term Cybersecurity. Abhishek Jain is supported in part by NSF CNS-1814919, NSF CAREER 1942789, Johns Hopkins University Catalyst award, AFOSR Award FA9550-19-1-0200, and research gifts from Ethereum, Stellar, Cisco.

References

[1] Surya Addanki, Kevin Garbe, Eli Jaffe, Rafail Ostrovsky, and Antigoni Polychroniadou. Prio+: Privacy preserving aggregate statistics via boolean shares. Cryptology ePrint Archive, Report 2021/576, 2021. https://eprint.iacr.org/2021/576.

[2] alex ozdemir. collaborative-zksnark. https://github.com/alex-ozdemir/collaborative-zksnark, 2022.

[3] arkworks contributors. `arkworks` zksnark ecosystem. https://arkworks.rs, 2022.

[4] László Babai, Lance Fortnow, and Carsten Lund. Non-deterministic exponential time has two-prover interactive protocols. In *31st FOCS*, pages 16–25. IEEE Computer Society Press, October 1990.

[5] Judit Bar-Ilan and Donald Beaver. Non-cryptographic fault-tolerant computing in constant number of rounds of interaction. In Piotr Rudnicki, editor, *Proceedings of the Eighth Annual ACM Symposium on Principles of Distributed Computing, Edmonton, Alberta, Canada, August 14-16, 1989*, pages 201–209. ACM, 1989.

[6] Gabrielle Beck, Aarushi Goel, Abhishek Jain, and Gabriel Kaptchuk. Order-C secure multiparty computation for highly repetitive circuits. In Anne Canteaut and François-Xavier Standaert, editors, *EUROCRYPT 2021, Part II*, volume 12697 of *LNCS*, pages 663–693. Springer, Heidelberg, October 2021.

[7] Eli Ben-Sasson, Alessandro Chiesa, Christina Garman, Matthew Green, Ian Miers, Eran Tromer, and Madars Virza. Zerocash: Decentralized anonymous payments from bitcoin. In *2014 IEEE Symposium on Security and Privacy*, pages 459–474. IEEE Computer Society Press, May 2014.

[8] Eli Ben-Sasson, Alessandro Chiesa, Matthew Green, Eran Tromer, and Madars Virza. Secure sampling of public parameters for succinct zero knowledge proofs. In *2015 IEEE Symposium on Security and Privacy*, pages 287–304, 2015.

[9] Eli Ben-Sasson, Alessandro Chiesa, Michael Riabzev, Nicholas Spooner, Madars Virza, and Nicholas P. Ward. Aurora: Transparent succinct arguments for R1CS. In Yuval Ishai and Vincent Rijmen, editors, *EUROCRYPT 2019, Part I*, volume 11476 of *LNCS*, pages 103–128. Springer, Heidelberg, May 2019.

[10] Nir Bitansky, Ran Canetti, Alessandro Chiesa, Shafi Goldwasser, Huijia Lin, Aviad Rubinstein, and Eran Tromer. The hunting of the SNARK. *J. Cryptol.*, 30(4):989–1066, 2017.

[11] Nir Bitansky, Ran Canetti, Alessandro Chiesa, and Eran Tromer. Recursive composition and bootstrapping for SNARKS and proof-carrying data. In Dan Boneh, Tim Roughgarden, and Joan Feigenbaum, editors, *45th ACM STOC*, pages 111–120. ACM Press, June 2013.

[12] Nir Bitansky and Alessandro Chiesa. Succinct arguments from multi-prover interactive proofs and their efficiency benefits. In Reihaneh Safavi-Naini and Ran Canetti, editors, *CRYPTO 2012*, volume 7417 of *LNCS*, pages 255–272. Springer, Heidelberg, August 2012.

[13] Alexander R. Block, Justin Holmgren, Alon Rosen, Ron D. Rothblum, and Pratik Soni. Public-coin zero-knowledge arguments with (almost) minimal time and space overheads. In Rafael Pass and Krzysztof Pietrzak, editors, *TCC 2020, Part II*, volume 12551 of *LNCS*, pages 168–197. Springer, Heidelberg, November 2020.

[14] Alexander R. Block, Justin Holmgren, Alon Rosen, Ron D. Rothblum, and Pratik Soni. Time- and space-efficient arguments from groups of unknown order. In Tal Malkin and Chris Peikert, editors, *CRYPTO 2021, Part IV*, volume 12828 of *LNCS*, pages 123–152, Virtual Event, August 2021. Springer, Heidelberg.

[15] Andrew J. Blumberg, Justin Thaler, Victor Vu, and Michael Walfish. Verifiable computation using multiple provers. Cryptology ePrint Archive,

Report 2014/846, 2014. https://eprint.iacr.org/2014/846.

[16] Dan Boneh, Elette Boyle, Henry Corrigan-Gibbs, Niv Gilboa, and Yuval Ishai. Zero-knowledge proofs on secret-shared data via fully linear PCPs. In Alexandra Boldyreva and Daniele Micciancio, editors, *CRYPTO 2019, Part III*, volume 11694 of *LNCS*, pages 67–97. Springer, Heidelberg, August 2019.

[17] Dan Boneh, Elette Boyle, Henry Corrigan-Gibbs, Niv Gilboa, and Yuval Ishai. Lightweight techniques for private heavy hitters. In *2021 IEEE Symposium on Security and Privacy*, pages 762–776. IEEE Computer Society Press, May 2021.

[18] Jonathan Bootle, Andrea Cerulli, Essam Ghadafi, Jens Groth, Mohammad Hajiabadi, and Sune K. Jakobsen. Linear-time zero-knowledge proofs for arithmetic circuit satisfiability. In Tsuyoshi Takagi and Thomas Peyrin, editors, *ASIACRYPT 2017, Part III*, volume 10626 of *LNCS*, pages 336–365. Springer, Heidelberg, December 2017.

[19] Jonathan Bootle, Andrea Cerulli, Jens Groth, Sune K. Jakobsen, and Mary Maller. Arya: Nearly linear-time zero-knowledge proofs for correct program execution. In Thomas Peyrin and Steven Galbraith, editors, *ASIACRYPT 2018, Part I*, volume 11272 of *LNCS*, pages 595–626. Springer, Heidelberg, December 2018.

[20] Jonathan Bootle, Alessandro Chiesa, and Jens Groth. Linear-time arguments with sublinear verification from tensor codes. In Rafael Pass and Krzysztof Pietrzak, editors, *TCC 2020, Part II*, volume 12551 of *LNCS*, pages 19–46. Springer, Heidelberg, November 2020.

[21] Jonathan Bootle, Alessandro Chiesa, and Siqi Liu. Zero-knowledge IOPs with linear-time prover and polylogarithmic-time verifier. In Orr Dunkelman and Stefan Dziembowski, editors, *EUROCRYPT 2022, Part II*, volume 13276 of *LNCS*, pages 275–304. Springer, Heidelberg, May / June 2022.

[22] Sean Bowe, Alessandro Chiesa, Matthew Green, Ian Miers, Pratyush Mishra, and Howard Wu. ZEXE: Enabling decentralized private computation. In *2020 IEEE Symposium on Security and Privacy*, pages 947–964. IEEE Computer Society Press, May 2020.

[23] Sean Bowe, Ariel Gabizon, and Matthew D. Green. A multi-party protocol for constructing the public parameters of the pinocchio zk-snark. In Aviv Zohar, Ittay Eyal, Vanessa Teague, Jeremy Clark, Andrea Bracciali, Federico Pintore, and Massimiliano Sala, editors, *Financial Cryptography and Data Security - FC 2018 International Workshops, BITCOIN, VOTING, and WTSC, Nieuwpoort, Curaçao, March 2, 2018, Revised Selected Papers*, volume 10958 of *Lecture Notes in Computer Science*, pages 64–77. Springer, 2018.

[24] Benedikt Bünz, Jonathan Bootle, Dan Boneh, Andrew Poelstra, Pieter Wuille, and Greg Maxwell. Bulletproofs: Short proofs for confidential transactions and more. In *2018 IEEE Symposium on Security and Privacy*, pages 315–334. IEEE Computer Society Press, May 2018.

[25] Vitalik Buterin. Some ways to use zk-snarks for privacy.

[26] Binyi Chen, Benedikt Bünz, Dan Boneh, and Zhenfei Zhang. HyperPlonk: Plonk with linear-time prover and high-degree custom gates. Cryptology ePrint Archive, Report 2022/1355, 2022. https://eprint.iacr.org/2022/1355.

[27] Alessandro Chiesa, Yuncong Hu, Mary Maller, Pratyush Mishra, Noah Vesely, and Nicholas P. Ward. Marlin: Preprocessing zkSNARKs with universal and updatable SRS. In Anne Canteaut and Yuval Ishai, editors, *EUROCRYPT 2020, Part I*, volume 12105 of *LNCS*, pages 738–768. Springer, Heidelberg, May 2020.

[28] Alessandro Chiesa, Ryan Lehmkuhl, Pratyush Mishra, and Yinuo Zhang. Eos: Efficient private delegation of zksnark provers. In *USENIX Security Symposium*. USENIX Association, 2023.

[29] Alessandro Chiesa, Dev Ojha, and Nicholas Spooner. Fractal: Post-quantum and transparent recursive proofs from holography. In Anne Canteaut and Yuval Ishai, editors, *EUROCRYPT 2020, Part I*, volume 12105 of *LNCS*, pages 769–793. Springer, Heidelberg, May 2020.

[30] Arka Rai Choudhuri, Aarushi Goel, Matthew Green, Abhishek Jain, and Gabriel Kaptchuk. Fluid MPC: Secure multiparty computation with dynamic participants. In Tal Malkin and Chris Peikert, editors, *CRYPTO 2021, Part II*, volume 12826 of *LNCS*, pages 94–123, Virtual Event, August 2021. Springer, Heidelberg.

[31] William R Claycomb and Alex Nicoll. Insider threats to cloud computing: Directions for new research challenges. In *2012 IEEE 36th annual com-*

puter software and applications conference, pages 387–394. IEEE, 2012.

[32] Graham Cormode, Michael Mitzenmacher, and Justin Thaler. Practical verified computation with streaming interactive proofs. In Shafi Goldwasser, editor, *ITCS 2012*, pages 90–112. ACM, January 2012.

[33] Ivan Damgård, Yuval Ishai, and Mikkel Krøigaard. Perfectly secure multiparty computation and the computational overhead of cryptography. In Henri Gilbert, editor, *EUROCRYPT 2010*, volume 6110 of *LNCS*, pages 445–465. Springer, Heidelberg, May / June 2010.

[34] Ivan Damgård, Yuval Ishai, Mikkel Krøigaard, Jesper Buus Nielsen, and Adam Smith. Scalable multiparty computation with nearly optimal work and resilience. In David Wagner, editor, *CRYPTO 2008*, volume 5157 of *LNCS*, pages 241–261. Springer, Heidelberg, August 2008.

[35] Zhiyong Fang, David Darais, Joseph P. Near, and Yupeng Zhang. Zero knowledge static program analysis. In Giovanni Vigna and Elaine Shi, editors, *ACM CCS 2021*, pages 2951–2967. ACM Press, November 2021.

[36] Matthew K. Franklin and Moti Yung. Communication complexity of secure computation (extended abstract). In *24th ACM STOC*, pages 699–710. ACM Press, May 1992.

[37] Ariel Gabizon, Zachary J. Williamson, and Oana Ciobotaru. PLONK: Permutations over lagrange-bases for oecumenical noninteractive arguments of knowledge. Cryptology ePrint Archive, Report 2019/953, 2019. https://eprint.iacr.org/2019/953.

[38] Sanjam Garg, Aarushi Goel, Abhishek Jain, Guru-Vamsi Policharla, and Sruthi Sekar. zkSaaS: Zero-knowledge snarks as a service. Cryptology ePrint Archive, Paper 2023/905, 2023.

[39] Daniel Genkin, Yuval Ishai, and Antigoni Polychroniadou. Efficient multi-party computation: From passive to active security via secure SIMD circuits. In Rosario Gennaro and Matthew J. B. Robshaw, editors, *CRYPTO 2015, Part II*, volume 9216 of *LNCS*, pages 721–741. Springer, Heidelberg, August 2015.

[40] Daniel Genkin, Yuval Ishai, Manoj Prabhakaran, Amit Sahai, and Eran Tromer. Circuits resilient to additive attacks with applications to secure computation. In David B. Shmoys, editor, *46th ACM STOC*, pages 495–504. ACM Press, May / June 2014.

[41] Shafi Goldwasser, Yael Tauman Kalai, and Guy N. Rothblum. Delegating computation: interactive proofs for muggles. In Richard E. Ladner and Cynthia Dwork, editors, *40th ACM STOC*, pages 113–122. ACM Press, May 2008.

[42] Alexander Golovnev, Jonathan Lee, Srinath Setty, Justin Thaler, and Riad S. Wahby. Brakedown: Linear-time and post-quantum SNARKs for R1CS. Cryptology ePrint Archive, Report 2021/1043, 2021. https://eprint.iacr.org/2021/1043.

[43] S. Dov Gordon, Daniel Starin, and Arkady Yerukhimovich. The more the merrier: Reducing the cost of large scale MPC. In Anne Canteaut and François-Xavier Standaert, editors, *EUROCRYPT 2021, Part II*, volume 12697 of *LNCS*, pages 694–723. Springer, Heidelberg, October 2021.

[44] Vipul Goyal, Antigoni Polychroniadou, and Yifan Song. Unconditional communication-efficient MPC via hall's marriage theorem. In Tal Malkin and Chris Peikert, editors, *CRYPTO 2021, Part II*, volume 12826 of *LNCS*, pages 275–304, Virtual Event, August 2021. Springer, Heidelberg.

[45] Vipul Goyal, Antigoni Polychroniadou, and Yifan Song. Sharing transformation and dishonest majority MPC with packed secret sharing. In Yevgeniy Dodis and Thomas Shrimpton, editors, *CRYPTO 2022, Part IV*, volume 13510 of *LNCS*, pages 3–32. Springer, Heidelberg, August 2022.

[46] Vipul Goyal, Antigoni Polychroniadou, and Yifan Song. Sharing transformation and dishonest majority MPC with packed secret sharing. Cryptology ePrint Archive, Report 2022/831, 2022. https://eprint.iacr.org/2022/831.

[47] Matthew Green, Mathias Hall-Andersen, Eric Hennenfent, Gabriel Kaptchuk, Benjamin Perez, and Gijs Van Laer. Efficient proofs of software exploitability for real-world processors. *Proc. Priv. Enhancing Technol.*, 2023(1):627–640, 2023.

[48] Jens Groth. On the size of pairing-based non-interactive arguments. In Marc Fischlin and Jean-Sébastien Coron, editors, *EUROCRYPT 2016, Part II*, volume 9666 of *LNCS*, pages 305–326. Springer, Heidelberg, May 2016.

[49] Jens Groth and Mary Maller. Snarky signatures: Minimal signatures of knowledge from simulation-extractable SNARKs. In Jonathan Katz and Hovav

Shacham, editors, *CRYPTO 2017, Part II*, volume 10402 of *LNCS*, pages 581–612. Springer, Heidelberg, August 2017.

[50] Paul Grubbs, Arasu Arun, Ye Zhang, Joseph Bonneau, and Michael Walfish. Zero-knowledge middleboxes. In *USENIX Security Symposium*, pages 4255–4272. USENIX Association, 2022.

[51] Justin Holmgren and Ron Rothblum. Delegating computations with (almost) minimal time and space overhead. In Mikkel Thorup, editor, *59th FOCS*, pages 124–135. IEEE Computer Society Press, October 2018.

[52] Harry A. Kalodner, Steven Goldfeder, Xiaoqi Chen, S. Matthew Weinberg, and Edward W. Felten. Arbitrum: Scalable, private smart contracts. In William Enck and Adrienne Porter Felt, editors, *USENIX Security 2018*, pages 1353–1370. USENIX Association, August 2018.

[53] Sanket Kanjalkar, Ye Zhang, Shreyas Gandlur, and Andrew Miller. Publicly auditable mpc-as-a-service with succinct verification and universal setup. In *IEEE European Symposium on Security and Privacy Workshops, EuroS&P 2021, Vienna, Austria, September 6-10, 2021*, pages 386–411. IEEE, 2021.

[54] Aniket Kate, Gregory M. Zaverucha, and Ian Goldberg. Constant-size commitments to polynomials and their applications. In Masayuki Abe, editor, *ASIACRYPT 2010*, volume 6477 of *LNCS*, pages 177–194. Springer, Heidelberg, December 2010.

[55] Markulf Kohlweiss, Mary Maller, Janno Siim, and Mikhail Volkhov. Snarky ceremonies. Cryptology ePrint Archive, Report 2021/219, 2021. https://eprint.iacr.org/2021/219.

[56] Abhiram Kothapalli, Elisaweta Masserova, and Bryan Parno. A direct construction for asymptotically optimal zksnarks. *IACR Cryptol. ePrint Arch.*, page 1318, 2020.

[57] Jonathan Lee. Dory: Efficient, transparent arguments for generalised inner products and polynomial commitments. In Kobbi Nissim and Brent Waters, editors, *TCC 2021, Part II*, volume 13043 of *LNCS*, pages 1–34. Springer, Heidelberg, November 2021.

[58] Seunghwa Lee, Hankyung Ko, Jihye Kim, and Hyunok Oh. vCNN: Verifiable convolutional neural network. Cryptology ePrint Archive, Report 2020/584, 2020. https://eprint.iacr.org/2020/584.

[59] Tianyi Liu, Xiang Xie, and Yupeng Zhang. zkCNN: Zero knowledge proofs for convolutional neural network predictions and accuracy. In Giovanni Vigna and Elaine Shi, editors, *ACM CCS 2021*, pages 2968–2985. ACM Press, November 2021.

[60] Assa Naveh and Eran Tromer. Photoproof: Cryptographic image authentication for any set of permissible transformations. In *IEEE Symposium on Security and Privacy, SP 2016, San Jose, CA, USA, May 22-26, 2016*, pages 255–271. IEEE Computer Society, 2016.

[61] Alex Ozdemir and Dan Boneh. Experimenting with collaborative zk-SNARKs: Zero-Knowledge proofs for distributed secrets. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 4291–4308, Boston, MA, August 2022. USENIX Association.

[62] Bryan Parno, Jon Howell, Craig Gentry, and Mariana Raykova. Pinocchio: Nearly practical verifiable computation. In *2013 IEEE Symposium on Security and Privacy, SP 2013, Berkeley, CA, USA, May 19-22, 2013*, pages 238–252. IEEE Computer Society, 2013.

[63] Alexey Pertsev, Roman Semenov, and Roman Storm. Tornado cash privacy solution version 1.4. 2019.

[64] Nicholas Pippenger. On the evaluation of powers and monomials. *SIAM J. Comput.*, 9(2):230–250, 1980.

[65] Deevashwer Rathee, Guru Vamsi Policharla, Tiancheng Xie, Ryan Cottone, and Dawn Song. Zebra: Anonymous credentials with practical on-chain verification and applications to kyc in defi. Cryptology ePrint Archive, Paper 2022/1286, 2022.

[66] Michael Rosenberg, Jacob White, Christina Garman, and Ian Miers. `zk-creds`: Flexible anonymous credentials from zkSNARKs and existing identity infrastructure. Cryptology ePrint Archive, Report 2022/878, 2022. https://eprint.iacr.org/2022/878.

[67] Berry Schoenmakers, Meilof Veeningen, and Niels de Vreede. Trinocchio: Privacy-preserving outsourcing by distributed verifiable computation. In Mark Manulis, Ahmad-Reza Sadeghi, and Steve Schneider, editors, *ACNS 16*, volume 9696 of *LNCS*, pages 346–366. Springer, Heidelberg, June 2016.

[68] Srinath Setty. Spartan: Efficient and general-purpose zkSNARKs without trusted setup. In Daniele Micciancio and Thomas Ristenpart, editors, *CRYPTO 2020, Part III*, volume 12172 of *LNCS*, pages 704–737. Springer, Heidelberg, August 2020.

[69] Adi Shamir. How to share a secret. *Communications of the Association for Computing Machinery*, 22(11):612–613, November 1979.

[70] Nigel P. Smart and Younes Talibi Alaoui. Distributing any elliptic curve based protocol. In Martin Albrecht, editor, *Cryptography and Coding - 17th IMA International Conference, IMACC 2019, Oxford, UK, December 16-18, 2019, Proceedings*, volume 11929 of *Lecture Notes in Computer Science*, pages 342–366. Springer, 2019.

[71] Justin Thaler. Time-optimal interactive proofs for circuit evaluation. In Ran Canetti and Juan A. Garay, editors, *CRYPTO 2013, Part II*, volume 8043 of *LNCS*, pages 71–89. Springer, Heidelberg, August 2013.

[72] Victor Vu, Srinath T. V. Setty, Andrew J. Blumberg, and Michael Walfish. A hybrid architecture for interactive verifiable computation. In *2013 IEEE Symposium on Security and Privacy, SP 2013, Berkeley, CA, USA, May 19-22, 2013*, pages 223–237. IEEE Computer Society, 2013.

[73] Riad S. Wahby, Max Howald, Siddharth Garg, Abhi Shelat, and Michael Walfish. Verifiable asics. In *IEEE Symposium on Security and Privacy, SP 2016, San Jose, CA, USA, May 22-26, 2016*, pages 759–778. IEEE Computer Society, 2016.

[74] Riad S. Wahby, Ioanna Tzialla, Abhi Shelat, Justin Thaler, and Michael Walfish. Doubly-efficient zksnarks without trusted setup. In *2018 IEEE Symposium on Security and Privacy, SP 2018, Proceedings, 21-23 May 2018, San Francisco, California, USA*, pages 926–943. IEEE Computer Society, 2018.

[75] Michael Walfish and Andrew J Blumberg. Verifying computations without reexecuting them. *Communications of the ACM*, 58(2):74–84, 2015.

[76] Chenkai Weng, Kang Yang, Xiang Xie, Jonathan Katz, and Xiao Wang. Mystique: Efficient conversions for zero-knowledge proofs with applications to machine learning. In Michael Bailey and Rachel Greenstadt, editors, *USENIX Security 2021*, pages 501–518. USENIX Association, August 2021.

[77] Howard Wu, Wenting Zheng, Alessandro Chiesa, Raluca Ada Popa, and Ion Stoica. DIZK: A distributed zero knowledge proof system. In William Enck and Adrienne Porter Felt, editors, *USENIX Security 2018*, pages 675–692. USENIX Association, August 2018.

[78] Tiancheng Xie, Jiaheng Zhang, Yupeng Zhang, Charalampos Papamanthou, and Dawn Song. Libra: Succinct zero-knowledge proofs with optimal prover computation. In Alexandra Boldyreva and Daniele Micciancio, editors, *CRYPTO 2019, Part III*, volume 11694 of *LNCS*, pages 733–764. Springer, Heidelberg, August 2019.

[79] Tiancheng Xie, Yupeng Zhang, and Dawn Song. Orion: Zero knowledge proof with linear prover time. In Yevgeniy Dodis and Thomas Shrimpton, editors, *CRYPTO 2022, Part IV*, volume 13510 of *LNCS*, pages 299–328. Springer, Heidelberg, August 2022.

[80] Jiaheng Zhang, Zhiyong Fang, Yupeng Zhang, and Dawn Song. Zero knowledge proofs for decision tree predictions and accuracy. In Jay Ligatti, Xinming Ou, Jonathan Katz, and Giovanni Vigna, editors, *ACM CCS 2020*, pages 2039–2053. ACM Press, November 2020.

[81] Jiaheng Zhang, Tianyi Liu, Weijie Wang, Yinuo Zhang, Dawn Song, Xiang Xie, and Yupeng Zhang. Doubly efficient interactive proofs for general arithmetic circuits with linear prover time. In Giovanni Vigna and Elaine Shi, editors, *ACM CCS 2021*, pages 159–177. ACM Press, November 2021.

[82] Yupeng Zhang, Daniel Genkin, Jonathan Katz, Dimitrios Papadopoulos, and Charalampos Papamanthou. vSQL: Verifying arbitrary SQL queries over dynamic outsourced databases. In *2017 IEEE Symposium on Security and Privacy*, pages 863–880. IEEE Computer Society Press, May 2017.

[83] Yupeng Zhang, Daniel Genkin, Jonathan Katz, Dimitrios Papadopoulos, and Charalampos Papamanthou. vram: Faster verifiable RAM with program-independent preprocessing. In *2018 IEEE Symposium on Security and Privacy, SP 2018, Proceedings, 21-23 May 2018, San Francisco, California, USA*, pages 908–925. IEEE Computer Society, 2018.

[84] ZkRollups. An incomplete guide to rollups. https://vitalik.ca/general/2021/01/ 05/rollup.html, 2021.