












FANNG-MPC: Framework for Artificial Neural Networks and Generic MPC

Najwa Aaraj¹, Abdelrahman Aly¹, Tim Güneysu², Chiara Marcolla¹,
Johannes Mono², Rogerio Paludo¹, Iván Santos-González¹, Mireia
Scholz¹, Eduardo Soria-Vazquez¹, Victor Sucasas¹ and Ajith Suresh¹

¹ Technology Innovation Institute, Abu Dhabi, UAE

`{firstname.lastname}@tii.ae`

² Ruhr University Bochum, Bochum, Germany

`{firstname.lastname}@rub.de`

Abstract. In this work, we introduce FANNG-MPC, a versatile secure multi-party computation framework capable to offer active security for privacy-preserving machine learning as a service (MLaaS). Derived from the now deprecated SCALE-MAMBA, FANNG is a data-oriented fork, featuring novel set of libraries and instructions for realizing private neural networks, effectively reviving the popular framework. To the best of our knowledge, FANNG is the first MPC framework to offer actively secure MLaaS in the dishonest majority setting.

FANNG goes beyond SCALE-MAMBA by decoupling offline and online phases and materializing the dealer model in software, enabling a separate set of entities to produce offline material. The framework incorporates database support, a new instruction set for pre-processed material, including garbled circuits and convolutional and matrix multiplication triples. FANNG also implements novel private comparison protocols and an optimized library supporting Neural Network functionality. All our theoretical claims are substantiated by an extensive evaluation using an open-sourced implementation, including the private inference of popular neural networks like LeNet and VGG16.

Keywords: Multi-Party Computation · Privacy-Preserving Machine Learning · Homomorphic Encryption · Neural Networks · MPC · FHE

1 Introduction

The rising relevance of AI and its rapid market penetration have sparked unprecedented interest in privacy-enhancing technologies (PETs) applied to machine learning (ML) [KM24]. This interest arises from the need to safeguard users' data, driven either by user concerns or obligations to regulations such as the General Data Protection Regulation (GDPR) in Europe and the California Consumer Privacy Act (CCPA) in the United States [Res24]. The privacy challenge in ML is twofold. Firstly, training AI models requires vast amounts of data, which may be distributed among various entities and subject to diverse privacy regulations. Secondly, the immense size of ML models makes deployment on the user side impractical, shifting AI towards cloud services, where users must send their data to the cloud. This poses a considerable privacy risk and can hinder the wide adoption of AI in scenarios involving sensitive data.

The privacy bottleneck can be mitigated through various PETs. Indeed, solutions based on data anonymization, generative models, or differential privacy have garnered attention from the research community, contributing to the coining of the term Privacy-Preserving Machine Learning (PPML) [CP21, NC23]. However, within the realm of PPML,

confidential computing has gained recent prominence. The concept involves performing data processing confidentially using secure Multi-Party Computation (MPC) and/or Fully Homomorphic Encryption (FHE), offering enhanced privacy guarantees. Several PPML frameworks have been proposed, incorporating MPC, FHE, or a combination of both. Unfortunately, the majority of these frameworks either employ generic constructions without ML-specific optimizations or utilize tailored protocols linked to specific trust settings, often confined to honest majority and semi-honest (i.e., passive) security, as depicted in Tab. 1.

In this work, we introduce FANNG-MPC, a novel MPC-based framework derived from the popular SCALE-MAMBA [ACC⁺21] framework. While maintaining the diverse set of protocols inherent in SCALE-MAMBA, FANNG introduces novel constructions tailored for PPML inference that operate in the dishonest majority setting with active security. Designed as a response to the common problem in PPML where many works offer only prototype implementations, FANNG aims to mitigate underestimation or neglect of challenges encountered in real-world deployments.

1.1 State-of-the-Art on PPML frameworks

In this section, we provide a brief overview of several important PPML frameworks that are based on MPC and HE techniques. While we will not go into the nuance of their protocol constructions, we will highlight their settings, techniques used, as well as the security model they adhere to. Given the extensive nature of the literature on this topic, we will focus on a selected few frameworks from each category, as listed in Tab. 1. For a more comprehensive understanding, we recommend referring to [KT21, CP21, NC23] for detailed information.

The field of PPML for two parties originated from the seminal work by Lindell and Pinkas [LP00], presenting a secure algorithm for data mining, particularly for decision trees. Subsequent research expanded to algorithms like k-means clustering, linear regression, and logistic regression, but these approaches suffer from high efficiency overheads and are primarily theoretical. Over the last two decades, significant advancements have occurred in the field, with practical implementations using MPC techniques. As illustrated in Tab. 1, the majority of the studies focus on scenarios involving a small number of parties, specifically 2, 3, and 4. While the 3- and 4-party protocols operate under an honest-majority assumption, all existing protocols in the dishonest-majority setting support only two parties, and sometimes use a trusted helper to improve efficiency. Furthermore, these two party protocols typically offer only passive security, with the exception of XONN [RSC⁺19], which can support active security through expensive cut-and-choose technique. However, XONN propose only a semi-honest security variant and presents general ideas for enhancing security without providing a concrete construction (see [RSC⁺19, §3.4]). Additionally, XONN’s construction is restricted to Binarized Neural Networks and relies on expensive Garbling techniques, which do not scale well to larger networks, such as the VGG16 model used in our work. Therefore, our emphasis in this work is on private ML inference in the two-party dishonest majority setting, with the goal of contributing to the closing of the gap in the literature on MPC-based PPML.

1.2 Our Contributions

In this work, we introduce FANNG-MPC, a versatile framework designed for efficient protocols using secure multi-party computation (MPC) techniques. Referred to as FANNG in short, our framework is an independent fork derived from SCALE-MAMBA [ACC⁺21]. It maintains the diverse set of protocols included in SCALE-MAMBA, which operates over various fields (\mathbb{F}_p), including binary. Concerning the threat model, FANNG accommodates both honest and dishonest majority settings, ensuring active security.

Table 1: Summary of PPML frameworks (only a representative subset) in the literature. Security: \circ - semi-honest, \ominus - malicious (abort), \odot - malicious (fair), \bullet - malicious (GOD). Notations: HE - Homomorphic Encryption, GC - Garbled Circuits, SS - Secret Sharing, OT - Oblivious Transfer, FSS - Function Secret Sharing.

Setting	Framework	Security	Techniques
2-party (dishonest majority)	CryptoNets [GDL ⁺ 16]	\circ	HE
	SecureML [MZ17]	\circ	HE+GC+SS
	MiniONN [LJLA17]	\circ	HE+GC+SS
	DeepSecure [RRK18]	\circ	GC
	Gazelle [JVC18]	\circ	HE+GC+SS
	XONN [RSC ⁺ 19]	\odot	GC
	CryptFlow2 [RRK ⁺ 20]	\circ	HE+OT+SS
	Delphi [MLS ⁺ 20]	\circ	HE+GC+SS
	ABY2.0 [PSSY21]	\circ	OT+GC+SS
	Cheetah [HjLHD22]	\circ	HE+OT+SS
2-party (+ helper)	Chameleon [RWT ⁺ 18]	\circ	OT+GC+SS
	Crypten [KVH ⁺ 21]	\circ	SS
	LLAMA [GKCG22]	\circ	FSS
3-party (honest majority)	ABY ³ [MR18]	\odot	GC+SS
	ASTRA [CCPS19]	\odot	SS
	SecureNN [WGC19]	\circ	SS
	SWIFT [KPPS21]	\bullet	SS
	Falcon [WTB ⁺ 21]	\odot	SS
4-party (honest majority)	Trident [CRS20]	\odot	GC+SS
	FLASH [BCPS20]	\bullet	SS
	SWIFT [KPPS21]	\bullet	SS
	Fantastic Four [DEK21]	\bullet	SS
	Tetrad [KPRS22]	\bullet	GC+SS
n-party	Cerebro [ZDC ⁺ 21]	\odot	HE+GC+SS
	MPClan [KPPS23]	\odot	SS

While FANNG serves as a general-purpose MPC framework supporting multiple protocols, its design is primarily focused on facilitating efficient PPML inference with fast online execution. In this approach, input-independent tasks, such as generating random correlated data, are performed by the MPC parties in an earlier phase called offline. These precomputed values are then used during the online phase, once the actual inputs become available. Fig. 1 depicts a scenario where the model owner and client act as data providers, demonstrating how FANNG securely executes ML inference using MPC servers based on the selected configuration. To the best of our knowledge, FANNG is the first framework exclusively supporting PPML inference in a dishonest majority setting with active security.

FANNG extends the functionality of SCALE-MAMBA by separating pre-processing from the online phase and introducing a dealer model [DHNO19, ST19]. The preprocessing model enables the generation of offline elements prior to online execution, facilitated by an I/O API connecting the MPC engine to a file system or a database. This feature facilitates the storage and retrieval of garbled circuits for comparisons, random masks for probabilistic truncation, Beaver and matrix/convolutional triples for efficient multiplications, and daBits for conversions. FANNG features an extended instruction set to accommodate the dealer model, along with a separate engine for Beaver and matrix triples generation based on BGV scheme [BGV14, CKR⁺20, MG23], executable by external dealers. The triples generated by the dealers can then be transformed into an MPC version using a *converter* (§3.2). Moreover, FANNG includes specialized implementations to enhance the performance of

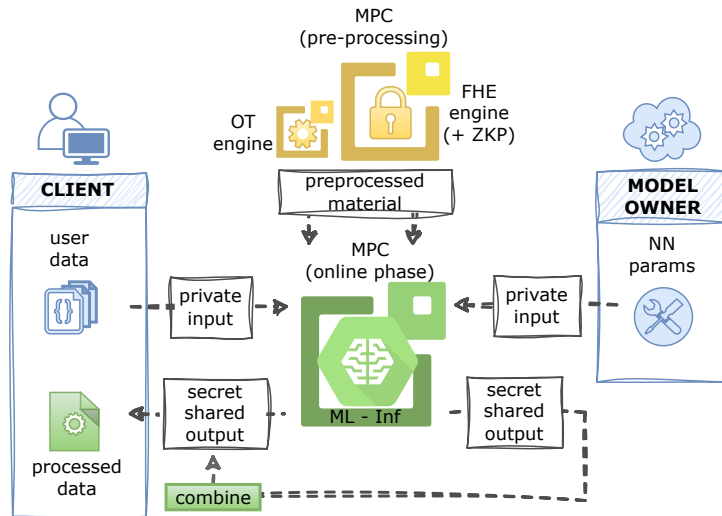


Figure 1: PPML inference between a client in possession of private data and a cloud server owning the model using the FANNG framework.

neural network (NN) operations. This encompasses convolutions, fully connected layers, and dedicated libraries for folding and normalization, ensuring optimal communication rounds.

Unlike a mere prototype, FANNG delves deeper into system-level details, identifies challenges, and addresses them using novel protocols and considerable engineering effort. Following the guidelines from [Eur14], the quality of our code corresponds to Technology Readiness Level 5 (TRL-5), and we aim to achieve TRL-7 in the future. This approach positions FANNG closer to a real-world implementation, facilitating more accurate performance results. For example, SCALE-MAMBA faces limitations in compiling large programs with the default optimizer for communication rounds (the `-O3` flag). Similarly, the MAMBA compiler within SCALE-MAMBA cannot process programs exceeding 2^{32} bytecodes, posing challenges for implementing large neural networks like VGG16 [SZ15], as considered in this work. FANNG addresses these issues through the application of novel engineering techniques.

To summarize, we have the following contributions:

1. Introduction of a novel MPC-based framework designed for private machine learning (ML) inference in a dishonest majority setting with active security.
2. Design and implementation of a dedicated *Dealer* Module, utilizing state-of-the-art FHE-based protocols to pre-process matrix triples, and the associated *Converter* Module.
3. Support for a dedicated *Preprocessing Unit* capable of handling various types of input-independent data, including support for offline garbling, and incorporating efficient storage mechanisms for persistent sharing of values.
4. Novel protocol for combining ReLU with truncations, resulting in improved instruction size and efficiency.
5. Specialized ML libraries for designed for efficient linear transformations and an enriched set of instructions.
6. Open-sourced implementation¹ with detailed evaluations, including private inference of large NNs like VGG16 [SZ15].

¹Available at <https://github.com/Crypto-TII/FANNG-MPC>.

2 Framework Design

In this section, we discuss the design of our FANNG-MPC framework. Our framework prioritizes PPML inference, emphasizing *active* security, although it supports any arbitrary MPC computation. Instead of merely designing a prototype, our focus extends to intricate system-level aspects such as instruction size, storage requirements, support for vectorization, and I/O handling (cf. §3.7). This approach allows us to emulate an MPC framework that closely resembles a real-world implementation. Our objective is to demonstrate the practicality of actively-secure MPC in a dishonest majority setting, a scenario still considered costly in the existing literature. Inherited from SCALE-MAMBA, FANNG operates over prime-order fields (\mathbb{F}_p) that encompass binary (\mathbb{F}_2), with plans for future work to include support for rings.

We chose SCALE-MAMBA as our baseline among existing MPC frameworks because it supports various protocols with *active security*, including full-threshold Secret Sharing (SS) [BCS20, KPR18, RST⁺21], and honest majority setups via Shamir SS, replicated SS, and generic monotone span programs [KRSW18, SW19]. We would like to stress that SCALE-MAMBA is the only framework in existence that allows us to combine those protocols with Boolean circuit evaluation via [HSS17] [HSS17] whilst providing active security. However, despite these features, SCALE-MAMBA falls short when it comes to enabling PPML inference due to several limitations:

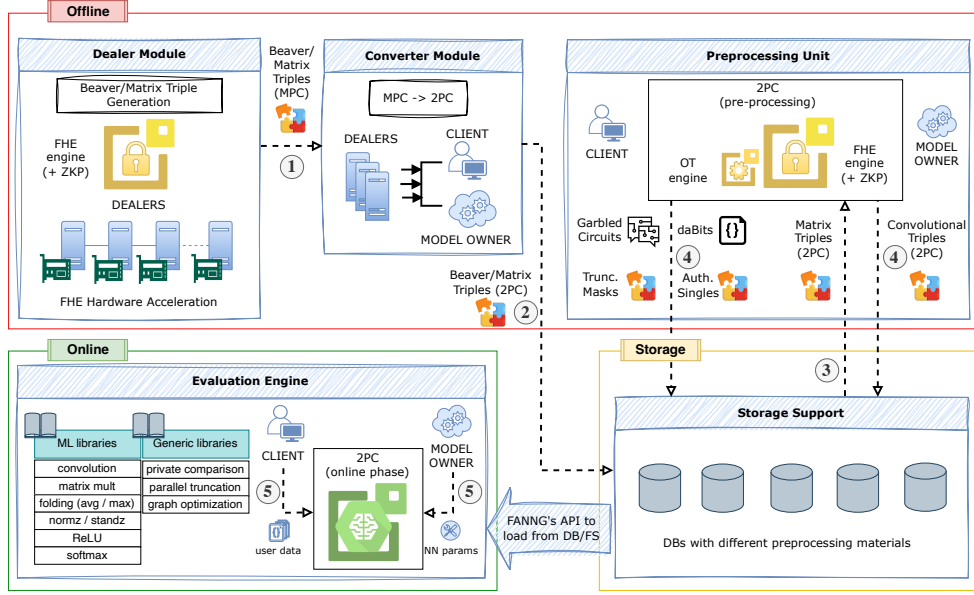
- The inability to compile large programs under the default optimizer (`-O3` flag) [ACC⁺21].
- MAMBA compiler incapability for programs beyond 2^{32} bytecodes, affecting large neural networks like VGG16 [SZ15].
- Absence of a decoupled *pre-processing* phase, which happens in parallel with the online execution.
- Limited I/O capabilities to console access by default. File storage support necessitates manual changes in the base code, indicating a lack of Controller-based I/O for storage and retrieval from files or databases.
- Inclusion of a restricted set of libraries for scientific operations, lacking specialized ML libraries and low-level instructions for ML-related operations.

Moreover, SCALE-MAMBA is no longer maintained with the last update in June 2022². Our primary objective is to enhance the functionality of this widely used MPC framework, ensuring compatibility with real-world setups and better suited for concrete machine learning applications, essentially revitalizing the framework.

Fig. 2 depicts the comprehensive design of FANNG, comprising primarily of five components. A brief overview of these components, in the context of our ML inference application (cf. Fig. 1), is provided next, with detailed descriptions provided in §3.

1. **Dealer Module (§3.1):** In MPC, outsourcing the pre-processing-phase computations (input-independent) to a trusted helper, a.k.a *commodity-based MPC* [DHNO19, ST19], can significantly enhance the efficiency of the online phase [RWT⁺18, KVH⁺21, GKCG22, KKP⁺24]. However, the existence of such a helper contradicts the main axiom of actively secure setting with dishonest majority, i.e. “the parties do not need to trust anyone”, and thus may not align with real deployments. To circumvent this, we incorporated a dealer module which replaces the trusted helper with a group of untrusted dealers. These dealers engage in an MPC protocol to generate correlated randomness, reducing the trust requirement from a single trusted helper to one party selected from many. Both the client and model owner can make the dealer selection, thus mitigating the trust concern.

²<https://nigelsmart.github.io/SCALE/>



① Dealers generate Beaver/Matrix Triples. ② Dealer-generated materials are converted into 2PC shares and stored in the DB. ③ Preprocessing Unit retrieves the Matrix Triples from the DB. ④ Preprocessed materials are stored into the DB. ⑤ Client submits an input query, and the Model Owner provides the NN model parameters for evaluation.

Figure 2: System design of FANNG-MPC framework for PPML inference.

The dealer module currently generates Beaver and matrix triples for fully connected and convolutional layers, addressing complex operations in ML model pre-processing. The design is flexible enough to accommodate multiple dealer modules, reflecting our anticipation of dedicated services replacing these modules in the future. For example, we are actively working on FPGA hardware acceleration to support matrix triple generation. This adaptable design enables the integration of various dedicated services for efficient pre-processing within FANNG.

2. **Converter Unit (§3.2):** The dealer module is designed to be adaptable in various scenarios involving different numbers of dealers and MPC servers. Hence, we need to convert the pre-processed material into a state that is compatible with this setup. The converter unit is specifically created to ease this transition, converting pre-processed materials from one type to another that is suitable for online evaluation involving a different set of parties. In this work, we focus on dealers in the dishonest majority setting and the subsequent conversions needed for online evaluation in two and three-party scenarios.
3. **Pre-processing Unit (§3.3):** This unit is responsible for generating input-independent data essential for the online evaluation of the MPC protocol. This approach has proven to significantly enhance the efficiency of the online phase, leading to practical runtimes [FLNW17, KPR18, Kel20, HHNZ19]. The unit currently covers Garbled Circuits (GCs), truncation masks, daBits, and authenticated singles. In the FANNG roadmap, we aim to extend dealer support for generating all these data types, except for authenticated singles and oblivious transfers.
4. **Storage Support (§3.4):** A modular component within FANNG, dedicated to persisting information shared among its various units. It revamps the legacy SCALE-MAMBA I/O by introducing a novel controller-based approach, enabling the stor-

age and retrieval of pre-processed material from files or databases, whilst maintaining backwards compatibility. Unlike its predecessor, it offers enhanced customization, allowing users to select the persistence mode via a configuration file. The current version of FANNG provides support for MySQL 8.0 and the File System (FS), with added PostgreSQL support in the Converter Unit, highlighting its adaptability to various database engines.

5. **Evaluation Engine (§3.5):** This engine is responsible for performing the protocol evaluation, private ML inference in our case, serving as the interface between the client and model owner. It orchestrates all other components in FANNG to facilitate the entire evaluation pipeline. This involves signaling dealers in the dealer module for data processing, performing its own preprocessing, utilizing storage support for storing the resulting data, and subsequently retrieving the stored preprocessed data for protocol evaluation by the MPC servers. Finally, it performs the evaluation using private inputs from both parties utilizing the preprocessed data.

3 Framework Details

In this section, we discuss the technical details of our framework’s components. We explore the integration of new functionalities and optimizations built upon SCALE-MAMBA, facilitating private evaluations of large neural networks like ResNet and VGG16. While we have categorized these new functionalities under distinct components, achieving a clear-cut separation is challenging, as many functionalities necessitate the collaborative functioning of multiple components.

Experimental setup: We have provided two different hardware test beds:

1. Our *General Purpose MPC* testbed includes 5 servers connected via Gigabit connections (1 Gbps) with a ping latency of $0.15ms$. Each server has 512 GB of RAM and Intel(R) Xeon(R) Silver 4208 @ 2.10GHz processors. They all share a common /home directory and are installed with /sbin/tc to simulate latency.
2. Our *Machine Learning* testbed comprises a single server equipped with 2TB of RAM and an Intel(R) Xeon(R) Gold 6250L CPU @ 3.90GHz processor.

Regarding communications, we conduct all our experimentation considering three relevant setups:

1. **Local:** All parties are emulated on the same machine with no communication cost. Albeit unrealistic, it is typically used as a baseline as it helps to evaluate the computation costs.
2. **Ping:** Parties run on different machines at point-to-point connection speed of $\approx 0.3ms$, suitable for highly dedicated setups such as in Triples-as-a-Service (TaaS) [ST19] with Dealers sharing the same data center [WPM⁺20, ZIC⁺21, RCF⁺21, PHJ⁺22].
3. **WAN:** Parties run on different machines, considering achievable common ping times for cloud service providers, which is $\approx 20ms$.

From an MPC perspective, we explore three settings, each with malicious security:

1. **Full Threshold (2p):** Two-party setting in the dishonest majority setting, tolerating at most 1 corruption.
2. **Full Threshold (3p):** Three-party setting in the dishonest majority setting, tolerating at most 2 corruptions.
3. **Shamir (3p):** Three-party setting in the honest majority setting, tolerating at most 1 corruption.

3.1 Dealer Module

In FANNG, the dealer module generates random Beaver, matrix and convolutional triples required for supporting multiplication, matrix multiplication and convolutions in PPML inference. We opted for Fully Homomorphic Encryption (FHE) [MSM⁺22] techniques, inspired by SCALE-MAMBA’s use of the FHE-based SPDZ protocol to generate traditional Beaver triples in the offline phase. Specifically, we adopt a levelled version of the BGV scheme [BGV14] to instantiate the FHE. Towards this, we implemented three state-of-the-art protocols, each serving a distinct purpose.

- Π_{Prep} : This protocol depends on leveled homomorphic encryption and serves as a crucial component for generating SPDZ-like matrix triples. We implemented the protocol by Chen et al. [CKR⁺20, Fig. 1], which avoids sacrificing a triple at the expense of an additional multiplicative depth.
- *FHE-based matrix multiplication*: To generate matrix triples instead of traditional Beaver triples, Π_{Prep} relies on an algorithm for homomorphically multiplying encrypted matrices. Towards this, we implemented the optimized protocol proposed by Mono and Güneysu [MG23], which allows the reuse of key-switching keys and eliminates constant multiplications.
- Π_{ZKPoK} : To achieve active security, Π_{Prep} uses zero-knowledge proof of knowledge (ZKPoK) techniques, which establish a global proof of knowledge for a set of ciphertexts. We implemented the approach proposed by Baum et al. [BCS20, Fig. 1], where the authors designed an n -prover protocol, providing the capability to prove the validity of the sum of n ciphertexts instead of proving each one individually.

FHE possesses an inherent characteristic where the error employed in encryption, essential for upholding FHE security properties, increases with each homomorphic operation. Notably, this growth becomes exponential when homomorphic multiplications are executed (see [MML⁺23] for a detailed analysis of error growth across operations in BGV). As a result, only a limited number of homomorphic operations can be carried out before the error level starts impacting the decryption operation. Hence, it is essential to evaluate error growth within the circuit and choose the FHE scheme parameters accordingly.

This section provides a detailed analysis of parameter estimation, focusing on homomorphic error, particularly within the most complex circuit across all our protocols—matrix multiplication depicted in [MG23, Fig. 6]. Although Mono and Güneysu [MG23] offer parameter estimation for BGV schemes, we observe that their protocol overlooks ZKPoK (Π_{ZKPoK}). Consequently, their analysis of matrix triple, starting with a fresh ciphertext, leads to an underestimation of the correct parameters and hence proves insufficient for our scenario.

We begin our analysis with a brief recap of all the necessary fundamental concepts. A detailed description of the BGV scheme and the related mathematical techniques is provided in §A.

3.1.1 4-Levelled BGV scheme

Let q be an integer product of 4 primes, i.e., $q = p_0 \cdot p_1 \cdot p_2 \cdot p_3$, with q_ℓ denoting $q_\ell = \prod_{j=0}^{\ell} p_j$, for any $\ell \leq 3$.

A BGV ciphertext is a vector of polynomials $\mathbf{c} \in \mathcal{R}_q^2$, where $\mathcal{R}_q = \mathbb{Z}_q[x]/\langle x^n + 1 \rangle$ and n is a power of two. At a high level, the BGV scheme relies on two key techniques:

- *Modulus-switching*: This procedure is essential for error reduction. Specifically, in our context, we switch from a ciphertext modulus q_ℓ to $q_{\ell-1}$, thereby reducing the error by $q_{\ell-1}/q_\ell = 1/p_\ell$ and introducing an error bounded by B_{scale} (see §A.2.3 for details).

- *Key-switching*: This technique is used to reduce the degree of a ciphertext polynomial after multiplication or change the key following a rotation. Various variants of the key-switching procedure exist. In our scenario, it initially increases the ciphertext modulus to $q_\ell \cdot P$ (where certain computations are performed) and subsequently applies modulus switching back to q_ℓ (see §A.2.4 for details).

For two ciphertexts \mathbf{c} and \mathbf{c}' with error bounds denoted by $\mathbf{B}_\mathbf{c}$ and $\mathbf{B}_{\mathbf{c}'}$, respectively, the error bound after their addition is $\mathbf{B}_\mathbf{c} + \mathbf{B}_{\mathbf{c}'}$, after their multiplication is $\mathbf{B}_\mathbf{c} \cdot \mathbf{B}_{\mathbf{c}'}$, and after a multiplication of \mathbf{c} by a scalar is $\mathbf{B}_{\text{const}} \cdot \mathbf{B}_\mathbf{c}$, where $\mathbf{B}_{\text{const}} \approx t\sqrt{n/12}$ [MML⁺23] (see §A.2 for details).

3.1.2 Noise analysis of matrix triple generation

To compute a valid and secure set of BGV parameters for the multiplication of matrices A and B , we analyze each stage of the Π_{Prep} protocol, focusing on the operations outlined by Chen et al. [CKR⁺20, Fig. 1]. Note that we instantiate matrix multiplication in [CKR⁺20] using the protocol from [MG23]. For consistency, we adopt the same notation as Chen et al. [CKR⁺20].

Top Modulus p_3 In Π_{Prep} protocol, the error of $\mathbf{c}_{A_{jk}}$ (and $\mathbf{c}_{B_{jk}}$) is bounded by $\mathbf{B}_{\text{clean}}^{\text{dishonest}}$ (see §A.2.6 for details). Before performing the matrix multiplication of matrices $\mathbf{c}_{A_{jk}}$ and $\mathbf{c}_{B_{jk}}$, we reduce the error down to a threshold \mathbf{B} , using the modulus switching procedure. Thus, the error is scaled by p_3 and the additional *modulus-switching error* $\mathbf{B}_{\text{scale}}$ (Equation 5) is added. Hence, we have

$$\frac{\mathbf{B}_{\text{clean}}^{\text{dishonest}}}{p_3} + \mathbf{B}_{\text{scale}} < \mathbf{B} \implies p_3 > \frac{\mathbf{B}_{\text{clean}}^{\text{dishonest}}}{\mathbf{B} - \mathbf{B}_{\text{scale}}}.$$

Middle Modulus p_2 During the FHE-based matrix multiplication, parties compute

$$\mathbf{c}_A \circledast \mathbf{c}_B = \sum_{\kappa=0}^{d-1} (\phi^\kappa(\mathbf{c}_A) \boxtimes \psi^\kappa(\mathbf{c}_B)),$$

where \boxtimes denotes homomorphic multiplication of two ciphertexts. As detailed in [MG23, Fig. 6], the computation of $\phi^\kappa(\mathbf{c}_A)$ involves addition of two *special* ciphertexts. These special ciphertexts are constructed by first rotating \mathbf{c}_A by κ positions, followed by multiplying the result of any rotation by a scalar, where $1 \leq \kappa \leq d$. Also, rotation operation involves key-switching procedure as well. On the other hand, $\psi^\kappa(\mathbf{c}_B)$ is obtained via simple rotation by κ positions.

While the rotations themselves do not influence the noise directly, switching the key back to the original adds key switching noise v_{ks} , which depends on the key switching method [MML⁺23]. In our case, we have $v_{\text{ks}} = \frac{\sqrt{3\omega \cdot \max(\tilde{q}_j)}}{P} \cdot \mathbf{B}_{\text{ks}} + \sqrt{k} \cdot \mathbf{B}_{\text{scale}}$ (cf. §A.2.4). Thus, if \mathbf{B} is the starting noise of \mathbf{c}_A and \mathbf{c}_B , then after the τ rotations, the bounded error of each ciphertext grows from \mathbf{B} to $\mathbf{B} + \tau \cdot v_{\text{ks}}$.

The next step for $\phi^\kappa(\mathbf{c}_A)$ involves a ciphertext-scalar multiplication, increasing the error to $\mathbf{B}_{\text{const}} \cdot (\mathbf{B} + \tau \cdot v_{\text{ks}})$. Similarly, the subsequent addition with a similar ciphertext doubles the error. A modulus switching is performed next to reduce the noise magnitude down to \mathbf{B} . Hence, the noise at this stage (from $\phi^\kappa(\mathbf{c}_A)$) is at most

$$\frac{2\mathbf{B}_{\text{const}} \cdot (\mathbf{B} + \tau \cdot (\frac{\sqrt{3\omega \cdot \max(\tilde{q}_j)}}{P} \cdot \mathbf{B}_{\text{ks}} + \sqrt{k} \cdot \mathbf{B}_{\text{scale}}))}{p_2} + \mathbf{B}_{\text{scale}}.$$

Since we want p_2 to be as small as possible, we use a larger \mathbf{B} and P such that,

$$p_2 > 2\mathbf{B}_{\text{const}} \cdot (\mathbf{B} + \tau \cdot \sqrt{k} \cdot \mathbf{B}_{\text{scale}}) / (\mathbf{B} - \mathbf{B}_{\text{scale}}).$$

Thus, we set $B \approx \alpha \cdot B_{\text{scale}}$, for $\alpha \geq 2$, and we have $p_2 \approx 2B_{\text{const}} \cdot (\alpha + \tau \cdot \sqrt{k})/(\alpha - 1)$.

Middle Modulus p_1 Note that the noise magnitude grows from B to $d \cdot B^2$ while computing $\sum_{\kappa=0}^{d-1} (\phi^\kappa(\mathbf{c}_A) \boxtimes \psi^\kappa(\mathbf{c}_B))$. Given that the product of two ciphertexts results in a 3-dimensional vector, key-switching is necessary. Additionally, to reduce noise to B , we require the modulus switching procedure. In FANNG, we employ hybrid key switching that allows for merging key switching with the modulus switching, enabling a direct switch to a smaller modulus, i.e., from $Q_1 = P \cdot q_1$ to q_0 decreasing the noise by $q_0/Q_1 = 1/(P \cdot p_1)$. In our case, the error before scaling down using the modulus switching is $\nu' = d \cdot B^2 \cdot P + \sqrt{2\omega} \cdot \max(\tilde{q}_j) \cdot B_{\text{ks}}$. Thus, the error after the modulus switching procedure is bounded by $\frac{1}{P \cdot p_1} \cdot \nu' + \sqrt{k+1} \cdot B_{\text{scale}}$ [MML⁺23, Sec. 3.2]. Since we want to reduce the noise back to B , we set

$$\frac{d \cdot B^2}{p_1} + \frac{\sqrt{2\omega} \cdot \max(\tilde{q}_j)}{P \cdot p_1} \cdot B_{\text{ks}} + \sqrt{k+1} \cdot B_{\text{scale}} < B. \quad (1)$$

For large values of P , Equation 1 becomes $\frac{d \cdot B^2}{p_1} + \sqrt{k+1} \cdot B_{\text{scale}} < B$, which (in the variable B) must have a positive discriminant. Namely $p_1 > 4d \cdot \sqrt{k+1} \cdot B_{\text{scale}}$, and thus, we can set $P \approx 10 \cdot \sqrt{3\omega} \cdot \max(\tilde{q}_j) \cdot (B_{\text{ks}}/B_{\text{scale}})$.

Bottom Modulus p_0 Note that modulus reduction is not required at level zero (cf. [MG23, Fig. 6]). However, since we execute `AddMacs(c)`, the error grows from B to B^2 . To ensure a correct distributed decryption, we require that the noise bounded by B^2 is smaller than $2 \cdot (1 + N \cdot 2^{\text{DD}_{\text{sec}}}) \cdot B^2$ (cf. Equation (7)). Thus, $p_0 > 2 \cdot (1 + N \cdot 2^{\text{DD}_{\text{sec}}}) \cdot (\alpha \cdot B_{\text{scale}})^2$.

All the parameters together Since we require q to be as small as possible, we set $\alpha = 2$. Moreover, we have only 2 parties in our case ($N = 2$) and $\tau \approx d$. Thus, we have:

$$\begin{aligned} p_0 &\approx 2^{\text{DD}_{\text{sec}}+4} \cdot B_{\text{scale}}^2, & p_1 &\approx 4 \cdot d \cdot \sqrt{k} \cdot B_{\text{scale}}, \\ p_2 &\approx 2 \cdot d \cdot \sqrt{k} \cdot B_{\text{const}}, & p_3 &\approx B_{\text{clean}}^{\text{dishonest}}/B_{\text{scale}}. \end{aligned}$$

Finally, setting our parameters as $t = 2^{128}$, $d = 64$, $w = 3$, $k = 5$ and $\text{DD}_{\text{sec}} = \text{ZK}_{\text{sec}} = 80$, we have $\log q \approx 760$ and security level $\lambda = 128$.

3.2 Converter Unit

This section describes FANNG's mechanism for transferring the data produced by a set of $N_{\mathcal{D}}$ dealers, denoted by the set \mathcal{D} (typically with $N_{\mathcal{D}} = |\mathcal{D}| > 2$) to the $N_{\mathcal{S}}$ MPC servers running the PPML inference. For ease of presentation, consider client \mathcal{C} and the model owner \mathcal{M} to be the MPC servers, i.e. $N_{\mathcal{S}} = 2$. When necessary to differentiate between various clients in a set \mathcal{C} with size $N_{\mathcal{C}}$, we denote them as $\{\mathcal{C}_j\}_{j \in [N_{\mathcal{C}}]}$. Also, κ denotes the computational security parameter.

General re-sharing strategy Consider a value $x \in \mathbb{F}_p$ additively secret-shared among the dealers $\{D_i\}_{i \in [N_{\mathcal{D}}]}$, i.e. $x = \sum_{i \in [N_{\mathcal{D}}]} x_i$ with D_i holding x_i . They aim to redistribute this value to a fresh sharing among the parties $\{\mathcal{M}, \mathcal{C}_j\}$. The naive method for achieving this involves each dealer D_i generating a 2-out-of-2 sharing of x_i and sending one share to each of \mathcal{M} and \mathcal{C}_j . This approach incurs a cost of $N_{\mathcal{D}} \cdot 2 \cdot p$ bits of communication, where p is the bit length of x . When $p > \kappa$, we can optimize the cost using a pseudo-random function $F : \{0, 1\}^\kappa \rightarrow \{0, 1\}^p$, as follows:

1. For $i \in [N_{\mathcal{D}}]$, dealer D_i does as follows:

- (a) Sample random keys $k_{M,C_j}^{(i)} \leftarrow \{0,1\}^\kappa$, for $j \in [N_C]$, where N_C is the bound on the number of clients that the model owner M expects.
- (b) Send $x_{M,C_j}^{(i)} = x_i - F(k_{M,C_j}^{(i)})$ to M , who defines its share of x as $x_{M,C_j} = \sum_{i \in [N_D]} x_{M,C_j}^{(i)}$.
- (c) When the specific identity of the client C_j becomes known, send the key $k_{M,C_j}^{(i)}$ to C_j .

2. C_j defines its share of x as $x_{C_j,M} = \sum_{i \in [N_D]} F(k_{M,C_j}^{(i)})$.

Note that the output shares satisfy the equation $x = x_{C_j,M} + x_{M,C_j}$. The approach above is described for only one value for simplicity, but the same key can be used for all the values re-shared to the same client, resulting in communication of $N_D \cdot (p + \kappa)$ bits.

The above strategy generalizes easily for N_S MPC servers, with $N_D \cdot p + N_D \cdot (N_S - 1) \cdot \kappa$ bits of communication for the optimized approach.

Re-sharing SPDZ values Currently in FANNG, dealers operate in a dishonest majority setting, using SPDZs Topgear [BCS20] for computation. In this context, every value x is associated with a message authentication code (MAC), denoted as $\text{MAC}(x)$. There exists a global MAC key, denoted as Δ and obviously generated by the dealers, which authenticates messages as $\text{MAC}(x) = x \cdot \Delta$. To simplify, the process of re-sharing a value from dealers to $\{M, C_j\}$ in FANNG involves re-sharing a triple $(x, \text{MAC}(x), \Delta)$. To achieve this, dealers employ the PRF-based approach outlined earlier for each of the three values. This approach is illustrated in Fig. 3 below.

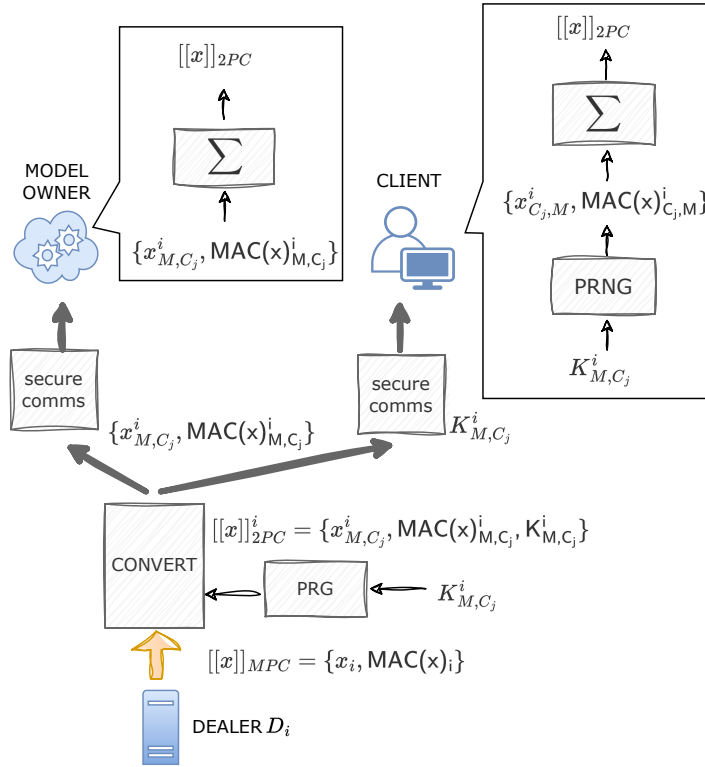


Figure 3: MPC-2PC Converter for SPDZ values in FANNG.

We remark that this approach allows a corrupted dealer in \mathcal{D} to introduce an *additive error* in the re-shared value by sending inconsistent values. Nevertheless, since we use additive secret-sharing, this kind of error could always be introduced by a corrupted party during reconstruction. Our use of information-theoretic MACs during the online phase effectively catches such errors and induces an abort, irrespective of how they were generated.

MAC key strategy Note that the dealers will use different MAC keys based on the trust relations among different clients. Specifically, if a client C_1 lacks trust in another client C_2 to potentially collude with the model owner M , C_1 cannot use the same MAC keys as C_2 . Hence, the dealers must take this into account when generating the pre-processing material. This is due to the fact that $\{M, C_2\}$ possess a 2-out-of-2 sharing of the MAC key, which will get compromised upon collusion. In the extreme case where no client trusts any other client to be non-colluding with M , the dealers should use a different MAC key for each $\{M, C_j\}$ pair.

Experimentation and Discussion To evaluate the performance of the converter unit, we utilized preprocessed data generated by three Dealers for PPML inference using the LeNet [LBBH98] architecture (see §4). The reported values were obtained from our *Machine Learning* testbed. For a single LeNet inference, converting the preprocessed shares to 2PC sharing semantics takes 13.4 seconds over a local network. However, this time is amortized across multiple instances, dropping to 69 seconds for 10 inferences and 595 seconds for 100 inferences. The converter relies on industry-standard technologies such as gRPC for communication and PostgreSQL for storage, with the primary overheads stemming from the limitations of these technologies.

3.3 Pre-processing Unit

This unit is responsible for performing all input-independent pre-processing tasks that are not carried out by the dealers. We dedicate this section mainly to provide details regarding moving the garbling operation in SCALE-MAMBA to the pre-processing stage, which significantly enhanced the online performance of the framework. We begin with the details of offline garbling.

3.3.1 Offline Garbling

For the distributed computation of garbling circuits (GCs) among n parties, SCALE-MAMBA incorporates the garbling schemes proposed by Hazay et al. [HSS17] and Wang et al. [WRK17]. Furthermore, it utilizes techniques from Zaphod [AOR⁺19] to facilitate interaction between Arithmetic (Full Threshold) and Boolean Circuits (Distributed GCs) through field conversion. Notably, the entire garbling procedure in SCALE-MAMBA occurs online, signifying that circuit garbling takes place just before the evaluation of the garbled circuit, in parallel with the online execution. This way, SCALE-MAMBA guarantees reactivity, enabling parties to decide on the protocol’s progression based on intermediate and public values.

The computational model based on the disassociation/independence of the offline and online phases is difficult to materialize in general-purpose frameworks like SCALE-MAMBA and MP-SPDZ, especially if there is no previous knowledge of the function. Moreover, there are several technical aspects that limit the adaptation of offline garbling in SCALE-MAMBA, especially when considering active security:

1. From a cryptographic perspective, the inputs from the online phase require masking with authenticated bits (aBits) generated via a chain of different Oblivious Transfer (OT) protocols that start from a set of *choicebits* selected by each party (cf. [KOS15] for details). The circuits themselves depend on similar processes to generate the keys

that are embedded in each circuit. In that sense, both need to come from the same set of choicebits.

2. From an engineering point of view, current versions of `SCALE-MAMBA` are not built to handle circuits offline. More specifically, there is no trivial way to parameterize the type and quantity of circuits needed online, as we can do with the triples (via the `-max` flag). Additionally, for architectural reasons, there are no mechanisms on `SCALE-MAMBA` that would allow us to trivially interact with them.

To decouple the garbling procedure from the online phase, `FANNG` introduces a set of new instructions capable of Garbling/Storage, Loading and Executing Gabled Circuits for the dishonest majority setting:

- `OGC(circuit_id, amount)`: Garbles a given `amount` of the specified circuit and stores the result to a database/file system.
- `LOADGC(circuit_id, amount)`: Loads specific `amount` of persisted `circuit_id` instances to memory.
- `EGC(circuit_id)`: Executes the next specific circuit on the pile, using the same interaction method as with `GC`, the instruction for garbling in `SCALE-MAMBA`.

To garble offline, the parties need to know the type and number of GCs they require beforehand. They can produce, store and load these circuits with the instructions above. `FANNG` also includes architectural components to support Database/File System connectivity for uploading a specified number of circuits of a given type. It is worth highlighting that GCs are processed offline using a specific set of *choicebits*. We then verify that the choicebits held by each client/online party for the circuits match with the ones being used to generate authenticated bits for the masks. In the case of PPML inference, GCs are generally used for private comparisons required in activation and pooling layers, but `FANNG` can handle any type of circuit.

Garbling Offline Data Flow Though `FANNG` treats the `SCALE-MAMBA`'s garbling functionality as a black box, moving the garbling process to offline requires manipulating the process necessary to create the keys embedded in the circuits. For instance, if a protocol requires `aBits`, `daBits`, and GCs altogether (for instance, in Aly et al. [ANSS22]), then the keys corresponding to all these materials should be generated from the same set of choicebits, as implemented in `SCALE-MAMBA`. Furthermore, in `SCALE-MAMBA`, a fresh set of choicebits is selected at every run. `FANNG` makes it possible to parameterize the choicebits, as illustrated in Fig. 4. This slight change in the flow above (replacing the fresh selection of choicebits), implies the persistence of the choicebits used to garble the circuits and the way they are consumed by the framework.

Note that `FANNG` currently supports offline garbling only by computing parties (MPC servers), not dealers. Additionally, for larger circuits, circuit sizes can be notably high (e.g. $\approx 10\text{MB}$ per party for float multiplication). As a result, the execution times of instructions like `LOADGC` heavily depend on the File System reading speed or Database response time. Therefore, we decided to design our DB support in a modular fashion, decoupled from the old I/O support in `SCALE-MAMBA`. New features introduced by `FANNG` for databases and file systems are specifically integrated into our modular data connectivity support.

Experimentation and Discussion We conducted experiments in this section using our *General Purpose MPC* testbed. We evaluated performance in a two-party scenario using the basic less-than-or-equal (`LTZEQ`) circuit [MRVW21], with a size of $\approx 800\text{KB}$, running in batches of up to 1000 circuits. As shown in Tab. 2, we obtain improvements for online runtime in the range $2.4\times\text{--}4.2\times$.

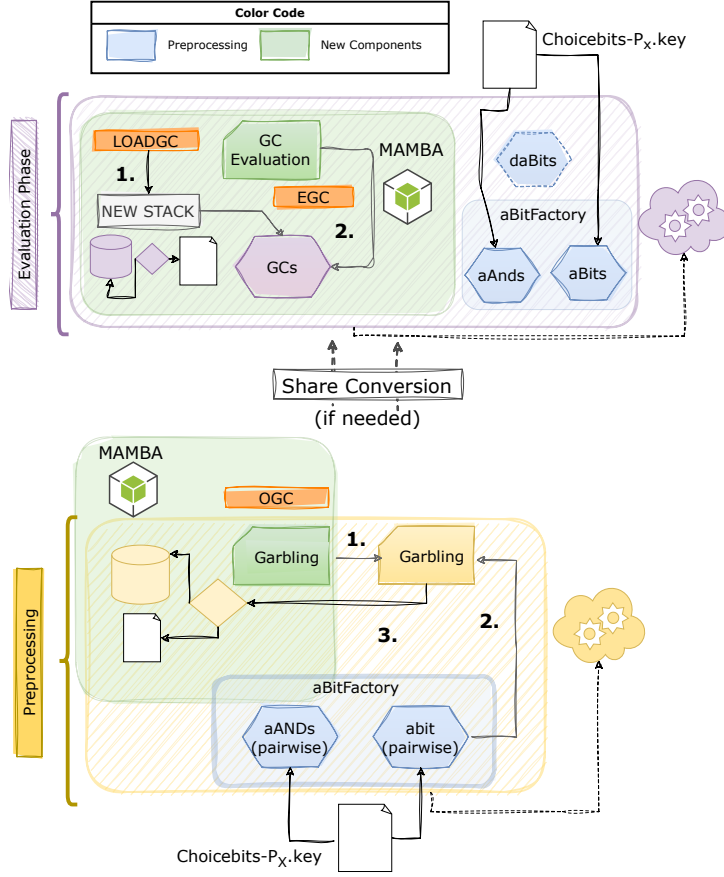


Figure 4: Offline Garbling in FANNG.

With respect to Loading and Storage times, while they can be amortized, they ultimately depend on the underlying user’s DB engine. We use a MySQL 8.0 Vanilla Dockerized installation. The timings in both cases are linked to the speed at which we can process two classic SQL commands, namely INSERT and SELECT. Advanced use cases may involve modern DB setups, such as in-memory DBs and other Big Data processing engines.

3.4 Storage Support

I/O is crucial for any modern application requiring persistence. Furthermore, real-life implementations demand flexibility in handling secret-shared/public inputs. While SCALE-MAMBA initially provided I/O capabilities, they were limited to console access, necessitating manual

Table 2: Offline Garbling Timings (seconds) per 1000 operations.

Stage	Offline Garbling			Online Garbling		
	Local	Ping	WAN	Local	Ping	WAN
Offline						
Garbling	2.91	4.7	86	==	==	==
Storage	17.86	17.86	17.86	==	==	==
Loading	14.88	14.88	14.88	==	==	==
Online						
Execution	0.97	2.1	62.6	3.89	6.83	148.66

code intervention for file system support [ACC+21]. FANNG addresses this requirement in a modular and adaptable way, via a novel Controller-Based IO. Furthermore, we introduced backward-compatible extensions to the existing I/O, incorporating *under the hood* flexible and maintainable components. This ensures current SCALE-MAMBA users can continue relying on the framework’s I/O, while providing future users with the option to leverage our novel storage support.

Legacy I/O SCALE-MAMBA I/O consists of two sections: i) instructions for private and public I/O invoked via MAMBA code, and ii) interfaces implementing I/O functionality directly in SCALE, with these interfaces being programmatically exchanged in the code. FANNG retains the instructions but broadens the set of interfaces, facilitating connections to file systems (FS) or databases (DBs) for storage (e.g., pre-processed materials generated by the framework) and loading during the online phase.

```
Storage_type = FileSystem
File_system_storage_directory = Data
```

Listing 1: I/O Configuration file tuned for File System access.

Controller-Based I/O FANNG adds a parametrizable I/O controller, customizable through a configuration file. Users can configure their instances to utilize any available I/O mediums via the controller. Currently, the framework supports two options: general FS and MySQL (1&2). This eliminates the need for users to recompile the framework for I/O changes. All database connections, including those used by the revamped Legacy I/O, rely on the controller. This centralized configuration enables users to manage the database settings in a single text file, regardless of the chosen I/O model. The Controller-Based I/O is employed across all new functionalities dedicated to producing or retrieving the pre-processed material.

```
Storage_type = MySQLDatabase
MySQL_url = tcp://my.dbserver.ae:3306
MySQL_user = my_user
MySQL_password = ****
MySQL_database = mpplib-database-dealer
```

Listing 2: I/O Configuration file tuned for MySQL access.

Currently, FANNG supports only MySQL 8.0. However, it is designed with flexibility, allowing seamless extension of both legacy and new I/O functionalities to other DB engines like SQLite and PostgreSQL.

3.5 Evaluation Engine

This section outlines FANNG’s contributions to efficiently evaluate online phase of protocols. FANNG introduces a novel way of integrating private comparison protocol with a state-of-the-art probabilistic truncation protocol to reduce communication rounds. Furthermore, FANNG introduces several libraries for implementing ML blocks, facilitating convolutional, fully connected and folding layers, among others.

3.5.1 Protocols for private comparisons

Activation functions are fundamental in NN models, with the Rectified Linear Unit (ReLU) being one of the most popular. Essentially, ReLU can be implemented through comparison and multiplication. FANNG integrates practical privacy-preserving comparisons, leveraging

contributions from the current SotA [ANSS22]. This involves applying mixed circuits to constructions introduced by Catrina and De Hoogh [CdH10], and Rabbit [MRVW21]. Our library includes an interface (`rabbit_sint`), returning the following for any $\langle x \rangle$, the secret-shares of $x \in \mathbb{Z}_{2^k}$ over prime-order field \mathbb{F}_p :

$$\text{LTZ}(\langle x \rangle) : \mathbb{Z}_p \rightarrow \{0, 1\} = \begin{cases} 1 & \text{if } x < 0, \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

The interface invokes constructions implemented as described in [ANSS22], performing boolean operations via Zaphod [AOR⁺19]. It can be parameterized with the following evaluation options:

- `rabbit_slack`: Integrates Zaphod [AOR⁺19] with the logic from Rabbit [MRVW21], providing statistical security for accelerated computing.
- `rabbit_list`: Implements the original Rabbit [MRVW21] with a rejection list.
- `rabbit_fp`: Similar to `rabbit_list` but assumes a bounded prime-order domain close to a power of 2.
- `rabbit_conv`: Utilizes the conversion circuits from SCALE-MAMBA to transform $\langle x \rangle$ on \mathbb{F}_p to \mathbb{Z}_{2^k} .
- `rabbit_less_than`: Takes 2 modulo \mathbb{Z}_{2^k} inputs and directly evaluates the binary circuit from Rabbit [MRVW21] using logical gates from SCALE-MAMBA.
- `dabits_ltz`: Instantiates the main contribution from [ANSS22], combining the original Catrina and de Hoogh [CdH10] construction with Boolean evaluation from Zaphod using `daBits` for random bit sampling.

The library is integrated into several NN modules within the FANNG compiler, including `relu_lib.py`, which implements several variations of ReLU’s. According to [ANSS22], `dabits_ltz` remains the fastest operation mode in the library, particularly when garbling is conducted offline, a capability supported by FANNG. Our results for private comparisons are presented in Tab. 3.

Table 3: Private comparisons using `rabbit_lib` in a 2 party setting. Time is measured in seconds per 1000 operations (*Online Garbling; †Offline Garbling).

Mode	Local	Ping	WAN
<code>rabbit_slack</code>	6.41	13.46	295.76
<code>rabbit_list</code>	120	457	33,365
<code>rabbit_fp</code>	64	138	18,315
<code>rabbit_conv</code>	12.8	19	300
<code>rabbit_less_than</code>	11.81	60.50	5,225
Default Mode			
<code>dabits_ltz</code> *	3.89	8.43	148.66
<code>dabits_ltz</code> †	0.97	2.07	62.52

The library also includes two VHDL versions of the underlying Boolean circuits, detailed in [ANSS22]. The difference lies in the presence or absence of XOR gates, with the former incurring an increase in the number of gates. While the former allows us to benefit from typical garbled circuit optimizations, the latter provides more parallelism and can be advantageous when employing replicated secret sharing for Boolean evaluation.

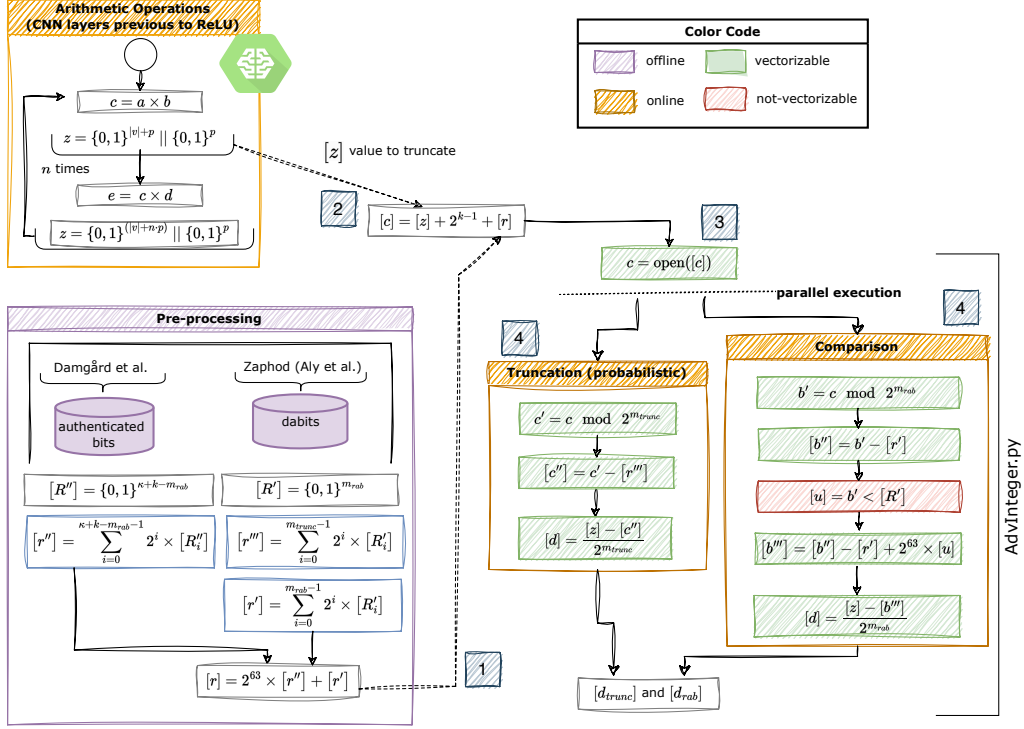


Figure 5: Protocol for batching secure truncation in parallel with private comparison.

3.5.2 Combining ReLU's with n Truncations

FANNG supports Fixed Point Arithmetic (FPA) in the same way as SCALE-MAMBA and MP-SPDZ, specifically implementing the approach outlined in [CS10]. The challenge of truncation in fixed point arithmetic is a widespread issue in ML applications (see for instance [AB24]). This section elaborates on the combination of a predetermined number of truncation operations with a comparison operation.

In FPA, for a fixed point value x represented with mantissa α_x and precision d , we have $x = \alpha_x \cdot 2^d$. Then, the multiplication of two fixed-point values x and y produces:

$$x \cdot y = \alpha_x \cdot \alpha_y \cdot 2^{2 \cdot d}. \quad (3)$$

To prevent the scaling factor from quickly saturating the domain space, it is necessary to truncate it by a factor of 2^d after each multiplication. The probabilistic truncation mechanism proposed in [CS10] is commonly used in the literature. It involves a single round of communication and utilizes pre-processed authenticated material, hereafter referred to as *truncation masks*.

In recent works (e.g., [RRK⁺20]), it was noted that ReLU and truncation can be effectively combined in a single call. The truncation protocol bears some similarities with the comparison protocol in [CdH10] and its extensions in [ANSS22]. Both protocols involve random sampling in the pre-processing phase and necessitate a much smaller word space compared to the domain size. Leveraging these observations, we present a protocol that concurrently performs truncation and the comparison protocol from [ANSS22]. Notably, this truncation addresses scaling issues arising from multiple multiplications. This enables the batching of truncations from multiple multiplications in a single execution, allowing parallel processing with private comparisons in the activation layers.

Fig. 5 illustrates parallel execution of comparison and truncation of value c resulted after n sequential multiplications, with k as the domain size and κ as the security param-

ter [ANSS22]. Pre-processing involves using contributions from [DFK⁺06] for authenticated bits (aBits) and Zaphod [AOR⁺19] for daBits. The main idea is that random sampling for truncation and comparison occurs in the same domain and over the same input. Therefore, we can use masked outputs from the comparison directly for truncation. The mask is utilized once for executing both protocols. Following this, we multiply the truncated value with the output bit from the comparison, similar to a traditional ReLU implementation. Notably, the generation of the 2^k secret shared bounded randomness [DFK⁺06] takes place during pre-processing in FANNG. On the other hand, SCALE-MAMBA does not separate this process from the online phase, resulting in a slower truncation process.

Concerning the number of batched truncation operations, the size of mantissa (v) plus $n \cdot d$ (representing the quantity of performed multiplications) is restricted by the value of k . In practice, FANNG and SCALE-MAMBA assume a word size of 64 bits. This upper bound be attributed to the limitations of the GC processor, which was designed to support 64-bit words. This limitation can be viewed as a trade-off between precision and the size of n or batched truncations.

Experimentation and Discussion: Consider a CNN setup where the network involves multiplication in the convolution layer, followed by another in the batch normalization layer. With quantization during normalization, if no truncation is applied, the precision expands to $3 \cdot d$ bits. In our experiments, we fixed d at 20 and restricted the represented values to not exceed 8 (3 bits), ensuring compatibility with 64-bit words in our experimentation.

Table 4: Performance of comparison in conjunction with truncation (`trunc_ltz`) in seconds per 1000 operations. ([†]Vectorized).

Mode	Optimized			Non-Optimized		
	Local	Ping	WAN	Local	Ping	WAN
Full Threshold (2p)						
[CdH10] + 1-T	1.61	2.71	114	4.66	18.1	793
[CdH10] + 2-T	1.68	3.15	135	4.69	19	814
<code>trunc_ltz(1-T)</code>	3.35	3.3	170	==	==	==
<code>trunc_ltz(2-T)</code>	4.3	3.37	190	==	==	==
<code>trunc_ltz(1-T)</code> [†]	0.84	2.1	62	==	==	==
<code>trunc_ltz(2-T)</code> [†]	0.86	2.2	62.7	==	==	==
Full Threshold (3p)						
[CdH10] + 1-T	2.69	4.46	116	8.4	23.1	798
[CdH10] + 2-T	3.09	5.09	136	8.56	23.2	820
<code>trunc_ltz(1-T)</code>	5.3	3.67	265	==	==	==
<code>trunc_ltz(2-T)</code>	8.6	3.77	276	==	==	==
<code>trunc_ltz(1-T)</code> [†]	1.34	2.44	103	==	==	==
<code>trunc_ltz(2-T)</code> [†]	1.35	2.9	103.8	==	==	==
Shamir (3p)						
[CdH10] + 1-T	1.37	3.93	115	5.52	17.02	792
[CdH10] + 2-T	1.61	4.48	136	6.02	17.07	814
<code>trunc_ltz(1-T)</code>	1.27	4.89	339	==	==	==
<code>trunc_ltz(2-T)</code>	1.5	5.13	349	==	==	==
<code>trunc_ltz(1-T)</code> [†]	1.23	4.1	306	==	==	==
<code>trunc_ltz(2-T)</code> [†]	1.25	4.2	308	==	==	==

To establish a comparable baseline, we provide equivalent running times and the number of instructions generated using the classical fixed-point truncation [CS10] and comparison [CdH10] protocols. Given that the use of circuit optimizers results in a substantial increase in the number of instructions for the baseline, we also refrain from employing them in our evaluations. In our setup, we utilized our *General Purpose MPC* testbed and evaluated results across Local, Ping, and WAN setups (cf. §3). We distinguish

between two types of executions: with and without vectorized inputs. Tab. 4 showcases the performance of the `trunc_ltz` operation, representing the evaluation of a comparison operation along with either one (1-T) or two (2-T) truncations.

We observe that vectorization is crucial for optimizing TCP-based communications, as large data structures can be segmented into TCP packets, maximizing the allowed packet size. This explains the performance degradation in the non-vectorized case when ping time increases. Additionally, as pointed out by Aly et al. [ANSS22], the network size somewhat constrains the effectiveness of circuit optimizers. Therefore, any practical scenario utilizing SCALE-MAMBA has to depend on non-optimized circuits, leading to limited performance as indicated by the times in **red**.

FANNG addresses this issue by providing a vectorized version of `trunc_ltz`, and the best timings are annotated in **green**. The results demonstrate that vectorization significantly improves both latency and throughput. For instance, our protocols can perform 1000 comparisons $13\times$ faster in the two-party full threshold (FT) setting over a WAN setup. In this context, we have also quantified the number of instructions generated by the compiler in Tab. 5.

Table 5: Number of instructions per ReLU Mode ([†]Vectorized).

Mode	Instructions	
	1-T	2-T
[CdH10] + [CS10]	2,638,000	3,028,000
<code>trunc_ltz</code>	480,000	482,000
<code>trunc_ltz</code> [†]	76,000	76,000

Compared to SCALE-MAMBA, we have significantly reduced the total number of instructions for compilation. Besides vectorization, this reduction is attributed to a series of optimizations implemented for generating aBits, batched truncation, and our `dabits_ltz` protocol. This reduction benefits RAM usage and compilation time, especially in average/large neural networks.

There is a performance gap between the evaluation of `trunc_ltz`, when executed on Shamir or FT. This difference arises from our reliance on offline garbling for FT, which is not possible on Shamir due to the absence of garbling in the evaluation of the Boolean circuit [AOR⁺19]. Lastly, transitioning from 1 to 2 batched truncations has a minimal impact on performance across all setups, a significant improvement considering the typical impact of truncation execution on quantized networks [KS22].

3.5.3 ML libraries

The framework includes specialized libraries to perform efficient linear transformations, specifically in convolutional and fully connected layers. FANNG also encompasses libraries for folding (average and max pooling), normalization and standardization, and output layers (e.g., softmax). The main contribution of these new libraries lies in their integration of FANNG’s novel functionalities, such as pre-processing of materials, and private comparison protocols. Moreover, these libraries are implemented with optimal communication rounds, eliminating the need for an optimizer during compilation.

3.6 Handling pre-processed materials

FANNG introduces a set of new instructions designed to handle the generation and loading of pre-processed materials for use in the online phase. In the following, we discuss details of the most relevant ones.

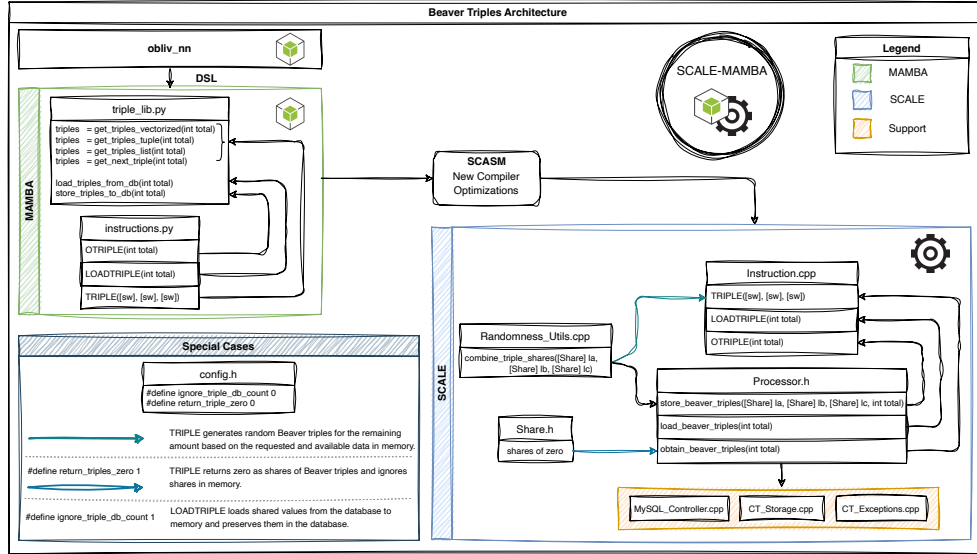


Figure 6: Architecture for random Beaver triples generation in FANNG framework.

Beaver Triples To enable preprocessing of Beaver triples, we provide two new instructions:

- **OTRIPLE(amount)**: Generates any specified `amount` of random Beaver triples using the existing triples factory and stores them in the database.
- **LOADTRIPLE(amount)**: Loads a specified amount of random Beaver triples from database into memory.

To encapsulate the instructions for preprocessing above including **TRIPLE**, FANNG provides a convenient library, `triples_lib.py`, encapsulating instructions, providing better usability and organization. The architecture related to this modification is depicted in Fig. 6, illustrating the process flow from the library to the relevant processes in `Processor.h`. While understanding the intricate details may not be necessary for end users, they can conveniently utilize these functionalities through the library as follows (Listing 3):

```
from Compiler.triples_lib import
    store_triples_to_db,
    load_triples_from_db,
    get_next_triple

store_triples_to_db(100)
load_triples_from_db(100)
next_triple = get_next_triple(100)
```

Listing 3: Preprocessing 100 random Beaver triples in FANNG.

Matrix Triples To enable the MPC servers to utilize the specialized convolutional and matrix triples [MG23, CKR⁺20], generated by the dealers, during the online phase, we need a mechanism to load them into memory once persisted. FANNG introduces two new instructions for this purpose.

- **LOADCT(type_id, amount)**: Loads any `amount` of Matrix Triple of specified `type_id` into memory.
- **CT_DYN(A,B,C, type_id)**: Extracts a Matrix Triple of specified `type_id` from memory and assigns it to the vectorized registers `A`, `B`, `C`.

daBits In order to use the daBits generated in the pre-processing, we introduce two new instructions:

- ODABIT(*amount*): Persists any specified *amount* of daBits using the pre-existing daBits factory.
- LOADDABIT(*amount*): Loads a specified *amount* of daBits into memory. Note that the *choicebits* used for generating the loaded daBits must be the same as those used for the Garbled Circuits and for FANNG execution.

We also modified existing DABIT instruction to now retrieve the next available daBits from memory. This *vectorizable* instruction can return multiple daBits if necessary. In cases where none are available, FANNG will resort to the daBits factory, a principle applicable to all instructions in FANNG that consume pre-process material.

Bounded Randomness Protocols, such as comparisons, are based on statistical security, and rely on masking bounded by some power of 2, say 2^k . SCALE-MAMBA generates masks programmatically, and for instance, a 40-bit mask necessitates 148 instruction lines per invocation in the bytecode. In applications such as machine learning, a substantial number of these masks are required, significantly affecting the amount of RAM needed for program compilation and loading into memory. To address this, we introduce three new instructions that not only transition the process entirely from the online phase but also reduce the number of instructions per invocation to one, independent of the size of the bound.

- OSRAND(*k*, *amount*): Generate and persist any specified *amount* of random masks bounded by 2^k using our DB Support.
- LOADSRAND(*k*, *amount*): Load into memory a specified *amount* of random masks bounded by 2^k .
- SRAND(*output*, *k*): Extract a random mask bounded by 2^k from memory and assign it to the register *amount*. In SCALE-MAMBA language, this vectorizable instruction can fill multiple registers in a single call.

Test Modes for Pre-processed Material: FANNG focuses on a Dealer/pre-processing model, offering users practical implementations of both. However, for development and simulation purposes, users may prefer to simulate these models. Following SCALE-MAMBA practices, we offer various testing modes in the `config.h` file. For example, the test mode for bounded randomness, daBits and GCs can be activated or deactivated as shown in Listing 4.

```

/* Ignores shares in memory for SRAND
   when set to 1. Used for testing to
   avoid consuming shares in DB. */
#define return_shares_zero 1

/* Ignores shares in memory for dabits.
   When set to 1, returns only 0's.
   */
#define return_dabit_zero 1

/* Ignores share counters for LOADCT,
 * LOADSRAND, and LOADGC when set to 1.
 */
#define ignore_share_db_count 1

```

Listing 4: Test Mode configurations in FANNG.

3.7 General Optimizations

In addition to the mentioned features, FANNG prioritizes usability by minimizing compiler-generated instructions, introducing novel functionality, and enhancing the online phase inherited from SCALE-MAMBA. We explore some of the most relevant items below:

- **Summation Instructions SUMS and SUMC:** Machine operations relying on matrices often involve constant summations. While local users can perform these operations, SCALE-MAMBA’s implementation involves instructions that can be significant for a large set of values $\{x_i, \dots, x_n\}$. This can lead to slower compilation times and increased RAM consumption. To mitigate this, we introduced two novel instructions, SUMS and SUMC, providing outputs of $\sum_{i=1}^n x_i$ for private and public inputs, respectively. These instructions simplify development and reduce the entropy of the instruction file, accelerating development time.
- **Communication Bottlenecks:** We observed a limitation in SCALE-MAMBA related to the number of elements it can send per message, specifically through the instructions opening \mathbb{F}_p inputs (START_OPEN and START_CLOSE). For instance, our experiments showed that it can only support up to 100 thousand elements per invocation, constraining the parallelization of instructions per round. To overcome this limitation, FANNG allows for a configurable number of SSL connections. When a user intends to open multiple elements in a single round, FANNG can navigate around this constraint by utilizing as many connections as configured or provided by the user.
- **Graph Theory Library:** We have extended the functionality of the framework beyond machine learning and towards generic MPC, incorporating a novel Graph Theory library. Currently, this library includes SotA methods in shortest path [AC22], with plans for additional expansion in the future.

4 Evaluation

This section provides details regarding PPML inference using our FANNG framework. We evaluate three different neural network (NN) architectures to accommodate varying complexities (see §B for additional architectural details).

- LeNet [LBBH98]: 5-layer network with 60K parameters, over MNIST dataset.
- A generic CNN: 8-layer network with 1.5 million parameters, over CIFAR10 dataset.
- VGG16 [SZ15]: 16-layer network with 37 million parameters, over CIFAR10 dataset.

Regarding model utility, we compare the output of each layer with a non-private PyTorch implementation and ensure a difference of no more than 6 bits (in the fractional part) in every layer. For the last 2-classifier layers when using CIFAR, we accept differences of up to 8 bits. We run the experiments with fractional part set to 20 bits.

The goal of FANNG is to decouple the preprocessing phase from the online phase. In this approach, random material for multiple instances is generated in a single batch (to take advantage of cost amortization) and stored in persistent memory. This setup is more representative of real-world deployments, so we chose to benchmark only the online phase. Timings for offline operations are provided in earlier sections. FANNG is, to our knowledge, the first framework to fully support PPML inference in the dishonest-majority setting with active security. For fairness, comparisons with protocols offering weaker security guarantees are not included. Our source code is publicly available at <https://github.com/Crypto-TII/FANNG-MPC>.

Tab. 6 provides our results for PPML inference using our *Machine Learning* testbed evaluated across Local, Ping, and WAN setups (cf. §3). The local setup allows us to evaluate the computational complexity of our protocols. In contrast, the WAN setup is better suited for analyzing performance over high-latency networks, such as the Internet. As compute demands have increased, organizations have shifted to using datacenters, where mutually distrusting parties can securely co-locate their servers in the same facility, connected by high-speed networks [WPM⁺20, ZIC⁺21, RCF⁺21, PHJ⁺22]. Our Ping setup models this environment by simulating a low-latency, high-speed network between MPC servers [WGC19, DEK21, HJLHD22]. In these tests, pre-processing was disabled using FANNG’s test mode support (cf. §3.6). This included simulating the generation of matrix and convolutional triples, GCs, and truncation masks by activating the relevant testing flags. Timings for generating other pre-processing elements, such as Beaver triples, daBits, and singles, were not included in the results. However, the timings reflect the entire online phase, including the time required to load the preprocessed materials from storage into memory.

Table 6: Timings for private ML inference using FANNG. Values are reported in seconds.

-	LeNet	CIFAR10 CNN	VGG16
Full-Threshold (2p)			
Local	10	109	534
Ping	20	516	1,159
WAN	410	12,175	18,074
Shamir (3p)			
Local	11	459	979
Ping	48	1,490	2,561
WAN	1,996	60,384	87,632
Full-Threshold (2p) - Without Activation			
Local	0.62	43	440
Ping	0.62	44	472
WAN	0.86	44	483

Although we achieve impressive runtimes for both Local and Ping setups, we note that the runtimes increase significantly over a WAN when transitioning from LeNet to the more complex VGG16. This is attributed to the absence of support for parallelizing Boolean circuit evaluations in SCALE-MAMBA affecting activation layers. Despite FANNG incorporating support for vectorization at various stages in ReLU computation, it relies on the Boolean circuit evaluation inherited as a black box from SCALE-MAMBA for activation functions. To be more specific, the sequential evaluation of garbled circuits in the case of full threshold and boolean circuits for Shamir contributes to this slowdown.

We plan to address this limitation in the future. However, to offer insight into its impact on our performance results, we have also included the evaluation results in Tab. 6 for the two-party full threshold setting after excluding the activation layers. In this scenario, we observe that the time complexity for Ping and WAN setups is comparable to that of a Local setup. This is reasonable, as the communication rounds for the three networks without activation layers are only 22 (LeNet), 44 (CNN for CIFAR10), and 577 (VGG16).

The performance gap between the full-threshold setting and Shamir is attributed to the absence of parallelization in the activation layer. In SCALE-MAMBA, Shamir utilizes replicated secret sharing over \mathbb{Z}_2 for Boolean circuit evaluation, a process inherited by FANNG. In contrast, full-threshold employs [HSS17] [HSS17] for Boolean circuits in both SCALE-MAMBA and FANNG, with the difference that FANNG can push the garbling phase to pre-processing as described in §3.3.1. The private comparison circuit over \mathbb{Z}_2 incurs 16 communication rounds, whereas the online phase in [HSS17] requires only 2 rounds.

These communication rounds are sequential per ReLU, making full-threshold significantly faster than Shamir for activation layers.

The similar performance of two networks trained on CIFAR10, despite a significant difference in size, can be attributed also to the dominance of activation layers. Specifically, the CIFAR10 CNN has 1.5 million parameters and 196,640 ReLUs, whereas the larger VGG16 has 37 million parameters but only 285,672 ReLUs.

Analytical Evaluation: Due to the absence of support for the parallelization of comparison circuits in FANNG, the current runtime obtained may not accurately represent the achievable performance. Rather, they constitute an overestimation, resulting in timings that are much higher than the potential achievable values, particularly for communication-dominant setups such as WAN. To address this limitation, we conduct a theoretical analysis by focusing on two key parameters: i) *Batch size*, representing the number of comparison circuits that FANNG could parallelize in its anticipated capability, and ii) *Overhead Factor*, indicating the increase in runtime when handling a batch of circuits in parallel compared to a single execution. The results are plotted in Fig. 7.

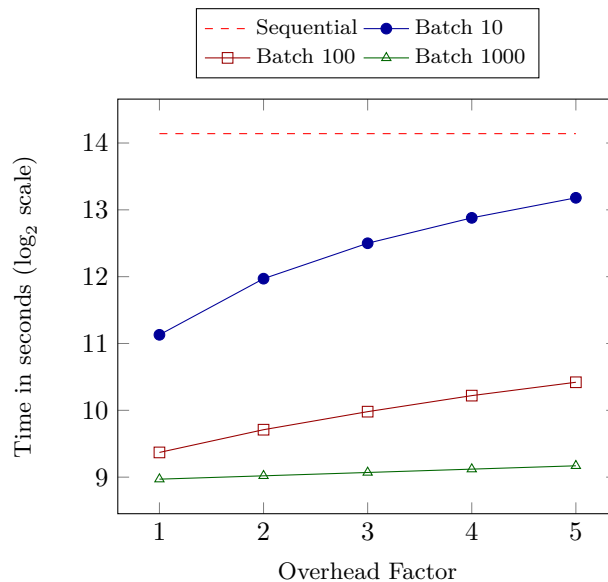


Figure 7: Analytical evaluation of VGG16 inference using FANNG over a WAN setup with batch sizes (Batch) $\in \{10, 100, 1000\}$ and Overhead Factor $\in \{1, 2, 3, 4, 5\}$. ‘Sequential’ denotes the baseline evaluation without parallelization.

As shown in Fig. 7, parallelizing the comparisons within activation layers can substantially enhance the runtime. For example, when the batch size is set to 100 and the overhead factor is set to 4, FANNG will require less than 20 minutes to execute a private inference on the VGG16 network over a WAN setup. This time can be further reduced to less than 10 minutes by using a higher batch size of 1000.

Summary The results highlight FANNG’s remarkable capability to achieve private inference for MNIST within seconds and for CIFAR10 within minutes, considering both Local and Ping times. This performance can be considered as state-of-the-art for Full-Threshold settings, as, to the best of the authors’ knowledge, no other implementation in this setting has provided better metrics. The potential for achieving comparable performance in WAN setups is evident by parallelizing activation layers, a goal set for future work. Furthermore, the evaluation, excluding ReLUs, reveals the potential to classify MNIST in under a second

and CIFAR10 in under a minute, emphasizing the need for ongoing efforts to enhance the efficiency of activation layers.

5 Conclusion & Future Work

In this work, we present FANNG-MPC, an MPC framework developed with a focus on private machine learning (ML) inference. FANNG extends the capabilities of the well-established SCALE-MAMBA framework, which supports various actively secure MPC protocols over fields. After identifying limitations in SCALE-MAMBA concerning ML inference, we introduce several innovations. These include dealer support for preprocessing and storage support to streamline the preprocessing phase. Our contributions, both in theory and engineering, are substantiated through a comprehensive evaluation that closely simulates real-world execution rather than a simple prototype. The results demonstrate the practicality of private ML inference within the actively secure setting in MPC with a dishonest majority.

In our future work, FANNG will evolve to provide a more capable pre-processing engine for private ML inference. While FANNG currently features a comprehensive instruction set for storing and loading all pre-processing materials, this functionality has not yet been implemented in the dealers. The current version of FANNG supports only matrix triple generation within the dealers, leaving the generation and storage of the remaining offline elements to be handled directly by the client and model owner.

In upcoming iterations, FANNG is set to extend its dealer support to include garbling, along with the generation of truncation masks, Beaver triples, and authenticated singles. Additionally, FANNG dealers will benefit from hardware acceleration through FPGA support. Moreover, as identified in our evaluations, the lack of parallelization for comparisons in SCALE-MAMBA significantly affects overall performance, and this issue will be addressed in the next iterations of FANNG. These enhancements will collectively fortify FANNG's capabilities in facilitating various aspects of privacy-preserving machine learning.

References

- [AB24] Wei Ao and Vishnu Naresh Boddeti. AutoFHE: Automated Adaption of CNNs for Efficient Evaluation over FHE. In *USENIX Security Symposium (USENIX Security)*, 2024.
- [AC22] Abdelrahman Aly and Sara Cleemput. A Fast, Practical and Simple Shortest Path Protocol for Multiparty Computation. In *European Symposium on Research in Computer Security (ESORICS)*, 2022.
- [ACC⁺21] Abdelrahman Aly, Karl Cong, Daniele Cozzo, Marcel Keller, Emmanuela Orsini, Dragos Rotaru, Oliver Scherer, Peter Scholl, Nigel P Smart, Titouan Tanguy, and Tim Wood. SCALE-MAMBA v1. 14: Documentation. *Documentation. pdf*, 2021. <https://homes.esat.kuleuven.be/~nsmart/SCALE/Documentation.pdf>.
- [ADPS16] Erdem Alkim, Léo Ducas, Thomas Pöppelmann, and Peter Schwabe. Post-quantum Key Exchange - A New Hope. In *USENIX Security Symposium (USENIX Security)*, 2016.
- [ANSS22] Abdelrahman Aly, Kashif Nawaz, Eugenio Salazar, and Victor Sucasas. Through the Looking-Glass: Benchmarking Secure Multi-party Computation Comparisons for ReLU 's. In *Cryptology and Network Security (CANS)*, 2022.

- [AOR⁺19] Abdelrahman Aly, Emmanuela Orsini, Dragos Rotaru, Nigel P. Smart, and Tim Wood. Zaphod: Efficiently Combining LSSS and Garbled Circuits in SCALE. In *Workshop on Encrypted Computing & Applied Homomorphic Cryptography (WAHC@CCS)*, 2019.
- [BCPS20] Megha Byali, Harsh Chaudhari, Arpita Patra, and Ajith Suresh. FLASH: Fast and Robust Framework for Privacy-preserving Machine Learning. *Proceedings of Privacy Enhancing Technologies (PoPETs)*, 2020.
- [BCS20] Carsten Baum, Daniele Cozzo, and Nigel P Smart. Using TopGear in Overdrive: A More Efficient ZKPoK for SPDZ. In *Selected Areas in Cryptography (SAC)*, 2020.
- [BGV14] Zvika Brakerski, Craig Gentry, and Vinod Vaikuntanathan. (Leveled) Fully Homomorphic Encryption without Bootstrapping. *ACM Transactions on Computer Theory*, 2014.
- [BMCM23] Beatrice Biasioli, Chiara Marcolla, Marco Calderini, and Johannes Mono. Improving and Automating BFV Parameters Selection: An Average-Case Approach. *Cryptology ePrint Archive*, 2023.
- [BV11] Zvika Brakerski and Vinod Vaikuntanathan. Fully Homomorphic Encryption from Ring-LWE and Security for Key Dependent Messages. In *Annual International Cryptology Conference (CRYPTO)*, 2011.
- [CCPS19] Harsh Chaudhari, Ashish Choudhury, Arpita Patra, and Ajith Suresh. AS-TRA: High Throughput 3PC over Rings with Application to Secure Prediction. In *ACM SIGSAC Conference on Cloud Computing Security Workshop (CCSW@CCS)*, 2019.
- [CdH10] Octavian Catrina and Sebastiaan de Hoogh. Improved Primitives for Secure Multiparty Integer Computation. In *Security and Cryptography for Networks (SCN)*, 2010.
- [CKR⁺20] Hao Chen, Miran Kim, Ilya Razenshteyn, Dragos Rotaru, Yongsoo Song, and Sameer Wagh. Maliciously Secure Matrix Multiplication with Applications to Private Deep Learning. In *International Conference on the Theory and Application of Cryptology and Information Security (ASIACRYPT)*, 2020.
- [CP21] José Cabrero-Holgueras and Sergio Pastrana. SoK: Privacy-Preserving Computation Techniques for Deep Learning. *Proceedings of Privacy Enhancing Technologies (PoPETs)*, 2021.
- [CRS20] Harsh Chaudhari, Rahul Rachuri, and Ajith Suresh. Trident: Efficient 4PC Framework for Privacy Preserving Machine Learning. In *Annual Network and Distributed System Security Symposium (NDSS)*, 2020.
- [CS10] Octavian Catrina and Amitabh Saxena. Secure Computation with Fixed-Point Numbers. In *Financial Cryptography and Data Security (FC)*, 2010.
- [CS16] Ana Costache and Nigel P Smart. Which Ring Based Somewhat Homomorphic Encryption Scheme is Best? In *The Cryptographers' Track at the RSA Conference (CT-RSA)*, 2016.
- [DEK21] Anders P. K. Dalskov, Daniel Escudero, and Marcel Keller. Fantastic Four: Honest-Majority Four-Party Secure Computation With Malicious Security. In *USENIX Security Symposium (USENIX Security)*, 2021.

- [DFK⁺06] Ivan Damgård, Matthias Fitz, Eike Kiltz, Jesper Buus Nielsen, and Tomas Toft. Unconditionally Secure Constant-Rounds Multi-party Computation for Equality, Comparison, Bits and Exponentiation. In *Theory of Cryptography Conference (TCC)*, 2006.
- [DHNO19] Ivan Damgård, Helene Haagh, Michael Nielsen, and Claudio Orlandi. Commodity-Based 2PC for Arithmetic Circuits. In *IMA International Conference on Cryptography and Coding (IMACC)*, 2019.
- [Eur14] European Association of Research & Technology Organisations. The TRL Scale as a Research & Innovation Policy Tool, EARTO Recommendations. https://www.earto.eu/wp-content/uploads/The_TRL_Scale_as_a_R_I_Policy_Tool_-_EARTO_Recommendations_-_Final.pdf, April 2014. Accessed: 2024-09-27.
- [FLNW17] Jun Furukawa, Yehuda Lindell, Ariel Nof, and Or Weinstein. High-Throughput Secure Three-Party Computation for Malicious Adversaries and an Honest Majority. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT)*, 2017.
- [GDL⁺16] Ran Gilad-Bachrach, Nathan Dowlin, Kim Laine, Kristin E. Lauter, Michael Naehrig, and John Wernsing. CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy. In *International Conference on Machine Learning (ICML)*, 2016.
- [GHS12] Craig Gentry, Shai Halevi, and Nigel P Smart. Homomorphic Evaluation of the AES Circuit. In *Annual International Cryptology Conference (CRYPTO)*, 2012.
- [GKCG22] Kanav Gupta, Deepak Kumaraswamy, Nishanth Chandran, and Divya Gupta. LLAMA: A Low Latency Math Library for Secure Inference. *Proceedings of Privacy Enhancing Technologies (PoPETs)*, 2022.
- [GM23] Andrea Di Giusto and Chiara Marcolla. Breaking the power-of-two barrier: noise estimation for BGV in NTT-friendly rings. *Cryptology ePrint Archive*, 2023.
- [HHNZ19] Marcella Hastings, Brett Hemenway, Daniel Noble, and Steve Zdancewic. SoK: General Purpose Compilers for Secure Multi-Party Computation. In *IEEE Symposium on Security and Privacy (IEEE S&P)*, 2019.
- [HjLHD22] Zhicong Huang, Wen jie Lu, Cheng Hong, and Jiansheng Ding. Cheetah: Lean and Fast Secure Two-Party Deep Neural Network Inference. In *USENIX Security Symposium (USENIX Security)*, 2022.
- [HSS17] Carmit Hazay, Peter Scholl, and Eduardo Soria-Vazquez. Low Cost Constant Round MPC Combining BMR and Oblivious Transfer. In *International Conference on the Theory and Application of Cryptology and Information Security (ASIACRYPT)*, 2017.
- [JVC18] Chiraag Juvekar, Vinod Vaikuntanathan, and Anantha Chandrakasan. GAZELLE: A Low Latency Framework for Secure Neural Network Inference. In *USENIX Security Symposium (USENIX Security)*, 2018.
- [Kel20] Marcel Keller. MP-SPDZ: A Versatile Framework for Multi-Party Computation. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2020.

- [KKP⁺24] Banashri Karmakar, Nishat Koti, Arpita Patra, Sikhar Patranabis, Protik Paul, and Divya Ravi. Asterisk: Super-fast MPC with a Friend. In *IEEE Symposium on Security and Privacy (IEEE S&P)*, 2024.
- [KM24] Jennifer King and Caroline Meinhardt. Rethinking Privacy in the AI Era: Policy Provocations for a Data-Centric World. *Stanford Institute for Human-Centered Artificial Intelligence*, 2024.
- [KOS15] Marcel Keller, Emmanuela Orsini, and Peter Scholl. Actively Secure OT Extension with Optimal Overhead. In *Annual International Cryptology Conference (CRYPTO)*, 2015.
- [KPPS21] Nishat Koti, Mahak Pancholi, Arpita Patra, and Ajith Suresh. SWIFT: Super-fast and Robust Privacy-Preserving Machine Learning. In *USENIX Security Symposium (USENIX Security)*, 2021.
- [KPPS23] Nishat Koti, Shravani Mahesh Patil, Arpita Patra, and Ajith Suresh. MPClan: Protocol Suite for Privacy-Conscious Computations. *Journal of Cryptology*, 2023.
- [KPR18] Marcel Keller, Valerio Pastro, and Dragos Rotaru. Overdrive: Making SPDZ Great Again. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT)*, 2018.
- [KPRS22] Nishat Koti, Arpita Patra, Rahul Rachuri, and Ajith Suresh. Tetrad: Actively Secure 4PC for Secure Training and Inference. In *Annual Network and Distributed System Security Symposium (NDSS)*, 2022.
- [KRSW18] Marcel Keller, Dragos Rotaru, Nigel P. Smart, and Tim Wood. Reducing Communication Channels in MPC. In *Security and Cryptography for Networks (SCN)*, 2018.
- [KS22] Marcel Keller and Ke Sun. Secure Quantized Training for Deep Learning. In *International Conference on Machine Learning (ICML)*, 2022.
- [KT21] Kwangjo Kim and Harry Chandra Tanuwidjaja. *Privacy-Preserving Deep Learning - A Comprehensive Survey*. Springer, 2021.
- [KVH⁺21] Brian Knott, Shobha Venkataraman, Awni Y. Hannun, Shubho Sengupta, Mark Ibrahim, and Laurens van der Maaten. CrypTen: Secure Multi-Party Computation Meets Machine Learning. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2021.
- [LBBH98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.
- [LJLA17] Jian Liu, Mika Juuti, Yao Lu, and N. Asokan. Oblivious Neural Network Predictions via MiniONN Transformations. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2017.
- [LP00] Yehuda Lindell and Benny Pinkas. Privacy Preserving Data Mining. In *Annual International Cryptology Conference (CRYPTO)*, 2000.
- [MG23] Johannes Mono and Tim Güneysu. Implementing and Optimizing Matrix Triples with Homomorphic Encryption. In *Asia Conference on Computer and Communications Security (ASIACCS)*, 2023.

- [MLS⁺20] Pratyush Mishra, Ryan Lehmkuhl, Akshayaram Srinivasan, Wenting Zheng, and Raluca Ada Popa. Delphi: A Cryptographic Inference Service for Neural Networks. In *USENIX Security Symposium (USENIX Security)*, 2020.
- [MML⁺23] Johannes Mono, Chiara Marcolla, Georg Land, Tim Güneysu, and Najwa Aaraj. Finding and Evaluating Parameters for BGV. In *International Conference on Cryptology in Africa (AFRICACRYPT)*, 2023.
- [MR18] Payman Mohassel and Peter Rindal. ABY³: A Mixed Protocol Framework for Machine Learning. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2018.
- [MRVW21] Eleftheria Makri, Dragos Rotaru, Frederik Vercauteren, and Sameer Wagh. Rabbit: Efficient Comparison for Secure Multi-Party Computation. In *Financial Cryptography and Data Security (FC)*, 2021.
- [MSM⁺22] Chiara Marcolla, Victor Sucasas, Marc Manzano, Riccardo Bassoli, Frank HP Fitzek, and Najwa Aaraj. Survey on Fully Homomorphic Encryption, Theory, and Applications. *Proceedings of the IEEE*, 2022.
- [MZ17] Payman Mohassel and Yupeng Zhang. SecureML: A System for Scalable Privacy-Preserving Machine Learning. In *IEEE Symposium on Security and Privacy (IEEE S&P)*, 2017.
- [NC23] Lucien K. L. Ng and Sherman S. M. Chow. SoK: Cryptographic Neural-Network Computation. In *IEEE Symposium on Security and Privacy (IEEE S&P)*, 2023.
- [PHJ⁺22] Rushi Patel, Pouya Haghi, Shweta Jain, Andriy Kot, Venkata Krishnan, Mayank Varia, and Martin C. Herbordt. COPA Use Case: Distributed Secure Joint Computation. In *IEEE International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, 2022.
- [PSSY21] Arpita Patra, Thomas Schneider, Ajith Suresh, and Hossein Yalame. ABY2.0: Improved Mixed-Protocol Secure Two-Party Computation. In *USENIX Security Symposium (USENIX Security)*, 2021.
- [RCF⁺21] Mark Russinovich, Manuel Costa, Cédric Fournet, David Chisnall, Antoine Delignat-Lavaud, Sylvan Clebsch, Kapil Vaswani, and Vikas Bhatia. Toward Confidential Cloud Computing. *Communications of the ACM*, 2021.
- [Res24] Precedence Research. Privacy-enhancing Computation Market Size, Share, Report 2033. <https://www.precedenceresearch.com/privacy-enhancing-computation-market>, 2024. Accessed: 05/04/2024.
- [RRK18] Bitva Darvish Rouhani, M. Sadegh Riazi, and Farinaz Koushanfar. Deepsecure: scalable provably-secure deep learning. In *Annual Design Automation Conference (DAC)*, 2018.
- [RRK⁺20] Deevashwer Rathee, Mayank Rathee, Nishant Kumar, Nishanth Chandran, Divya Gupta, Aseem Rastogi, and Rahul Sharma. CrypTFlow2: Practical 2-Party Secure Inference. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2020.
- [RSC⁺19] M. Sadegh Riazi, Mohammad Samragh, Hao Chen, Kim Laine, Kristin E. Lauter, and Farinaz Koushanfar. XONN: XNOR-based Oblivious Deep Neural Network Inference. In *USENIX Security Symposium (USENIX Security)*, 2019.

- [RST⁺21] Dragos Rotaru, Nigel P. Smart, Titouan Tanguy, Frederik Vercauteren, and Tim Wood. Actively Secure Setup for SPDZ. *Journal of Cryptology*, 2021.
- [RWT⁺18] M. Sadegh Riazi, Christian Weinert, Oleksandr Tkachenko, Ebrahim M. Songhori, Thomas Schneider, and Farinaz Koushanfar. Chameleon: A Hybrid Secure Computation Framework for Machine Learning Applications. In *Asia Conference on Computer and Communications Security (ASIACCS)*, 2018.
- [ST19] Nigel P. Smart and Titouan Tanguy. TaaS: Commodity MPC via Triples-as-a-Service. In *ACM SIGSAC Conference on Cloud Computing Security Workshop (CCSW@CCS)*, 2019.
- [SW19] Nigel P. Smart and Tim Wood. Error Detection in Monotone Span Programs with Application to Communication-Efficient Multi-party Computation. In *The Cryptographers' Track at the RSA Conference (CT-RSA)*, 2019.
- [SZ15] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations (ICLR)*, 2015.
- [WGC19] Sameer Wagh, Divya Gupta, and Nishanth Chandran. SecureNN: 3-Party Secure Computation for Neural Network Training. *Proceedings of Privacy Enhancing Technologies (PoPETs)*, 2019.
- [WPM⁺20] Pierre-Francois Wolfe, Rushi Patel, Robert Munafo, Mayank Varia, and Martin C. Herbordt. Secret Sharing MPC on FPGAs in the Datacenter. In *International Conference on Field-Programmable Logic and Applications (FPL)*, 2020.
- [WRK17] Xiao Wang, Samuel Ranellucci, and Jonathan Katz. Global-Scale Secure Multiparty Computation. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2017.
- [WTB⁺21] Sameer Wagh, Shruti Tople, Fabrice Benhamouda, Eyal Kushilevitz, Prateek Mittal, and Tal Rabin. Falcon: Honest-Majority Maliciously Secure Framework for Private Deep Learning. *Proceedings of Privacy Enhancing Technologies (PoPETs)*, 2021.
- [ZDC⁺21] Wenting Zheng, Ryan Deng, Weikeng Chen, Raluca Ada Popa, Aurojit Panda, and Ion Stoica. Cerebro: A Platform for Multi-Party Cryptographic Collaborative Learning. In *USENIX Security Symposium (USENIX Security)*, 2021.
- [ZIC⁺21] Michael Zink, David E. Irwin, Emmanuel Cecchet, Hakan Saplakoglu, Orran Krieger, Martin C. Herbordt, Michael Daitzman, Peter Desnoyers, Miriam Leiser, and Suranga Handagala. The Open Cloud Testbed (OCT): A Platform for Research into new Cloud Technologies. In *IEEE International Conference on Cloud Networking (CloudNet)*, 2021.

A FHE parameters

In this section, we will provide a detailed description of the BGV scheme and all the mathematical notions necessary to understand the noise analysis provided in §3.1.

A.1 Mathematical background

We denote by \mathcal{R} the polynomial ring $\mathbb{Z}[x]/\langle x^n + 1 \rangle$, where n is a power of two and by $\mathcal{R}_p = \mathbb{Z}_p[x]/\langle x^n + 1 \rangle$, where p is an integer. We indicate by t and q the plaintext and the ciphertext modulus, where $t \equiv 1 \pmod{2n}$ and q a chain of primes, such that $q = q_{L-1}$ and $q_\ell = \prod_{j=0}^{\ell} p_j$ where $p_j \equiv 1 \pmod{2n}$ and $\ell \leq L-1$ [GHS12]. Moreover, if $x \in \mathbb{R}$, we write $[x]_p \in [-p/2, p/2)$ for the centered representative of $x \pmod{p}$.

We denote the RLWE error distribution by χ_e and the secret key distribution by χ_s . If χ is a probabilistic distribution and $a \in \mathcal{R}$ is a random polynomial, we write $a \leftarrow \chi$ when sampling each coefficient independently from χ .

We also define two parameters, ZK_{sec} and DD_{sec} , related to the simulation-based security of the ZKPoK protocol [BCS20]. ZK_{sec} measures the statistical distance between the coefficients of the ring elements (represented as polynomials) in an honest ZKPoK transcript and one generated through simulation. Likewise, DD_{sec} denotes the statistical distance between the distributions of honest and simulated transcripts in the distributed decryption protocol.

A.1.1 Canonical embedding and norms

Let $a \in \mathcal{R}$ be a polynomial. We recall that the *infinity norm* of a is defined as $\|a\|_\infty = \max\{|a_i| : 0 \leq i \leq n-1\}$.

The *canonical embedding* of a is the vector obtained by evaluating a at all primitive $2n$ -th roots of unity. The *canonical embedding norm* of a is defined as

$$\|a\|^{can} = \max_{j \in \mathbb{Z}_m^*} |a(\zeta^j)|,$$

where ζ is a fixed complex primitive $2n$ -th root of unity.

Let us consider a random polynomial $a \in \mathcal{R}$ where each coefficient is sampled independently from one of the following zero-mean distributions:

- $\mathcal{DG}(\sigma^2)$, the discrete Gaussian distribution with standard deviation σ .
- \mathcal{B}_k , the centered binomial distribution of width k .
- \mathcal{U}_q , the uniform distribution over \mathbb{Z}_q .

Then, the random variable $a(\zeta)$ is well approximated by centred Gaussian distribution with variance nV_a , where V_a is the variance of each coefficient in a [GM23]. Moreover, we can bound the canonical norm of a as $\|a\|^{can} \leq D\sqrt{nV_a}$ with probability greater or equal to $1 - ne^{-D^2}$ [BMCM23]. Thus, we set $D = 8$, so the probability of failure is limited to 2^{-76} .

To compute $\|a\|^{can}$, we have to study the variance V_a of each coefficient of a . Specifically,

$$V_a = \begin{cases} \sigma^2 & \text{if } a \leftarrow \mathcal{DG}(\sigma^2) \\ k/2 & \text{if } a \leftarrow \mathcal{B}_k \\ q^2/12 & \text{if } a \leftarrow \mathcal{U}_q \end{cases} \quad (4)$$

In the following χ_e is the discrete Gaussian distribution $\mathcal{DG}(\sigma^2)$, and, as in [ADPS16], we approximate $\mathcal{DG}(\sigma^2)$ with a centered binomial distribution \mathcal{B}_k with $k = 21$. Thus, the variance of each element of the error vector is $V_e = 21/2 = 10.5$ and its standard deviation is 3.24. Moreover, we set $\chi_s = \mathcal{B}_1$, thus $V_s = 0.5$.

A.2 BGV scheme

The three basic algorithms of BGV are as follows:

- KeyGen(λ): Key generation algorithm samples $s \leftarrow \mathcal{B}_1$, $a \leftarrow \mathcal{U}_q$ and $e_{sk} \leftarrow \mathcal{DG}(\sigma^2)$, outputs the secret key \mathbf{sk} and the public key \mathbf{pk} , where

$$\mathbf{sk} = s \text{ and } \mathbf{pk} = (b, a) = [(-a \cdot s + te_{sk}, a)]_{q_\ell}.$$

- Enc $_{\mathbf{pk}}(m)$: Encryption algorithm takes as input a plaintext $m \in \mathcal{R}_t$ and the public key \mathbf{pk} . It samples $u \leftarrow \mathcal{B}_1$, $e_0, e_1 \leftarrow \mathcal{DG}(\sigma^2)$ and outputs the $\mathbf{c} = (\mathbf{c}, \ell, \nu)$, where

$$\mathbf{c} = (c_0, c_1) = [(b \cdot u + te_0 + m, a \cdot u + te_1)]_q,$$

is a ciphertext, ℓ denotes the level and ν the *critical quantity* of \mathbf{c} (see below).

- Dec $_{\mathbf{sk}}(\mathbf{c})$: Decryption algorithm takes as input the secret key \mathbf{sk} and the ciphertext $\mathbf{c} = (c_0, c_1)$ and outputs

$$m = [[c_0 + c_1 \cdot s]_{q_\ell}]_t.$$

A.2.1 Ciphertext noise

Let $\mathbf{c} = (\mathbf{c}, \ell, \nu)$ be the *extended ciphertext*. The *critical quantity* ν of \mathbf{c} (for the associated level ℓ) is defined as the polynomial $\nu = [c_0 + c_1 \cdot s]_{q_\ell}$, and it determines whether \mathbf{c} can be correctly decrypted [CS16]. Specifically, if the error does not wrap around the modulus q_ℓ , namely $\|\nu\|^{can} < q_\ell/2$, the decryption algorithm works. Otherwise, the plaintext cannot be recovered due to excessive noise growth.

To understand the error growth and analyze the critical quantity for any homomorphic operation in the BGV scheme, we refer the readers to [MML⁺23].

A.2.2 Homomorphic operations

Let $\mathbf{c} = (\mathbf{c}, \ell, \nu)$ and $\mathbf{c}' = (\mathbf{c}', \ell, \nu')$ be two extended ciphertexts at the same level ℓ . Let $\alpha \in \mathcal{R}_t$ be a constant polynomial. Then,

- Add(\mathbf{c}, \mathbf{c}'): Addition algorithm outputs

$$([(c_0 + c'_0, c_1 + c'_1)]_{q_\ell}, \ell, \nu + \nu').$$

- Mul(\mathbf{c}, \mathbf{c}'): Multiplication algorithm outputs

$$([(c_0 \cdot c'_0, c_0 \cdot c'_1 + c_1 \cdot c'_0, c_1 \cdot c'_1)]_{q_\ell}, \ell, \nu \cdot \nu').$$

Note that the output of the multiplication is a vector $\mathbf{d} = (d_0, d_1, d_2) \in \mathcal{R}_{q_\ell}^3$. To convert the ciphertext \mathbf{d} back to a ciphertext $\bar{\mathbf{c}} = (\bar{c}_0, \bar{c}_1) \in \mathcal{R}_{q_\ell}^2$ we use a relinearization procedure called key-switching.

A.2.3 Modulus switching

The *modulus switching* procedure allows sacrificing one (or more) of the primes p_i that compose the ciphertext moduli q to obtain a noise reduction. As mentioned before, in our case, we switch from a ciphertext modulus q_ℓ to $q_{\ell-1}$. Let $\mathbf{c} = (\mathbf{c}, \ell, \nu)$, then

- ModSwitch(\mathbf{c}): Modulus switching algorithm sets $\delta = t[-ct^{-1}]_{p_\ell}$ and outputs $([\frac{1}{p_\ell}(\mathbf{c} + \delta)]_{q_{\ell-1}}, \ell - 1, \nu_{ms})$,

where [MML⁺23] $\|\nu_{ms}\|^{can} \leq \frac{1}{p_\ell} \|\nu\|^{can} + B_{scale}$, with

$$B_{scale} = Dt \sqrt{\frac{n}{12}(1 + nV_s)}. \quad (5)$$

A.2.4 Key switching

The *key-switching* technique is used for either reducing the degree of a ciphertext polynomial, usually the output of a multiplication, or changing the key after a rotation. There are different variants of the key-switching procedure. In this work, we used the Hybrid-RNS which combines the RNS adaptations of BV [BV11] and GHS [GHS12] variant. The BV variant decomposes d_2 with respect to a base \mathbf{b} to reduce the error growth and the GHS variant switches to a bigger ciphertext modulus $Q_\ell = q_\ell P$. Then, key switching takes place in \mathcal{R}_{Q_ℓ} and, by modulus switching back down to q_ℓ , the error is reduced again. As a trade-off, we have to make sure that our RLWE instances are secure with respect to Q_ℓ .

However, in the Hybrid-RNS variant, instead of decomposing with respect to each single RNS prime, we group the primes into ω chunks. Specifically, the modulus $q = p_0 \cdots p_{L-1}$ is split in a smaller numbers $\tilde{q}_j = \prod_{i=1}^k r_i^{(j)}$ of k elements by gathering the $r_i^{(j)}$ in ω chunks:

$$q_\ell = p_0 \cdots p_\ell = r_1^{(1)} \cdots r_k^{(1)} \cdots r_1^{(\omega)} \cdots r_k^{(\omega)} = \tilde{q}_1 \cdots \tilde{q}_\omega$$

Hence, we do not apply the decomposition to the base \mathbf{b} but to the base \tilde{q}_j . Thus, we define \mathcal{D} and \mathcal{P} as

$$\mathcal{D}(\alpha) = \left(\left[\alpha \left(\frac{q_\ell}{\tilde{q}_1} \right)^{-1} \right]_{\tilde{q}_1}, \dots, \left[\alpha \left(\frac{q_\ell}{\tilde{q}_\omega} \right)^{-1} \right]_{\tilde{q}_\omega} \right)$$

$$\mathcal{P}(\beta) = \left(\left[\beta \frac{q_\ell}{\tilde{q}_1} \right]_{q_\ell}, \dots, \left[\beta \frac{q_\ell}{\tilde{q}_\omega} \right]_{q_\ell} \right).$$

Thus, if $\mathbf{d} = (d_0, d_1, d_2) \in \mathcal{R}_{q_\ell}^3$ is the output of multiplication, we have to extend d_2 from each \tilde{q}_j to Q_ℓ . So, the key switching algorithms are defined as

- KeySwitchGen(s): Sample $\mathbf{a} \leftarrow \mathcal{U}_{Q_\ell}^\omega$, $\mathbf{e} \leftarrow \chi_e^\omega$. Output key switching key

$$\mathbf{ks} = (\mathbf{ks}_0, \mathbf{ks}_1) \equiv (-\mathbf{a} \cdot s + t\mathbf{e} + P\mathcal{P}(s^2), \mathbf{a}) \pmod{Q_\ell}$$

- KeySwitch($\mathbf{ks}, \mathfrak{d}$) Compute:

$$\mathbf{c}' \equiv (Pd_0 + \langle \mathcal{D}(d_2), \mathbf{ks}_0 \rangle, Pd_1 + \langle \mathcal{D}(d_2), \mathbf{ks}_1 \rangle) \pmod{Q_\ell}.$$

Set $\boldsymbol{\delta} = t[-\mathbf{c}'t^{-1}]_P$, modulus switch back and output $([\frac{1}{P}(\mathbf{c}' + \boldsymbol{\delta})]_{q_\ell}, \ell, \nu_{\mathbf{ks}})$.

The division is done considering either $P \approx \sqrt[\omega]{q}$ if the $r_i^{(j)}$ has the same size or $P \approx \tilde{q}_i$ supposing that \tilde{q}_i is the biggest among the \tilde{q}_j . So $P = \prod_{j=1}^k P_j$.

Note that, before scaling down with the modulus switching, the noise is $\nu' = \nu P + \sqrt{\omega(\ell+1)} \max(\tilde{q}_j) \mathbf{B}_{\mathbf{ks}}$, where [MML⁺23]

$$\mathbf{B}_{\mathbf{ks}} = Dtn\sqrt{V_e/12}. \quad (6)$$

Thus, the Hybrid-RNS key switching noise after the modulus switching is bounded by $\frac{q_\ell}{Q_\ell} \|\nu'\|^{can} + \mathbf{B}_{\text{scale}}$, that is,

$$\|\nu_{\mathbf{ks}}\|^{can} \leq \|\nu\|^{can} + \sqrt{\omega(\ell+1)} \frac{\max(\tilde{q}_j)}{P} \mathbf{B}_{\mathbf{ks}} + \sqrt{k} \mathbf{B}_{\text{scale}}.$$

A.2.5 Distributed Decryption

The SPDZ offline phase utilizes a form of distributed decryption, which is also supported in the BGV scheme. A secret key $s \in \mathcal{R}_q$ can be additively distributed among N parties

by assigning each party a value s_i , such that $s = s_1 + \dots + s_N$. As explained in [BCS20], the BGV parameter growth using distributed decryption. Indeed, in the usual BGV case, for the bottom modulus $p_0 = q_0$, we do not apply either the key switching or the modulus switching afterwards. Thus, to ensure correct decryption, we require that $\|\nu\|^{can} < q_0/2$. Instead, in the case of distributed decryption, we have

$$q_0 > 2(1 + N2^{DD_{\text{sec}}}) \cdot \|\nu\|^{can}. \quad (7)$$

See [BCS20] for more details.

A.2.6 Dishonest Encryption

In the BGV scheme, we have that the noise after a fresh encryption is bounded by [MML⁺23] $\|[c_0 + c_1 \cdot s]_{q_e}\|^{can} \leq D\sqrt{nV_{m+te_{sk}u+e_1s+e_0}}$. However, using the ZKPoK protocol, we are only able to guarantee that [BCS20] $\|2\sum_{i=1}^N m_i\|_{\infty} \leq N \cdot 2^{\text{ZK}_{\text{sec}}+2}t/2$ and $\|2\sum_{i=1}^N e_{j,i}\|_{\infty} \leq N \cdot 2^{\text{ZK}_{\text{sec}}+2}\rho_j$ where $\rho_{sk} = 1$ and for $\rho_0 = \rho_1 = 21$. Thus,

$$\begin{aligned} \|c_0 - c_1 \cdot s\|^{can} &\leq \sum_{i=1}^N \|2 \cdot m_i\|^{can} + t(\|2 \cdot e_{sk,i}\|^{can} \|u\|^{can} \\ &\quad + \|2 \cdot e_{1,i}\|^{can} \cdot \|s\|^{can} + \|2 \cdot e_{0,i}\|^{can}). \end{aligned}$$

Since $\|a\|^{can} \leq n\|a\|_{\infty}$, we have

$$\begin{aligned} \|c_0 - c_1 \cdot s\|^{can} &\leq t \cdot n \cdot N \cdot 2^{\text{ZK}_{\text{sec}}+1} \\ &\quad + D \cdot t \cdot n \cdot N \cdot 2^{\text{ZK}_{\text{sec}}+2} \cdot \rho_{sk} \sqrt{nV_u} \\ &\quad + D \cdot t \cdot n \cdot N \cdot 2^{\text{ZK}_{\text{sec}}+2} \cdot \rho_1 \sqrt{nV_s} \\ &\quad + t \cdot n \cdot N \cdot 2^{\text{ZK}_{\text{sec}}+2} \cdot \rho_0. \end{aligned}$$

Because $V_u = V_s = 1/2$, we set

$$\text{B}_{\text{clean}}^{\text{dishonest}} \approx t \cdot n \cdot N \cdot 2^{\text{ZK}_{\text{sec}}+2} (21 + 11D\sqrt{2n}). \quad (8)$$

B Network Architectures

Layer	Input Size	Description	Output
Convolution	$32 \times 32 \times 1$	Window size 5×5 , Stride (1, 1), Padding (0, 0), output channels 6	$28 \times 28 \times 6$
ReLU Activation	$28 \times 28 \times 6$	ReLU(\cdot) on each input	$28 \times 28 \times 6$
Max Pooling	$28 \times 28 \times 6$	Window size 2×2 , Stride (2, 2)	$14 \times 14 \times 6$
Convolution	$14 \times 14 \times 6$	Window size 5×5 , Stride (1, 1), Padding (0, 0), output channels 16	$10 \times 10 \times 16$
ReLU Activation	$10 \times 10 \times 16$	ReLU(\cdot) on each input	$10 \times 10 \times 16$
Max Pooling	$10 \times 10 \times 16$	Window size 2×2 , Stride (2, 2)	$5 \times 5 \times 16$
Convolution	$5 \times 5 \times 16$	Window size 5×5 , Stride (1, 1), Padding (0, 0), output channels 120	$1 \times 1 \times 120$
ReLU Activation	$1 \times 1 \times 120$	ReLU(\cdot) on each input	$1 \times 1 \times 120$
Fully Connected Layer	120	Fully connected layer	84
ReLU Activation	84	ReLU(\cdot) on each input	84
Fully Connected Layer	84	Fully connected layer	10

Figure 8: LeNet network architecture [LBBH98] for training over MNIST dataset.

Layer	Input Size	Description	Output
Convolution	$32 \times 32 \times 3$	Window size 3×3 , Stride (1, 1), Padding (1, 1), output channels 32	$32 \times 32 \times 32$
Batch Normalization	$32 \times 32 \times 32$	BN(\cdot) on each input	$32 \times 32 \times 32$
ReLU Activation	$32 \times 32 \times 32$	ReLU(\cdot) on each input	$32 \times 32 \times 32$
Convolution	$32 \times 32 \times 32$	Window size 3×3 , Stride (1, 1), Padding (1, 1), output channels 64	$32 \times 32 \times 64$
Batch Normalization	$32 \times 32 \times 64$	BN(\cdot) on each input	$32 \times 32 \times 64$
ReLU Activation	$32 \times 32 \times 64$	ReLU(\cdot) on each input	$32 \times 32 \times 64$
Max Pooling	$32 \times 32 \times 64$	Window size 2×2 , Stride (2, 2)	$16 \times 16 \times 64$
Convolution	$16 \times 16 \times 64$	Window size 3×3 , Stride (1, 1), Padding (1, 1), output channels 128	$16 \times 16 \times 128$
Batch Normalization	$16 \times 16 \times 128$	BN(\cdot) on each input	$16 \times 16 \times 128$
ReLU Activation	$16 \times 16 \times 128$	ReLU(\cdot) on each input	$16 \times 16 \times 128$
Convolution	$16 \times 16 \times 128$	Window size 3×3 , Stride (1, 1), Padding (1, 1), output channels 128	$16 \times 16 \times 128$
Batch Normalization	$16 \times 16 \times 128$	BN(\cdot) on each input	$16 \times 16 \times 128$
ReLU Activation	$16 \times 16 \times 128$	ReLU(\cdot) on each input	$16 \times 16 \times 128$
Max Pooling	$16 \times 16 \times 128$	Window size 2×2 , Stride (2, 2)	$8 \times 8 \times 128$
Convolution	$8 \times 8 \times 128$	Window size 3×3 , Stride (1, 1), Padding (1, 1), output channels 256	$8 \times 8 \times 256$
Batch Normalization	$8 \times 8 \times 256$	BN(\cdot) on each input	$8 \times 8 \times 256$
ReLU Activation	$8 \times 8 \times 256$	ReLU(\cdot) on each input	$8 \times 8 \times 256$
Convolution	$8 \times 8 \times 256$	Window size 3×3 , Stride (1, 1), Padding (1, 1), output channels 256	$8 \times 8 \times 256$
Batch Normalization	$8 \times 8 \times 256$	BN(\cdot) on each input	$8 \times 8 \times 256$
ReLU Activation	$8 \times 8 \times 256$	ReLU(\cdot) on each input	$8 \times 8 \times 256$
Max Pooling	$8 \times 8 \times 256$	Window size 2×2 , Stride (2, 2)	$4 \times 4 \times 256$
Fully Connected Layer	4096	Fully connected layer	32
ReLU Activation	32	ReLU(\cdot) on each input	32
Fully Connected Layer	32	Fully connected layer	10

Figure 9: CNN network architecture for private classification over CIFAR-10 dataset. It contains 1.5 million paramerters

Layer	Input Size	Description	Output
Convolution	$32 \times 32 \times 3$	Window size 3×3 , Stride $(1, 1)$, Padding $(1, 1)$, output channels 64	$32 \times 32 \times 64$
ReLU Activation	$32 \times 32 \times 64$	ReLU(\cdot) on each input	$32 \times 32 \times 64$
Convolution	$32 \times 32 \times 64$	Window size 3×3 , Stride $(1, 1)$, Padding $(1, 1)$, output channels 64	$32 \times 32 \times 64$
ReLU Activation	$32 \times 32 \times 64$	ReLU(\cdot) on each input	$32 \times 32 \times 64$
Max Pooling	$32 \times 32 \times 64$	Window size 2×2 , Stride $(2, 2)$	$16 \times 16 \times 64$
Convolution	$16 \times 16 \times 64$	Window size 3×3 , Stride $(1, 1)$, Padding $(1, 1)$, output channels 128	$16 \times 16 \times 128$
ReLU Activation	$16 \times 16 \times 128$	ReLU(\cdot) on each input	$16 \times 16 \times 128$
Convolution	$16 \times 16 \times 128$	Window size 3×3 , Stride $(1, 1)$, Padding $(1, 1)$, output channels 128	$16 \times 16 \times 128$
ReLU Activation	$16 \times 16 \times 128$	ReLU(\cdot) on each input	$16 \times 16 \times 128$
Max Pooling	$16 \times 16 \times 128$	Window size 2×2 , Stride $(2, 2)$	$8 \times 8 \times 128$
Convolution	$8 \times 8 \times 128$	Window size 3×3 , Stride $(1, 1)$, Padding $(1, 1)$, output channels 256	$8 \times 8 \times 256$
ReLU Activation	$8 \times 8 \times 256$	ReLU(\cdot) on each input	$8 \times 8 \times 256$
Convolution	$8 \times 8 \times 256$	Window size 3×3 , Stride $(1, 1)$, Padding $(1, 1)$, output channels 256	$8 \times 8 \times 256$
ReLU Activation	$8 \times 8 \times 256$	ReLU(\cdot) on each input	$8 \times 8 \times 256$
Convolution	$8 \times 8 \times 256$	Window size 3×3 , Stride $(1, 1)$, Padding $(1, 1)$, output channels 256	$8 \times 8 \times 256$
ReLU Activation	$8 \times 8 \times 256$	ReLU(\cdot) on each input	$8 \times 8 \times 256$
Max Pooling	$8 \times 8 \times 256$	Window size 2×2 , Stride $(2, 2)$	$4 \times 4 \times 256$
Convolution	$4 \times 4 \times 256$	Window size 3×3 , Stride $(1, 1)$, Padding $(1, 1)$, output channels 512	$4 \times 4 \times 512$
ReLU Activation	$4 \times 4 \times 512$	ReLU(\cdot) on each input	$4 \times 4 \times 512$
Convolution	$4 \times 4 \times 512$	Window size 3×3 , Stride $(1, 1)$, Padding $(1, 1)$, output channels 512	$4 \times 4 \times 512$
ReLU Activation	$4 \times 4 \times 512$	ReLU(\cdot) on each input	$4 \times 4 \times 512$
Convolution	$4 \times 4 \times 512$	Window size 3×3 , Stride $(1, 1)$, Padding $(1, 1)$, output channels 512	$4 \times 4 \times 512$
ReLU Activation	$4 \times 4 \times 512$	ReLU(\cdot) on each input	$4 \times 4 \times 512$
Max Pooling	$4 \times 4 \times 512$	Window size 2×2 , Stride $(2, 2)$	$2 \times 2 \times 512$
Convolution	$2 \times 2 \times 512$	Window size 3×3 , Stride $(1, 1)$, Padding $(1, 1)$, output channels 512	$2 \times 2 \times 512$
ReLU Activation	$2 \times 2 \times 512$	ReLU(\cdot) on each input	$2 \times 2 \times 512$
Convolution	$2 \times 2 \times 512$	Window size 3×3 , Stride $(1, 1)$, Padding $(1, 1)$, output channels 512	$2 \times 2 \times 512$
ReLU Activation	$2 \times 2 \times 512$	ReLU(\cdot) on each input	$2 \times 2 \times 512$
Convolution	$2 \times 2 \times 512$	Window size 3×3 , Stride $(1, 1)$, Padding $(1, 1)$, output channels 512	$2 \times 2 \times 512$
ReLU Activation	$2 \times 2 \times 512$	ReLU(\cdot) on each input	$2 \times 2 \times 512$
Max Pooling	$2 \times 2 \times 512$	Window size 2×2 , Stride $(2, 2)$	$1 \times 1 \times 512$
Fully Connected Layer	512	Fully connected layer	4096
ReLU Activation	4096	ReLU(\cdot) on each input	4096
Fully Connected Layer	4096	Fully connected layer	4096
ReLU Activation	4096	ReLU(\cdot) on each input	4096
Fully Connected Layer	4096	Fully connected layer	1000
ReLU Activation	1000	ReLU(\cdot) on each input	1000

Figure 10: VGG16 network architecture [SZ15] for training over CIFAR-10 dataset. It contains 37 million parameters.