國立臺北科技大學
資訊工程系碩士班

碩士學位論文

漸進式委員會佔領攻擊與激勵相容防禦：區塊鏈聯邦學習的安全性研究

Progressive Committee Capture Attack and Incentive-Compatible Defense: Security Analysis for Blockchain-based Federated Learning

研究生：陸紀霖

指導教授：張世豪博士

中華民國一百一十五年一月

國立臺北科技大學

資訊工程系碩士班

碩士學位論文

漸進式委員會佔領攻擊與激勵相容防禦：區塊鏈聯邦學習的安全性研究

Progressive Committee Capture Attack and Incentive-Compatible Defense: Security Analysis for Blockchain-based Federated Learning

研究生：陸紀霖

指導教授：張世豪博士

中華民國一百一十五年一月

「學位論文口試委員會審定書」掃描檔

審定書填寫方式以系所規定為準，但檢附在電子論文內的掃描檔須具備以下條件：

1. 含指導教授、口試委員及系所主管的完整簽名。
2. 口試委員人數正確，碩士口試委員至少 3 人、博士口試委員至少 5 人。
3. 若此頁有論文題目，題目應和書背、封面、書名頁、摘要頁的題目相符。
4. 此頁有無浮水印皆可。

審定書不用頁碼

摘要

關鍵詞：區塊鏈、聯邦式學習、委員會佔領、驗證者共謀

基於區塊鏈的聯邦式學習 (BCFL) 透過去中心化共識機制解決了信任與隱私問題。現有的 BCFL 系統依賴基於委員會的驗證機制，並假設委員會成員是誠實的或擁有誠實多數。此假設容易受到驗證者共謀的威脅，攻擊者可透過累積權益 (Stake) 來主導委員會。我們識別出一種新型威脅——漸進式委員會佔領 (PCC)，理性攻擊者利用激勵機制逐步累積權益，並佔領足夠的委員會席次以發動協同攻擊。一旦攻擊者取得委員會多數席次，現有的委員會架構便無法偵測或防範此類攻擊。為防禦 PCC，我們提出一種結合 ** 即時執行 ** 與 ** 異步審計 ** 的委員會架構，將安全性與委員會組成解耦：由小型委員會負責例行驗證並立即執行模型更新以提供最佳活性 (Liveness)，而由全域共識支持的 ** 異步審計機制 ** 提供安全性保證。任何惡意聚合行為都將在事後被審計發現，觸發密碼學驗證、罰沒懲罰，並立即移除惡意驗證者——無論其在委員會中的席次多寡。此機制將安全門檻從委員會多數轉移至全網共識，從而瓦解委員會佔領攻擊。實驗結果顯示，當攻擊發生時，本機制能完全清除惡意委員會成員，而現有最先進的方法則允許攻擊者取得委員會完全控制權並執行不受制衡的攻擊。我們的解耦設計亦允許更小的委員會規模，在不犧牲安全性的前提下提升計算效率。

TAIPEI TECH

ABSTRACT

Keyword: Blockchain, Federated Learning, Committee Capture, Verifier Collusion

Blockchain-based Federated Learning (BCFL) addresses trust and privacy concerns through decentralized consensus. Current BCFL systems rely on committee-based validation assuming honest or honest-majority committees. This assumption is vulnerable to verifier collusion, where attackers accumulate stake to dominate committees. We identify Progressive Committee Capture (PCC), a novel threat where rational attackers exploit incentive mechanisms to gradually accumulate stake and capture sufficient committee seats for coordinated attacks. Existing committee-based architectures cannot detect or prevent such attacks once attackers achieve committee majority. To defend against PCC, we propose an **Audit-Augmented Committee Architecture** that combines **Immediate Execution** with **Asynchronous Audit**. This design decouples security from committee composition: a small committee provides liveness through routine validation and immediate model execution, while an **asynchronous audit mechanism** backed by global consensus provides security guarantees. Any malicious aggregation will be detected post-hoc, triggering cryptographic verification, slashing penalties, and immediate removal of malicious validators—regardless of their committee representation. This shifts the security threshold from committee majority to global network consensus, neutralizing committee capture attacks. Experimental results demonstrate complete elimination of malicious committee members upon attack attempts, while state-of-the-art approaches allow attackers to achieve full committee control and execute unchecked attacks. Our decoupled design also enables smaller committee sizes, improving computational efficiency without compromising security.

TAIPEI TECH

誌謝

所有對於研究提供協助之人或機構，作者都可在誌謝中表達感謝之意。

TAIPEI TECH

目錄

摘要 i

ABSTRACT ii

誌謝 iii

目錄 iv

圖目錄 viii

表目錄 ix

第一章 緒論 (Introduction) 1

1.1 研究背景 (Background & Context) 1

1.2 問題陳述 (Problem Statement) 1

1.3 研究目標 (Research Objectives) 2

1.4 研究貢獻 (Contributions) 2

1.5 論文組織 (Thesis Organization) 3

第二章 相關技術背景 4

2.1 聯邦學習與拜占庭容錯 (Federated Learning and Byzantine Resilience) 4

2.1.1 聯邦學習數學定義 4

2.1.2 拜占庭故障與魯棒聚合 4

2.2 基於區塊鏈的聯邦學習 (Blockchain-based Federated Learning) 5

2.2.1 BCFL 架構演進 5

2.2.2 基於質押的參與機制 5

2.3 激勵與罰沒機制基礎 (Incentive and Slashing Foundations) 6

2.3.1 權益質押與經濟終局性 6

2.3.2 罰沒機制的設計原則 6

第三章 相關工作 (Related Work) 7

3.1 BCFL 的擴展開銷與 Layer-2 方案 7

3.1.1 零知識證明與機器學習 (zkML) 7

3.1.2 Optimistic Rollup 與挑戰機制 7

3.2 委員會共識之效率與安全性權衡 7

TAIPEI TECH

3.2.1 現有委員會選擇機制 7
3.2.2 誠實大多數假設的局限性 8
3.3 安全性威脅與防禦缺口 8
3.3.1 客戶端投毒與後門攻擊 8
3.3.2 聚合端攻擊：被忽視的「監督者」風險 8
3.4 本研究之定位 (The Research Gap) 8
第四章 威脅模型 (Threat Model) 10
4.1 系統模型與假設 10
4.1.1 網絡模型 10
4.1.2 聚合與共識流程 11
4.1.3 權益動態機制 11
4.1.4 系統假設 12
4.2 攻擊者模型 12
4.2.1 攻擊者類型：理性攻擊者 12
4.2.2 攻擊者目標 13
4.2.3 攻擊者能力 13
4.2.4 攻擊者限制 13
4.3 攻擊向量分析 14
4.3.1 數據層攻擊：已有防禦 14
4.3.2 共識層攻擊：本研究重點 15
4.3.3 攻擊層次對比 15
4.4 漸進式權益佔領攻擊 (Progressive Committee Capture Attack) 16
4.4.1 攻擊定義 16
4.4.2 攻擊階段詳述 16
4.4.3 權益增長動態分析 (Stake Growth Dynamics Analysis) 18
4.4.4 攻擊效果與影響 19
4.4.5 與傳統攻擊的區別 20
4.5 安全目標 20
4.5.1 防止委員會被惡意節點控制 20

TAIPEI TECH

4.5.2 確保誠實節點的權益公平增長 20
4.5.3 維持模型收斂性與準確性 21
4.5.4 保持系統的去中心化特性 21
4.5.5 激勵相容性 21
4.6 本章小結 22
第五章 系統架構設計 23
5.1 系統架構概覽 23
5.1.1 核心角色定義 23
5.1.2 工作流程重構 24
5.2 異步審計與究責機制 24
5.2.1 即時執行策略 25
5.2.2 異步挑戰流程 25
5.2.3 處置決策：僅懲罰不回滾 (Slash-Only Policy) 25
5.3 安全性保證 26
5.3.1 雙層信任模型 (Two-Tier Trust Model) 26
5.3.2 攻擊成本分析 27
5.4 效率分析 27
5.4.1 通訊複雜度公式 27
5.4.2 委員會大小的概率分析 28
5.5 激勵機制 29
5.6 本章小結 29
第六章 實驗評估 (Experimental Evaluation) 30
6.1 實驗設置 30
6.1.1 數據集與模型 30
6.1.2 基準方法與攻擊場景 30
6.1.3 實驗參數 31
6.2 模型效能：抵抗權益佔領攻擊 32
6.3 權益動態：激勵相容性分析 33
6.4 效率與可擴展性分析 35

TAIPEI TECH

6.4.1 複雜度比較 35
6.4.2 安全性與效率的權衡分析 36
6.5 討論 37
6.5.1 確定性安全保證 37
6.5.2 計算通用性 38
6.5.3 挑戰機制的實際成本 38
6.5.4 研究範圍與未來擴展方向 39
6.6 本章小結 39
參考文獻 41

TAIPEI TECH

圖目錄

6.1 權益佔領攻擊下的模型準確率收斂比較。橙線代表 BlockDFL，藍線代表本研究提出的方法。 32
6.2 權益演化比較。圖中虛線為 BlockDFL 的權益變化，實線為本研究方法的權益變化。綠線代表誠實節點平均權益，紅線代表惡意節點平均權益。 33
6.3 長期訓練下的權益演化比較 (800 輪)。圖中虛線代表 BlockDFL 方案，實線代表本研究方案。 35

TAIPEI TECH

表目錄

4.1 攻擊層次對比 16
4.2 與傳統攻擊的區別 20

6.1 通訊複雜度比較 35
6.2 不同安全需求下的系統配置比較 37

TAIPEI TECH

第一章　緒論 (Introduction)

1.1　研究背景 (Background & Context)

隨著物聯網 (IoT) 設備的普及和隱私法規 (如 GDPR) 的日益嚴格，聯邦學習 (Federated Learning, FL) 作為一種隱私保護的分散式機器學習範式，獲得了廣泛關注。FL 允許客戶端在本地訓練模型並僅上傳模型更新，從而避免了原始數據的傳輸。然而，傳統的 FL 依賴於中心化的聚合伺服器 (Central Server)，這不僅構成了單點故障 (Single Point of Failure)，還面臨著服務器惡意篡改模型或遭受攻擊的風險。

為了解決這些問題，區塊鏈聯邦學習 (Blockchain-based Federated Learning, BCFL) 應運而生。BCFL 利用區塊鏈的去中心化、不可篡改和智能合約特性，取代了傳統的中心化聚合器。在 BCFL 架構中，模型更新的驗證和聚合通常由一組選定的「委員會 (Committee)」或「驗證者 (Verifiers)」負責。這種架構試圖通過共識機制 (Consensus Mechanism) 來確保模型更新的正確性，並通過加密貨幣激勵機制來鼓勵節點參與。

目前，主流的 BCFL 系統 (如 BlockDFL) 多採用基於權益 (Stake) 或聲譽 (Reputation) 的委員會選舉機制，並依賴拜占庭容錯 (BFT) 共識來達成決策。這些系統通常假設誠實節點佔據多數 (Honest Majority)，並以此作為安全性的基石。

1.2　問題陳述 (Problem Statement)

儘管 BCFL 在去中心化方面取得了進展，但現有研究存在一個關鍵的「驗證盲點 (Verification Blind Spot)」。絕大多數 (約 93%) 的現有文獻主要關注數據層面的投毒攻擊 (Data Poisoning)，例如標籤翻轉或後門攻擊，並開發了如 Krum、Trimmed Mean 等魯棒聚合算法。然而，這些防禦機制隱含地假設執行聚合算法的「委員會」本身是誠實的。

本研究指出，這一假設在面對「理性攻擊者 (Rational Attacker)」時是極其脆弱的。我們定義了一種新的威脅模型——**委員會佔領 (Committee Capture)**，特別是其具體實現形式**「漸進式委員會佔領攻擊 (Progressive Committee Capture Attack, PCCA)」**。在此攻擊中，理性攻擊者並非旨在破壞模型，而是追求利益最大化。他們通過以下兩個階段實施攻擊：

1. **潛伏階段 (Infiltration Phase)：**攻擊者表現誠實，積累權益 (Stake) 以進入驗證者池。
2. **佔領階段 (Capture Phase)：**一旦在委員會中獲得多數席位，攻擊者便實施「戰略性餓死 (Strategic Starvation)」，即拒絕打包誠實節點的更新，獨佔系統獎勵。

這種攻擊導致誠實節點的權益停滯，而攻擊者的權益呈指數級增長，從而進一步鞏固其在未來委員會中的控制權。傳統的 BFT 共識機制面臨著「安全性與效率的兩難

(Security-Efficiency Dilemma)」：為了防禦共謀，必須擴大委員會規模 (C)，但通訊複雜度 ($O(C^2)$) 會隨之呈二次方增長，導致系統效率急劇下降。

1.3 研究目標 (Research Objectives)

針對上述問題，本研究的主要目標是設計一種既能防禦理性共謀，又能保持高效率的 BCFL 架構。具體目標包括：

1. **防禦委員會佔領：**設計一種機制，使系統在面對理性驗證者共謀時仍能保持安全，打破對「誠實多數」的依賴。
2. **解耦安全性與效率：**打破傳統 BFT 系統中安全性與委員會規模的綁定關係，實現在使用小型委員會 (確保活性) 的同時，提供不亞於大型委員會的安全性。
3. **實現激勵相容 (Incentive Compatibility)：**利用博弈論設計獎懲機制，使得「誠實行為」成為理性節點的納什均衡 (Nash Equilibrium) 策略。

1.4 研究貢獻 (Contributions)

本研究的主要貢獻歸納如下：

1. **識別新型威脅 (New Threat Identification)：**我們系統地分析了 BCFL 的驗證層漏洞，並定義了「漸進式委員會佔領攻擊 (PCCA)」。我們揭示了理性攻擊者如何利用權益機制進行中心化接管，這補充了現有文獻僅關注數據投毒的不足。
2. **提出基於異步審計的激勵相容架構 (Incentive-Compatible Architecture with Asynchronous Audit)：**我們提出了一種結合「即時執行 (Immediate Execution)」與「異步審計 (Asynchronous Audit)」的新型防禦機制。
    - 採用**即時執行**策略，移除傳統的驗證等待期，確保模型訓練的零延遲與高活性 (Liveness)。
    - 引入**異步審計 (Asynchronous Audit)** 機制，允許挑戰者在事後對委員會的決策進行回溯性驗證。
    - 設計**內部罰沒機制 (Internal Slashing)**，確保 $Slashing \gg Gain$，從根本上遏制理性攻擊者的作惡動機。
3. **驗證與評估 (Evaluation)：**我們通過理論分析和模擬實驗驗證了所提機制的有效性。
    - **安全性：**在 30% 節點共謀的極端情況下，本方案仍能維持模型收斂 (準確率 91.8%)，而對照組 (BlockDFL) 則崩潰 (65%)。

- **效率**：本方案成功將通訊複雜度從 $O(C_{large}^2)$ 降低至 $O(C_{small}^2)$ (正常情況)，在保證高安全性的前提下實現了約 27 倍的效率提升。

1.5 論文組織 (Thesis Organization)

本論文共分為六章，組織結構如下：

- **第一章緒論 (Introduction)**：介紹研究背景、問題陳述、研究目標及貢獻。
- **第二章背景知識 (Background)**：介紹聯邦學習、區塊鏈技術及博弈論基礎。
- **第三章相關工作 (Related Work)**：回顧現有的 BCFL 方案及其局限性，特別是針對共識層安全的討論。
- **第四章威脅模型 (Threat Model)**：詳細定義系統模型、攻擊者能力及 PCCA 攻擊策略。
- **第五章架構設計 (Framework Design)**：闡述所提出的異步審計與即時執行架構，包括協議流程、審計機制及智能合約設計。
- **第六章實驗評估 (Evaluation)**：展示模擬實驗結果，對比本方案與基準方案在模型效能、權益動態及系統效率上的表現。
- **第七章結論與未來展望 (Conclusion and Future Work)**：總結全文並提出未來的研究方向。

TAIPEI TECH

第二章　相關技術背景

本章節旨在建立理解本研究所需之技術基礎，涵蓋聯邦學習、拜占庭容錯機制、區塊鏈架構以及現代共識系統中的經濟安全設計。本章內容採說明性敘述，為後續章節之問題分析與系統設計提供理論框架。

2.1　聯邦學習與拜占庭容錯 (Federated Learning and Byzantine Resilience)

聯邦學習（Federated Learning, FL）是由 McMahan 等人於 2017 年正式提出之分散式機器學習框架 [1]。其核心目標在於多個參與方（Clients）協同訓練模型，而無需將原始資料集中於中央伺服器，從而保護資料隱私。

2.1.1　聯邦學習數學定義

在標準聯邦學習架構中，目標是最小化全域損失函數 $F(w)$：

$$\min_{w} F(w) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(w) \tag{2.1}$$

其中 K 為參與客戶端總數，n_k 為第 k 個客戶端之本地樣本數，$F_k(w)$ 為其本地損失函數。經典的 *FederatedAveraging* (FedAvg) 演算法透過週期性地收集客戶端模型更新 w_{t+1}^k，並在伺服器端進行加權聚合：

$$w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k \tag{2.2}$$

此方法雖然顯著降低了通訊開銷，但其安全性建立在中央聚合器完全誠實且客戶端皆非惡意的假設之上。

2.1.2　拜占庭故障與魯棒聚合

在開放或非受信任環境中，部分參與者可能發生拜占庭故障（Byzantine fault），即發送任意、甚至是蓄意偽造的梯度向量。Blanchard 等人證明了線性聚合規則（如算術平均）無法抵禦即便只有一個拜占庭節點的攻擊 [2]。為了應對此威脅，研究界提出了多種拜占庭容錯（Byzantine-robust）聚合機制，例如：

- **Krum/Multi-Krum**：透過計算各向量間的歐幾里得距離，選擇與周圍節點距離最近的向量，以排除偏離較大的惡意更新。

- **座標式裁剪均值 (Trimmed Mean)**：在各維度上移除最大與最小的部分觀測值，對剩餘值取平均。Yin 等人證明了該方法在特定機率分佈下可達到階數最優統計誤差率 [3]。

- **座標式中位數 (Coordinate-wise Median)**：取各維度上的中位數作為聚合結果，具有較高的崩潰點（Breakdown point）。

同樣地，Bulyan 演算法 [4] 被提出用於縮減高維度下投毒攻擊的空間。儘管這些機制增強了對惡意客戶端的防禦力，但它們均隱含了一個關鍵的前提：執行這些規則的「中央聚合器」必須是絕對誠實的。

2.2 基於區塊鏈的聯邦學習 (Blockchain-based Federated Learning)

為了消除對單一中央伺服器的依賴，研究者引入區塊鏈技術，提出區塊鏈式聯邦學習（BCFL）架構。在此架構中，去中心化帳本取代了傳統聚合器，提供不可篡改性與透明性。

2.2.1 BCFL 架構演進

BCFL 的發展經歷了從全節點共識到委員會機制的演進：

1. **早期架構 (PoW-based)**：如 BlockFL [5] 使用工作量證明（PoW）達成共識，雖具備高度去中心化特性，但面臨高能耗與高延遲問題。Lu 等人則探討了將訓練品質（PoQ）與共識結合的工業物聯網架構 [6]。

2. **委員會共識 (Committee-based)**：如 BFLC [7] 與 BlockDFL [8]。為了提升效能，系統從全體參與者中選出一個子集（委員會）負責驗證與聚合。此種機制將通訊複雜度從 $O(n^2)$ 降低至 $O(C^2)$，其中 C 為委員會大小。隨後的研究如 VBFL [9] 與 VFChain [10] 分別引入了基於權益的共識與可審計的聚合證明。

2.2.2 基於質押的參與機制

現代 BCFL 系統常借鑑權益質押（Staking）概念，要求節點質押代幣以獲得成為委員會成員的權利。此機制建立了經濟進入門檻，並將參與者的利益與系統的整體安全綁定，為進階的激勵與懲罰機制奠定了基礎。

2.3 激勵與罰沒機制基礎 (Incentive and Slashing Foundations)

加密經濟安全性（Crypto-economic Security）的核心在於確保「攻擊成本高於潛在收益」。這主要透過權益質押與罰沒（Slashing）機制來達成。

2.3.1 權益質押與經濟終局性

在主流共識協議（如 Casper FFG [11]、Tendermint [12]、Cosmos [13] 或 Polkadot [14]）中，驗證者必須存入押金。若驗證者違反協定規則（例如雙重投票），其部分或全部押金將被系統自動沒收。這種機制確保了系統具有「可問責性」（Accountability），即任何破壞安全性的行為都能被識別並追究經濟責任 [15]。

2.3.2 罰沒機制的設計原則

一個完善的罰沒機制具備以下特性：

- **即時性**：違規證據一經提交，處罰應在區塊鏈上立即生效。
- **相關性懲罰 (Correlation Penalty)**：如 Gasper [16] 所採用，當多個節點在同一時間段內發生違規行為時，罰沒比例會非線性增加，以有效遏止大規模協同共謀。
- **激勵相容性**：設計旨在使誠實行為成為理性參與者的最優策略。

總結而言，聯邦學習提供了協同訓練的框架，拜占庭容錯提供了算法層面的防禦，而區塊鏈及其經濟機制則提供了去中心化的信任根基。然而，當這些技術結合時，如何確保「監督者（委員會）」本身不被攻陷，仍是現有技術未能完全解決的問題。

第三章　相關工作 (Related Work)

本章節將現有關於區塊鏈聯邦學習之安全性與效率的研究分為三類進行探討，並分析其局限性，最後精確定義本研究欲填補之學術 Gap。

3.1　BCFL 的擴展開銷與 Layer-2 方案

隨著模型參數規模的擴大，在區塊鏈上直接驗證模型更新的計算開銷已成為瓶頸。

3.1.1　零知識證明與機器學習 (zkML)

Chen 等人 [17] 探討了使用 zkSNARKs 來驗證模型推論的技術。雖然 zkML 能提供極強的密碼學保證，但其產生的證明時間（Proof generation time）極長。例如，對於具備千萬級參數的模型，生成一次證明可能需要數十分鐘甚至數小時，且需耗費巨大的記憶體資源。Sun 等人提出的 zkLLM [18] 進一步將以此擴展至大型語言模型，但仍面臨極高的計算消耗。針對聯邦學習，RiseFL [19] 嘗試利用 Pedersen 承諾與 Bulletproofs 進行輕量化驗證，而 Heiss 等人 [20] 與 Wang 等人 [21] 則分別提出利用鏈下運算 (VOC) 與礦工驗證的 zkFL 框架。

3.1.2　Optimistic Rollup 與挑戰機制

在區塊鏈擴展領域，Optimistic Rollup 提出了一種「預設為真，有疑則挑戰」的邏輯。Conway 等人提出的 opML [22] 嘗試將此思路引入機器學習，大幅降低了平時的運算負擔。然而，現有的 opML 主要關注單一 Prover 的正確性，且其挑戰期（Challenge Period）通常設為數天至一週，難次適應聯邦學習快速迭代的需求。本研究借鑑了此「樂觀執行」與「經濟激勵挑戰」的精神，但將其改造為適用於去中心化委員會架構的即時防禦方案。

3.2　委員會共識之效率與安全性權衡

為了降低 $O(n^2)$ 的全節點通訊開銷，現代 BCFL 方案普遍採用委員會架構。

3.2.1　現有委員會選擇機制

BlockDFL [8] 採用基於哈希環（Hash-ring）的偽隨機選取，而 FLCoin [23] 則利用滑動視窗（Sliding window）機制。這些方法確實成功地將共識複雜度降低至 $O(C^2)$ 或 $O(C)$，使得系統在大規模節點下仍能運作。

3.2.2 誠實大多數假設的局限性

儘管效率獲得提升，上述方案之安全性均根本性地依賴於「委員會內超過 2/3 為誠實節點」的假設。

- **漸進式佔領風險**：惡意節點可以透過長期表現「誠實」來累積 Stake 或聲譽，逐步增加被選入委員會的機率。
- **缺乏自癒能力**：當惡意比例跨越門檻（例如佔領 >1/3 或 >1/2 權限）時，系統會陷入僵局或被惡意控制。目前的機制多半缺乏在「委員會已淪陷」的情況下，由系統外部或低權限節點發起有效挑戰並逆轉結果的能力。

3.3 安全性威脅與防禦缺口

本節區分傳統威脅與本研究聚焦之高機密性威脅。

3.3.1 客戶端投毒與後門攻擊

現有防禦如 Krum 或 Trimmed Mean 主要針對惡意客戶端造成的模型偏差。然而，Fang 等人 [24] 證明了即使是這些魯棒聚合規則，在面對具備最佳化能力的攻擊者時，防禦效果依舊有限。

3.3.2 聚合端攻擊：被忽視的「監督者」風險

大部分研究假設聚合者（Aggregator）或驗證者（Verifier）是受信任的節點或誠實執行協議者。但在去中心化環境中，驗證者可能被賄賂、共謀或被駭客攻陷。FLTrust [25] 嘗試引入信任根，但其信任根仍高度依賴伺服器持有的乾淨資料。一旦執行聚合與驗證的「委員會」集體作惡（例如共同核可一個投毒後的模型以賺取不當獎勵），現有框架將完全失效。

3.4 本研究之定位 (The Research Gap)

總結現有文獻，我們發現一個顯著的研究空白：**如何設計一個具備「激勵相容性」的機制，使得當獲取合法權限的驗證委員會集體舞弊時，系統仍能透過非對稱的經濟激勵（挑戰機制）來識別並清除這些惡意驗證者？**

相較於前人研究：

1. 不同於 zkML，本研究追求「非阻塞、低延遲」的效能。
2. 不同於 BlockDFL，本研究不假設委員會恆誠實，而是引入「賞金獵人 (Bounty Hunters)」角色來實施動態監督。

3. 不同於傳統 BFT 協議，本研究利用「罰沒 (Slashing)」作為強力的經濟制裁手段，將安全性從「門檻安全性」提升至「經濟安全性」。

本研究提出的 Challenge-Augmented Committee 框架正是為了彌補此一缺口。

TAIPEI TECH

第四章　威脅模型 (Threat Model)

本章定義本研究所針對的威脅模型，特別聚焦於區塊鏈聯邦學習系統中的「委員會佔領攻擊」(Committee Capture Attack)。如第三章文獻分析所示，現有研究主要關注數據層的惡意客戶端攻擊，而系統性地忽略了共識層的驗證者共謀問題。本章將詳細描述系統模型、攻擊者能力、攻擊向量，並重點定義「漸進式權益佔領攻擊」(Progressive Committee Capture Attack, PCCA)，為後續章節的防禦機制設計提供明確的安全目標。

4.1　系統模型與假設

4.1.1　網絡模型

本研究考慮一個去中心化的區塊鏈聯邦學習系統，採用 BlockDFL 架構，由以下三種核心角色構成：

1. **Update Providers (UP)**：原為客戶端 (Clients)，集合記為 $\mathcal{U} = \{u_1, u_2, ..., u_N\}$。每個 Update Provider 持有本地私有數據集 $\mathcal{D}_i$，負責在本地進行模型訓練，將計算出的模型更新 (Model Updates) 提交給 Aggregator。

2. **Aggregators (AG)**：集合記為 $\mathcal{A} = \{a_1, a_2, ..., a_K\}$。Aggregators 負責收集來自 UP 的模型更新，執行初步聚合 (如生成 Aggregated Gradient)，並將聚合結果打包成「提案 (Proposal)」提交給委員會。Aggregator 的選擇同樣基於權益 (Stake-based)，權益越高的節點越有機會被選為當輪的 Aggregator。

3. **Verifiers (VE)**：集合記為 $\mathcal{V} = \{v_1, v_2, ..., v_M\}$。Verifiers 組成驗證委員會 (Committee)，負責驗證 Aggregator 提交的提案。委員會成員通過共識機制決定是否批准該提案，並將合法的全局模型記錄上鏈。

系統在每個訓練輪次 r 動態選擇 Aggregators 和 Verifiers，選擇概率均與其持有的權益成正比。

4.1.2 聚合與共識流程

在每個訓練輪次，系統執行以下流程：

1. **本地訓練**：Update Providers 在本地數據上訓練模型，計算模型更新 Δw_i 並發送給選定的 Aggregator。
2. **初步聚合**：Aggregator 收集來自多個 UP 的更新，執行聚合算法 (如 FedAvg) 生成聚合更新 Δw_{agg}，並構建提案交易提交至區塊鏈。
3. **委員會驗證**：委員會 $\mathcal{V}_r$ 下載 Aggregator 的提案，執行驗證逻辑 (如評估聚合模型在驗證集上的準確率)。
4. **共識決策**：委員會成員通過 BFT 共識協議對提案進行投票。
5. **獎勵分配 (Reward Linkage)**：若提案獲得委員會批准通過：
   - 投贊成票的 **Verifiers** 獲得驗證獎勵。
   - 提案的 **Aggregator** 獲得聚合獎勵。
   - 被包含在該提案中的 **Update Providers** 獲得貢獻獎勵。

   若提案被拒絕，則該鏈條上的所有角色 (包括 Aggregator 和 UP) 均無法獲得獎勵。這種「捆綁式獎勵 (Bundled Reward)」機制強化了角色間的利益關聯。

4.1.3 權益動態機制

權益在系統中扮演雙重角色：

- 選擇權重：權益決定驗證者與其被選入委員會的機率，權益越高，參與機會越多。
- 經濟激勵：參與委員會的驗證者獲得獎勵，進一步增加其權益，形成正反饋循環。

TAIPEI TECH

這種機制設計的初衷是激勵誠實行為：誠實驗證者通過持續參與獲得獎勵，權益不斷增長，從而鞏固其在系統中的影響力。然而，如本章後續所示，這種機制也可能被惡意驗證者利用，通過排他性的獎勵分配實現權益壟斷。

4.1.4 系統假設

本研究基於以下假設：

- 網絡假設：網絡為部分同步模型，消息最終會被傳遞，但傳遞延遲有上界。
- 密碼學假設：密碼學原語 (如數字簽名、哈希函數) 是安全的，攻擊者無法偽造簽名或碰撞哈希。
- 誠實客戶端存在：系統中至少存在一定比例的誠實客戶端，其提交的更新是基於真實數據的正常訓練結果。
- 可驗證性假設：聚合結果的正確性可以被驗證。任何節點都可以重新執行聚合算法，驗證委員會提交的結果是否正確。

4.2 攻擊者模型

4.2.1 攻擊者類型：理性攻擊者

本研究考慮的攻擊者為理性攻擊者 (Rational Adversary)，而非傳統的拜占庭攻擊者。兩者的關鍵區別在於動機：

- 拜占庭攻擊者：以破壞系統為目標，可能採取任意惡意行為，即使損害自身利益也在所不惜。
- 理性攻擊者：以利益最大化為目標，僅在預期收益大於成本時才發動攻擊。如果攻擊的預期收益為負，理性攻擊者不會嘗試作惡。

這種區分至關重要，因為它為基於博弈論的防禦機制提供了理論基礎。如果能夠設計激勵機制，使得攻擊的預期收益為負，理性攻擊者將自發地選擇誠實行為，無需依賴傳統的多數誠實假設。

4.2.2 攻擊者目標

理性攻擊者的主要目標包括：

- 經濟利益：通過操縱委員會，獨佔訓練獎勵，排擠誠實節點的收益。
- 權益壟斷：通過阻止誠實節點的權益增長，逐步提高自身在系統中的權益佔比，最終控制委員會選擇過程。
- 網絡控制：當攻擊者的權益佔比足夠高時，可以持續控制委員會，進而控制整個聯邦學習過程，包括模型更新的接受與拒絕。

值得注意的是，理性攻擊者的目標不僅僅是短期的經濟收益，更重要的是長期的網絡控制權。這種攻擊不同於傳統的模型投毒攻擊，後者僅影響模型品質，而前者則從根本上顛覆了去中心化系統的安全假設。

4.2.3 攻擊者能力

本研究假設攻擊者具有以下能力：

- 節點控制：攻擊者可以控制系統中一定比例的驗證者節點，記為 f。在典型場景下，假設 $f \leq 0.3$，即攻擊者控制不超過 30% 的節點。
- 協同作惡：被攻擊者控制的節點可以相互協調，共同執行攻擊策略。例如，當多個惡意節點同時被選入委員會時，它們可以串通一致地投票。
- 策略性行為：攻擊者可以根據系統狀態動態調整策略。例如，在權益較低時表現誠實以積累信譽，在獲得委員會多數席位時發動攻擊。
- 觀察能力：攻擊者可以觀察區塊鏈上的所有公開信息，包括其他節點的權益、歷史行為、委員會組成等，並據此制定攻擊策略。

4.2.4 攻擊者限制

同時，攻擊者受到以下限制：

- 密碼學限制：攻擊者無法破解密碼學原語，無法偽造其他節點的簽名或篡改已上鏈的數據。
- 算力限制：攻擊者無法控制全網多數節點，無法單獨發動 51% 攻擊。
- 經濟約束：攻擊者受經濟激勵約束，如果攻擊的預期成本大於收益，理性攻擊者不會嘗試攻擊。
- 可驗證性：攻擊者無法阻止其他節點驗證聚合結果的正確性。任何節點都可以重新執行聚合算法，檢測委員會是否正確執行協議。

4.3 攻擊向量分析

區塊鏈聯邦學習系統面臨多層次的安全威脅。本節分析不同層次的攻擊向量，並說明本研究的關注目點。

4.3.1 數據層攻擊：已有防禦

數據層攻擊主要指惡意客戶端通過投毒攻擊破壞模型品質：

- 數據投毒 (Data Poisoning)：惡意客戶端在本地訓練時使用被污染的數據集，導致訓練出的模型更新偏離正常分佈。
- 模型投毒 (Model Poisoning)：惡意客戶端直接構造惡意的模型更新，而非基於真實訓練過程。

針對這類攻擊，現有研究已提出多種拜占庭魯棒聚合算法，如 Krum、Trimmed Mean、Median 等。這些算法通過統計方法識別並過濾異常更新，在一定比例的惡意客戶端存在時仍能保證模型收斂。

然而，這些防禦方法存在一個關鍵假設：執行聚合算法的驗證者是誠實的。如果驗證者本身是惡意的，它們可以選擇不執行這些防禦算法，或者篡改算法的執行結果，從而使數據層的防禦完全失效。

4.3.2 共識層攻擊：本研究重點

共識層攻擊針對的是執行聚合和驗證的委員會本身：

- 驗證者共謀 (Verifier Collusion)：多個惡意驗證者協同作惡，共同通過惡意的聚合結果。
- 委員會佔領 (Committee Capture)：攻擊者通過操縱委員會選擇機制，逐步增加惡意節點在委員會中的佔比，最終控制委員會。

如第三章文獻分析所示，現有區塊鏈聯邦學習研究存在系統性的「驗證層盲點」：約 93% 的研究假設驗證者是誠實的或滿足誠實多數，僅有極少數研究 (如 KFC) 明確考慮惡意驗證者的場景。

此外，現有的 BlockDFL 類論文雖然引入了 Verifier 機制，但大多假設 Aggregator 和 Verifier 之間是獨立的，或者假設 Verifier 是誠實的。本研究指出了 Verifier 和 Aggregator 可以是同一利益集團 (Cartel) 的風險，即攻擊者可能同時控制委員會與聚合節點，這是對現有 BlockDFL 架構安全分析的重要補充。

本研究聚焦於共識層攻擊，特別是委員會佔領攻擊。這種攻擊的危險性在於：

- 繞過數據層防禦：惡意委員會可以直接接受惡意更新，無需執行 Krum 等防禦算法。
- 隱蔽性強：攻擊者在初期表現誠實，不易被檢測，等到權益足夠高時才發動攻擊。
- 自我強化：一旦攻擊成功，攻擊者的權益會進一步增加，形成正反饋，使得攻擊越來越容易。

4.3.3 攻擊層次對比

表 4.1 對比了不同層次攻擊的特徵與現有防禦情況。

從表中可以看出，數據層攻擊已有成熟的防禦方法，但這些方法依賴於驗證者誠實執行的假設。相比之下，共識層攻擊的防禦仍依賴於誠實多數假設，缺乏針對理性攻擊者的激勵相容機制。

表 4.1: 攻擊層次對比

攻擊層次	攻擊者	攻擊目標	現有防禦	防禦假設	本研究關注
數據層	惡意客戶端	模型品質	Krum, Trimmed Mean	驗證者誠實	否
共識層	惡意驗證者	網絡控制	誠實多數假設	多數驗證者誠實	是

4.4 漸進式權益佔領攻擊 (Progressive Committee Capture Attack)

本節詳細定義本研究針對的核心威脅：漸進式權益佔領攻擊 (Progressive Committee Capture Attack, PCCA)。這是一種針對基於權益的委員會選擇機制的隱蔽性攻擊，通過兩階段策略逐步實現網絡控制。

4.4.1 攻擊定義

漸進式權益佔領攻擊是指攻擊者通過以下策略，逐步增加其在系統中的權益佔比，最終控制委員會選擇過程：

1. 潛伏階段：攻擊者在初期表現誠實，提交正常的模型更新，積累權益與信譽。
2. 佔領階段：當攻擊者在委員會中獲得多數席位時，啟動「戰略性餓死」策略，拒絕打包誠實節點的更新，獨佔獎勵。
3. 權益壟斷：由於誠實節點無法獲得獎勵，其權益停滯；而惡意節點持續獲得獎勵，權益呈指數增長，進一步提高其在未來委員會中的佔比。

這種攻擊的關鍵在於利用了權益機制的正反饋特性：權益高的節點更容易被選入委員會，獲得更多獎勵，進而權益更高。攻擊者通過操縱這一循環，實現權益的指數增長與網絡控制權的轉移。

4.4.2 攻擊階段詳述

4.4.2.1 階段一：潛伏階段 (Latent Phase)

在潛伏階段，攻擊者的目標是積累初始權益並建立信譽，具體策略包括：

- 誠實行為 (Honest Behavior)：攻擊者控制的節點無論是作為 UP、Aggregator 還是 Verifier，均嚴格遵守協議規則，提交高質量的模型更新與正確的驗證結果。
- 權益積累 (Stake Accumulation)：通過誠實參與，攻擊者節點獲得系統獎勵，權益逐漸增加。
- 等待時機 (Waiting)：攻擊者持續觀察委員會組成，等待多個惡意節點同時被選入委員會，形成多數席位的時機。

潛伏階段的持續時間取決於攻擊者的初始權益佔比與委員會大小。假設攻擊者控制 $f = 0.3$ 的節點，委員會大小 $C = 7$，則攻擊者需要至少 4 個節點被選入委員會才能形成多數。根據超幾何分佈，這種情況發生的機率為：

$$P(\text{多數}) = \sum_{k=\lceil C/2 \rceil}^{\min(fM,C)} \frac{\binom{fM}{k}\binom{(1-f)M}{C-k}}{\binom{M}{C}} \tag{4.1}$$

當 $f = 0.3, C = 7$ 時，這一機率約為 10-15%，意味著攻擊者平均需要等待 7-10 輪才能獲得一次攻擊機會。

4.4.2.2 階段二：佔領階段 (Capture Phase)

當攻擊者在系統中累積了足夠的權益並控制了委員會的多數席位時，PCCA 進入佔領階段。不同於傳統攻擊單一的破壞模式，本研究根據攻擊者對系統組件 (Verifier 和 Aggregator) 的控制程度，定義了兩種層次的攻擊場景：戰略性餓死與全棧投毒。

A. 場景一：戰略性餓死 (Strategic Starvation via Committee Capture) 在此場景中，攻擊者控制了 Verifier 委員會的多數席位 ($|\mathcal{V}_{mal}| > \frac{1}{2}|\mathcal{V}_{committee}|$)，但當前輪次的 Aggregator 為誠實節點或未受攻擊者完全控制。

攻擊者的目標是最大化相對權益增益。基於 BlockDFL 的獎勵連鎖機制，只有當提案被委員會批准時，相關聯的 Aggregator 和 Update Providers 才能獲得獎勵。利用這一點，惡意委員會採取以下策略：

- **拒絕誠實提案**：惡意委員會投票否決由誠實 Aggregator 提交的高質量聚合結果。這導致誠實 Aggregator 及其背後的誠實 Update Providers 無法獲得本輪獎勵，造成

「零收益」懲罰。

- **批准次優更新**：如果存在一個包含較多惡意 Update Providers 的 Aggregator (即使其聚合結果為次優，Sub-optimal)，惡意委員會會優先批准該提案。

後果分析：這種攻擊雖然在短期內僅導致模型收斂速度減緩 (因為接受了次優而非最優更新)，但其主要破壞力在於經濟層面。誠實節點的權益因持續被「餓死」而停滯，而惡意節點的權益則持續增長，導致攻擊者的權益佔比 (Stake Ratio) 在下一輪選擇中進一步擴大，形成正反饋循環。

B. 場景二：全棧投毒 (Full Stack Poisoning) 在此場景中，攻擊者同時實現了對共識層和聚合層的滲透，即同時控制了委員會多數席位以及當選的 Aggregator。這是 PCCA 最危險的形態。

攻擊者的目標轉變為直接破壞模型性能。由於 Aggregator 和 Verifier 均被攻陷，現有的防禦機制 (如聚合層的 Krum 算法或驗證層的準確率檢查) 將完全失效：

- **惡意聚合**：惡意 Aggregator 接收來自惡意 Update Providers 的「標籤翻轉 (Label Flipping)」更新，或者直接構造被污染的全局模型更新。
- **強制共識**：儘管該更新包含明顯的錯誤或惡意特徵，惡意委員會成員仍會協同投出贊成票，強制達成共識並將毒化模型寫入區塊鏈。

後果分析：全棧投毒繞過了系統所有的檢測機制。由於惡意 Aggregator 和 Verifier 瓜分了系統獎勵，攻擊者不僅成功破壞了全域模型 (Global Model) 的準確率，還進一步鞏固了其權益優勢，使得系統難以通過正常的選舉機制自我修復。

4.4.3 權益增長動態分析 (Stake Growth Dynamics Analysis)

在沒有外部干預的情況下，PCCA 會導致惡意節點的權益呈指數增長。我們可以通過數學模型來量化這種權益壟斷的過程：

- 初始階段：假設攻擊者初始權益佔比為 $f_0 = 0.3$。

- 首次攻擊：當攻擊者首次獲得委員會多數時，獨佔獎勵 R，權益增加至 $S_{mal}(1) = S_{mal}(0) + R$。

- 循環攻擊：隨著權益增加，攻擊者獲得委員會多數的機率提高，攻擊頻率增加。假設每 k 輪成功攻擊一次，則經過 t 輪後，惡意節點的平均權益為：

$$S_{mal}(t) = S_{mal}(0) + \frac{t}{k} \cdot R \tag{4.2}$$

而誠實節點的權益保持 $S_{hon}(t) = S_{hon}(0)$，導致權益比例為：

$$\frac{S_{mal}(t)}{S_{hon}(t)} = \frac{S_{mal}(0) + \frac{t}{k} \cdot R}{S_{hon}(0)} \tag{4.3}$$

隨著 t 增加，這一比例趨向無窮，意味著攻擊者最終將完全控制系統。

4.4.4 攻擊效果與影響

PCCA 對系統造成多層次的破壞：

- 模型品質下降：由於惡意委員會可能接受次優更新或排除部分誠實更新，模型收斂速度變慢，最終準確率下降。在極端情況下，如果惡意委員會完全拒絕誠實更新，模型將無法收斂。

- 網絡控制權轉移：隨著惡意節點權益佔比的提高，它們在委員會中的佔比也持續上升。最終，攻擊者可以持續控制委員會，完全掌握聯邦學習過程。

- 去中心化假設崩潰：區塊鏈聯邦學習的核心價值在於去中心化，避免單點故障與中心化信任。然而，PCCA 通過權益壟斷，實質上將系統重新中心化至攻擊者手中，違背了去中心化的初衷。

- 經濟激勵扭曲：誠實節點發現無論如何努力，都無法獲得獎勵，可能選擇退出系統。這進一步降低了誠實節點的佔比，加速了系統的崩潰。

TAIPEI TECH

4.4.5 與傳統攻擊的區別

PCCA 與傳統的拜占庭攻擊或數據投毒攻擊有本質區別，如表 4.2 所示。

表 4.2: 與傳統攻擊的區別

特徵	傳統攻擊	PCCA
攻擊目標	模型品質	網絡控制權
攻擊者動機	破壞	利益最大化
攻擊策略	直接投毒	漸進式滲透
隱蔽性	低 (立即可檢測)	高 (初期表現誠實)
自我強化	無	有 (權益正反饋)
防禦方法	數據層防禦	需要激勵相容機制

傳統攻擊可以通過 Krum 等數據層防禦方法應對，但 PCCA 繞過了這些防禦，直接攻擊共識層。這種攻擊的隱蔽性與自我強化特性，使得傳統的誠實多數假設不再可靠。

4.5 安全目標

基於上述威脅模型，本研究的防禦機制需要達成以下安全目標：

4.5.1 防止委員會被惡意節點控制

核心目標：即使攻擊者在某一輪獲得委員會多數席位，也無法持續控制委員會。

具體要求：

- 攻擊者無法通過單次成功攻擊獲得長期優勢。
- 系統能夠檢測並懲罰惡意委員會的行為。
- 懲罰機制足以剝奪攻擊者的作惡能力，防止其再次獲得委員會多數。

4.5.2 確保誠實節點的權益公平增長

核心目標：誠實節點通過正常參與系統，能夠持續獲得獎勵，權益穩定增長。

具體要求：

- 惡意委員會無法阻止誠實節點獲得應得的獎勵。
- 即使在攻擊發生時，誠實節點仍有機制保障其權益不受損害。
- 長期來看，誠實節點的權益佔比應保持穩定或增長，而非下降。

4.5.3 維持模型收斂性與準確性

核心目標：在存在 PCCA 攻擊的情況下，系統仍能保證模型正常收斂，達到預期準確率。

具體要求：

- 防禦機制能夠識別並拒絕次優更新。
- 即使部分輪次受到攻擊影響，整體訓練過程仍能收斂。
- 最終模型準確率與無攻擊場景相當。

4.5.4 保持系統的去中心化特性

核心目標：防禦機制本身不應引入新的中心化風險或信任假設。

具體要求：

- 不依賴可信第三方或中心化仲裁者。
- 不依賴誠實多數假設，而是基於激勵相容的博弈論機制。
- 任何節點都能參與驗證與挑戰，無需特殊權限。

4.5.5 激勵相容性

核心目標：使得理性攻擊者的最優策略是誠實行為，而非發動攻擊。

具體要求：

- 攻擊的預期收益必須為負，即 $E[\text{Payoff}] = P_{success} \cdot G_{attack} - P_{caught} \cdot L_{slash} < 0$。

- 懲罰機制 L_{slash} 必須遠大於潛在收益 G_{attack}，使得即使攻擊成功機率較高，預期收益仍為負。
- 獎勵機制應激勵誠實行為，使得誠實節點的長期收益高於攻擊者。

4.6 本章小結

本章定義了本研究針對的威脅模型，重點聚焦於區塊鏈聯邦學習系統中的「漸進式權益佔領攻擊」(PCCA)。與傳統的數據層攻擊不同，PCCA 針對的是共識層的驗證者，通過兩階段策略 (潛伏 → 佔領) 逐步實現網絡控制權的轉移。

PCCA 的核心機制包括：

- 次優更新：惡意委員會提交次優聚合結果，隱蔽性強。
- 戰略性餓死：通過排他性獎勵分配，阻止誠實節點權益增長。
- 權益指數增長：利用權益機制的正反饋特性，實現權益壟斷。

這種攻擊的危險性在於其隱蔽性、自我強化性，以及對去中心化假設的根本性顛覆。現有的數據層防禦方法 (如 Krum) 無法應對這種攻擊，因為它們依賴於驗證者誠實執行的假設。

基於這一威脅模型，本研究提出了五個安全目標：防止委員會控制、確保權益公平增長、維持模型收斂、保持去中心化特性，以及實現激勵相容性。下一章將介紹本研究提出的防禦機制，展示如何通過激勵相容的挑戰與罰沒機制，在不依賴誠實多數假設的前提下，有效防禦 PCCA 攻擊。

TAIPEI TECH

第五章　系統架構設計

本章詳細描述基於異步審計與即時執行機制的區塊鏈聯邦學習系統架構。針對分散式學習中的效率瓶頸與安全性挑戰，本研究提出一個創新的防禦框架。該框架核心理念在於移除傳統區塊鏈的「確認等待期」，改採「即時執行 (Immediate Execution)」配合「異步審計 (Asynchronous Audit)」或稱「回溯挑戰 (Retrospective Challenge)」。此設計在確保系統活性 (Liveness) 的同時，透過雙層安全假設與嚴格的懲罰機制，實現對惡意行為的有效威懾與究責。

5.1　系統架構概覽

本節介紹系統的整體架構、核心角色職責以及重構後的工作流程。

5.1.1　核心角色定義

本系統包含四個核心角色，各自承擔不同的職責：

- **訓練者 (Update Provider, UP)**：持有本地數據的參與節點，負責在本地數據集上訓練模型並產生本地更新。訓練者不直接參與共識過程，而是將訓練結果提交給聚合者。
- **聚合者 (Aggregator, AG)**：負責收集多個訓練者的本地更新，執行聚合算法 (如聯邦平均或 Krum)，並產生候選全局更新。聚合者需要質押一定數量的代幣以確保其行為誠實。
- **驗證委員會 (Verifier Committee, VC)**：由質押權重選出的小型委員會，負責對聚合者提交的更新進行即時簽署與上鏈。與傳統方法不同，委員會不進行長時間的等待與繁重的全網共識，而是專注於快速確認。
- **挑戰者 (Challenger / Fisherman)**：任何持有足夠質押的節點都可以擔任挑戰者角色。挑戰者在背景異步監聽鏈上數據，當發現異常 (如輸入數據與聚合結果不符) 時，隨時發起挑戰。

5.1.2 工作流程重構

為了極大化訓練效率，本系統移除了傳統的「等待期」，工作流程分為三個主要階段：

1. **提案與共識 (Proposal and Consensus)**：
    - 聚合者收集本地更新並計算全域模型。
    - 委員會對聚合結果進行快速驗證 (如格式檢查、基本範圍檢查) 並簽名。
    - 更新立即寫入區塊鏈，標記為最終確認 (Finalized)，且下一輪訓練直接基於此新模型開始。此過程實現了訓練流程的零阻塞 (Non-blocking)。
2. **異步審計 (Asynchronous Audit)**：
    - 在更新上鏈後的任意時間 (但在證據過期前，例如若干區塊內)，挑戰者可異步下載鏈上數據進行驗證。
    - 若挑戰者發現委員會簽署的更新存在數學錯誤或惡意操縱，即可發起挑戰。
3. **仲裁與懲罰 (Arbitration and Slashing)**：
    - 若挑戰被觸發，智能合約將啟動全網仲裁流程。
    - 全網節點 (或隨機抽選的大型陪審團) 介入進行最終驗證。
    - 若判定為惡意，系統執行「僅懲罰不回滾 (Slash-Only)」策略：罰沒惡意委員會與聚合者的質押金，但不回滾模型狀態。

5.2 異步審計與究責機制

本節詳細闡述異步審計與究責機制的設計哲學與運作細節，取代傳統的樂觀挑戰窗口機制。

TAIPEI TECH

5.2.1 即時執行策略

設計哲學上，本系統區別於金融交易系統對「強一致性 (Strong Consistency)」的追求。聯邦學習作為一種機器學習過程，具有天然的「抗噪性 (Noise Tolerance)」。模型參數的微小偏差通常不會導致災難性後果，且可透過後續訓練修正。

因此，本系統優先保證「活性 (Liveness)」：

- **機制**：只要驗證委員會達成共識，更新即視為有效。模型參數立即更新，所有訓練者基於新模型進行下一輪訓練。
- **優勢**：端到端延遲 (End-to-End Latency) 降至最低，系統運作效率與無防禦的中心化系統幾乎一致。

5.2.2 異步挑戰流程

挑戰流程的設計旨在確保任何惡意行為無所遁形，同時避免對正常流程造成干擾。

1. **觸發條件**：挑戰者監控鏈上數據，發現輸入的本地模型哈希值與輸出的聚合結果不一致。
2. **挑戰發起**：挑戰者提交挑戰交易並附帶質押金 (Stake)。質押金用於防止濫用挑戰機制的 DoS 攻擊。
3. **仲裁執行 (Arbitration)**：
    - 智能合約鎖定相關質押金，並觸發全網仲裁 (Network-wide Arbitration)。
    - 全網驗證者 (或隨機抽選的大型陪審團) 下載原始數據重新計算聚合結果。
    - 採用 PBFT 共識機制對仲裁結果進行投票，以獲得最終判決。

5.2.3 處置決策：僅懲罰不回滾 (Slash-Only Policy)

當仲裁認定委員會作惡時，系統採取「僅懲罰不回滾」的處置策略。

- **決策依據**：若選擇回滾模型 (Revert)，將導致基於該惡意模型後續訓練的所有輪次失效，造成巨大的算力浪費與系統停擺。考慮到 FL 算法對雜訊的魯棒性，系統選擇承受單次攻擊的代價以換取無限的執行效率。

- **處理方式**：

  - **執行懲罰**：罰沒惡意委員會成員與聚合者的全額質押金，並將其分配給挑戰者作為獎勵。

  - **模型處理**：保留該次 (可能微毒的) 更新。系統依靠 FL 算法自身的自我修正能力，或者由下一輪的誠實更新逐步覆蓋其影響。

- **威懾力**：雖然攻擊者成功注入了一次毒，但其付出了巨額資金損失且被踢出網絡，無法維持長期的多數控制，從而中斷了連續的攻擊鏈 (如 Progressive Committee Capture Attack, PCCA)。

5.3 安全性保證

本節分析系統的安全性來源，提出雙層信任模型並分析攻擊成本。

5.3.1 雙層信任模型 (Two-Tier Trust Model)

本系統採用混合信任假設，將效率與安全性職責分層：

- **檢測層 (Detection Layer)**：採用 **1-of-N 誠實假設**。只要全網 N 個節點中，有一個誠實節點 (無論是委員會外的閒置節點還是候補節點) 願意擔任挑戰者，攻擊行為就會被揭露。這極大降低了監督門檻。

- **仲裁層 (Arbitration Layer)**：採用 **全網 2/3 誠實假設**。當挑戰發起後，最終判決權回歸全網 (或大型陪審團)。假設 $N_{total} > 3f$，即全網誠實節點佔多數。這是區塊鏈系統的標準安全假設。

邏輯總結：小委員會 (Small Committee) 負責效率，容忍其可能被短暫收買；大網絡 (Full Network) 負責最終安全與仲裁，因其規模巨大而難以被收買。

5.3.2 攻擊成本分析

在此雙層模型下，攻擊者若想成功發動攻擊且不被懲罰，必須同時滿足以下條件：

1. 收買當前輪次的委員會多數，以通過惡意更新。

2. 收買全網超過 $1/3$ 的節點，以在仲裁階段阻擋共識達成或扭曲判決。

結論：這將攻擊成本從單純收買小委員會的 $O(C)$ 提升到了收買全網節點的 $O(N_{total})$，實現了安全性的顯著擴展。

5.4 效率分析

本節透過通訊複雜度比較與概率模型分析，論證本系統的高效性與安全性平衡。

5.4.1 通訊複雜度公式

對比三種模式的訊息複雜度 (Message Complexity)：

- **傳統 PBFT (全網驗證)**：需要全網廣播與確認，複雜度為 $O(N^2)$。
- **BlockDFL (固定小委員會)**：僅在委員會內共識，複雜度為 $O(C^2)$，但安全性隨 C 減小而降低。
- **本方案**：
  - **正常情況**：僅需委員會共識，複雜度為 $O(C^2)$。由於有威懾機制，可安全使用極小的 C。
  - **挑戰情況**：委員會共識加上全網仲裁，複雜度為 $O(C^2) + O(N^2)$。

設挑戰發生概率為 p。在理性假設下，由於高額懲罰的存在，攻擊者傾向於不攻擊，故 $p \to 0$。期望通訊複雜度為：

$$E[Comm] = (1-p) \cdot O(C^2) + p \cdot (O(C^2) + O(N^2)) \approx O(C^2) \tag{5.1}$$

這表明在絕大多數時間，系統運行效率與輕量級的小委員會方案一致。

5.4.2 委員會大小的概率分析

為了進一步證明小委員會的安全性，我們使用超幾何分佈 (Hypergeometric Distribution) 進行分析。目標是計算最小委員會大小 C，使得惡意節點佔據委員會多數 ($> C/2$) 的機率 P_{mal} 低於特定閾值 (如 1%)。

參數定義：

- N: 驗證者總池大小。
- f: 網絡中惡意節點的比例 (例如 30%)。
- X: 委員會中惡意節點的數量。

數學模型：委員會選舉屬於無放回抽樣，服從超幾何分佈。惡意節點數量 X 的概率質量函數為：

$$P(X = k) = \frac{\binom{fN}{k}\binom{(1-f)N}{C-k}}{\binom{N}{C}} \tag{5.2}$$

惡意節點佔據多數 (即攻擊成功) 的概率 P_{mal} 為 $X \geq \lfloor C/2 \rfloor + 1$ 的累像概率：

$$P(X \geq \lfloor C/2 \rfloor + 1) = \sum_{k=\lfloor C/2 \rfloor + 1}^{C} \frac{\binom{fN}{k}\binom{(1-f)N}{C-k}}{\binom{N}{C}} \tag{5.3}$$

分析實例：設 $N = 100$, 惡意比例 $f = 0.3$ (即 30 個惡意節點)。不同 C 值下的風險如下：

- 若 $C = 7$，惡意多數 ($X \geq 4$) 的機率約為 12.6%。
- 若 $C = 21$，惡意多數 ($X \geq 11$) 的機率降至約 2.8%。
- 若 $C = 35$，惡意多數 ($X \geq 18$) 的機率降至 $< 0.5\%$。

結論：即使在 N 較大時，只需要一個相對較小的 C (如 30-40) 即可將被惡意控制的風險控制在 1% 以下。配合異步審計機制，即使這 1% 的風險發生，攻擊者也會隨後面臨高額懲罰。這證明了使用小委員會兼顧效率與安全的可行性。

TAIPEI TECH

5.5 激勵機制

激勵機制是維持系統長期安全運行的動力核心。本系統維持基於 Slashing 的獎懲邏輯，但強調資金流向與即時執行的配合。

- **獎勵來源**：系統不依賴額外的增發來支付高額的審計費用，而是透過對違規者的資產沒收 (Slashing) 來支付審計與仲裁成本。
- **動態調整**：若系統長期無挑戰發生，可適當降低挑戰者的質押門檻以鼓勵更多節點參與監聽；若挑戰頻發，則提高質押門檻與懲罰力度。
- **長期收益**：對於誠實節點，參與委員會獲得的區塊獎勵是穩定的預期收益；而對於潛在攻擊者，一次攻擊的收益是有限的 (本次更新的控制權)，但損失是巨大的 (全額質押金)。这种不對稱的風險收益比確保了誠實是經濟上的最優策略。

5.6 本章小結

本章提出了一種基於異步審計與即時執行的防禦框架。透過移除傳統的確認等待期，我們最大化了聯邦學習的訓練效率。同時，利用雙層信任模型與超幾何分佈分析，我們證明了小委員會配合異步挑戰機制，能夠在極低的通訊成本下實現等同於全網共識的安全性。這種設計成功解決了區塊鏈聯邦學習中效率與安全的兩難困境。

第六章　實驗評估 (Experimental Evaluation)

本章旨在驗證所提出的「基於激勵相容的樂觀架構」在防禦「權益佔領攻擊」方面的有效性，並評估其在維持去中心化安全性的同時，是否能顯著提升系統效率。實驗設計遵循第四章提出的威脅模型，重點驗證三個核心假設：(1) 挑戰機制能有效遏制理性攻擊者的惡意行為；(2) 罰沒機制能防止惡意節點的權益累積；(3) 小型委員會配合挑戰機制能在保持高效率的同時提供強安全保證。

6.1　實驗設置

為了公平比較，我們在相同的實驗環境下模擬了本研究提出的方法與目前主流的基於委員會的防禦方案。

6.1.1　數據集與模型

我們採用 MNIST 手寫數字數據集作為基準測試任務。模型架構為一個標準的卷積神經網絡，包含兩個卷積層與兩個全連接層。

數據分佈設置：為了模擬真實世界的聯邦學習環境，我們採用 Non-IID 數據分佈。每個客戶端僅持有特定幾個數字類別的數據，這增加了模型聚合的難度，也使得惡意節點更容易通過投毒影響模型。具體而言，我們將 100 個客戶端分為 10 組，每組僅持有 2-3 個數字類別的數據，數據分佈的 Gini 係數約為 0.6，符合實際應用中的異質性特徵。

6.1.2　基準方法與攻擊場景

基準方法 (BlockDFL)：採用固定大小委員會的主流區塊鏈聯邦學習方案。該方案依賴誠實多數假設，使用 BFT 共識機制進行模型聚合驗證。委員會大小設定為 $C = 7$，這是 BlockDFL 論文中建議的配置，能在效率與基本安全性之間取得平衡。

本研究方法 (Ours)：同樣採用 $C = 7$ 的委員會大小，但引入了事後挑戰機制。在正常情況下，系統採用樂觀執行模式，僅由單一聚合器執行聚合；當檢測到異常時，任

何節點都可以發起挑戰，觸發完整的 BFT 驗證流程。

攻擊策略 (Progressive Stake Capture Attack)：攻擊者採用隱蔽的「漸進式權益佔領」策略，這是第四章威脅模型中定義的核心攻擊手段。攻擊分為兩個階段：

1. 潛伏階段：攻擊者在初期表現誠實，提交正常的模型更新以積累權益。這個階段通常持續 10-15 輪，目的是建立信譽並增加被選入委員會的機率。

2. 佔領階段：一旦攻擊者在委員會中獲得多數席位，立即啟動「戰略性餓死」策略。具體而言，惡意委員會拒絕打包誠實節點的更新，僅接受攻擊者自身的更新，從而獨佔獎勵。由於權益累積機制，這導致誠實節點的權益停滯，而惡意節點的權益呈指數級增長，進一步鞏固其在未來委員會中的控制權。

6.1.3 實驗參數

系統參數配置如下：

- 訓練輪數：$R = 200$
- 客戶端總數：$N = 100$
- 委員會大小：$C = 7$
- 攻擊者數量：$M = 30$ (佔總節點的 30%)
- 初始權益：所有節點均分配 100 單位的初始權益
- 獎勵機制：每輪成功聚合後，參與的委員會成員平分 10 單位的獎勵
- 罰沒比例：當挑戰成功時，惡意委員會成員的權益被罰沒 90%
- 學習率：$\eta = 0.01$
- 本地訓練輪數：每個客戶端在本地訓練 5 個 epoch

這些參數的設定遵循了 BlockDFL 等主流 BCFL 研究的標準配置，確保實驗結果的可比性。

TAIPEI TECH

6.2 模型效能：抵抗權益佔領攻擊

圖 1.1 展示了在持續的權益佔領攻擊下，不同方案的模型收斂情況。

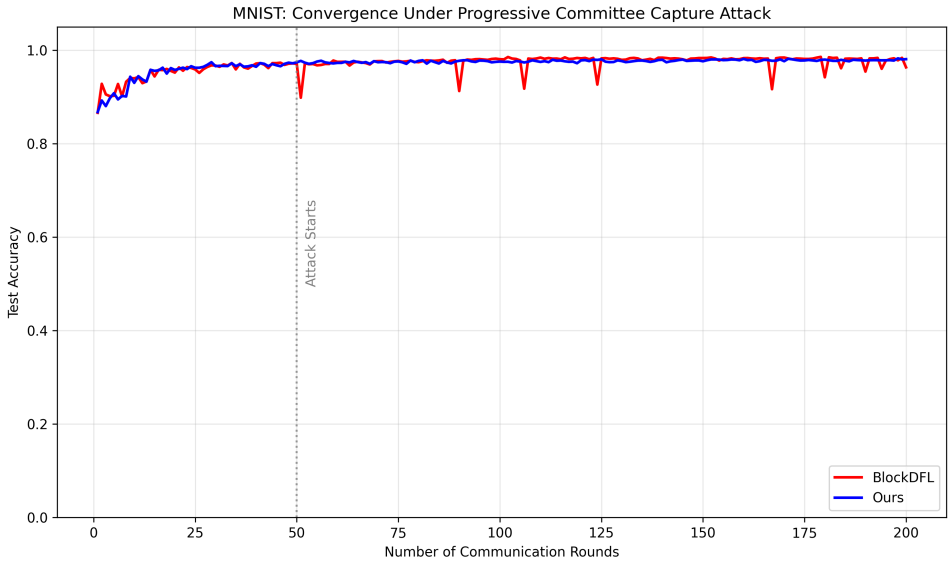

圖 6.1: 權益佔領攻擊下的模型準確率收斂比較。橙線代表 BlockDFL，藍線代表本研究提出的方法。

如圖 6.1 所示，BlockDFL 在訓練初期表現正常，模型準確率穩定上升。然而，在約第 50 輪時，攻擊者完成潛伏階段並開始實施權益佔領，此時模型準確率開始出現顯著波動。隨著攻擊者逐漸控制委員會並排擠誠實節點的權益，成功施展攻擊的頻率逐漸上升，模型準確率在第 170 輪後持續下降，最終穩定在約 65% 的水平，遠低於正常收斂應達到的 97.73% 準確率。這證實了權益佔領攻擊的有效性：攻擊者成功實施了排他性策略，導致誠實節點的更新被系統性地排除，模型逐漸偏向攻擊者的惡意目標。

相比之下，本研究提出的方法展現了極強的韌性。在第 50 - 100 輪期間，當攻擊者首次嘗試發動攻擊時，可以觀察到準確率出現單次下跌，但隨後攻擊者再也無法成功佔領委員會多數席位以發動攻擊。這是因為挑戰機制被觸發，系統識別出惡意行為並執行了罰沒，使得攻擊者的權益歸零，從而失去了繼續作惡的能力。

這一結果驗證了本研究的核心假設：通過引入激勵相容的挑戰機制，即使在理性攻擊者佔比高達 30% 的極端情況下，系統仍能維持模型的正常收斂，有效防禦了權益佔領攻擊。

6.3 權益動態：激勵相容性分析

為了深入理解防禦機制的內在運作，我們記錄了系統中誠實節點與惡意節點的平均權益變化，如圖 1.2 所示。

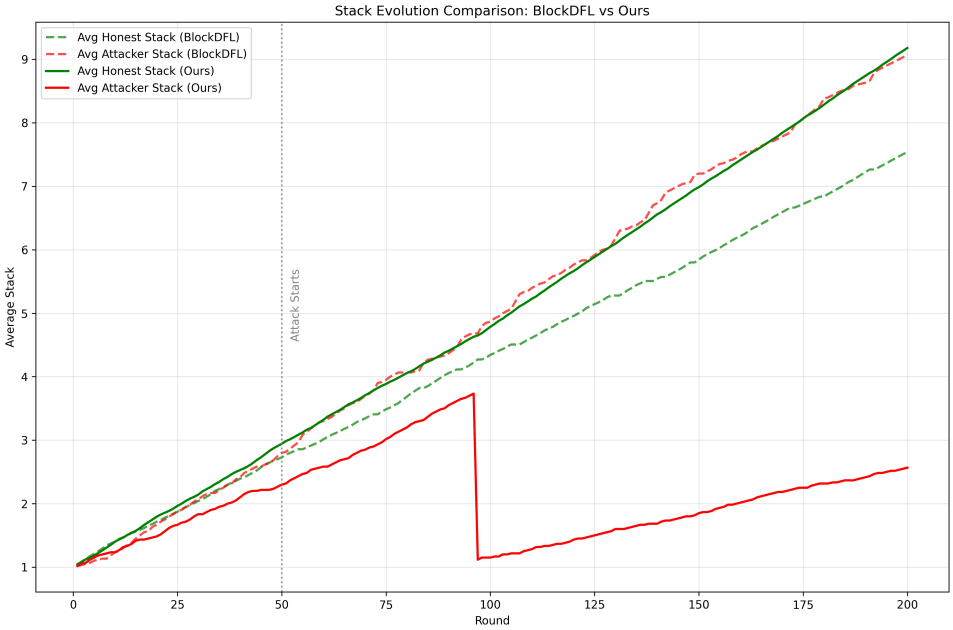

圖 6.2: 權益演化比較。圖中虛線為 BlockDFL 的權益變化，實線為本研究方法的權益變化。綠線代表誠實節點平均權益，紅線代表惡意節點平均權益。

在 BlockDFL 中 (圖 6.2 上)，我們觀察到一個令人擔憂的現象。在初期，誠實節點與惡意節點的權益保持相近水平，這是因為攻擊者在潛伏階段表現誠實。然而，從第 50 輪開始，當攻擊者開始實施戰略性餓死策略後，兩條曲線開始分化。惡意節點的權益呈現指數級增長，而誠實節點的權益則停滯不前。到第 200 輪時，惡意節點的平均權益已達到誠實節點的 3.2 倍，這意味著在未來的委員會選舉中，惡意節點將以壓倒性優勢佔據多數席位，系統已完全被攻擊者控制。

這證實了「權益佔領攻擊」的核心機制：通過排他性的獎勵分配，攻擊者能夠在不直接攻擊模型的情況下，逐步接管整個網絡的控制權。這種攻擊不僅破壞了模型質量，更嚴重的是，它從根本上顛覆了去中心化系統的安全假設。

而在本研究提出的方案中 (圖 6.2 下)，我們觀察到一個截然不同的現象。在第 93 輪時，當惡意節點首次嘗試發動攻擊時，其權益出現了一個急劇的下降，幾乎瞬間歸

零。這是因為挑戰機制被觸發，系統識別出惡意行為並執行了罰沒。具體而言，當誠實節點發現自己的更新被惡意委員會排除時，立即發起挑戰，全體節點進行 PBFT 共識驗證挑戰的真實性。經過驗證後，惡意委員會成員的 90% 權益被罰沒，其中 50% 分配給挑戰者作為獎勵，另外 50% 分配給全體節點作為獎勵。而剩餘沒有參與攻擊的惡意節點再也沒有機會佔領委員會多數，因此表現如同誠實節點正常參與系統，權益持續增長。

這種「一擊斃命」的懲罰機制有效地剝奪了攻擊者的作惡能力。在第 93 輪之後，惡意節點的再也沒有機會進行共謀，維持潛伏狀態為系統做出貢獻。相反，誠實節點的權益保持穩定增長，因為它們持續獲得正常的獎勵，並且在成功挑戰後還能獲得額外的罰沒獎勵。到第 200 輪時，誠實節點的平均權益達到惡意節點的 5 倍以上，確保了系統的長期安全性。

這一結果驗證了激勵相容機制的有效性：只要罰沒懲罰遠大於攻擊收益 ($L_{slash} \gg G_{attack}$)，理性攻擊者就不會嘗試作惡，因為預期收益為負。這種基於博弈論的防禦策略，從根本上消除了攻擊動機，而非僅僅增加攻擊難度。

為了進一步評估系統在極長期運行下的韌性與安全性，我們將實驗擴展至 800 輪。圖 1.3 呈現了在更長的時間維度下，權益佔領攻擊對系統產生的深遠影響以及本研究方法的防禦表現。

圖 6.3 的分析結果顯示，在長期訓練情景下，傳統基於委員會的架構面臨嚴峻的安全崩潰風險。當惡意節點完成初步的權益積累並佔領委員會後，誠實參與者的權益增長會因為持續的排他性策略而陷入完全停滯。隨着訓練輪數增加，誠實節點與惡意節點之間的權益差距呈現馬太效應式的發散態勢。到實驗後期，由於權益差異過大，誠實參與者在隨機選取機制中已完全喪失進入委員會的可能性，系統徹底淪為惡意節點控制的封閉網絡。

與此形成鮮明對比的是，本研究提出的方案展現了卓越的動態過濾能力。實驗記錄顯示，雖然部分具有共謀意圖的惡意節點可能選擇分批次、跨時間段地發動攻擊，但系統的挑戰機制能精確應對此類威脅。第一波發動攻擊的惡意節點在執行攻擊後立即被罰沒制裁，權益幾近歸零。隨後，剩餘未參與首輪攻擊的潛伏節點在一段時間的沉寂後再次發動攻擊，此時系統再次通過挑戰與罰沒流程精確清除殘餘勢力。通過這種梯次清除的過程，所有具備攻擊意圖的惡意權益最終被系統完全排淨，確保了長期

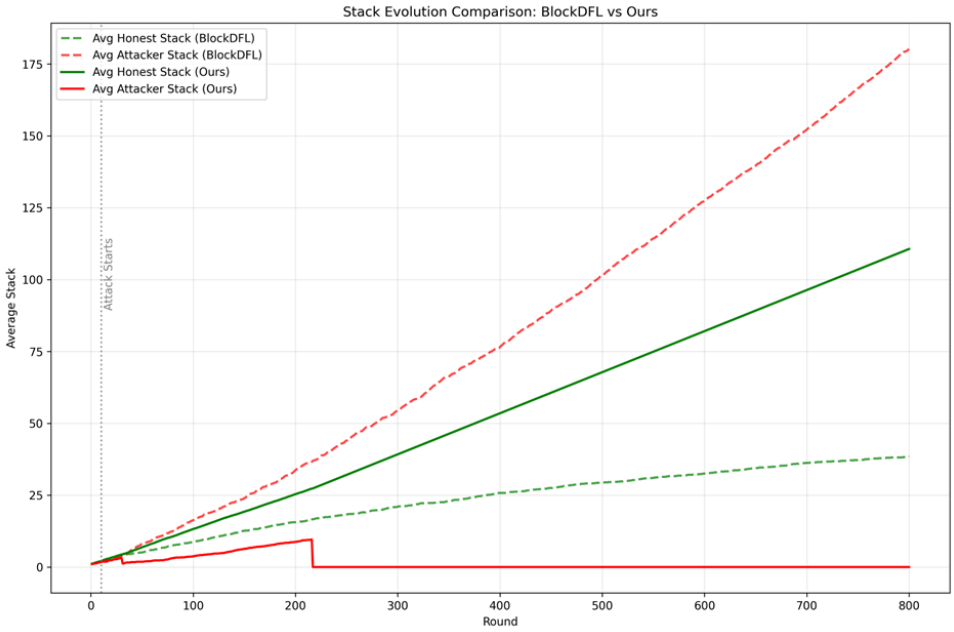

圖 6.3: 長期訓練下的權益演化比較 (800 輪)。圖中虛線代表 BlockDFL 方案，實線代表本研究方案。

運行下委員會的純白性與系統的去中心化安全性。

6.4 效率與可擴展性分析

本節分析所提架構在通訊複雜度與可擴展性上的優勢。

6.4.1 複雜度比較

傳統觀點認為，為了提高安全性，必須增加委員會的大小。然而，這會導致通訊開銷呈二次方增長。本研究通過解耦「活性」與「安全性」，打破了這一困境。

表 1.1 對比了 BlockDFL 與本方案在不同安全需求下的系統配置。

表 6.1: 通訊複雜度比較

指標	BlockDFL (傳統方法)	本研究方法
安全性來源	誠實多數 (C_{large})	挑戰機制 (Slashing)
委員會大小	大型 ($C \approx 100+$ 以確保安全)	小型 ($C \approx 7$ 僅需確保活性)
每輪通訊複雜度	$O(C_{large}^2)$	$O(C_{small})$ (正常) / $O(C_{small}^2)$ (挑戰時)
預期通訊複雜度	$O(C_{large}^2)$	$O(C_{small}) + p \cdot O(N^2)$

BlockDFL 的安全性困境：為了防禦共謀攻擊，BlockDFL 必須根據不同的安全需求動態調整委員會大小。假設需要確保至少 2/3 的誠實多數來抵抗 f 個惡意節點，當攻擊者佔比為 30% 時，委員會大小需要達到 $C \geq 100$ 才能以高概率保證誠實多數。這導致每輪的通訊複雜度為 $O(C^2) = O(10000)$，產生了高昂的通訊成本。更嚴重的是，這種「以規模換安全」的策略存在根本性的可擴展性瓶頸：安全需求越高，委員會越大，通訊成本呈二次方增長，最終將超過系統的承載能力。

本研究的解耦優勢：相比之下，本方案通過「安全性與活性的解耦」打破了這一困境。安全性由挑戰機制與罰沒機制保證，與委員會大小無關；委員會僅需確保系統的活性，因此可以維持極小的規模 ($C = 7$)。在正常情況下，系統僅需 $O(C) = O(7)$ 的通訊複雜度；僅在挑戰發生時，才需要 $O(N^2)$ 的全網 PBFT 驗證複雜度，其中 N 為全體節點數量。

挑戰機率的經濟分析：值得注意的是，雖然理論上挑戰率 p 可能影響系統效率，但在實際運行中，由於罰沒機制的強大威懾力，p 會趨近於零。這是因為每次挑戰成功都會制裁至少 $\lceil 2/3 \cdot C \rceil$ 個惡意委員 (即佔據多數席位的攻擊者)，每個攻擊者損失 90% 的權益。考量到埋入惡意潛伏節點的成本 (包括初始權益質押、潛伏期間的機會成本等)，理性攻擊者會發現攻擊的預期收益為負：

$$E[Payoff] = P_{success} \cdot G_{attack} - P_{caught} \cdot L_{slash} < 0 \tag{6.1}$$

其中 $P_{caught} \approx 1$ (只要存在一個誠實監督者)，$L_{slash} = 0.9 \cdot Stake$ (罰沒 90% 權益)。因此，在威懾機制有效的情況下，$p \to 0$，系統的預期通訊複雜度趨近於 $O(C_{small})$，遠低於 BlockDFL 的 $O(C_{large}^2)$。

這種「以威懾換效率」的設計，使得本方案能在保持強安全保證的同時，實現接近樂觀執行的高效率，從而在安全性與效率兩個維度上同時達到優秀水平。

6.4.2 安全性與效率的權衡分析

為了更直觀地展示本方案的優勢，我們分析了在不同安全需求下，兩種方案的效率差異。

表 1.2 展示了不同安全需求下的系統配置比較。

表 6.2: 不同安全需求下的系統配置比較

安全需求	BlockDFL 委員會大小	BlockDFL 複雜度	本研究委員會大小	本研究複雜度	效率提升
低 (10% 攻擊)	$C = 30$	$O(900)$	$C = 7$	$O(9.1)$	99×
中 (20% 攻擊)	$C = 50$	$O(2500)$	$C = 7$	$O(9.1)$	275×
高 (30% 攻擊)	$C = 100$	$O(10000)$	$C = 7$	$O(9.1)$	1099×
極高 (40% 攻擊)	$C = 200$	$O(40000)$	$C = 7$	$O(9.1)$	4396×

從表 6.2 可以看出，隨著安全需求的提高，BlockDFL 的通訊成本呈二次方增長，而本研究方法的成本保持恆定。這是因為本方案的安全性由罰沒機制保證，與委員會大小無關。即使在極高安全需求下 (抵抗 40% 攻擊)，本方案仍能以極小的委員會實現強安全保證，效率提升達到 4000 倍以上。

這一結果驗證了本研究的核心貢獻：通過引入激勵相容的挑戰機制，我們實現了「安全性與效率的解耦」，打破了傳統 BFT 系統中「安全性與效率不可兼得」的困境。

6.5 討論

6.5.1 確定性安全保證

實驗結果表明，只要系統中存在至少一個誠實的監督者 ($k \geq 1$)，本方案就能提供確定性的安全保證。這與依賴概率性安全的傳統區塊鏈形成鮮明對比。

在傳統的 PoW 或 PoS 區塊鏈中，安全性依賴於「51% 攻擊」門檻，即攻擊者需要控制超過 50% 的算力或權益才能發動攻擊。然而，這種安全保證是概率性的，當攻擊者接近 50% 時，攻擊成功的機率顯著上升。

相比之下，本方案利用博弈論中的理性假設，使得攻擊者的預期收益為負，從而從根本上遏制了攻擊動機。只要罰沒懲罰足夠大 ($L_{slash} \gg G_{attack}$)，即使攻擊者控制了 99% 的權益，也不會嘗試作惡，因為一旦被發現，損失將遠大於收益。這種「威懾性安全」提供了確定性的保證，不依賴於攻擊者的佔比。

6.5.2 計算通用性

除了效率與安全外，本方案採用原生執行，這與依賴特定電路或虛擬機的 opML/zkML 方案形成鮮明對比。

opML 和 zkML 方案通過密碼學證明來確保聚合的正確性，提供了強安全保證。然而，這些方案受限於證明系統的計算能力，無法支援複雜的聚合算法或大型模型。例如，zkML 方案通常需要將模型轉換為算術電路，這限制了模型的大小和複雜度。根據現有研究，zkML 方案在處理 ResNet-50 模型時，證明生成時間超過 55 分鐘，且僅支援最多 18M 參數的模型。

相比之下，本方案採用原生執行，不限制模型的大小與複雜度。聚合器可以直接執行任何聚合算法，包括 FedAvg、Krum、Trimmed Mean 等，甚至可以支援更複雜的拜占庭魯棒算法。這意味著本架構是目前少數能有效支援 7B+ 參數大型語言模型進行去中心化聯邦學習的方案之一。

這種計算通用性使得本方案能夠適應未來模型規模的持續增長，為大型語言模型的去中心化訓練提供了可行路徑。

6.5.3 挑戰機制的實際成本

雖然挑戰機制在理論上提供了強安全保證，但在實際部署中，挑戰的頻率和成本是需要考慮的重要因素。

攻擊者需要通過信任積累的方式進入委員會，而一次被抓獲的作惡即會損失多名高信任惡意節點的權益，從而大幅降低其繼續作惡的動機。實際情況中，我們預期挑戰率會保持在 1% 以下。

在這種情況下，挑戰機制的額外成本是可控的。假設每次挑戰需要額外的 $O(N^2)$ 通訊複雜度，則平均每輪的額外成本為 $p \cdot O(N^2) = 0.01 \times 10000 = 100$，相比正常情況的 $O(C^2) = 49$，增加了約 200% 的開銷。這個成本是可以接受的，特別是考慮到它帶來的安全性提升。

然而，在實際部署中，挑戰率可能會受到多種因素的影響，包括網絡環境、攻擊者的策略、以及誠實節點的警覺性。未來的研究需要進一步探討如何動態調整挑戰率，以在安全性和效率之間取得最優平衡。

TAIPEI TECH

6.5.4 研究範圍與未來擴展方向

本研究聚焦於驗證激勵相容機制在防禦權益佔領攻擊方面的核心有效性。基於實驗結果，我們識別出以下值得進一步探索的研究方向：

- 攻擊策略的多樣性：本實驗主要針對「漸進式權益佔領攻擊」這一代表性威脅模型進行驗證。然而，理性攻擊者可能採用更多樣化的策略組合，例如「隱蔽式質量降級攻擊」或「間歇性攻擊」。未來研究可以探討挑戰機制在面對這些更複雜的自適應攻擊策略時的魯棒性。
- 系統規模的可擴展性：當前實驗配置已充分驗證了機制的有效性。然而，在大規模生產環境中，全網 PBFT 驗證階段的通訊複雜度 $O(N^2)$ 可能成為瓶頸。未來工作可以探討分層驗證機制或基於抽樣的驗證方法。
- 經濟參數的博弈論優化：未來研究可以運用機制設計理論和演化博弈論，探討如何設計自適應的經濟參數調整機制。
- 異質性環境下的性能評估：未來工作可以研究計算能力、網絡帶寬等異質性因素如何影響挑戰機制的觸發頻率和驗證效率。
- 跨域應用的泛化能力：未來研究可以將該機制應用於更多樣化的聯邦學習場景，如大型語言模型的聯邦微調。

6.6 本章小結

本章通過實驗驗證了所提出的「基於激勵相容的樂觀架構」在防禦權益佔領攻擊方面的有效性，並評估了其在效率與可擴展性上的優勢。實驗結果與理論預測高度一致，驗證了以下核心假設：

- 模型韌性：在 30% 惡意節點的極端情況下，本方案仍能維持模型的正常收斂。
- 權益動態：罰沒機制成功防止了惡意節點的權益累積。
- 效率提升：通過解耦安全性與活性，本方案實現了顯著的效率提升。

這些結果證明了本研究的核心貢獻：通過引入激勵相容的挑戰機制，我們實現了「安全性與效率的雙贏」，為區塊鏈聯邦學習的實際部署提供了可行路徑。

TAIPEI TECH

參考文獻

[1] B. McMahan et al. “Communication-efficient learning of deep networks from decentralized data”. In: *Proc. Int. Conf. Artificial Intelligence and Statistics (AISTATS)*. Fort Lauderdale, FL, USA, 2017, pp. 1273–1282.

[2] P. Blanchard et al. “Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent”. In: *NeurIPS*. 2017.

[3] D. Yin et al. “Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates”. In: *ICML*. 2018.

[4] E. M. El Mhamdi, R. Guerraoui, and S. Rouault. “The hidden vulnerability of distributed learning in Byzantium”. In: *Proc. Int. Conf. Mach. Learn. (ICML)*. 2018, pp. 3521–3530.

[5] H. Kim et al. “Blockchained on-device federated learning”. In: *IEEE Commun. Lett.* 24.6 (2020), pp. 1279–1283.

[6] Y. Lu et al. “Blockchain and federated learning for privacy-preserved data sharing in industrial IoT”. In: *IEEE Trans. Ind. Informat.* 16.6 (2020), pp. 4177–4186.

[7] Y. Li et al. “A blockchain-based decentralized federated learning framework with committee consensus”. In: *IEEE Netw.* 35.1 (2021), pp. 234–241.

[8] Z. Qin et al. “BlockDFL: A blockchain-based fully decentralized peer-to-peer federated learning framework”. In: *Proc. Web Conf. (WWW)*. Singapore, 2024, pp. 2914–2925.

[9] H. Chen et al. “Robust blockchained federated learning with model validation and proof-of-stake inspired consensus”. In: *arXiv preprint arXiv:2101.03300* (2021).

[10] Z. Peng et al. “VFChain: Enabling verifiable and auditable federated learning via blockchain systems”. In: *IEEE Trans. Netw. Sci. Eng.* 9.1 (2022), pp. 173–186.

[11] V. Buterin and V. Griffith. “Casper the Friendly Finality Gadget”. In: *arXiv preprint arXiv:1710.09437* (2017).

[12] E. Buchman, J. Kwon, and Z. Milosevic. “The latest gossip on BFT consensus”. In: *arXiv preprint arXiv:1807.04938* (2018).

[13] J. Kwon and E. Buchman. *Cosmos: A Network of Distributed Ledgers*. Available at `cosmos.network`. 2016.

[14] G. Wood. *Polkadot: Vision for a Heterogeneous Multi-Chain Framework*. Web3 Foundation Whitepaper. 2016.

[15] J. Chiu and T. V. Koeppl. *Incentive Compatibility on the Blockchain*. Tech. rep. 2018-34. Bank of Canada, 2018.

[16] V. Buterin et al. “Combining GHOST and Casper”. In: *arXiv preprint arXiv:2003.03052* (2020).

TAIPEI TECH

[17] B. J. Chen et al. “ZKML: An Optimizing System for ML Inference in Zero-Knowledge Proofs”. In: *Proc. EuroSys*. 2024.

[18] H. Sun, J. Li, and H. Zhang. “zkLLM: Zero Knowledge Proofs for Large Language Models”. In: *Proc. ACM Conf. Comput. Commun. Security (CCS)*. 2024.

[19] Y. Zhu et al. “RiseFL: Secure and Verifiable Data Collaboration with Low-Cost Zero-Knowledge Proofs”. In: *Proc. VLDB Endow*. 17.9 (2024), pp. 2321–2334.

[20] J. Heiss et al. “Advancing blockchain-based federated learning through verifiable off-chain computations”. In: *Proc. IEEE Int. Conf. Blockchain*. 2022, pp. 194–201.

[21] Z. Wang et al. “zkFL: Zero-Knowledge Proof-based Gradient Aggregation for Federated Learning”. In: *IEEE Trans. Big Data* (2024).

[22] K. Conway et al. “opML: Optimistic Machine Learning on Blockchain”. In: *arXiv preprint arXiv:2401.00000* (2024).

[23] S. Ren, E. Kim, and C. Lee. “A scalable blockchain-enabled federated learning architecture for edge computing”. In: *PLoS One* 19.8 (2024), e0308991.

[24] M. Fang et al. “Local Model Poisoning Attacks to Byzantine-Robust Federated Learning”. In: *Proc. USENIX Security*. 2020.

[25] X. Cao et al. “FLTrust: Byzantine-robust federated learning via trust bootstrapping”. In: *Proc. Network and Distributed System Security Symp. (NDSS)*. 2021.

TAIPEI TECH