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

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——漸進式委員會佔領攻擊 (PCCA)，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 PCCA，我們提出一種審計驅動型委員會 BlockDFL (AC-BlockDFL)，將安全性與委員會組成解耦：由小型委員會負責例行驗證以提供活性 (Liveness)，而由全域共識支持的挑戰機制提供安全性保證。任何惡意聚合行為都將觸發密碼學驗證與罰沒懲罰，立即沒收參與共謀的委員會成員之全額質押。此機制將安全門檻從委員會多數轉移至全網共識，打破攻擊者依賴的權益累積正反饋循環。在 2000 輪的長期模擬實驗中，本機制將攻擊發生次數從 107 次壓制至 5 次，相較於現有方法實現超過 20 倍的攻擊抑制效果。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運算效率。

ABSTRACT

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

Blockchain-based Federated Learning (BCFL) addresses trust and privacy concerns through decentralized consensus mechanisms. Existing BCFL systems rely on committee-based validation architectures, assuming that committee members are honest or possess an honest majority. This assumption is vulnerable to verifier collusion, where attackers can dominate committees by accumulating stake. We identify a novel threat called Progressive Committee Capture Attack (PCCA), in which rational attackers exploit incentive mechanisms to gradually accumulate stake and occupy sufficient committee seats to launch coordinated attacks. Once an attacker obtains a committee majority, existing architectures fail to detect or prevent such attacks. To defend against PCCA, we propose Audit-driven Committee BlockDFL (AC-BlockDFL), which decouples security from committee composition: a small committee handles routine validation to provide liveness, while a challenge mechanism supported by global consensus ensures security guarantees. Any malicious aggregation behavior triggers cryptographic verification and slashing penalties, resulting in the immediate confiscation of all staked assets from the colluding committee members. This mechanism shifts the security threshold from a committee majority to global consensus, breaking the positive feedback loop of stake accumulation relied upon by attackers. In a long-term simulation experiment of 2000 rounds, this mechanism suppressed the number of successful attacks from 107 to 5, achieving over 20 times the attack suppression effect compared to existing methods. Our decoupled design also allows for smaller committee sizes, enhancing computational efficiency without compromising security.

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。

目錄

圖目錄

表目錄

第一章　緒論

隨著人工智慧與分散式運算技術的蓬勃發展，聯邦學習 (Federated Learning) [**mcmahan2017communication**] 與區塊鏈技術的深度整合，催生了區塊鏈賦能的聯邦學習 (Blockchain-based Federated Learning, BCFL) 這一創新技術典範。此技術路徑之所以受到學界與產業界的高度重視，根本原因在於其成功回應了多方互不信任情境下協作機器學習的核心挑戰。在低軌衛星網路 (LEO) [**pokhrel2021blockchain**, **wu2024sharded**, **elmahallawy2025decentralized**]、車聯網 (V2X) [**liu2021blockchain**, **pokhrel2020autonomous**] 以及工業物聯網 (IIoT) [**lu2020blockchain**, **qu2020decentralized**] 等實際應用場景中，BCFL 展現了其獨特且難以替代的技術價值。以 LEO 衛星星座為代表的太空人工智慧應用場景尤具說明性意義，在此類場景中星地通訊窗口通常僅維持約五分鐘，且下行頻寬受限於每秒 8 Mbps 左右的水準 [**wu2024sharded**]，這種嚴苛的通訊條件使得傳統依賴地面站進行模型聚合的訓練方案在技術上幾乎無法實現。BCFL 透過在異質衛星營運商之間建立去中心化的信任基礎設施，成功將模型收斂所需時間縮減達三十小時之譜 [**elmahallawy2025decentralized**]，充分展現了此技術路徑在極端環境下的應用潛力。類似地，在工業 4.0 的發展背景下，BCFL 使得協作工廠能夠在完全不洩露各自商業機密的前提下共同進行預測性維護模型的訓練，實驗資料顯示此種架構可將通訊開銷較傳統集中式方案降低約 41% [**lu2020blockchain**]。上述多元應用場景共同呈現出三項關鍵特徵：缺乏可信賴的中心化協調者、運算與通訊資源高度受限、以及訓練資料在統計分布上的顯著異質性，正是這些特徵的交織作用，使得 BCFL 逐步確立其作為通用去中心化學習架構首選方案的地位。

然而，當 BCFL 系統嘗試邁向大規模實際部署時，卻面臨著嚴峻的效能瓶頸，此問題在學術文獻中通常被稱為「可擴展性兩難困境」。究其根源，目前絕大多數 BCFL 系統採用實用拜占庭容錯協議 (Practical Byzantine Fault Tolerance, PBFT) [**castro1999practical**] 或其各種變體作為底層共識機制，而此類協議固有的 $O(N^2)$ 訊息複雜度特性，導致系統效能隨著參與節點數量的增加而呈現急劇下降的趨勢。FLCoin [**ren2024scalable**] 的實證研究提供了具體的量化證據：當參與節點數量達到一百個時，

單一輪次的共識過程所產生的訊息交換量將突破兩萬條，共識延遲因而攀升至二十五秒以上，此延遲水準已經達到甚至超過模型訓練本身所需時間的量級，對於要求即時響應的應用場景而言構成了根本性的障礙。在車聯網 (IoV) 的應用場景中，Chai 等人 [**chai2021hierarchical**] 進一步指出，傳統區塊鏈架構（如比特幣或以太坊）所依賴的高頻寬資訊交互與龐大算力需求，與車輛節點的高移動性特徵存在本質上的衝突。在不穩定的連線環境下，要求車載單元執行高強度的全網共識驗證，將導致系統無法滿足即時知識共享的低延遲需求。這對於需要頻繁進行模型更新的車載智慧系統而言顯然是無法接受的效能表現。除了通訊效能問題之外，區塊鏈節點對於儲存空間的高需求同樣構成嚴峻挑戰，例如比特幣全節點需要約 200 GB 的儲存空間，而以太坊更是超過 465 GB，這與邊緣運算設備通常僅具備 KB 至 MB 等級記憶體容量的現實形成了尖銳的衝突 [**fedchain2024**]。效能與資源的這種雙重束縛，使得要求所有節點參與驗證的傳統全節點架構在實際工業部署情境中越來越難以為繼，迫切需要創新的技術方案來突破這些根本性的限制。

正是為了化解上述可擴展性挑戰，學術界近年來將研究重心轉向「委員會機制 (Committee Mechanism)」的探索與發展，此機制的核心設計理念是將原本由全體節點共同承擔的驗證責任，集中委派給一個規模相對較小的驗證者委員會來執行。這種架構轉變的根本邏輯在於，透過大幅減少實際參與共識決策過程的節點數量，可以顯著降低系統的通訊複雜度與運算負擔。目前學界提出的主流委員會選拔機制包含數種不同的技術路徑：基於雜湊環結構的隨機抽樣方法 [**qin2024blockdfl**]、依據持有代幣數量或權益進行加權選舉的方法 [**li2021blockchain**, **ren2024scalable**]、以及利用可驗證隨機函數 (VRF) 實現的 Sortition 機制 [**shayan2021biscotti**, **weng2021deepchain**, **gilad2017algorand**]。委員會機制的導入確實為系統效能帶來了立竿見影的改善效果，FLCoin [**ren2024scalable**] 的研究報告顯示，透過採用滑動窗口選舉策略，系統成功將通訊開銷降低了 90%，並實現了 5.7 倍的訓練速度提升；BFLC [**li2021blockchain**] 則藉由委員會驗證機制將共識延遲穩定控制在三秒以內。這些最佳化成果成功地將通訊複雜度從全網共識的 $O(N^2)$ 降至與委員會規模 C 相關的 $O(C^2)$ 甚至 $O(C)$ 等級，極大提升了系統的吞吐量與響應速度。然而，這種為追求效率而將驗證權力集中於少數委員

會成員的架構轉變，也在無形中引入了全新的安全攻擊面，對系統的長期安全性與穩定性構成了潛在威脅，而這些威脅至今尚未獲得學術界的充分探討。

深入剖析上述安全隱憂可以發現，現有委員會防禦機制普遍存在一個容易被研究者忽視的結構性弱點，即對「誠實多數假設」的過度依賴。儘管當前的區塊鏈聯邦學習系統已經廣泛導入 Krum、Trimmed Mean 或 Median 等具備拜占庭強健性的聚合演算法 [**blanchard2017machine**, **yin2018byzantine**]，但這些演算法能夠發揮預期防禦效果的根本前提是：執行聚合運算的委員會成員中，誠實節點必須穩定維持佔多數的地位。當前安全性研究的主要關注焦點集中在惡意客戶端上傳毒化梯度的資料層攻擊，卻在很大程度上忽略了一種更為隱蔽的威脅形式，即委員會成員之間可能形成的策略性共謀行為。一旦攻擊者成功控制委員會中超過三分之二的席位，便能夠串聯起來繞過所有現有的資料層強健聚合邏輯，甚至可以任意偽造聚合結果而無需擔心遭到系統的懲罰或追責。正如 FedBlock [**nguyen2024fedblock**] 所指出，當任何參與者都可能成為驗證者時，系統不能僅仰賴誠實多數假設，而必須具備主動檢測並隔離惡意驗證者的能力，然而這項關鍵機制在現有的區塊鏈聯邦學習架構中普遍付之闕如，這揭示了當前安全防禦體系在共識決策層面的本質脆弱性。

上述分析揭示了當前研究領域中存在的一個關鍵「研究缺口」：現有的 BCFL 系統普遍缺乏能夠應對「漸進式委員會佔領攻擊 (Progressive Committee Capture Attack, PCCA)」的自我修復與防禦機制。在 PCCA 攻擊模式中，對手並非採取直接的暴力破壞策略，而是實施一種被稱為「策略性餓死 (Strategic Starvation)」的隱蔽手法，具體而言，攻擊者在成功掌控委員會之後，會優先處理與自身經濟利益相關的模型更新請求，同時系統性地拒絕為誠實參與者提供驗證服務，藉此操縱系統的獎勵分配機制與權益動態變化。值得特別關注的是，現有的各種解決方案大多僅聚焦於特定的防禦維度，例如基於同態加密 (Homomorphic Encryption, HE) [**zhang2020batchcrypt**, **miao2022privacy**] 或權益證明 (Proof of Stake, PoS) [**chen2021robust**] 的技術方案，雖然在隱私保護或准入控制等特定面向上確實有所突破，卻無法在委員會本身已經淪陷而不再可信的情況下，繼續保證模型聚合更新的正確性以及系統資源分配的公平性。由於缺乏事後的「可追溯審計能力」與「有效經濟威懾手段」，一旦誠實多數假設在某一輪次中被攻破，現有系

統便缺乏任何機制來識別並懲罰惡意行為者，使得攻擊者得以在後續輪次中持續佔據優勢地位。因此，如何將系統的安全性保障與對共識節點集體誠實性的依賴進行有效解耦，已成為實現真正意義上去中心化人工智慧平台所必須跨越的關鍵技術門檻。

針對這一嚴峻的技術挑戰，本論文提出一套名為「審計驅動型委員會 BlockDFL (Audit-driven Committee BlockDFL, AC-BlockDFL)」的創新防禦體系，旨在為區塊鏈聯邦學習系統建構一道動態且具備自我修復能力的安全屏障。本研究的核心設計理念在於，透過引入異步審計機制與內部罰沒協議，賦予現有委員會架構檢測並汰除惡意驗證者的能力。這種設計使得系統能夠將「活性 (Liveness)」與「安全性 (Security)」這兩項往往相互牽制的系統屬性進行有效的解耦處理。在本架構下，即使委員會在特定輪次中並非完全可信甚至已經被惡意節點所捕獲，系統仍然能夠透過分散部署於全網的挑戰者網路，對委員會所做出的每一項決策進行事後審計與追責，進而檢舉並懲罰任何偽造或誤導性的聚合結果。這種創新機制不僅大幅提升了系統面對策略性攻擊時的韌性，更為去中心化環境下的模型協作訓練提供了一層額外且獨立的安全性保障，確保了系統在追求效率提升的同時，不會犧牲去中心化架構所應堅守的核心安全原則。

進一步而言，本研究在理論分析、協議設計與實驗驗證三個相互支撐的維度上均做出了具體而紮實的學術貢獻。在理論分析層面，本研究首次對漸進式委員會佔領攻擊 (PCCA) 模式進行了形式化的數學定義，並透過系統性的模擬實驗量化評估了該攻擊對系統長期激勵相容性所造成的深層破壞效應，藉此揭露了傳統防禦機制在面對具有演化特性的策略性攻擊時所存在的根本局限。基於上述理論發現，本研究進一步結合博弈論模型設計了一套精密的內部罰沒協議 (Internal Slashing Protocol)，此協議透過精心設計的經濟懲罰機制，確保任何節點進行審計驗證所需付出的成本始終低於其從事惡意行為可能獲取的收益，從而促使誠實行為成為所有理性參與者在長期重複博弈下的唯一納什均衡策略 [**chiu2018incentive**]。這些理論層面的創新貢獻在後續的大規模模擬實驗中獲得了充分的實證支持。實驗資料顯示，即使在 30% 節點進行惡意共謀的極端對抗環境下，採用 AC-BlockDFL 架構的系統仍能將模型最終準確率穩定維持在 98.6% 以上的水準，同時在達成相同安全性保障指標的前提下將通訊開銷降低了 44.4%，並將系統的最低不可用率從 20% 大幅壓制至 5% 以下，顯著提升了系統在對抗

性環境中的強健性與運作效率。

本論文後續章節的組織架構安排如下。在本緒論之後，第二章將系統性地奠定本研究所需的理論基礎，內容涵蓋聯邦學習的核心概念、區塊鏈共識機制的運作原理、以及拜占庭容錯技術的數學基礎，並針對現有去中心化聯邦學習領域的相關文獻進行批判性的回顧與評述。以第二章建立的知識基礎為起點，第三章將詳細定義本研究所採用的系統模型與威脅模型，並針對 PCCA 攻擊者的行為特徵、能力邊界與策略空間進行嚴謹的形式化描述。在完成理論模型的建構之後，第四章將提出 AC-BlockDFL 架構的核心設計方案，完整闡明具體的協議運作流程、關鍵演算法的設計細節、以及支撐整體架構的安全性證明邏輯。緊接著，第五章將透過一系列精心設計的針對性實驗與全面的資料比較分析，在多種不同的對抗場景設定下驗證本架構所具備的效能優勢與系統韌性。最終，第六章將總結本研究的主要發現與學術貢獻，並就本機制在未來分散式人工智慧生態系統中的應用潛力與可能的演進方向進行前瞻性的探討。

第二章　背景知識與相關研究

本章旨在建立理解區塊鏈聯邦學習委員會安全性所需的理論基礎與技術背景。首先，本章將從聯邦學習的核心價值出發，闡明中央化架構面臨的信任困境，進而說明區塊鏈技術如何作為去中心化信任的基礎設施。接著，本章將介紹拜占庭容錯理論的基本原理，為理解委員會共識機制的安全性閾值提供數學基礎。在此基礎上，本章將探討區塊鏈聯邦學習如何從全節點共識演進至委員會架構，並分析委員會規模與安全性之間的權衡關係。隨後，本章將詳細介紹本研究採用的基準系統模型 BlockDFL 之委員會架構，其內容涵蓋角色定義、運作流程與獎勵機制。最後，本章將回顧現有驗證方法的局限性，指出當前研究在面對策略性攻擊者時的盲區，從而定位本研究欲填補的關鍵缺口。

2.1　聯邦學習與去中心化信任需求

2.1.1　聯邦學習的核心機制

聯邦學習是一種分散式機器學習典範，其核心創新在於實現「資料不動、模型動」的訓練機制 [**mcmahan2017communication**]。在傳統的集中式機器學習中，所有訓練資料必須彙集至中央伺服器進行處理，這種做法在面對隱私敏資料或資料傳輸成本高昂的場景時顯得力不從心。聯邦學習透過將訓練過程分散至資料所在的終端裝置，僅將模型更新而非原始資料上傳至伺服器進行聚合，從根本上改變了資料與運算的關係。這種架構使得醫療機構能夠在不共享病患紀錄的前提下協同訓練診斷模型，金融機構能夠在不揭露客戶交易資料的情況下建立風險評估系統，行動裝置製造商能夠利用數百萬用戶的使用習慣最佳化輸入法預測，而無需將敏感的打字內容上傳至雲端。

聯邦學習的標準訓練流程可概括為四個階段的迭代循環。在每一輪訓練中，中央伺服器首先將當前的全域模型參數分發給選定的客戶端；各客戶端隨後在本地私有資料上執行若干輪梯度下降，產生反映本地資料特性的模型更新；

客戶端將這些更新上傳至伺服器；伺服器執行聚合演算法（最常見的是 FedAvg [**mcmahan2017communication**]）將各客戶端的更新整合為新的全域模型。此循環持續進行直至模型收斂或達到預設的訓練輪數。值得注意的是，聯邦學習面對的資料分布通常具有高度異質性：不同客戶端持有的資料量可能相差懸殊，資料的類別分布也往往呈現顯著差異，這種非獨立同分布（Non-IID）的特性為模型訓練與安全防護帶來了獨特的挑戰 [**kairouz2021advances**]。

2.1.2 中央化架構的信任困境

儘管聯邦學習在資料隱私保護上取得了重要進展，其標準架構仍存在一個根本性的信任假設：所有參與者必須信任中央聚合伺服器會誠實地執行聚合運算並正確地分發結果。然而，在缺乏有效驗證機制的情況下，這項假設構成了系統的單點脆弱性。中央伺服器可能因遭受攻擊、內部人員惡意行為或系統故障而偏離預期行為，而客戶端對此幾乎無從察覺，更遑論採取補救措施。

中央化架構面臨的信任風險可歸納為三個層面。第一個層面是聚合正確性的不可驗證性。當伺服器宣稱某一全域模型是由特定客戶端更新聚合而成時，客戶端無法獨立驗證此宣稱的真實性。伺服器可能執行選擇性聚合，意即僅納入部分客戶端的更新而排除具體，或者直接篡改聚合結果以植入後門。研究已證實，透過精心設計的模型修改，攻擊者可在不顯著影響主任務效能的情況下，使模型對特定輸入產生預設的錯誤輸出 [**bagdasaryan2020how**]。第二個層面是單點故障風險。中央伺服器一旦因攻擊、硬體故障或網路問題而離線，整體訓練流程即刻中斷，且由於缺乏分散式的狀態同步機制，系統難以從中間狀態恢復。第三個層面是隱私保護的局限性。儘管聯邦學習避免了原始資料的直接傳輸，研究表明惡意的聚合伺服器仍可能透過分析客戶端提交的模型更新，推論出關於訓練資料的敏感資訊 [**zhu2019deep**]。

這些信任風險在跨組織協作的場景中尤為突出。當多個相互獨立甚至存在競爭關係的機構希望聯合訓練模型時，由任何單一機構擔任中央聚合者都難以獲得其他參與者的充分信任。即便引入第三方作為中立的聚合服務提供者，仍無法從根本上消除對

該第三方誠實性的依賴。這種信任困境限制了聯邦學習在高價值、高敏感場景中的應用潛力，也促使研究者開始探索去中心化的替代方案。

2.1.3 區塊鏈作為去中心化信任基礎設施

區塊鏈技術具備不可篡改性、透明性與去中心化這三項核心特性，恰好對應了中央化聯邦學習面臨的信任困境，使其成為建構去中心化聯邦學習系統的理想基礎設施。不可篡改性源於區塊鏈的鏈式雜湊結構：每個區塊包含前一區塊的雜湊值，任何對歷史資料的修改都將導致後續所有區塊的雜湊值連鎖變化，從而被網路中的其他節點立即偵測。這項特性確保了一旦聚合結果被記錄於區塊鏈，便無法在事後被悄然篡改。透明性則意味著所有被記錄的交易與狀態變更對全體參與者可見，客戶端可以驗證自己的更新是否被納入聚合，也可以審計歷史聚合過程是否遵循預定的規則。去中心化消除了對單一實體的信任依賴：區塊鏈網路由眾多獨立節點共同維護，即使部分節點失效或行為惡意，只要誠實節點佔據多數，系統仍能正確運作。

將區塊鏈整合至聯邦學習架構，可從多個層面強化系統的可信賴性。在聚合正確性方面，智能合約可編碼確定性的聚合規則，確保聚合過程按照預定邏輯執行，而非依賴聚合者的自我約束。聚合結果連同參與者資訊被記錄於區塊鏈，形成永久可查的審計軌跡。在系統可用性方面，區塊鏈的分散式架構天然具備容錯能力：即使部分節點離線，其他節點仍可維持系統運作，避免了中央伺服器故障導致的全面停擺。在激勵對齊方面，區塊鏈原生的代幣機制可用於設計精細的獎懲制度，對誠實貢獻者給予獎勵，對惡意行為者施加經濟懲罰，從而在博弈論意義上引導參與者趨向誠實行為。

然而，區塊鏈並非萬能的信任解決方案。區塊鏈共識機制本身需要假設惡意節點不超過特定比例；以拜占庭容錯協議為例，此閾值通常設定為三分之一。當攻擊者控制的節點超過此閾值時，區塊鏈的安全性保證將不再成立。此外，區塊鏈的共識過程涉及大量的節點間通訊，其延遲與頻寬成本可能與聯邦學習對快速迭代的需求產生張力。這些考量促使研究者發展出委員會架構等效率最佳化方案，但也隨之引入了新的安全性議題。下一節將首先介紹拜占庭容錯的理論基礎，為理解這些安全性議題提供

必要的背景知識。

2.2 拜占庭容錯的理論基礎

2.2.1 拜占庭將軍問題與容錯閾值

拜占庭將軍問題背景 拜占庭將軍問題由 Lamport、Shostak 與 Pease 於 1982 年正式提出 [**lamport1982byzantine**]，是分散式系統容錯理論的基石。問題的設定源自一個軍事隱喻：拜占庭帝國的數支軍隊包圍敵城，各軍由一位將軍指揮，將軍們僅能透過信使相互通訊。然而，部分將軍可能是叛徒，他們會刻意傳遞錯誤訊息以阻撓忠誠將軍達成一致決策。問題的核心在於：如何設計一個協議，使得所有忠誠將軍能就「進攻」或「撤退」達成共識，即使存在叛徒試圖破壞協調？此問題的形式化定義包含兩個交互一致性條件：所有忠誠節點必須就相同的值達成共識（一致性），且若發起者是誠實的，則共識結果必須是發起者提出的值（正確性）。

數學限制與 Quorum 交叉原理 拜占庭將軍問題存在一個根本性的數學限制：在僅使用口頭訊息的情況下，問題可解若且唯若誠實節點超過總數的三分之二。換言之，若系統中有 N 個節點，最多只能容忍 f 個拜占庭節點，其中 $N \geq 3f + 1$。此限制可透過最簡單的三節點、一叛徒場景直觀理解。考慮指揮官向兩位副官發送命令的情境：若指揮官是叛徒，他可能向副官 A 發送「進攻」，向副官 B 發送「撤退」；當兩位誠實副官相互交換收到的命令時，各自都會發現矛盾。然而，若副官 B 是叛徒而指揮官誠實，副官 B 可能向副官 A 謊稱「指揮官說撤退」。關鍵的洞察在於：從副官 A 的視角來看，這兩種情境完全無法區分，因為在兩種情況下，他都收到來自指揮官的「進攻」訊息，以及來自 B 聲稱收到「撤退」的訊息。任何確定性演算法在此情境下都必然失敗，這從根本上限制了拜占庭容錯系統的設計空間。

此三分之一閾值的數學根源在於 Quorum 交叉原理。為確保任何決策都獲得足夠的誠實節點背書，系統需要收集至少 $2f + 1$ 個節點的確認。由於最多 f 個節點可能

是惡意的，$2f+1$ 個確認中必然包含至少 $f+1$ 個來自誠實節點。任意兩個大小為 $2f+1$ 的節點群體，其交集至少包含 $f+1$ 個節點，這確保了至少有一個誠實節點見證了兩次決策，從而防止系統對同一問題做出矛盾的決定。將節點總數代入約束條件 $N \geq 2f+1+f$，即得 $N \geq 3f+1$。

2.2.2 實用拜占庭容錯協議的核心概念

拜占庭將軍問題的早期解法雖然在理論上可行，但其指數級的通訊複雜度使其僅具學術意義。1999 年，Castro 與 Liskov 提出實用拜占庭容錯協議（Practical Byzantine Fault Tolerance, PBFT）[**castro1999practical**]，首次將 BFT 共識的通訊複雜度降至多項式級別 $O(N^2)$，使其在實際系統中可行。PBFT 的設計目標是在部分同步網路模型下，以合理的效能代價換取對任意惡意行為的容錯能力。

PBFT 協議的運作依賴 $N = 3f+1$ 個副本節點，其中一個被指定為主節點（Primary），負責為客戶端請求分配序號並發起共識。協議透過三個階段達成共識：預準備（Pre-prepare）、準備（Prepare）與提交（Commit）。在預準備階段，主節點將請求連同分配的序號廣播給所有副本；在準備階段，收到預準備訊息的副本向其他所有副本廣播準備訊息，當某副本收集到 $2f$ 個匹配的準備訊息時，表明系統中有足夠多的節點認可此請求的序號分配；在提交階段，進入準備狀態的副本廣播提交訊息，當收集到 $2f+1$ 個提交訊息時，副本確信此請求已被系統接受，可以執行並回覆客戶端。三階段設計的核心目的是確保即使主節點是惡意的，也無法導致誠實節點對請求順序產生分歧。

PBFT 的通訊複雜度為 $O(N^2)$，這是因為在準備與提交階段，每個節點都需要向其他所有節點發送訊息。以 $N=7$（可容忍 2 個拜占庭節點）的配置為例，每輪共識約需交換 100 則訊息；當節點數增至 $N=22$（可容忍 7 個拜占庭節點）時，訊息數增至約 900 則。這種二次方增長限制了 PBFT 在大規模網路中的直接應用，實務部署通常限制在 10 至 20 個節點的規模。後續研究如 HotStuff [**yin2019hotstuff**] 透過流水線化設計與閾值簽章技術，將複雜度進一步降至 $O(N)$，但在本研究關注的許可制聯盟鏈場景中，

節點數量通常在 PBFT 可承受的範圍內。

2.2.3 BFT 共識在區塊鏈聯邦學習中的角色

在區塊鏈聯邦學習系統中，拜占庭容錯共識扮演著確保聚合結果正確性的關鍵角色。與傳統區塊鏈應用（如加密貨幣交易）不同，聯邦學習的「交易」是模型更新，而「帳本狀態」是全域模型參數。當負責聚合的節點可能被攻陷或本身即為惡意參與者時，系統需要一個機制來驗證聚合結果的正確性，並在多個可能存在分歧的結果中達成共識。BFT 協議正是為此目的而設計：它確保只要惡意節點不超過總數的三分之一，系統就能就聚合結果達成一致，且該結果必然反映誠實多數的判斷。

然而，將 BFT 共識直接應用於大規模聯邦學習系統面臨顯著的效率挑戰。聯邦學習通常涉及數十至數百個參與者，若所有參與者都參與每一輪的 BFT 共識，$O(n^2)$ 的通訊複雜度將成為嚴重的效能瓶頸。更重要的是，聯邦學習需要頻繁迭代（典型的訓練過程可能包含數百至數千輪），因此每輪都執行完整的全網共識將導致訓練時間大幅延長。這種效率與安全性之間的張力，促使研究者發展出委員會架構：由一個小型的代表性子集執行共識，以較低的通訊成本達成近似的安全保證。下一節將詳細探討這種架構演進及其伴隨的安全性權衡。

2.3 區塊鏈聯邦學習的委員會架構演進

2.3.1 從全節點到委員會的效率驅動

區塊鏈聯邦學習的早期研究嘗試將傳統 BFT 共識直接應用於全體參與者，但很快便遭遇了可擴展性的瓶頸。以 BFLC [**li2021blockchain**] 的實驗配置為例，當參與者數量達到 20 個時，採用完整 PBFT 共識的每輪延遲已超過 100 毫秒；若將參與者擴展至數百個規模，共識延遲將增長至秒級甚至更長，這對於需要快速迭代的聯邦學習訓練而言顯然無法接受。更根本的問題在於通訊頻寬的消耗：每輪共識中，每個節點都需要接收並處理來自其他所有節點的訊息，當節點數量增加時，網路負擔呈平方級增長，

這在頻寬受限的邊緣運算環境中尤為致命。

委員會架構的核心理念是將共識責任委派給一個規模遠小於全網的代表性子集。令全網節點數為 N，委員會規模為 C，其中 $C \ll N$。委員會內部執行 BFT 共識的通訊成本為 $O(C^2)$，委員會決策結果廣播至全網的成本為 $O(N)$，總通訊成本為 $O(C^2+N)$。當 C 維持在較小的常數（如 7 至 20）時，此成本近似於線性 $O(N)$，相較於全節點 PBFT 的 $O(N^2)$ 實現了數量級的改善。FLCoin [**ren2024scalable**] 的實驗資料印證了這一分析：在 500 個節點的網路中，採用規模為 100 的滑動視窗委員會，相較於全節點共識可減少約 90% 的通訊開銷，共識延遲維持在 3 秒以內。

委員會架構的效率優勢使其迅速成為區塊鏈聯邦學習的主流設計典範。然而，這種效率的提升並非沒有代價：系統的安全性不再由全網的誠實多數保證，而是取決於委員會的組成是否可信。若攻擊者能夠控制委員會中超過三分之一的席位，便可操控共識結果，通過惡意的聚合提案或拒絕誠實的提案。這種從「全網安全」到「委員會安全」的轉變，將安全性分析的焦點從「全網惡意節點比例」轉移至「委員會選舉機制的抗操控能力」。

2.3.2 委員會選舉機制的設計空間

委員會選舉機制決定了哪些節點將被選入委員會，其設計直接影響系統的安全性與公平性。現有方案大致可分為四種取向：隨機選擇、權益導向、聲譽導向與貢獻導向，各有其優勢與潛在風險。

隨機選擇機制透過密碼學隨機數決定委員會組成，其核心優勢在於不可預測性：攻擊者無法提前知曉哪些節點將被選中，因而難以針對性地部署攻擊。RapidChain [**zamani2018rapidchain**] 採用分散式隨機數生成協議，確保選舉結果對所有參與者而言都是不可預測且可驗證的。然而，純粹的隨機選擇可能將惡意或低品質的節點選入委員會，且無法反映節點過去的行為表現。權益導向機制將選中機率與節點持有的權益（stake）掛鉤，持有越多權益的節點越可能被選入委員會。這種設計的理論基礎是經濟激勵對齊：高權益節點若行為惡意將面臨更大的經濟損失，因此傾向誠實。以太坊 2.0

的驗證者選舉即採用此機制 [**gasper**]。然而，權益導向可能導致「富者愈富」的中心化傾向，且無法防範願意承受經濟損失的攻擊者。

聲譽導向機制根據節點的歷史行為表現運算聲譽分數，高聲譽者優先被選入委員會。BESIFL [**chen2021robust**] 追蹤各節點提交更新的品質，將聲譽作為委員會選舉的權重。此機制能有效過濾曾有惡意行為記錄的節點，但也面臨兩項挑戰：新加入者缺乏歷史記錄，可能陷入「冷啟動」困境；更重要的是，策略性攻擊者可透過長期的誠實行為累積聲譽，待時機成熟後再發動攻擊。貢獻導向機制以節點對聯邦學習的實質貢獻（如訓練資料量、模型品質）作為選舉依據。FLCoin [**ren2024scalable**] 的滑動視窗機制即屬此類：節點透過提交有效的模型更新獲得「份額」，在視窗內持有份額的節點組成委員會。這種設計與聯邦學習的目標直接對齊，但貢獻指標可能被博弈；例如，攻擊者可先提交高品質更新以獲取委員會席位，再利用此席位通過惡意提案。

在實務應用中，為了在安全性、效能與經濟效用之間取得最佳平衡，先進的區塊鏈架構往往傾向於採用混合型的選舉設計。以 BlockDFL [**qin2024blockdfl**] 為例，該系統整合了「權益加權」與「確定性隨機選擇」（Stake-weighted Deterministic Random Selection）兩大理念，其選舉結果由前一個區塊的雜湊值決定。這種設計透過雜湊函數的確定性確保了選舉結果對全網透明且可驗證，同時藉由權益加權機制引入經濟激勵，旨在引導參與者維持長期的誠實行為。然而，這種設計雖然兼顧了不可預測性與作業效率，卻也繼承了權益導向系統的固有風險，即攻擊者可能透過持續累積權益來逐步提升其在委員會中的獲選機率。這種從權益累積轉化為影響力集中的現象，使得系統面臨「潛伏滲透」的威脅，亦即惡意節點在獲取足以左右共識的地位前，極易透過偽裝誠實行為來規避基於歷史表現的防禦機制。

2.3.3 委員會規模與安全性的權衡分析

委員會規模的選擇涉及效率與安全性之間的核心權衡，較小的委員會帶來更低的通訊成本與更快的共識速度，但也更容易被攻擊者滲透，而較大的委員會提供更強的安全保證，卻犧牲了效率優勢。理解這一權衡需要從機率論的角度分析委員會被攻破

的風險，而超幾何分佈為此提供了精確的數學工具。當從 N 個節點（其中 fN 個為惡意節點）中隨機選取 C 個組成委員會時，由於委員會成員的選擇是一個無放回抽樣過程，委員會中恰有 k 個惡意節點的機率遵循超幾何分佈，其機率質量函數可表示為：

$$P(X = k) = \frac{\binom{fN}{k}\binom{(1-f)N}{C-k}}{\binom{N}{C}} \tag{2.1}$$

其中 $\binom{N}{m}$ 表示二項式係數，代表從 N 個元素中選擇 m 個元素的方式數量。這個公式的分子部分運算了選擇 k 個惡意節點和 $C-k$ 個誠實節點的所有可能組合方式，而分母則是從 N 個節點中選擇 C 個節點的總組合數。

對於採用 PBFT 共識的委員會而言，安全性分析需要區分兩種不同的威脅閾值。第一種閾值為三分之一，當惡意節點超過委員會的三分之一時，攻擊者能夠阻止委員會達成任何共識，因為 PBFT 協議要求至少 $2f+1$ 個節點同意才能通過提案，這種攻擊形式本質上是一種拒絕服務攻擊，雖然無法注入惡意內容，但能夠癱瘓系統的正常運作。第二種閾值為三分之二，這是更為嚴重的威脅情境，當惡意節點佔據委員會超過三分之二的席位時，攻擊者不僅能夠阻止誠實提案通過，更能夠強制通過惡意提案，完全控制委員會的決策結果。在本研究所關注的漸進式委員會佔領攻擊情境中，攻擊者的目標正是達成後者，透過控制委員會來通過有利於自身的提案並排除誠實節點的更新，因此後續的風險分析將以三分之二作為委員會被惡意控制的臨界閾值。

基於上述分析，委員會被惡意控制的風險機率 P_{mal} 可以表示為惡意節點數量達到或超過 $\lfloor 2C/3 \rfloor + 1$ 的累積機率：

$$P_{mal} = P(X \geq \lfloor 2C/3 \rfloor + 1) = \sum_{k=\lfloor 2C/3 \rfloor + 1}^{C} \frac{\binom{fN}{k}\binom{(1-f)N}{C-k}}{\binom{N}{C}} \tag{2.2}$$

為了具體理解這個機率模型的實際意涵，以下考察一個具代表性的數值案例。假設驗證者總池規模 $N = 100$，網路中惡意節點的比例 $f = 0.3$，即存在 30 個惡意節點和 70 個誠實節點，這是一個相對極端的假設，因為 30% 的惡意比例已經接近大多數拜占庭容錯系統所能容忍的上限。在這種條件下，不同委員會規模所對應的被惡意控制風險

呈現出明顯的差異。當委員會規模 $C = 5$ 時，惡意節點需要至少佔據 4 個席位才能達到控制閾值，透過超幾何分佈的運算，這種情況發生的機率約為 2.74%。當委員會規模增加到 $C = 7$ 時，惡意節點需要至少 5 個席位才能控制，此時風險機率約為 2.42%，略有下降但改善不明顯。

當委員會規模進一步增加到 $C = 9$ 時，情況出現了顯著變化，此時惡意節點需要佔據至少 7 個席位才能達到三分之二的控制閾值，而這種情況發生的機率驟降至約 0.28%，這個數值已經低於許多實際系統所設定的風險容忍度。繼續增加委員會規模，當 $C = 11$ 時風險進一步降至約 0.25%，當 $C = 13$ 時則降至約 0.21%。這些資料揭示了兩個重要的洞察。其一，委員會規模與安全性之間並非線性關係，存在一個「甜蜜點」區間（約 $C = 9$ 至 $C = 13$），在此區間內增加委員會規模能夠帶來顯著的安全性提升。其二，即使在相當高的惡意節點比例下，只需要一個規模適中的委員會就能將被惡意控制的風險壓制到相當低的水平。然而，委員會規模的增加直接推高了共識的通訊成本，由於 PBFT 的通訊複雜度為 $O(C^2)$，$C = 13$ 的委員會其內部通訊量約為 $C = 5$ 的 6.76 倍，這種成本增長在頻繁迭代的聯邦學習場景中尤為顯著。

上述分析揭示了委員會架構在傳統設計典範下面臨的根本性困境：在「門檻安全性」的框架內，系統設計者被迫在效率與安全性之間做出取捨，無法同時最大化兩者。然而，這種困境的存在源於一個隱含的假設，即安全性的保障必須依賴於「降低委員會被攻破的機率」。若能改變安全性的實現方式本身，使其不再受制於機率運算，則委員會規模與安全性之間的強耦合關係便有望被解構。這一觀察為第**??** 章所提出的審計驅動型委員會 BlockDFL 提供了理論切入點。

2.3.4 基準系統模型：BlockDFL 委員會架構

為深入分析區塊鏈聯邦學習中委員會機制的安全性，本研究採用 BlockDFL [**qin2024blockdfl**] 作為基準系統模型。BlockDFL 於 2024 年發表於 WWW 會議，代表當前完全去中心化點對點聯邦學習架構的最新進展。該系統透過角色分離、權益加權選舉與拜占庭容錯共識的結合，在效率與安全性之間取得了當前文獻中的最佳平衡。本

節將詳細介紹其系統架構、運作流程與獎勵機制，作為後續威脅分析的基礎框架。

2.3.4.1 系統角色與職責定義

BlockDFL 採用角色分離的設計理念，將參與者依據其在每輪訓練中承擔的職責劃分為三種角色：更新提供者 (Update Provider)、聚合者 (Aggregator) 與驗證者 (Validator)。這種分工模式源於一個核心洞察，在去中心化環境中，若由單一節點同時負責訓練、聚合與驗證，將難以建立有效的制衡機制。透過將這三項職責分派給不同的參與者群體，系統得以在各環節引入相互監督，降低單點惡意行為對全域模型的影響。更新提供者構成系統中的多數參與者，其職責是利用本地私有資料執行模型訓練，並將訓練所得的模型更新提交給聚合者。由於訓練資料始終保留在本地裝置，更新提供者的隱私得以保護，這體現了聯邦學習「資料不動、模型動」的核心價值。

聚合者的職責則是收集來自多個更新提供者的本地更新，執行篩選與聚合運算，將結果打包為全域更新提案並提交給驗證者。每輪訓練中可能有多個聚合者同時運作，各自獨立收集更新並提交競爭性的提案，這種設計避免了單一聚合者壟斷聚合權的風險。驗證者組成委員會，負責評估各聚合者提交的提案品質，並透過拜占庭容錯共識機制選出最佳提案寫入區塊鏈。驗證者的數量通常遠小於更新提供者，以維持共識效率。角色的分配並非固定不變，而是在每輪訓練開始時根據上一區塊的雜湊值與各參與者的權益重新決定。具體而言，區塊雜湊被映射至一個雜湊環，每個參與者依據其權益大小佔據環上相應比例的空間。系統依序從雜湊環上選出聚合者與驗證者，未被選中者則成為更新提供者。

這種機制確保了角色分配的確定性與不可預測性，給定相同的區塊雜湊與權益分布，角色分配結果完全確定，便於驗證；另一方面，由於區塊雜湊在區塊產生前無法預知，參與者難以提前操控自身角色。更重要的是，權益加權的設計使得持有較多權益的參與者更有可能被選為聚合者或驗證者，這反映了系統對「高權益者傾向誠實」的信任假設。值得注意的是，權益不僅影響角色分配的機率，更在聚合過程中扮演關鍵角色。當聚合者收集更新時，來自高權益節點的更新不僅有更高機率被納入聚合候選集，在最終的加權聚合中也被賦予更高的權重，這意味著高權益節點對全域模型演化

方向的影響力顯著大於低權益節點。這種設計強化了權益與影響力之間的正相關關係，也為後續討論的權益累積攻擊提供了經濟基礎。

2.3.4.2 訓練輪次的運作流程

BlockDFL 的每輪訓練遵循一個結構化的流程，從角色分配開始，經由本地訓練、聚合、驗證與共識，最終完成全域模型更新與獎勵分配。訓練輪次始於角色分配階段，當新一輪開始時，所有參與者根據最新區塊的雜湊值與當前權益分布，確定性地運算出本輪的角色分配結果。由於運算過程僅依賴公開可驗證的資訊，任何參與者皆可獨立驗證角色分配的正確性，無需依賴中央協調者。角色確定後，更新提供者隨即進入本地訓練階段，在各自的私有資料集上執行若干輪隨機梯度下降，產生本地模型更新並將其發送給聚合者。這些更新僅包含模型參數的變化量，而非原始訓練資料，從而在協作學習與隱私保護之間取得平衡。

聚合階段是 BlockDFL 架構的關鍵環節，其設計直接影響了系統對惡意更新的抵抗能力以及權益機制的運作方式。每個聚合者獨立收集來自更新提供者的本地更新，當收集數量達到預設閾值後，開始執行篩選與聚合程序。篩選的目的在於過濾潛在的惡意更新，其過程分為兩個步驟：首先，聚合者依據更新提供者的權益進行加權隨機抽樣，權益較高的節點有更高機率被納入聚合候選集，這種設計基於「高權益節點更傾向誠實」的假設，同時也創造了一個正反饋機制，使得權益較高的節點更容易影響模型演化；其次，透過本地推論測試評估各候選更新的品質，排除表現異常者。通過篩選的更新隨後進行權益加權聚合，即各更新在聚合過程中的權重正比於其提供者的權益佔比，這種設計使得高權益節點對全域模型的影響力進一步放大。聚合完成後，聚合者將結果打包為全域更新提案，簽署後提交給驗證者委員會。

值得注意的是，由於多個聚合者同時運作且各自獨立收集更新，同一輪中將產生多個競爭性的提案，這為後續的驗證階段提供了選擇空間。驗證與共識階段決定了哪個提案將被接受並寫入區塊鏈。當驗證者委員會收到足夠數量的提案後，驗證程序啟動。每位驗證者獨立使用 Krum 演算法 [**blanchard2017machine**] 對所有提案進行評分，Krum 分數較低者代表與其他提案的整體距離較小，被視為品質較高。基於評分結果，

驗證者對各提案進行投票，僅當某提案的 Krum 分數優於三分之二以上的其他提案時，驗證者才會投下贊成票。投票過程遵循簡化的 PBFT 協議，當某提案獲得超過三分之二驗證者的贊成票時，該提案被正式接受。委員會的領導者隨即將接受的提案連同相關資訊打包成新區塊，廣播至全網。所有參與者收到新區塊後，依據其中的全域更新同步更新本地模型，完成本輪訓練。

2.3.4.3 獎勵機制與激勵設計

BlockDFL 的獎勵機制遵循「有貢獻才有回報」的設計原則，旨在解決分散式系統中普遍存在的搭便車問題。在傳統聯邦學習中，無論參與者是否真正貢獻高品質的訓練成果，皆可獲得最終全域模型的使用權，這削弱了誠實參與的激勵。BlockDFL 透過將權益獎勵與「被選中」直接綁定，確保只有對本輪全域模型更新有實質貢獻的參與者才能獲得回報，從而建立起正向的激勵結構。具體而言，當某一提案通過委員會共識並被寫入區塊鏈時，系統將權益增量分配給三類參與者。第一類是提交該提案的聚合者，其承擔了收集更新、執行篩選與聚合運算的工作，並承受提案可能未被選中的風險。第二類是本地更新被納入該提案的更新提供者，他們貢獻了訓練運算資源與本地資料的價值。第三類是對該提案投下贊成票的驗證者，他們執行了驗證運算並參與了共識決策。這三類參與者均分本輪的權益獎勵，其身份被明確記錄於區塊之中，確保獎勵分配的透明與可驗證。

相對地，未對本輪全域模型更新做出貢獻的參與者則不獲得任何獎勵。這包括提案未被選中的其他聚合者、本地更新未被納入獲選提案的更新提供者、以及對獲選提案投下反對票或未參與投票的驗證者。這種設計創造了明確的激勵導向，聚合者有動機提交高品質的提案以提高被選中的機率，更新提供者有動機提交優質的本地更新以增加被納入提案的可能性，驗證者則有動機投票支持真正優質的提案，因為只有投票與最終結果一致時才能獲得獎勵。然而，此獎勵機制在激勵誠實行為的同時，也創造了一個具有正反饋特性的動態系統。當參與者獲得權益獎勵時，其總權益增加，這直接提升了該參與者在未來輪次被選為聚合者或驗證者的機率，進而增加其獲得更多獎勵的機會。

這種正反饋循環透過前述的權益加權機制進一步強化，高權益節點不僅更容易被選為關鍵角色，其作為更新提供者時提交的更新也更容易被納入聚合，並在聚合過程中獲得更高權重。BlockDFL 的設計者預期此正反饋將使持續誠實貢獻的參與者逐漸累積優勢，而惡意參與者的影響力則相對削弱 [**qin2024blockdfl**]。這項預期建立在一個關鍵假設之上，獲得高權益的參與者必然是長期誠實貢獻者。然而，若攻擊者能夠在潛伏期間偽裝成誠實參與者並成功累積權益，此正反饋機制反而可能成為攻擊者鞏固優勢的工具。更值得關注的是，權益在聚合過程中的雙重作用，既影響被選中機率又決定聚合權重，使得高權益攻擊者不僅能更頻繁地參與決策，更能在參與時施加更大的影響力，這種複合效應顯著放大了權益累積對系統安全性的潛在威脅。

2.3.4.4 本研究採用此模型的理由

本研究選擇 BlockDFL 作為基準系統模型，基於以下四項考量。首先，BlockDFL 代表了區塊鏈聯邦學習委員會架構的最新技術水準，其於 2024 年發表於頂級網路研討會，融合了角色分離、權益加權選舉、雙層評分機制與拜占庭容錯共識等多項先進設計，具有高度的代表性。其次，BlockDFL 的系統定義清晰完整，論文詳細說明了角色職責、運作流程與獎勵規則，這為形式化的威脅分析提供了堅實基礎。相較於部分僅提供概念性描述的研究，BlockDFL 的明確定義使得安全性分析能夠建立在具體的系統行為之上，而非抽象的假設。第三，BlockDFL 的委員會機制與獎勵設計具有廣泛的通用性。儘管不同 BCFL 系統在具體實現上存在差異，但「小型委員會執行共識」與「權益驅動的角色選舉」已成為此領域的主流設計典範。因此，針對 BlockDFL 所發現的安全問題與提出的防禦機制，在原理上可推廣至採用類似架構的其他系統。最後，BlockDFL 已有公開的實驗資料與效能基準，這為本研究後續的防禦機制評估提供了可比較的參照點。綜合以上考量，BlockDFL 是本研究進行威脅建模與防禦設計的理想分析對象。

2.4 現有驗證方法與其局限

2.4.1 密碼學驗證方法的運算瓶頸

零知識機器學習（Zero-Knowledge Machine Learning, zkML）代表了密碼學驗證方法在機器學習領域的前沿探索 [**chen2024zkml**]。其核心理念是將機器學習運算轉換為算術電路，並生成零知識證明，使驗證者無需重新執行運算即可確認結果的正確性。這種方法在理論上提供了最強的安全保證：證明的正確性完全依賴密碼學假設，無需信任任何參與方。若能將 zkML 應用於聯邦學習的聚合驗證，系統將能在不揭露個別更新內容的前提下，證明聚合結果確實是由指定的本地更新按照預定規則運算而得。

然而，zkML 面臨嚴峻的運算效能挑戰，這使其在當前技術條件下難以應用於實際的聯邦學習系統。將神經網路運算轉換為算術電路的過程會產生大量的多項式約束，約束數量隨模型複雜度急劇膨脹。根據現有基準測試，即使是相對簡單的 LeNet [**lecun1998gradient**] 模型，其約束數量也可達數億級別；對於 ResNet-18 [**he2016deep**] 等級的模型，證明生成時間需要近一分鐘；而對於 VGG16 [**simonyan2015very**] 或更大規模的模型，證明生成可能耗時數十分鐘甚至數小時，且需要數百 GB 乃至 TB 級別的記憶體 [**chen2024zkml**]。考慮到聯邦學習通常需要數百至數千輪迭代，每輪都執行如此耗時的證明生成顯然不切實際。

更關鍵的局限在於 zkML 難以支援拜占庭容錯聚合演算法。Krum [**blanchard2017machine**] 與 Multi-Krum 等防禦性聚合方法需要運算所有客戶端更新之間的成對距離，這在零知識電路中會產生 $O(N^2 \cdot d)$ 的約束爆炸，其中 N 為客戶端數量，d 為模型參數維度。排序與中位數運算在算術電路中同樣極度昂貴。現有的 zkFL 方案如 RiseFL [**zhu2024risefl**] 僅能支援 L2-norm 有效性檢查等簡單驗證，而無法實現完整的拜占庭容錯聚合驗證。這意味著即使克服了效能瓶頸，zkML 仍無法為採用 Krum 等防禦機制的系統（如 BlockDFL）提供聚合正確性的密碼學證明。

2.4.2 樂觀執行方法的架構限制

樂觀機器學習（Optimistic Machine Learning, opML）採用與 zkML 截然不同的設計哲學：預設所有運算結果都是正確的，僅在有參與者提出質疑時才啟動驗證程序 [**conway2024opml**]。這種「樂觀執行」的模式大幅降低了正常情況下的運算成本，因為絕大多數時候驗證程序不會被觸發。當爭議發生時，系統透過互動式的二分協議（Bisection Protocol）逐步縮小爭議範圍，最終定位至單一運算步驟，由鏈上的欺詐證明虛擬機（Fraud Proof Virtual Machine, FPVM）進行仲裁。ORA Protocol 已展示此方法可支援 LLaMA 2 [**touvron2023llama**] 等數十億參數規模的模型在以太坊上運行 [**ora2024opml**]。

然而，opML 的架構設計與聯邦學習的需求存在根本性的衝突。首先是挑戰期的問題。為確保驗證者有充足時間偵測並提交欺詐證明，主流的樂觀執行系統如 Optimism [**optimism2024rollup**] 與 Arbitrum [**kalodner2018arbitrum**] 採用長達一週的挑戰期。這種設計對於區塊鏈交易的最終確認或許可以接受，但對於需要快速迭代的聯邦學習訓練而言則完全不可行。若每輪聚合都需等待一週才能確認，數百輪的訓練將耗時數年。即使大幅縮短挑戰期，仍會顯著拖慢訓練進度，且可能因驗證者反應時間不足而削弱安全保證。

其次，opML 的信任模型與 BCFL 的多驗證者場景存在落差。opML 建立在「AnyTrust」假設之上：只要存在至少一個誠實的驗證者願意監控並挑戰錯誤結果，系統就是安全的。這本質上是一個單一提交者與單一挑戰者之間互相對抗的兩方爭議模型。然而，BCFL 的委員會共識涉及多個驗證者對多個提案的集體決策，這種多方參與的結構難以直接套用 opML 的爭議解決框架。此外，opML 的設計假設運算輸入（如模型更新）是公開可見的，以便驗證者能夠重新執行運算並發現錯誤。這與聯邦學習對更新隱私的保護需求存在張力。

2.4.3 委員會驗證方法的安全假設

相較於密碼學方法與樂觀執行方法，基於委員會共識的驗證方法在效率與實用性之間取得了較佳的平衡，因而成為當前 BCFL 系統的主流選擇。這類方法的核心思想是由一個小型委員會代全網執行驗證與共識，透過拜占庭容錯協議確保只要委員會中的誠實成員佔據多數，驗證結果就是可信的。前文介紹的 BlockDFL 即屬此類，其他代表性系統包括 FLCoin [**ren2024scalable**] 與 BFLC [**li2021blockchain**]。

FLCoin 提出基於滑動視窗的動態委員會機制，將聯邦學習的貢獻歷史作為委員會成員資格的依據。節點透過提交有效的模型更新獲得「份額」，在固定大小的滑動視窗內持有份額的節點組成當輪委員會。這種設計使委員會組成與 FL 目標直接對齊，且透過視窗的滑動實現成員的動態更替。FLCoin 的安全性分析顯示，在全網惡意節點比例不超過 25% 且視窗大小為 100 的條件下，委員會安全的機率可達 98.4% [**ren2024scalable**]。然而，論文並未深入分析惡意節點透過策略性參與逐步累積份額的可能性，其實驗也假設無惡意節點參與，未驗證對抗性累積策略的防禦效果。

BFLC 開創性地將委員會共識引入 BCFL，採用聲譽機制決定委員會組成 [**li2021blockchain**]。系統追蹤各節點提交更新的品質歷史，高聲譽者優先被選入委員會。委員會成員使用 K-fold 交叉驗證評估提交的模型更新，透過共識決定是否接受。這種設計能有效過濾曾有不良記錄的節點，但也面臨冷啟動問題：新加入者缺乏歷史記錄，難以建立初始信任。更重要的是，後續研究明確指出 BFLC「容易被惡意節點混入委員會，從而導致系統偏差」[**qin2024blockdfl**]，當惡意節點佔據委員會半數席位時即可發動攻擊。

綜觀這些委員會驗證方法，它們共享一個根本性的安全假設：委員會的誠實多數。無論採用隨機抽樣、權益加權、聲譽評分或貢獻歷史等任何選舉機制，所有方案都假設某種形式的誠實多數能夠在委員會層級得到維持。這一假設在面對靜態的、比例固定的惡意節點時或許成立，但當攻擊者採取動態的、策略性的行為時，其有效性便值得商榷。下一小節將進一步分析這種靜態假設的盲區。

2.4.4 系統性局限：靜態分析的盲區

回顧上述三類驗證方法，可以發現一個貫穿其中的系統性局限：現有的安全性分析幾乎都建立在「攻擊者資源固定」的靜態假設之上。zkML 的安全性證明假設攻擊者的運算能力無法突破密碼學難題；opML 假設至少存在一個持續監控的誠實驗證者；委員會方法則假設惡意節點在全網中的比例 α 是一個固定的常數，並據此運算委員會被攻破的機率。這種「快照式」的分析視角忽略了一個關鍵的動態因素：理性的攻擊者會根據系統狀態調整其策略，而非機械地執行固定行為。

以委員會方法為例，第 **??** 節的超幾何分佈分析假設每輪選舉都是從固定的惡意節點比例中獨立抽樣。然而，若系統的獎勵機制存在正反饋特性（例如在 BlockDFL 架構中，獲得獎勵的節點將累積更多權益，進而提高未來被選中的機率），則各輪選舉之間並非獨立事件。攻擊者可利用此特性採取「耐心策略」：在初期表現完全誠實以累積權益與聲譽，僅在其控制的節點佔據委員會優勢時才發動攻擊。這種行為模式無法被靜態的機率分析所捕捉。

更根本的問題在於，現有分析未區分「惡意節點的存在比例」與「惡意節點的有效影響力」。在權益加權的選舉機制中，一個持有 10% 權益的惡意節點，其對委員會組成的影響力可能遠大於十個各持有 1% 權益的惡意節點。若攻擊者能夠透過合法途徑（累積獎勵、建立聲譽）集中權益於少數節點，即使其控制的節點數量佔全網比例不高，仍可能在委員會層級取得超額的影響力。FLCoin 的分析雖然考慮了惡意節點比例與委員會安全機率的關係，但未分析惡意節點比例本身可能隨時間演變的動態過程。BlockDFL 假設「持有大量權益的參與者傾向誠實」，但未充分論證此假設在面對長期潛伏的策略性攻擊者時是否仍然成立。

表 **??** 總結了三類驗證方法在安全性、效率與通用性三個維度上的特性。可以看出，沒有任何一種方法能夠同時滿足所有需求：zkML 提供最強的安全保證但效率過低且無法支援複雜聚合；opML 效率較高但挑戰期過長且架構不適用於多驗證者場景；委員會方法在效率與實用性上表現最佳，但其安全性依賴可能被違反的誠實多數假設。這種「安全性-效率-通用性」的三難困境，構成了本領域研究的核心挑戰。

表 2.1: BCFL 驗證方法比較

方法類別	安全性基礎	主要效率瓶頸	對 BCFL 的適用性
zkML	密碼學證明，無需信任假設	證明生成耗時數分鐘至數小時	低：無法支援 Krum 等複雜聚合
opML	經濟激勵與 AnyTrust 假設	挑戰期長達數天	低：架構不適用於多驗證者場景
委員會共識	委員會誠實多數假設	共識通訊成本 $O(c^2)$	高：主流選擇，但假設可能被違反

2.5 研究缺口：從激勵設計到安全隱患

前述分析揭示了現有 BCFL 驗證方法的共同盲區：靜態的安全性假設無法應對動態的攻擊策略。本節將聚焦於當前最具實用性的委員會驗證方法，深入探討其獎勵機制如何在激勵誠實行為的同時，也為策略性攻擊者創造了可利用的漏洞。這一分析將指向本研究欲填補的關鍵缺口，並為第三章的威脅模型建構奠定基礎。

2.5.1 獎勵機制的正反饋特性

第 **??** 節詳細介紹了 BlockDFL 的獎勵機制：當某一提案通過委員會共識時，提交該提案的聚合者、更新被納入的更新提供者、以及投贊成票的驗證者共同分享權益獎勵。這種設計的初衷是解決搭便車問題，確保只有實質貢獻者才能獲得回報。然而，從系統動態的角度審視，此機制創造了一個具有正反饋特性的循環結構。

正反饋的運作邏輯可描述如下：節點獲得權益獎勵後，其總權益增加；由於角色選舉採用權益加權機制，權益增加直接提升該節點在未來輪次被選為聚合者或驗證者的機率；作為聚合者或驗證者，節點更有機會參與被接受的提案，從而獲得更多獎勵。這種「獎勵 → 權益增加 → 選中機率提升 → 更多獎勵」的循環，在數學上構成一個正反饋系統。BlockDFL 的設計者預期此特性將使誠實參與者逐漸累積優勢，因為他們持續提交高品質更新並誠實驗證，從而獲得穩定的獎勵流入。相對地，惡意參與者若因提交低品質更新或投票與最終結果不一致而錯失獎勵，其相對權益份額將逐漸稀釋。

然而，這一預期建立在一個隱含假設之上：系統能夠有效區分誠實行為與惡意行

為，並據此差異化地分配獎勵。問題在於，當惡意節點選擇「偽裝誠實」，意即在攻擊發動前完全模仿誠實節點的行為，系統便無法將其與真正的誠實節點區分開來。在這種情況下，惡意節點同樣能夠獲得獎勵、累積權益、提升影響力，正反饋機制反而成為攻擊者滲透系統的工具。

2.5.2 策略性攻擊者的潛在威脅

基於上述分析，可以設想一種策略性的攻擊模式：攻擊者控制的節點在初期階段完全表現為誠實參與者，正常參與訓練、提交高品質更新、在驗證時投票支持真正優質的提案。透過這種「潛伏」行為，攻擊節點逐步累積權益與聲譽，提高其被選為聚合者或驗證者的機率。當攻擊者評估其控制的節點已在委員會中佔據足夠優勢時（例如在某一輪選舉中恰好佔據超過三分之一的驗證者席位），才發動實際攻擊。

這種攻擊模式的危險性在於其隱蔽性與自我強化性。在潛伏階段，攻擊節點的行為與誠實節點完全一致，現有的異常偵測機制無從識別。更重要的是，一旦攻擊者在某一輪成功控制委員會，他們可以操控共識結果，將獎勵僅分配給自己控制的節點（透過接受包含攻擊節點更新的提案、拒絕僅包含誠實節點更新的提案），同時確保攻擊節點作為「投贊成票的驗證者」獲得驗證獎勵。這種操控使攻擊者的權益份額在成功攻擊後加速增長，進一步提高其在未來輪次控制委員會的機率，形成惡性循環。

現有文獻已零星地意識到委員會滲透的風險。FedBlock 在其未來展望中指出：「目前版本中，智能合約以隨機方式選取客戶端作為驗證者，但此選擇標準可能並非最佳」，並警告「驗證者本身亦可能是惡意的」[**nguyen2024fedblock**]。BFLC 的後續評論指出該系統「容易被惡意節點混入委員會」[**qin2024blockdfl**]。然而，這些觀察多停留在定性描述的層次，尚未發展出系統性的威脅模型來形式化分析此類攻擊的具體機制、成功條件與潛在危害。特別是，現有研究未充分探討獎勵機制的正反饋特性如何與委員會選舉機制交互作用，使得攻擊者能夠透過「合法」途徑逐步擴大影響力。

2.5.3 本研究的切入點

綜合本章的分析，可以明確界定當前 BCFL 委員會安全性研究的核心缺口：

第一，**缺乏針對策略性攻擊者的形式化威脅模型**。現有安全性分析假設惡意節點比例固定且行為模式單一，未考慮攻擊者可能採取的動態策略，如長期潛伏、選擇性攻擊時機、以及利用正反饋機制累積優勢。

第二，**缺乏對獎勵機制安全性意涵的深入分析**。現有研究將獎勵機制視為激勵誠實行為的工具，但未系統性地審視其正反饋特性可能被攻擊者利用的風險。

第三，**缺乏能夠偵測與抵禦漸進式滲透的防禦機制**。無論是隨機選擇、權益加權或聲譽導向，現有委員會選舉機制都未針對「透過合法途徑累積影響力」的攻擊模式設計相應的防護措施。

針對上述缺口，本研究將在後續章節展開以下工作。第三章將形式化定義「漸進式委員會佔領攻擊」（Progressive Committee Capture Attack, PCCA）的威脅模型，基於第 **??** 節建立的 BlockDFL 系統模型，分析攻擊者如何利用獎勵機制的正反饋特性逐步滲透委員會，並量化攻擊在不同參數設定下的成功機率與所需時間。第四章將提出針對 PCCA 的防禦架構設計，透過打破正反饋循環與引入動態挑戰機制，在維持系統效率的前提下提升對策略性攻擊者的抵抗能力。第五章將透過實驗評估所提出方法的有效性，驗證其在多種攻擊場景下的防護效能。

第三章　威脅模型

基於第二章所建立的委員會架構系統模型，本章將深入剖析該架構在面對理性攻擊者時所呈現的安全脆弱性。本章的核心任務在於定義並分析「漸進式委員會佔領攻擊」(Progressive Committee Capture Attack, PCCA)，這是一種專門針對權益機制設計缺陷的隱蔽性攻擊手法。透過揭示攻擊者如何利用權益機制內建的正反饋特性逐步實現網路控制權的轉移，本章為後續章節的防禦機制設計提供明確的安全目標與理論基礎。值得特別強調的是，這種攻擊與傳統的模型投毒攻擊存在本質性差異，其危險性並非體現在對單一模型品質的破壞，而是從根本上顛覆了去中心化系統的安全假設，能夠將表面上維持去中心化形態的聯邦學習系統，實質上重新集權化至攻擊者手中。

3.1　攻擊者模型

3.1.1　攻擊者類型：理性攻擊者

本研究所考慮的攻擊者屬於理性攻擊者 (Rational Adversary) 範疇，這與傳統區塊鏈安全研究中常見的拜占庭攻擊者存在本質性差異。拜占庭攻擊者的行為動機往往是純粹的破壞性，他們可能採取任意惡意行為來癱瘓系統，即使這些行為會導致自身利益受損也在所不惜，這種攻擊模型源自於分散式系統理論中對最壞情況的假設。然而，在實際的區塊鏈應用場景中，攻擊者往往具有明確的經濟動機而非單純追求破壞，他們的行為模式遵循經濟理性原則，首要目標是利益最大化。這意味著理性攻擊者會仔細評估每次攻擊行為的預期收益與成本，只有當預期收益明顯大於成本時才會採取行動，而如果能夠透過機制設計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，無需依賴傳統的誠實多數假設。這種區分為基於博弈論的防禦機制提供了理論基礎，也是本研究設計激勵相容機制的關鍵前提。

理性攻擊者的目標體系呈現出多層次性與長期性的特徵，這種複雜的目標結構使得攻擊行為更加隱蔽且難以偵測。在最直接的層面，攻擊者追求經濟利益的最大化，

具體表現為透過操縱委員會來獨佔訓練獎勵，將誠實節點排除在獎勵分配機制之外。然而，這種短期經濟收益只是攻擊者目標體系的表層，更深層的目標在於權益壟斷與網路控制，透過系統性地阻止誠實節點的權益增長，攻擊者能夠逐步提高自身在整個系統中的權益佔比，這種權益佔比的提升會直接轉化為在委員會選擇過程中的優勢地位。當攻擊者的權益佔比達到某個臨界點後，他們將能夠更頻繁地控制委員會的組成，進而掌握聯邦學習過程中的關鍵決策權，包括決定哪些模型更新會被接受、哪些會被拒絕。這種從經濟收益到網路控制的轉變體現了攻擊者策略的長期性與系統性，也是 PCCA 攻擊之所以危險的根本原因，因為它並非僅僅影響模型品質，而是從根本上顛覆了去中心化系統的權力結構。

3.1.2 攻擊者能力與限制

在能力方面，本研究假設攻擊者能夠控制系統中一定比例的驗證者節點，這個比例記為 f，在典型的威脅場景下我們假設 $f \leq 0.3$，即攻擊者控制的惡意節點總數 M 滿足 $M = f \times N$，其中 N 為全網節點總數。這意味著在典型的威脅場景下攻擊者最多控制全網 30% 的節點。這個假設並非任意設定，而是基於實際區塊鏈系統中攻擊者資源有限的現實考量，因為控制更高比例的節點需要投入大量的經濟資源與協調成本。被攻擊者控制的這些節點並非孤立運作，而是能夠相互協調並共同執行精心設計的攻擊策略，例如當多個惡意節點同時被選入同一個委員會時，它們可以串通一致地投票，形成協同作惡的局面。更值得注意的是，攻擊者具備策略性調整能力，能夠根據系統的動態狀態靈活改變行為模式，在權益積累的早期階段可能完全表現誠實以建立信譽並累積資源，而一旦獲得委員會的多數席位便會立即切換至攻擊模式。此外，攻擊者擁有完整的觀察能力，可以追蹤區塊鏈上的所有公開資訊，包括其他節點的權益分布、歷史行為記錄、委員會組成變化等，並基於這些資訊進行精確的策略規劃。

然而，攻擊者的能力並非無限，其行為同時受到多個維度的約束，這些約束為防禦機制的設計提供了重要的切入點。從密碼學角度來看，攻擊者無法突破系統所採用的密碼學原語，這意味著他們既無法偽造其他節點的數位簽章，也無法篡改已經寫入

區塊鏈的歷史資料，區塊鏈的不可篡改性為系統提供了可靠的審計基礎。在網路控制層面，攻擊者的節點數量受到經濟成本的限制，無法達到發動傳統 51% 攻擊所需的絕對多數，這使得攻擊者必須採用更為精細的策略來實現其目標。更關鍵的是，理性攻擊者的行為受到經濟激勵的根本性約束，如果精心設計的防禦機制能夠確保攻擊的預期成本大於潛在收益，那麼理性攻擊者將不會嘗試發動攻擊。此外，系統的可驗證性特徵為防禦提供了重要基礎：攻擊者無法阻止其他節點獨立驗證聚合結果的正確性，任何參與者都可以重新執行聚合演算法並檢測委員會是否正確遵守協議規則，這種透明性與可驗證性為後續設計挑戰機制奠定了技術可行性基礎。

3.2 攻擊向量分析

區塊鏈聯邦學習系統作為一個多層次的複雜架構，其安全威脅同樣呈現出層次化的特徵，不同層次的攻擊具有截然不同的目標、手法與防禦需求。本節的目標是系統性地分析不同層次的攻擊向量，釐清各層防禦的現狀與局限，進而明確本研究的關注焦點。這種層次化的分析框架不僅有助於理解 PCCA 攻擊的獨特性，也能揭示現有研究在安全分析上存在的系統性盲點，為後續的防禦機制設計提供清晰的問題定位。

3.2.1 資料層攻擊：已有防禦

資料層攻擊主要針對聯邦學習的訓練階段，透過污染訓練資料或模型更新來破壞最終模型的品質，這類攻擊在聯邦學習安全研究中已經得到廣泛的關注與深入的探討。具體而言，惡意客戶端可能採用資料投毒 (Data Poisoning) 手段，在本地訓練時刻意使用被污染的資料集，導致產生的模型更新偏離正常分布，從而影響全域模型的收斂方向。另一種更直接的攻擊方式是模型投毒 (Model Poisoning)，惡意客戶端不經過真實的訓練過程，而是直接構造精心設計的惡意模型更新向量，這些更新可能包含後門觸發器或導向特定的錯誤分類行為。針對這類資料層威脅，現有的聯邦學習研究已經發展出相對成熟的防禦框架，其中最具代表性的是拜占庭強健聚合演算法，如 Krum [**blanchard2017machine**]、Trimmed Mean [**yin2018byzantine**]、Median 等方法，這

些演算法的核心思想是利用統計學方法識別並過濾異常的模型更新，即使在存在一定比例惡意客戶端的情況下，仍能保證全域模型朝著正確的方向收斂。

然而，這些看似完備的防禦方法實際上建立在一個關鍵但往往被忽視的假設之上：執行這些防禦演算法的驗證者本身是誠實的。這個假設在傳統的中心化聯邦學習場景中或許是合理的，因為中心化伺服器的可信度通常由組織層面的信任保證，但在去中心化的區塊鏈聯邦學習系統中，驗證者同樣是由網路中的普通節點擔任，並沒有任何外部的信任背書。如果驗證者本身受到攻擊者控制，他們完全可以選擇不執行這些拜占庭強健演算法，或者更隱蔽地篡改演算法的執行結果，宣稱執行了防禦措施但實際上接受了惡意更新。在這種情況下，無論資料層的防禦演算法設計得多麼精妙，都將完全失去效力。這揭示了一個根本性的問題：資料層防禦的有效性依賴於共識層的安全性，如果共識層本身被攻陷，資料層的所有防線都將不攻自破，這種層次間的依賴關係構成了現有防禦體系的結構性弱點。

3.2.2 共識層攻擊：本研究重點

相較於已經得到充分研究的資料層攻擊，針對共識層的攻擊則構成了本研究的核心關注對象，這類攻擊的目標不是訓練資料或模型更新本身，而是負責執行聚合和驗證工作的委員會機制。驗證者共謀 (Verifier Collusion) 是這類攻擊的典型形式，多個惡意驗證者可以透過事先協調，在投票環節協同作惡，共同通過明顯包含錯誤或惡意特徵的聚合結果。更具威脅性的是委員會佔領 (Committee Capture) 攻擊，攻擊者不滿足於偶然的共謀機會，而是試圖系統性地操縱委員會選擇機制，逐步增加惡意節點在委員會中的席位佔比，最終實現對委員會的持續性控制。如第 **??** 章的文獻回顧所揭示的，現有區塊鏈聯邦學習研究在這個層面存在系統性的「驗證層盲點」，統計資料顯示約 93% 的相關研究在設計系統時隱含地假設驗證者是誠實的或者至少滿足誠實多數的條件，僅有極少數研究明確考慮了惡意驗證者可能存在的場景並嘗試設計相應的防禦機制。

更值得關注的是，即使在引入了驗證者機制的 BlockDFL 類系統中，大多數研究

仍然假設聚合者和驗證者之間在利益上是相互獨立的，或者至少驗證者群體內部維持著誠實多數。本研究指出了一個被普遍忽視的風險：驗證者和聚合者完全可能形成利益集團 (Cartel)，攻擊者可以同時滲透委員會與聚合節點，形成從上游到下游的完整控制鏈。這種「全棧控制」的風險是對現有 BlockDFL 架構安全分析的重要補充，也是 PCCA 攻擊得以成功的關鍵條件之一。共識層攻擊之所以比資料層攻擊更加危險，在於其具有三個顯著特徵：首先是防禦繞過能力，一旦委員會被惡意節點控制，所有的資料層防禦機制都可以被直接忽略或篡改；其次是隱蔽性，攻擊者在權益積累的早期階段可以完全表現誠實，不會觸發任何異常檢測機制；第三是自我強化特性，一旦攻擊成功，攻擊者將獲得更多獎勵，導致其權益進一步增加，形成正反饋循環。

3.2.3 攻擊層次對比

為了更清晰地呈現不同層次攻擊的特徵差異與防禦現狀，表 **??** 提供了系統性的對比分析，這種對比有助於理解本研究選擇聚焦於共識層攻擊的理論依據。

表 3.1: 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
資料層	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	網路控制	誠實多數假設	多數驗證者誠實	是

從表中可以清楚地看到，資料層攻擊已經發展出相對完善的防禦方法體系，但這些方法的有效性建立在驗證者誠實執行協議的假設之上，相比之下，共識層攻擊的防禦仍然停留在依賴誠實多數假設的階段，缺乏針對理性攻擊者的激勵相容機制。這種防禦上的不對稱性正是本研究需要填補的關鍵空白。更深層次地看，資料層防禦與共識層防禦之間存在著依賴關係：前者的有效性完全取決於後者的可靠性，因此即使投入再多的研究資源去最佳化資料層的拜占庭強健演算法，如果不能從根本上解決共識層的安全問題，整個防禦體系仍然建立在不穩固的基礎之上。這種認識促使本研究將焦點放在共識層的安全性分析與防禦機制設計上，而非繼續在資料層防禦的技術細節上進行增量式的改進。

3.3 漸進式委員會佔領攻擊

本節將詳細定義本研究針對的核心威脅：漸進式委員會佔領攻擊 (Progressive Committee Capture Attack, PCCA)。這是一種專門針對基於權益的委員會選擇機制的隱蔽性攻擊手法，其獨特之處在於透過精心設計的兩階段策略，利用權益機制內在的正反饋特性，實現從小規模滲透到顯著優勢地位的漸進式轉變，最終建立起對委員會決策的持續性影響力。

3.3.1 攻擊定義與核心機制

PCCA 的本質是一種針對權益衍生系統的經濟攻擊，其核心在於利用「權益-選舉-獎勵-權益」這一閉環機制中存在的正反饋特性。在正常運作的權益證明系統中，節點的權益決定了其被選入委員會的機率，而成功參與委員會工作又會獲得獎勵從而增加權益，這種設計的初衷是激勵節點誠實參與，但攻擊者可以將這一機制轉化為累積優勢的工具。PCCA 的攻擊策略分為兩個明確的階段：在潛伏階段，攻擊者控制的節點完全遵守協議規則，表現得與誠實節點無異，目的是積累初始權益並建立良好的信譽記錄，這個階段的持續時間取決於攻擊者的初始資源與委員會的隨機選擇結果。攻擊者會持續觀察系統狀態，等待一個關鍵的時機窗口：當多個惡意節點恰好同時被選入同一個委員會，且其席位數超過委員會總席位的三分之二時，攻擊便進入第二階段。

在佔領階段，攻擊者利用在委員會中的多數優勢，啟動「戰略性餓死」(Strategic Starvation) 策略，這種策略的核心不是直接破壞模型品質，而是透過操縱投票結果來控制獎勵分配。具體而言，惡意委員會會系統性地拒絕由誠實節點主導的聚合提案，即使這些提案包含高品質的模型更新，由於區塊鏈聯邦學習系統通常採用「提案-投票-獎勵」的連動機制，被拒絕的提案意味著相關的聚合者和更新提供者都無法獲得本輪獎勵。透過持續執行這種排他性策略，惡意節點能夠獲得相對於誠實節點更高比例的系統獎勵，逐步擴大其權益優勢。隨著攻擊者權益佔比的提升，其在未來委員會選舉中獲得多數席位的機率也會相應提高，形成自我強化的正反饋循環。演算法 **??** 以形式化

的方式呈現了 PCCA 的決策邏輯，清晰展示了攻擊者如何根據當前控制比例動態調整其行為模式。

Algorithm 1 High-Level Strategy of Progressive Committee Capture Attack (PCCA)

Require: Current Committee $\mathcal{V}$, Adversary Controlled Nodes $\mathcal{C}_{adv}$
Ensure: Action for the current round
Check Phase: Calculate control ratio $\rho = \frac{|\mathcal{V} \cap \mathcal{C}_{adv}|}{|\mathcal{V}|}$
if $\rho \leq 2/3$ **then**
State 1: Shadow Mode (Lurking)
    Follow the protocol honestly to accumulate stake and await majority.
else
State 2: Capture Mode (Occupying)
    if Aggregator is Adversarial then
        Full Stack Poisoning: Force approve malicious proposal.
    else
        Strategic Starvation: Force reject honest proposal.
    end if
end if

攻擊者在每一輪開始時都會運算其在當前委員會中的控制比例 ρ，這個比例決定了攻擊者採取的行為模式。當控制比例未超過三分之二時，攻擊者進入「影子模式」，嚴格遵守協議規則以避免暴露身份並持續積累權益，一旦控制比例超越臨界值，攻擊者立即切換至「佔領模式」。此時的具體策略取決於當輪聚合者的身份：如果聚合者本身也受攻擊者控制，那麼整個提案-驗證鏈條都在攻擊者掌握之中，此時可以執行更激進的「全棧投毒」策略，直接將包含惡意內容的模型更新寫入區塊鏈；如果聚合者為誠實節點，攻擊者則採用相對保守的「戰略性餓死」策略，透過拒絕誠實提案來實現經濟層面的打擊，同時避免在技術層面留下明顯的攻擊痕跡。

3.3.2 攻擊階段詳述

3.3.2.1 階段一：潛伏階段

潛伏階段是 PCCA 攻擊成功的關鍵前提，其核心目標是在不引起任何懷疑的情況下，為後續的佔領階段創造必要條件。在這個階段，攻擊者面臨的主要挑戰是如何在誠實行為與權益積累之間取得平衡，由於委員會的選擇基於權益加權的隨機抽樣，攻擊者的初始權益佔比直接決定了其節點被選入委員會的機率，進而影響多個惡意節點

同時入選的可能性。假設攻擊者控制全網 $f = 0.3$ 的節點，而委員會大小為 $C = 7$，那麼要形成超過三分之二的多數優勢，至少需要 5 個惡意節點同時被選中。根據第 **??** 節的超幾何分布分析，這種情況發生的機率約為 2.4%，這意味著攻擊者平均需要等待約 42 輪才能獲得一次發動攻擊的機會，這種低頻率的攻擊窗口使得潛伏階段可能持續相當長的時間。

在這漫長的等待期間，攻擊者必須維持完美的誠實表現以避免被識別為可疑節點。當攻擊者控制的節點被選為更新提供者時，它們會基於本地資料集進行真實的模型訓練，提交符合協議規範的高品質更新；當被選為聚合者時，它們會正確執行聚合演算法，包括運行 Krum 等拜占庭強健機制來過濾異常更新；當被選為驗證者時，它們會認真驗證聚合結果的正確性，對誠實的提案投贊成票，對存在問題的提案投反對票。這種全方位的誠實表現不僅能夠幫助攻擊者積累權益，更重要的是建立起良好的歷史記錄，使得其他節點和監督機制都將其視為可信的誠實參與者。潛伏階段的持續時間是彈性的，攻擊者會根據權益積累的速度與委員會組成的隨機結果動態調整策略，在確保安全的前提下耐心等待最佳的攻擊時機。

3.3.2.2 階段二：佔領階段

當攻擊者在系統中累積了足夠的權益並成功控制了某一輪委員會的超過三分之二席位時，PCCA 進入最關鍵的佔領階段。與傳統攻擊採取單一破壞模式不同，PCCA 在佔領階段展現出高度的策略彈性，根據攻擊者對系統不同組件的控制程度採取不同層次的攻擊手法，這種分層策略設計使得攻擊既能最大化經濟收益，又能根據實際情況控制暴露風險。

場景一：戰略性餓死 (Strategic Starvation via Committee Capture) 在第一種場景中，攻擊者成功控制了驗證者委員會的絕對多數席位，但當輪的聚合者角色仍由誠實節點擔任或未完全受攻擊者控制，這種非對稱的控制狀態為攻擊者提供了一種獨特的攻擊機會。其核心策略是透過操縱投票結果來重新分配系統的經濟激勵，基於 BlockDFL 架構中普遍採用的獎勵連鎖機制，只有當聚合提案獲得委員會的批准並成功寫入區塊鏈

時，相關的聚合者和更新提供者才能獲得本輪的獎勵分配。攻擊者正是利用這一機制設計的關鍵環節，透過控制委員會的投票權來決定誰能獲得獎勵、誰將被排除在外。惡意委員會會採取系統性的差別對待策略：對於由誠實聚合者提交的聚合提案，即使這些提案基於高品質的模型更新並且聚合過程完全正確，惡意委員會仍然會協同投出反對票，使其無法達到所需的三分之二多數支持。

戰略性餓死攻擊的破壞力主要體現在經濟層面而非技術層面，這種攻擊的精妙之處在於其高度的隱蔽性。從模型品質的角度看，由於系統仍然接受了某種形式的模型更新，訓練過程並未完全停滯，只是收斂速度相對放緩，這使得攻擊行為不易被外部觀察者識別為明顯的惡意行為。然而，從經濟激勵的角度看，這種攻擊造成了顯著的後果：誠實節點發現無論自己多麼努力地訓練模型、提交高品質更新，最終都會在委員會投票環節被系統性地排除，無法獲得應得的經濟回報。這種「付出努力但得不到回報」的狀態會導致兩種效應：一方面，誠實節點因為無法獲得獎勵而使其權益增長停滯，在未來的委員會選舉中其被選中的機率相對下降；另一方面，惡意節點透過獲取更高比例的獎勵實現權益的相對增長，其在下一輪委員會中的佔比優勢進一步擴大。這種馬太效應形成了正反饋循環，使得攻擊者的優勢隨時間推移而不斷鞏固。

場景二：全棧投毒 (Full Stack Poisoning)　第二種場景代表了 PCCA 攻擊的最極端形態，攻擊者不僅控制了委員會的絕對多數，同時也成功滲透了當輪的聚合者角色，這種「全棧控制」狀態意味著從模型聚合到結果驗證的整個流程都處於攻擊者的掌控之下，系統原本設計的多層防禦機制完全失效。在這種情況下，攻擊者的目標從經濟打擊轉向直接的技術破壞，透過向區塊鏈中注入惡意的模型更新來破壞全域模型的性能。全棧投毒攻擊的執行過程展現了多層防禦失效的連鎖反應：在聚合層面，惡意聚合者可以選擇性地接收來自惡意更新提供者的投毒更新，這些更新可能採用標籤翻轉 (Label Flipping)、梯度反轉或後門注入等多種投毒技術；在驗證層面，由惡意委員會對這個明顯包含問題的聚合結果進行投票表決，即使任何具備運算能力的節點都可以重新執行聚合演算法並發現結果的異常，但由於委員會成員超過三分之二都是惡意的，他們會協同投出贊成票，強制使該提案達到共識所需的支持門檻。

全棧投毒攻擊的後果是全方位的，從模型品質角度看，被污染的更新一旦寫入區塊鏈並被全網採用，將直接導致全域模型的準確率大幅下降，在某些精心設計的後門攻擊場景下，模型甚至可能在特定輸入下表現出完全違背預期的行為。從經濟層面看，由於惡意聚合者和惡意更新提供者瓜分了本輪的全部獎勵，攻擊者不僅成功破壞了模型，還進一步鞏固了其經濟優勢，使得系統越來越難以透過正常的選舉機制實現自我恢復。值得強調的是，全棧投毒場景的出現揭示了一個被廣泛忽視的系統性風險：在現有的 BlockDFL 架構中，聚合者和驗證者雖然在協議設計上被視為相互制約的獨立角色，但在實際攻擊場景下，它們完全可能被同一利益集團所控制形成合謀關係，這是對現有安全分析框架的重要挑戰。

3.3.3 權益增長動態分析

為了更精確地理解 PCCA 攻擊的長期影響，我們需要建立權益演化的數學模型，量化分析在沒有外部干預的情況下攻擊者的權益佔比如何隨時間推移而變化。假設系統初始狀態下，攻擊者控制的節點總權益為 $S_{mal}(0)$，誠實節點的總權益為 $S_{hon}(0)$，攻擊者的初始權益佔比為 $f_0 = \frac{S_{mal}(0)}{S_{mal}(0)+S_{hon}(0)} = 0.3$。在潛伏階段，雙方的權益都保持正常增長，攻擊者透過誠實參與獲得獎勵，權益佔比維持在初始水平附近。關鍵的轉折點出現在攻擊者首次獲得委員會超過三分之二席位的時刻，此時戰略性餓死策略開始生效。

然而，值得特別注意的是，即使在佔領階段，惡意節點的權益增長也並非呈現指數式的無限擴張。這是因為在 BlockDFL 的獎勵機制中，誠實節點仍然能夠透過擔任更新提供者角色獲得部分獎勵，即使他們的提案被惡意委員會拒絕，他們作為被選中提案的更新提供者時仍可分得相應的獎勵份額。這種機制設計意指惡意節點無法完全壟斷系統的全部獎勵，而是會與誠實節點形成一種動態的權益分配平衡。具體而言，假設系統每輪分配的總獎勵為 R，惡意委員會雖然能夠透過操縱投票使獎勵更傾向於流向惡意節點，但誠實節點作為更新提供者的貢獻仍會獲得部分補償，這使得雙方的權益比例會趨向於某個穩定的比值而非無限分化。

基於上述分析，權益演化的動態可以表示為一個有界的增長模型，其核心在於領先者優勢在穩態下的收斂特性。定義 α 為優勢係數 (Advantage Coefficient)，用以衡量惡意節點在成功控制委員會時能夠獲得的獎勵比例優勢。由於誠實節點仍能透過擔任更新提供者角色獲得部分獎勵，這個係數 α 具有上界，通常在 1.1 至 1.2 之間，這意味著惡意節點每輪獲得的獎勵大約是誠實節點的 1.1 到 1.2 倍，而非無限倍數。這種有界的優勢係數導致雙方的權益比例會收斂到一個穩定值：

$$\lim_{t\to\infty} \frac{S_{mal}(t)}{S_{hon}(t)} = \alpha \cdot \frac{S_{mal}(0)}{S_{hon}(0)} \tag{3.1}$$

從系統動力學的角度看，這是一個具有穩定平衡點的正反饋系統，而非傳統理解中的發散系統。這種平衡的存在並不意味著 PCCA 攻擊不具威脅性，相反地，它揭示了攻擊的另一種危險形態：攻擊者能夠建立並維持一種持久的「領先者優勢」，即使無法實現完全壟斷，也能長期保持對系統治理的顯著影響力。這種持續性的權益優勢使得攻擊者能夠更頻繁地控制委員會，形成一種「常態化」的治理失衡狀態，而現有系統缺乏打破這種平衡的內在機制。

3.3.4 攻擊效果與影響

PCCA 攻擊對區塊鏈聯邦學習系統造成的破壞是多維度且層層遞進的，其影響範圍涵蓋了技術性能、經濟激勵、系統治理等多個關鍵層面。在模型品質層面，即使攻擊者採取相對溫和的戰略性餓死策略，系統的訓練效能也會受到明顯影響。由於惡意委員會傾向於批准次優更新而拒絕最優更新，每一輪訓練對全域模型的改進幅度都會小於正常情況，導致收斂速度顯著放緩。在某些情況下，如果被批准的次優更新與全域模型的最佳改進方向存在較大偏差，甚至可能出現訓練震盪或陷入局部最優的情況。在全棧投毒場景下，模型品質的損害更加直接和嚴重，被注入的惡意更新可能包含精心設計的後門觸發器或針對特定類別的偏差，使得模型在大部分正常輸入上表現正常，但在特定條件下產生攻擊者預期的錯誤行為。

從網路治理權的角度看，PCCA 實現了權力結構的顯著傾斜。在攻擊的初期階段，系統表面上仍然維持著去中心化的形態，委員會的組成看起來是透過隨機選舉產生的，各個節點都有機會參與。但隨著攻擊者權益佔比的持續上升並穩定在優勢水平，這種表面上的去中心化逐漸演變為實質上的寡頭主導。當攻擊者的權益佔比穩定在較高水平時，他們獲得委員會多數席位的機率將顯著高於隨機分布的預期值，意味著從統計意義上他們能夠在更多輪次中控制委員會，形成一種「軟性壟斷」的治理狀態。這種從分散到集中的權力轉移過程，雖然不會達到完全壟斷的程度，但已經足以嚴重損害區塊鏈系統的核心價值主張。

在經濟激勵層面，PCCA 造成了激勵機制的扭曲與部分失靈。對於誠實節點而言，他們會發現一個令人沮喪的現實：即使投入大量運算資源進行本地訓練、提交高品質的模型更新，在委員會投票環節仍然面臨被系統性排斥的風險，獲得的經濟回報明顯低於預期。這種「付出與回報不成比例」的狀態會逐漸瓦解誠實節點的參與動機，理性的節點會進行成本效益分析，當持續的低回報無法覆蓋參與系統所需的運算成本、網路成本和時間成本時，部分節點可能選擇降低參與程度或退出系統。這種節點流失會形成另一層負面效應：誠實節點的退出會進一步提高惡意節點的相對權益佔比，使得系統更容易被控制，這又會加速更多誠實節點的離開，形成一種緩慢但持續的惡性循環。

3.3.5 與傳統攻擊的區別

為了更清晰地凸顯 PCCA 攻擊的獨特性與威脅性，表 **??** 提供了與傳統拜占庭攻擊和資料投毒攻擊的系統性對比，這種對比有助於理解為何現有的防禦機制難以有效應對 PCCA。

從攻擊目標來看，傳統的資料投毒或模型投毒攻擊主要關注破壞機器學習模型的性能指標，例如降低分類準確率、植入後門、造成特定類別的誤判等，這類攻擊的影響主要局限在機器學習的技術層面，即使攻擊成功，系統的治理結構和參與者組成並不會發生根本改變。相比之下，PCCA 的目標是奪取系統的治理權，控制決定模型演化

表 3.2: 與傳統攻擊的區別

特徵	傳統攻擊	PCCA
攻擊目標	模型品質	網路控制權
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	資料層防禦	需要激勵相容機制

方向的委員會機制，一旦攻擊成功，攻擊者不僅能夠影響模型品質，更能決定哪些節點可以參與、哪些提案會被接受，實質上控制了系統的未來走向。從攻擊者動機角度來看，傳統拜占庭攻擊者的行為模式往往基於最壞情況假設，他們可能出於意識形態、惡意競爭或純粹的破壞慾望而發動攻擊，即使這些行為會導致自身經濟利益受損也在所不惜；而 PCCA 則建立在理性經濟人的假設之上，攻擊者的每一步行動都經過精心運算，目標是最大化長期的經濟收益，這種基於理性的攻擊模型更貼近現實世界中的威脅場景。

從攻擊策略的時間維度來看，傳統攻擊通常採取直接而迅速的方式，惡意節點從一開始就提交明顯異常的更新或投票，試圖在短時間內對系統造成最大破壞，這種「一次性」的攻擊模式雖然可能在短期內造成嚴重影響，但也使得攻擊行為容易被檢測系統識別。PCCA 則採用漸進式的長期策略，攻擊者願意在潛伏階段投入大量時間和資源來建立信譽，只在時機成熟時才發動攻擊，這種耐心的策略使得攻擊具有極強的隱蔽性，因為在攻擊的大部分時間裡，惡意節點的行為與誠實節點完全無法區分。更關鍵的是，PCCA 具有傳統攻擊所不具備的自我強化特性：傳統攻擊即使成功也不會改變攻擊者與誠實節點之間的力量對比，下一輪攻擊仍然面臨同樣的難度；但 PCCA 每成功一次，攻擊者的相對權益優勢就會擴大，未來攻擊的成功率也隨之提高，形成滾雪球效應。這種正反饋機制使得系統一旦開始被滲透，就會沿著權益失衡的軌道持續發展，直到達到某個穩定的不平衡狀態。

從防禦策略的角度來看，傳統攻擊已經發展出相對成熟的應對方法，主要集中在資料層面的統計檢測與過濾，Krum、Trimmed Mean、Median 等拜占庭強健聚合演算法能夠有效識別並排除異常的模型更新，這些方法的有效性已經在大量實驗中得到驗證。

然而，PCCA 攻擊完全繞過了這些資料層防禦，因為它直接攻擊的是執行這些防禦演算法的驗證者本身，當驗證者被攻陷後，無論資料層的防禦設計得多麼精妙，都可以被選擇性地忽略或篡改。這揭示了一個層次化的依賴關係：資料層防禦的有效性完全依賴於共識層的安全性。要應對 PCCA，需要從根本上改變防禦思路，不能再依賴誠實多數假設，而是必須設計激勵相容的機制，使得理性攻擊者發現誠實行為才是其利益最大化的最優策略，這需要引入經濟懲罰、挑戰驗證等新的防禦維度，構建一個多層次的安全框架。

3.4 安全目標

基於前述對 PCCA 攻擊機制與其深遠影響的剖析，本節將進一步定義防禦體系所必須達成的核心安全目標。這些目標不僅旨在阻斷漸進式佔領攻擊的演進路徑，更致力於在動態變化的對抗環境中，維護系統去中心化的純粹性與經濟激勵的合理性。為避免在解決安全威脅的同時引入新的集權風險或效率瓶頸，本研究將從防止控制權持續、確保權益增長公平、維持模型效能穩定、保障治理去中心化以及建構激勵相容體系五個關鍵維度，系統化地勾勒出防禦機制的設計藍圖與驗證基準。

3.4.1 防止委員會被惡意節點持續控制

防禦機制的首要任務在於打破 PCCA 攻擊所依賴的自我強化循環，阻斷惡意節點將單次偶發的委員會優勢轉化為長期性系統控制權的途徑。為了達成此一目標，系統必須建立一套多維度的行為監控與回饋機制，從動態審計的角度即時偵測委員會的投票模式，特別是針對系統性拒絕高品質提案或盲目承認異常更新的偏差行為進行預警。一旦偵測到潛在的合謀跡象，防禦體系將啟動對應的經濟制裁流程，透過罰沒機制（Slashing）對作惡者施以重懲，沒收其在系統中質押的權益，使其付出的作惡代價遠超任何潛在的短期獲利。在此過程中，懲罰的執行必須摒棄對單一可信第三方的依賴，轉而透過去中心化的挑戰-驗證框架來實現。透過賦予網路上任何權益持有者提出質疑並觸發重新驗證的權利，系統能將原本隱蔽的委員會作惡行為置於透明的監督環境下，

大幅提升攻擊者的串謀難度與暴露成本。這種從被動檢測到主動威懾的轉變，是確保系統長期維持公平競爭與安全性的重要基石。

3.4.2 確保誠實節點的權益公平增長

在防禦 PCCA 攻擊的架構中，確保誠實參與者的權益獲得公平且穩定的增長，與抑制惡意侵佔同樣具有舉足輕重的地位。傳統系統在遭受戰略性餓死攻擊時，誠實節點往往因提案被惡意委員會駁回而面臨獎勵歸零的窘境，進而導致其權益佔比在動態博弈中不斷縮減。為了修正這種激勵扭曲，本研究提出了一套基於貢獻證明的彈性獎勵分配方案，旨在打破「全有或全無」的投票獎勵鏈條。具體而言，系統引入了基於歷史聲譽與運算貢獻的備選補償通道，即使節點提交的優質提案暫時被惡意委員會拒絕，只要該提案在後續的審計或挑戰階段被證實為優於當前鏈上結果，參與提交的資料提供者仍能獲得其應得的激勵份額。這種設計不僅保障了誠實節點在面對攻擊時的經濟生存空間，更能有效抵銷惡意節點累積權益的馬太效應。透過長期且平滑的激勵回饋機制，系統能維持參與者生態的多樣性，防止網路因資源過度集中而淪為寡頭控制的工具，從而從經濟層面深化去中心化系統的強健性。

3.4.3 維持模型收斂性與準確性

儘管本研究側重於防範治理權層面的權益佔領，但確保聯邦學習模型的收斂性與預測準確性，始終是安全防禦體系不可動搖的技術底線。在動態變化的惡意環境中，防禦機制必須具備精準識別並過濾次優或惡意梯度更新的強健性，這要求驗證層級不再僅僅是對聚合結果進行形式上的背書，而是要引入實質性的品質稽查。透過在驗證者委員會中嵌入基於小規模公開資料集的基準測試，或採用多重聚合一致性檢查，系統能夠有效阻斷全棧投毒攻擊對全域權重的直接污染。當聚合提案展現出與歷史趨勢嚴重背離的特徵或在基準測試中表現異常時，系統將自動觸發冷卻期或擴散驗證流程，要求更多獨立驗證者介入審核，從而緩解單一輪次被佔領所帶來的負面影響。長期而言，這種分層過濾機制確保了模型能夠在充滿雜訊與敵意的網路環境中，穩定地朝向

最優解收斂，最終達到的模型品質應能與完全信賴環境下的表現相抗衡，維持系統作為機器學習基礎設施的實用價值。

3.4.4 保持系統的去中心化特性

安全目標的追求不應以犧牲區塊鏈系統的核心靈魂即去中心化為代價，因此防禦機制的設計必須嚴格遵循無須許可且非特權化的原則。在本研究的框架下，防禦職能被分散於全網的參與者手中，而非寄予於少數所謂的可信節點或具備仲裁權限的超級委員會。任何形式的安全介入，從異常檢測到挑戰發起，再到最終的鏈上裁判，均需透過去中心化的協議規則自動執行，並輔以可驗證運算與偽隨機函數等密碼學工具來保證程序的公正透明。這種設計確保了防禦機制本身不會成為潛在的單點故障或淪為權力濫用的工具，任何具有持有權益的節點，無論其權益規模大小，皆能依循預設的挑戰規則對可疑決策發起質疑。透過去除對中心化實體的依賴，系統能夠在面對拜占庭攻擊或理性串謀時，展現出自發性的抵抗力與自我修復能力，真正實現在無信任假設下的端到端安全，保障區塊鏈去中心化治理結構的完整性與純粹性。

3.4.5 激勵相容性

激勵相容性（Incentive Compatibility）構成了本研究安全框架的理論基盤，其實質是在承認節點經濟理性前提下，透過機制設計引導博弈均衡趨向於誠實行為。在對抗理性攻擊者時，單純的技術限制往往不足以應對層出不窮的新穎策略，必須從經濟底層建立起「誠實即最優」的穩態結構。從定量分析的角度來看，攻擊行為的預期收益必須始終保持為負值，即滿足數學不等式 $E[\text{Payoff}] = P_{\text{success}} \cdot G_{\text{attack}} - P_{\text{caught}} \cdot L_{\text{slash}} < 0$。此處的 P_{success} 與 G_{attack} 代表了攻擊的成功機率與潛在回報，而 P_{caught} 與 L_{slash} 則分別權衡了作惡暴露的風險與隨之產生的經濟損失。為了維持此一不等式的恆定成立，本研究所設計的防禦體系從提高檢測效率、強化罰沒力度以及分散單次收益三個層次，動態調配博弈參數。這種多維度的經濟威懾不僅增加了攻擊者的機會成本，更使其在長期的期望收益運算中得出作惡不如誠實的結論，從而實現了系統安全與個體理性之間的

完美共振。

3.5 本章小結

本章系統性地構建了針對區塊鏈聯邦學習委員會架構的威脅模型，核心聚焦於一種新型的共識層攻擊：漸進式委員會佔領攻擊 (Progressive Committee Capture Attack, PCCA)。與傳統的資料層投毒攻擊著眼於破壞模型品質不同，PCCA 的目標在於透過經濟手段逐步奪取系統的治理權，最終實現對整個網路的持續性影響力。這種攻擊之所以危險，不僅在於其隱蔽性和自我強化特性，更在於它揭示了現有區塊鏈聯邦學習研究中普遍存在的一個系統性盲點：絕大多數研究在設計驗證機制時隱含地假設驗證者是誠實的或至少滿足誠實多數，但這個假設在去中心化環境下並沒有可靠的保證機制。

本章首先定義了理性攻擊者模型，明確了攻擊者以利益最大化而非單純破壞為目標的行為特徵，這種攻擊者模型更貼近現實世界中的威脅場景。在此基礎上，我們詳細剖析了 PCCA 的兩階段攻擊策略：在潛伏階段，攻擊者透過完美的誠實表現積累權益與信譽，耐心等待多個惡意節點同時被選入委員會的時機窗口；一旦獲得超過三分之二的席位優勢，攻擊立即進入佔領階段，根據對系統組件的控制程度採取戰略性餓死或全棧投毒策略。前者透過系統性地拒絕誠實提案來阻止誠實節點獲得獎勵，後者則在同時控制聚合者和驗證者的情況下直接注入惡意更新。

透過權益增長動態分析，我們釐清了 PCCA 攻擊的實際影響特徵。值得注意的是，由於誠實節點仍能透過擔任更新提供者角色獲得部分獎勵，惡意節點的權益增長並非呈現指數式的無限擴張，而是會與誠實節點形成一種動態平衡，惡意節點的權益通常穩定在誠實節點的 1.1 至 1.2 倍左右。這種平衡的存在並不意味著攻擊不具威脅性，相反地，它揭示了攻擊者能夠建立並維持一種持久的「領先者優勢」，形成常態化的治理失衡狀態。基於這一威脅分析，本章提出了五個層次化的安全目標：防止委員會持續控制、確保誠實節點權益公平增長、維持模型收斂性與準確性、保持系統去中心化特性，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，展示如何透過審計驅動型委員會 BlockDFL，在不依賴誠實多數假設的前提下構建激勵相容的防禦體系，

實現上述安全目標。

第四章　系統架構設計

區塊鏈聯邦學習系統在邁向大規模部署的過程中，始終面臨效率與安全性之間難以調和的張力。傳統拜占庭容錯共識機制固然能夠提供堅實的安全保證，但其伴隨的 $O(N^2)$ 通訊成本卻與機器學習場景中頻繁迭代更新的需求產生根本性衝突。第 **??** 章的威脅分析已經揭示了現有委員會機制在面對理性攻擊者時的結構性缺陷：小規模委員會雖然顯著降低了通訊複雜度，但其固有的集中化特性為攻擊者提供了透過漸進式權益累積逐步掌控驗證權力的可乘之機，而現有防禦機制對「誠實多數假設」的過度依賴更使得系統缺乏對策略性攻擊者的有效威懾手段。為突破這一困境，本章提出「審計驅動型委員會 BlockDFL」(Audit-driven Committee BlockDFL, AC-BlockDFL)，該架構建立在第 **??** 所定義的 BlockDFL 委員會模型之上，透過引入異步審計機制與內部罰沒協議，實現從傳統「門檻安全性」向「經濟安全性」的典範轉移。

本架構的核心設計哲學源於對區塊鏈系統狀態最終性需求與聯邦學習迭代訓練特性之間關係的重新審視。金融交易系統對每一筆交易都要求即時且不可逆的正確性保證，因為任何錯誤都可能導致資產的永久損失，這種特性迫使傳統區塊鏈系統必須在每次狀態變更前達成全網共識。然而，AC-BlockDFL 認識到聯邦學習的訓練過程具備多輪迭代的特性，單一輪次中的偏差可透過後續訓練逐步修正，這為將安全性驗證從同步的阻塞式流程轉變為異步的非阻塞式審計機制提供了設計空間。在此基礎上，AC-BlockDFL 透過引入經濟懲罰機制從根本上重塑了攻擊者的理性決策空間，使得任何試圖操縱委員會共識的行為都將面臨遠超其潛在收益的經濟損失，從而在博弈論層面消除了發動攻擊的經濟誘因。

本章的結構安排如下：第 **??** 節概述 AC-BlockDFL 的系統架構與元件互動關係，闡明挑戰者角色與鏈下儲存整合如何嵌入現有的委員會流程；第 **??** 節描述運作協議的逐步執行流程；第 **??** 節深入探討異步審計與挑戰機制的運作原理，涵蓋挑戰觸發邏輯、內生動態質押模型，以及狀態最終性與不回滾策略的設計考量；第 **??** 節以形式化的定理與證明論證雙層信任模型所提供的安全保障；第 **??** 節透過通訊、運算與儲存三個維度的開銷分析，量化評估本架構的效率優勢。

4.1 系統架構概覽

審計驅動型委員會 BlockDFL 的設計目標在於建立一個兼具經濟安全性與高執行效率的去中心化學習平台，而這一目標的實現建立在對現有 BlockDFL 架構的繼承與創新之上。如第 **??** 節所詳述，BlockDFL 透過角色分離的設計理念，將參與者劃分為更新提供者、聚合者與驗證者三種角色，並透過權益加權的隨機選舉機制決定每輪的角色分配。AC-BlockDFL 完整保留了 BlockDFL 的訓練流程與角色定義，包括更新提供者的本地訓練職責、聚合者的提案生成流程，以及驗證委員會的 Krum [**blanchard2017machine**] 評分與 PBFT 共識機制，這些經過驗證的設計元素構成了本架構運作的基礎框架。

AC-BlockDFL 的核心創新體現在三個層面的架構擴展。第一個層面是引入第四種角色，即「挑戰者」(Challenger)，以及與之配套的異步審計機制。這一角色的設計遵循開放准入原則，任何網路參與者只要願意質押規定數額的代幣，即可在該輪次中承擔挑戰者的監督職責，這種開放式的准入設計確保了監督權力不會集中於少數節點之手，從而避免了在解決委員會信任問題的同時引入新的中心化風險。挑戰者的核心職責在於持續監聽鏈上資料，獨立重新執行 Krum 演算法的運算，並將運算結果與委員會選定的全域更新進行比對，由於 Krum 演算法是一個完全確定性的數學運算，給定相同的輸入必然產生相同的輸出，因此委員會無法透過資訊不對稱來掩蓋其惡意行為。

第二個層面是鏈下儲存架構的整合。考量到模型更新的資料量通常遠大於一般區塊鏈交易，AC-BlockDFL 將沉重的模型梯度與權重資料儲存於星際檔案系統 (InterPlanetary File System, IPFS) 之上，僅將資料的內容識別符 (Content Identifier, CID) 與相關元資料記錄於鏈上。這種設計將鏈上儲存複雜度從 $O(\text{ModelSize})$ 降至 $O(\text{HashSize})$，有效緩解了區塊鏈帳本膨脹的問題。為確保審計期間的資料可用性，參與節點在異步審計窗口 (Challenge Window) 的存續期間內持續釘選 (pin) 相關的 IPFS 資料，待審計窗口關閉且未發生挑戰後，節點即可解除釘選以釋放儲存空間。這種基於生命週期管理的儲存策略，在審計所需的資料可用性與長期儲存成本之間取得了合理的平衡。

第三個層面是「先執行後審計」的安全性保障模式。在 BlockDFL 中，委員會的決策即為最終決策，系統缺乏對委員會潛在惡意行為的事後追責能力。而在 AC-BlockDFL 中，委員會達成共識後系統立即執行模型更新，但同時開啟了一個異步的審計窗口，允許挑戰者對委員會的決策進行事後驗證。圖 **??** 展示了 AC-BlockDFL 的完整運作流程，清晰呈現了挑戰機制如何嵌入現有的委員會共識流程。這種設計使得系統能夠在絕大多數正常情況下以最小的通訊開銷快速完成模型更新，同時保留了在檢測到異常行為時啟動全網仲裁的能力，本質上將監督權力從少數委員會成員民主化到了整個網路。

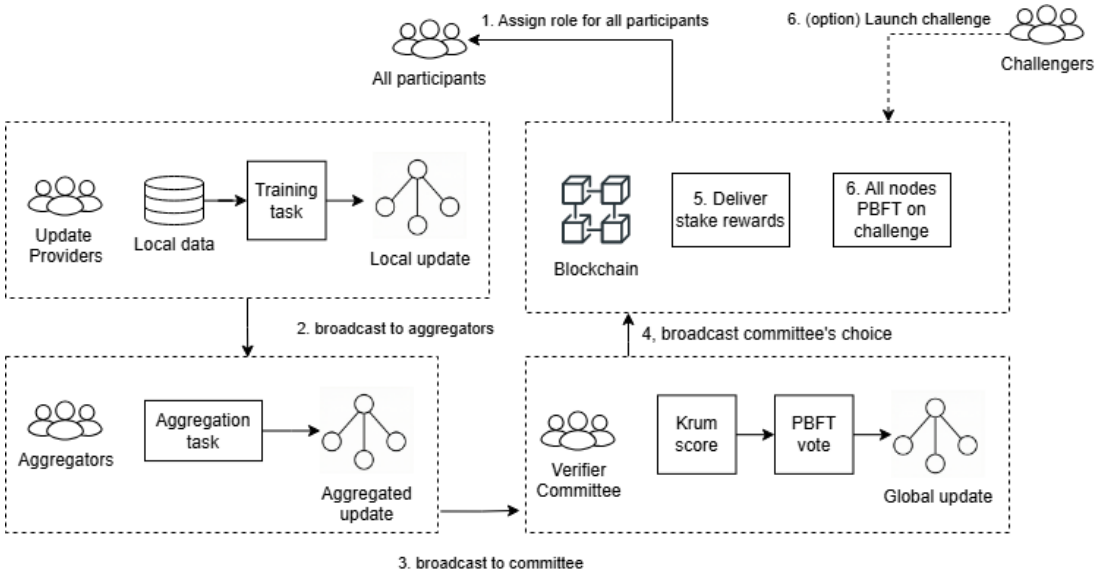

圖 4.1: Audit-driven Committee BlockDFL (AC-BlockDFL) 系統架構與工作流程圖

4.2 運作協議流程

本節描述 AC-BlockDFL 在正常運作情境下的逐步執行流程，即系統未遭受攻擊時的「快樂路徑」(Happy Path)。演算法 **??** 以形式化的方式呈現了從角色分配到模型更新的完整協議，其核心特徵在於委員會達成共識後立即提交全域模型更新，無需等待任何額外的確認期，這種設計選擇體現了對系統「活性」(Liveness) 的優先保障。

協議的第一階段為角色分配。當新一輪訓練開始時，所有參與者根據最新區塊的

Algorithm 2 AC-BlockDFL Execution Protocol (Instant Update)

Require: Current Round r, Total Stake Weighted Nodes $\mathcal{N}$
Ensure: Updated Global Model w_{r+1}

Phase 1 — Role Assignment:
Blockchain selects $\mathcal{V}$ (Validator), $\mathcal{A}$ (Aggregator), $\mathcal{U}$ (Update Provider) from $\mathcal{N}$ based on stake-weighted randomness derived from previous block hash.
Phase 2 — Training & Off-chain Storage:
Each $u \in \mathcal{U}$ trains locally using w_r, broadcasts updates to $\mathcal{A}$.
Each $a \in \mathcal{A}$ aggregates updates into proposal p_a.
Each a uploads p_a to IPFS $\rightarrow$ obtains Content Identifier CID_a.
Each a submits CID_a and metadata to $\mathcal{V}$ via on-chain transaction.
Phase 3 — On-chain Consensus & Instant Update:
$\mathcal{V}$ retrieves proposals from IPFS using $\{\text{CID}_a\}$, verifies data availability.
$\mathcal{V}$ runs Krum scoring on all proposals $\{p_a\}$.
$\mathcal{V}$ votes on the best proposal via PBFT.
Commit w_{r+1} to blockchain **immediately**.
Record winning CID^* and voter identities on-chain.
Distribute rewards to contributing $\mathcal{U}, \mathcal{A}, \mathcal{V}$.
Phase 4 — Audit Window Opens:
Asynchronous challenge period begins (see Section **??**).
Participating nodes pin relevant IPFS data for the duration of the challenge window.

雜湊值與當前權益分布，確定性地運算出本輪的角色分配結果，此運算過程僅依賴公開可驗證的鏈上資訊，任何參與者皆可獨立驗證角色分配的正確性而無需依賴中央協調者。第二階段涵蓋本地訓練與鏈下儲存，更新提供者在各自的私有資料集上執行模型訓練並將本地更新發送給聚合者，聚合者完成篩選與聚合運算後，將提案上傳至 IPFS 以取得內容識別符，隨後透過鏈上交易將此識別符與相關元資料提交給驗證委員會。這種將沉重資料負載置於鏈下的設計，確保了區塊鏈帳本僅記錄輕量級的雜湊參照，從而大幅降低鏈上儲存壓力。

第三階段是鏈上共識與即時更新，構成整個協議的核心環節。驗證委員會的成員透過鏈上記錄的 CID 從 IPFS 取得各聚合提案的完整內容，在確認資料可用性後執行 Krum 演算法進行提案評分，並依據評分結果透過 PBFT 協議進行投票表決。當某提案獲得超過三分之二驗證者的贊成票時，該提案被正式接受，對應的全域模型更新立即寫入區塊鏈，獲勝提案的 CID 與投票者身份同步記錄於鏈上，為後續可能的審計提供完整的可追溯資訊。獎勵隨即分配給對本輪全域模型更新有實質貢獻的更新提供者、聚合者與驗證者。第四階段標誌著異步審計窗口的開啟，系統進入背景監督狀態，參與節點在此期間持續釘選相關的 IPFS 資料以確保審計所需的資料可用性，挑戰機制的

具體運作邏輯將在下一節詳述。

4.3 異步審計與挑戰機制

異步審計機制是 AC-BlockDFL 架構中最具創新性的設計要素，其核心理念在於將傳統區塊鏈系統中同步驗證與即時執行之間的緊密耦合關係予以解構。傳統的拜占庭容錯系統要求在每次狀態變更之前必須達成全網共識，這種「悲觀併發控制」的設計哲學雖然能夠提供強大的即時正確性保證，卻也導致了系統吞吐量與延遲性能的嚴重退化。AC-BlockDFL 則採用了「樂觀執行」的設計哲學，允許系統在委員會達成共識後立即更新模型，而將嚴格的正確性驗證推遲到異步的背景審計流程中進行。這種設計使得系統能夠在不犧牲長期安全性的前提下最大化執行效率，其安全邊際來自於經濟懲罰機制對理性攻擊者的威懾效果。本節將依序闡述挑戰流程的觸發邏輯、質押金額的內生動態定價模型，以及不回滾策略的設計考量。

4.3.1 挑戰觸發邏輯

挑戰流程的觸發條件建立在運算確定性與資料透明性的深度解耦之上，確保了即使驗證委員會的決策受到操縱，這些行為也能夠被具備監督意願的網路參與者所揭露。具體而言，挑戰者透過持續追蹤區塊鏈上記錄的內容識別符（CID）參照，自 IPFS 分散式儲存系統中檢索出對應輪次的所有聚合提案完整內容，並於本地端獨立重新執行 Krum 演算法。由於 Krum 演算法在本質上是嚴格確定性的數學處理流程，給定相同的輸入提案集，其運算結果具備唯一的客觀性與一致性。因此，若委員會選定的最優提案與挑戰者根據鏈上記錄運算出的理論結果存在任何偏差，該行為都將被精確識別。這種依賴密碼學證據與數學運算而非主觀判斷的驗證模式，為系統提供了強大的事後審計能力，使得委員會的技術性操縱或運算失誤在透明的區塊鏈環境下無所遁形。

演算法 **??** 以形式化的方式詳細展示了異步挑戰機制的執行邏輯。當挑戰者偵測到委員會決策存在瑕疵時，需提交一筆包含預設數額質押金的挑戰交易以正式啟動仲裁流程。此處的質押機制在設計上具備雙重且深遠的含義：其一在於建立一道防禦屏障，

藉由提高發動挑戰的經濟門檻來防止惡意節點透過大量無效或虛假的挑戰請求耗盡系統資源，達成對全網仲裁機制的拒絕服務攻擊；其二則在於為全網參與監督提供必要的經濟激勵與保障，避免防禦機制在無人參與的情況下流於形式。考量到全網節點在仲裁過程中必須執行複雜的驗證運算並負擔資料傳輸成本，若缺乏合理的報酬機制，將導致參與者因搭便車心理而喪失監督積極性。質押金的設定確保了仲裁過程的資源消耗能獲得補償，不僅增加了挑戰者發起請求的審慎度，更透過罰沒與獎勵的動態分配機制，解決了去中心化監督中普遍存在的動機缺失問題。仲裁一旦觸發，全網驗證者將透過 IPFS 取得提案資料並獨立執行 Krum 評分，最終經由 PBFT 協議對決策偏離與否達成共識。若挑戰成立，系統將對惡意委員會成員執行全額罰沒，而已經提交的模型更新則遵循第 **??** 節所述之不回滾策略保持不變。

Algorithm 3 Asynchronous Challenge Mechanism (Slash-Only)

Require: Challengers $\mathcal{C}$, On-chain CID references, IPFS data store
Ensure: Punishment for Malicious Committee Acts

for each Challenger $ch \in \mathcal{C}$ **do**
    ch retrieves all proposal CIDs from on-chain records.
    ch downloads full proposal data $\{p_a\}$ from IPFS using CIDs.
    ch re-executes Krum algorithm on $\{p_a\}$.
    if ch detects outcome mismatch with committed w_{r+1} **then**
        ch posts **Challenge Transaction** with deposit $D_{\text{challenge}}$.
        Arbitration Triggered: All nodes download IPFS data via CID and re-verify.
        if Malicious Consensus Confirmed by $> 2/3$ of network **then**
            Slash full stake of colluding committee members $\mathcal{V}_{\text{mal}}$.
            Reward Challenger ch from slashed funds.
            Distribute remaining slashed funds to honest participants.
            // Note: Model w_{r+1} is NOT reverted (see Section **??***).*
        else
            Forfeit Challenger ch's deposit $D_{\text{challenge}}$.
        end if
        Exit Loop.
    end if
end for

罰沒資金的分配遵循強化激勵相容性的設計原則。成功發起挑戰的挑戰者將獲得罰沒金額中相當可觀的比例作為獎勵，這種設計確保了監督活動在經濟上具有吸引力，從而維持足夠數量的節點願意投入資源進行持續審計。剩餘的罰沒資金則分配給在被攻擊輪次中提供了正確更新的訓練者以及積極參與仲裁過程的驗證節點，這種廣泛的獎勵分配機制不僅補償了誠實節點因系統遭受攻擊而承受的潛在損失，更重要的是創

造了一種集體監督的文化，使得每個參與者都有動力關注系統的整體治理健康狀況。

4.3.2 內生動態質押模型

質押金額的定價機制直接關係到挑戰機制能否在不同經濟環境下持續有效運作，是維繫系統經濟安全性的核心基石。根據去中心化系統的設計原則，所有核心參數都應當源自系統內部的可驗證資訊，而非仰賴外部基礎設施的資料饋送，因為任何形式的外部依賴都可能成為潛在的攻擊面或單點故障來源。以鏈上預言機引入代幣對法幣匯率來動態調整質押門檻為例，此種做法雖然在邏輯上直觀，卻顯著增加了系統的外部依賴程度，更使得質押機制本身暴露於預言機操控與報價延遲的風險之下。基於這一考量，AC-BlockDFL 的質押定價完全錨定於系統內部的經濟活動指標，透過將懲罰力度定義為當輪區塊獎勵的倍數，確保無論代幣的法幣計價如何波動，攻擊行為的相對經濟損失始終顯著高於其潛在的相對經濟收益，從而在無需外部資料源的情況下實現了機制的自我穩定性。

這一定價策略的關鍵在於準確辨識出攻擊者的根本經濟動機：在理性博弈假設下，發動委員會佔領攻擊所能獲取的最大即時利益，歸根結柢就是當輪的區塊獎勵 R_{round}。既然攻擊收益以 R_{round} 為嚴格上界，那麼只要將懲罰規模同樣以 R_{round} 為基準並乘以足夠大的倍數，便能在系統內部建立起一套自洽且穩定的經濟威懾結構。據此，本研究將惡意委員會成員遭受罰沒時的懲罰金額 D_{slash} 定義為：

$$D_{\text{slash}} = \lambda \times R_{\text{round}}, \quad \lambda \gg 1 \tag{4.1}$$

其中 λ 為懲罰倍數參數，其取值需確保即使攻擊者成功壟斷整輪的全部獎勵，罰沒損失仍遠超其潛在收益。在實驗配置中，每位驗證者的初始質押設定為 100 單位，而單輪獎勵 R_{round} 面向驗證者的分配約為 1.0 單位，故 λ 的實質取值約為 100，意味著一次罰沒所造成的損失相當於一百輪正常獎勵的總和。面對如此極端不對稱的風險收益結構，理性攻擊者在進行成本效益評估時，必然會得出攻擊的預期淨收益為負的結論，作惡

的經濟誘因也因此從根本上被消除。

挑戰者提交挑戰交易時所需質押的門檻 $D_{\text{challenge}}$ 同樣遵循內生定價的設計邏輯，其數值的設定需要同時滿足經濟可持續性與准入可及性這兩項看似對立的約束條件。就經濟可持續性而言，當挑戰失敗而質押被沒收時，沒收的資金應足以補償全網參與仲裁的運算成本，以防止惡意節點透過發起大量無效挑戰來消耗全網資源。設全網參與仲裁的節點數量為 N_{arb}，每個節點執行一次完整 Krum 驗證運算的邊際成本為 ϵ（其單位可理解為單次運算所需的計算資源成本或資源調度 Gas 費），則門檻需滿足 $D_{\text{challenge}} \geq N_{\text{arb}} \cdot \epsilon$ 的基本邊界條件。在實務上，$D_{\text{challenge}}$ 可進一步設定為 $\alpha \times R_{\text{round}}$，其中 α 為平衡監督積極性與抗攻擊能力的調整係數。由於 R_{round} 隨網路經濟活動規模自然增減，當獎勵池增大時，攻擊成本與挑戰成本將同步實現線性擴展，確保了懲罰力度始終能夠覆蓋潛在的區塊獎勵竊取收益，實現了質押機制在不同市場環境下的自適應調節。

4.3.3 狀態最終性與不回滾策略

當仲裁確認委員會存在惡意行為時，AC-BlockDFL 採用「僅懲罰不回滾」的處置策略，即對惡意節點執行經濟罰沒但不撤銷已經提交的模型更新。這一設計選擇並非基於對機器學習模型固有強健性的樂觀假設，而是主要出於分散式系統穩定性與區塊鏈狀態最終性的考量。

從區塊鏈系統設計的角度而言，狀態回滾與帳本不可篡改性這一核心原則之間存在根本性衝突。區塊鏈的價值主張建立在每一個經共識確認的區塊都具備最終性 (Finality) 的保證之上，一旦允許因事後審計結果而回滾歷史區塊的狀態，便為長程攻擊 (Long-Range Attack) 等利用歷史重寫的攻擊向量開啟了空間。在聯邦學習的場景中，全網仲裁的時間延遲意味著當仲裁最終判定某個早期輪次存在問題時，該輪次之後可能已經累積了數十甚至數百個後續區塊，撤銷這些區塊將摧毀所有中間交易的最終性，對系統的可信賴性造成災難性的打擊。

從工程實踐的角度來看，全域模型狀態的回滾涉及極高的協調複雜度。回滾操作

要求所有網路節點同步恢復至歷史狀態，並在此基礎上重新執行從被攻擊輪次開始的所有後續訓練，這不僅會造成大量運算資源的浪費，更需要設計複雜的分散式協調協議來確保所有節點一致地完成狀態回溯。考慮到聯邦學習通常需要經歷數百甚至數千個訓練輪次，回滾操作的沉沒成本與協調難度都將隨著被攻擊輪次與當前輪次之間的距離增長而急劇攀升，使其在實務上幾乎不可行。

因此，AC-BlockDFL 採用「前向修正」(Forward Correction) 的策略取代歷史回滾：透過對惡意行為者施加嚴厲的經濟懲罰來消除未來的攻擊誘因，同時依賴後續輪次中誠實節點的正常訓練來漸進式地修正模型軌跡的偏差。這種處置方式在本質上是一種工程權衡，以接受單次偶發性的模型精度波動為代價，換取帳本不可篡改性的完整保全與系統運作連續性的穩定維持。第 **??** 章的實驗結果將從實證角度展示，在罰沒機制有效威懾理性攻擊者的前提下，系統在長期訓練過程中能夠自然收斂至與無攻擊環境相當的模型品質水準。

4.4 安全性分析

AC-BlockDFL 架構的安全性建立在一個精心設計的雙層信任模型之上，該模型透過將不同層級的安全職責分配給不同的參與者群體，成功地在維持高效率的同時提供了等同於全網共識的安全保障。傳統的區塊鏈系統通常採用單一層級的信任假設，要求每一次狀態變更都必須經過全網共識的嚴格驗證，這種設計雖然能夠提供強大的安全保證，但其高昂的通訊成本使其難以應用於需要頻繁更新的場景。AC-BlockDFL 透過引入分層信任的概念，使得系統在正常情況下以小委員會的效率運行，而在異常情況下能夠迅速升級至全網共識的安全等級。本節將透過三個形式化的安全性定理及其證明，嚴謹地論證此雙層信任模型所提供的安全保障。

4.4.1 檢測層：1-of-N 誠實假設

雙層信任模型的第一層是檢測層，其採用了極為寬鬆的「1-of-N 誠實假設」。這個假設的含義是：只要全網 N 個參與節點中存在至少一個誠實節點願意擔任挑戰者的角

色，任何委員會層級的惡意行為就能夠被成功揭露。這種假設的寬鬆程度遠超過傳統拜占庭容錯系統所要求的「三分之二誠實節點」條件，因為它僅需要單一誠實節點的存在而非多數誠實節點的協調行動。從概率角度而言，在一個擁有數百或數千個參與者的大型網路中，所有節點同時選擇沉默或串謀的可能性極其微小。檢測層的設計巧妙地利用了區塊鏈系統的資料透明性特質：由於所有聚合提案的 CID 都被記錄在鏈上且對應的完整資料可透過 IPFS 公開存取，任何節點都能夠獨立地重新執行驗證運算，攻擊者即使成功控制了當前輪次的整個委員會，也無法阻止其他節點發現異常。

定理 4.1 (檢測完備性)**.** *令 $\mathcal{V}_r$ 為第 r 輪的驗證委員會，Krum$(\{p_a\})$ 為對所有聚合提案執行 Krum 演算法所得的確定性正確結果，w_{r+1} 為委員會實際選定並寫入區塊鏈的全域更新。若 $w_{r+1} \neq$ Krum$(\{p_a\})$，且全網 N 個節點中存在至少一個誠實節點 c^* 願意擔任挑戰者角色，則此偏離行為必然被偵測。*

證明. 此定理的證明建立在 Krum 演算法的確定性特質與區塊鏈資料的公開可驗證性之上。Krum 演算法的運算過程完全由其輸入決定：給定同一組聚合提案 $\{p_a\}$，任何執行者無論身份與位置，都將得到唯一且一致的輸出結果 Krum$(\{p_a\})$。在 AC-BlockDFL 的協議設計中，所有聚合提案的 CID 在委員會共識階段即被記錄於區塊鏈，對應的完整提案資料可透過 IPFS 公開存取，任何持有區塊鏈帳本的節點都能透過 CID 取得這些資料。因此，誠實挑戰者 c^* 可以從 IPFS 下載與委員會完全相同的輸入集合 $\{p_a\}$，在本地獨立執行 Krum 演算法，所得結果必然為 Krum$(\{p_a\})$。當 c^* 將此結果與委員會實際選定的 w_{r+1} 進行比對時，若兩者不一致，則 c^* 即可確認委員會的決策存在偏離，並據此發起挑戰交易。由於 c^* 的驗證過程僅依賴公開可存取的鏈上 CID 參照、IPFS 資料與確定性演算法，委員會無法透過隱藏資訊或製造歧義來規避偵測。因此，只要存在至少一個誠實且具備質押能力的挑戰者，任何偏離正確 Krum 結果的委員會決策都必然被偵測。 □

此定理的實務意涵在於，攻擊者若希望其惡意決策不被偵測，唯一的途徑是確保全網沒有任何一個誠實節點願意擔任挑戰者，這在大規模網路中幾乎不可能實現。相較於傳統 BFT 系統要求三分之二誠實節點的嚴格條件，1-of-N 假設將偵測門檻降至理

論最低限度，極大地擴展了安全性的適用範圍。

4.4.2 仲裁層：全網三分之二誠實假設

雙層信任模型的第二層是仲裁層，其採用了「全網三分之二誠實假設」。當挑戰被發起並進入仲裁階段後，最終的判決權力從小委員會回歸到全網範圍，要求網路中誠實節點的數量必須超過總節點數的三分之二，即 $N_{\text{total}} > 3f$。這是幾乎所有拜占庭容錯共識協議的標準假設，也是區塊鏈系統普遍依賴的安全基礎。在仲裁階段，所有參與驗證的節點透過 IPFS 下載相關提案資料並重新執行 Krum 運算，隨後透過 PBFT 協議對挑戰的正當性進行投票，只有當超過三分之二的節點確認委員會確實存在錯誤時，挑戰才會被判定為成立。

定理 4.2 (懲罰確定性)**.** *令全網節點總數為 N_{total}，其中惡意節點數量 f 滿足 $N_{total} > 3f$。若挑戰者依據定理* **??** *成功偵測到委員會的惡意決策並提交了有效的挑戰交易，則此惡意行為必然在仲裁階段被確認，且參與共謀的委員會成員必然遭受質押金的全額罰沒。*

證明. 仲裁過程的核心是全網範圍的 PBFT 共識。當挑戰交易被提交後，智能合約自動從鏈上調取該輪次的所有提案 CID，並要求全網驗證節點從 IPFS 下載完整提案資料後獨立重新執行 Krum 演算法。由於 Krum 的確定性特質（如定理 **??** 的證明所述），所有誠實驗證節點將得到一致的正確結果 $\text{Krum}(\{p_a\})$，並能據此判斷 w_{r+1} 是否偏離正確值。在 $N_{\text{total}} > 3f$ 的假設下，至少有 $N_{\text{total}} - f > 2N_{\text{total}}/3$ 個誠實節點參與仲裁投票，這些誠實節點基於相同的確定性運算結果將一致地投票確認委員會決策存在偏離。由於 PBFT 協議要求超過三分之二的贊成票即可達成共識，而誠實節點的數量已超過此門檻，因此仲裁共識必然成立。共識達成後，智能合約自動執行預定義的罰沒邏輯，沒收所有在該輪次中對偏離結果投贊成票的委員會成員之全額質押金，此過程由智能合約的確定性執行保證，不受任何外部干預。 □

定理 **??** 與定理 **??** 的結合構成了 AC-BlockDFL 安全性保障的完整邏輯鏈：前者確保惡意行為「必然被發現」，後者確保被發現的惡意行為「必然受到懲罰」。這兩層保障

的疊加效果是，攻擊者在發動攻擊之前即可預見其行為將面臨偵測與懲罰的雙重確定性後果，這種確定性正是經濟安全性得以成立的邏輯前提。

4.4.3 攻擊成本的形式化分析

基於前述兩層信任機制所提供的安全保障，本節將進一步形式化地分析攻擊者若試圖在 AC-BlockDFL 架構中發動一次獲利且確保不受懲罰的攻擊，所需跨越的關鍵經濟門檻。在第 **??** 節所建立的內生動態質押模型中，攻擊成本不再被視為一個靜態且固定的數值，而是被賦予了與系統整體質押規模及內部經濟指標 R_{round} 緊密掛鉤的動態屬性。這種設計的核心哲學在於建立一套自適應的平衡機制，確保無論代幣的市場價格如何劇烈波動，攻擊者所面臨的潛在罰沒損失始終能與其可能的非法獲利保持量級上的顯著落差。透過將安全成本內生化於區塊鏈自身的經濟循環中，AC-BlockDFL 成功在無需依賴外部資料源的前提下，從經濟維度建立起一道難以逾越的准入障礙，使得攻擊行為在理性博弈的框架下因期望收益轉負而變得無效。

理解這一形式化分析的前提，在於精確區分攻擊過程中兩種性質截然不同的經濟負擔及其對應的實現難度，這兩者共同構成了系統分層防禦的理論邊界。首先，委員會的佔領在本質上屬於一個具有強烈不確定性的機率性事件，雖然攻擊者可以透過策略性地累積更高的權益佔比，來提升其在特定輪次選舉中獲得超過三分之二席位的機率，但其最終能否在目標輪次實際取得驗證主導權，始終受制於具備密碼學可驗證隨機性的選舉結果。相比之下，仲裁階段的規避則是一個純粹的確定性資本門檻問題，根據定理 **??** 與定理 **??** 的結論，只要攻擊者無法在物理層面阻止全網 PBFT 仲裁共識的達成，其惡意行為便必然會在審計窗口期內被揭露並遭受罰沒。而要阻止共識，攻擊者必須控制網路總投票權的三分之一以上，這構成了一個堅實的物理界限。

定理 4.3 (無懲罰攻擊的資本門檻)**.** *在 AC-BlockDFL 架構中，攻擊者若要完成一次惡意委員會決策且完全規避隨後而來的經濟懲罰，其在全網中必須掌握的權益資本需滿足以下下界：*

$$\text{Cost}_{\text{total}} \geq \frac{1}{3} N \cdot \bar{s} \tag{4.2}$$

其中 N 代表全網節點的總體規模，$\bar{s}$ 則為所有參與節點的平均質押額數。此外，即便攻擊者滿足上述資本條件而具備了阻斷仲裁活性之能力，其仍須在委員會隨機選舉中以機率性的方式獲得超過三分之二的席位，方能實際對當前輪次的共識結果產生實質影響。

證明. 攻擊者若要達成「攻擊成功且不受懲罰」的最終目標，必須同時克服兩個不同層級的安全防線。在委員會層級，攻擊者必須在目標輪次的隨機選舉中恰好分配到超過三分之二的驗證者席位，此條件的實現機率由權益佔比決定，具有天然的隨機與不確定性，攻擊者僅能透過增加權益來提升勝算，卻無法將其轉化為必然。在全網層級，根據定理 **??**，一旦挑戰被任何誠實節點發起並經全網驗證確認，所有惡意節點的質押將被全額罰沒。遵循拜占庭容錯理論，PBFT 協議要求獲得超過三分之二的權重同意方能達成共識，因此攻擊者若要逃避懲罰，唯一途徑是控制全網至少 $\lceil N/3 \rceil$ 的投票權重，以破壞仲裁共識的活性，使其無法執行罰沒。此確定性的資本門檻與全網規模成線性正比，其量級為 $O(N)$。由於規避懲罰是發動獲利攻擊的邏輯前提，故總體攻擊成本的下界必然為 $\frac{1}{3}N \cdot \bar{s}$。 □

此定理所揭示的深層意涵在於，AC-BlockDFL 透過異步審計與全網仲裁機制，成功將攻擊者面對的經濟門檻從局部的委員會規模提升至全域的網路規模。在缺乏事後追責機制的傳統 BlockDFL 架構中，攻擊者的作惡成本僅取決於其控制小型委員會所需的資源，其複雜度始終維持在 $O(C)$ 的水平，這使得具備一定資本實力的攻擊者能在較短時間內完成對系統的佔領。AC-BlockDFL 則迫使攻擊者在發動具體攻擊之前，必須先解決如何抗衡全網三分之二誠實多數的問題，其成本量級發生了跨越式的躍升。考慮到實際應用中全網節點數 N 通常遠大於委員會規模 C（例如在本研究的基準實驗配置中，$N = 100$ 而 $C = 7$），這種分層的安全防線為系統提供了極強的強健性。

值得進一步討論的是，即便攻擊者在財力上足以支撐全網三分之一以上的權益佔比，其所面對的仍然不是一條通往非法收益的坦途。控制全網三分之一的節點雖然提供了阻止懲罰執行的確定性能力，卻無法保證在每一輪的委員會選舉中都能如願獲得主導地位，後者依然受到不可預測的隨機分佈約束。換言之，攻擊者需要投入大量的

沉沒成本來確保即使作惡也不會損失本金，但這筆高昂的資本投入僅換取了一個不確定的攻擊機會，這種「高昂確定性投入換取微小不確定收益」的結構，在博弈論層面極大地削弱了理性參與者的作惡衝動。由此可見，AC-BlockDFL 成功地在維持執行效率的同時，利用全網的累計權益為小型委員會的運行提供了深度的安全保障。

4.4.4 激勵相容性的博弈論分析

基於前述安全性定理，本節運用博弈論的分析框架論證 AC-BlockDFL 的經濟懲罰機制如何使誠實行為成為所有理性參與者的最優策略。第 **??** 章在安全目標中提出了激勵相容性的數學條件，要求攻擊者的預期收益必須為負值，本節將在 AC-BlockDFL 的具體架構參數下展開這一分析。

對於理性攻擊者而言，其決策問題可建模為單次博弈的期望收益運算。設攻擊者成功控制委員會後所能獲得的單輪最大經濟收益為 G_{attack}，被全額罰沒的質押金損失為 L_{slash}，則攻擊的預期收益可表示為：

$$E[\text{Payoff}] = P_{\text{success}} \cdot G_{\text{attack}} - P_{\text{caught}} \cdot L_{\text{slash}} \tag{4.3}$$

其中 P_{success} 為攻擊者在特定輪次成功控制委員會的機率，P_{caught} 為惡意行為被偵測並受到懲罰的機率。定理 **??** 與定理 **??** 的結合表明，在 1-of-N 誠實假設與全網三分之二誠實假設同時成立的條件下，P_{caught} 趨近於 1。需要注意的是，P_{success} 衡量的是攻擊者在委員會選舉中獲得多數席位的機率，P_{caught} 衡量的是惡意決策被偵測的機率，兩者分屬不同層面的事件：攻擊者只有在 P_{success} 對應的條件實現時才能發動攻擊，而一旦攻擊發動，P_{caught} 趨近於 1 確保其必然面臨懲罰。因此式 (**??**) 可簡化為：

$$E[\text{Payoff}] = P_{\text{success}} \cdot (G_{\text{attack}} - L_{\text{slash}}) \tag{4.4}$$

激勵相容性的充分條件由此清晰浮現：只要 $L_{\text{slash}} > G_{\text{attack}}$，則無論攻擊成功機率

P_{success} 取何值，預期收益都嚴格為負。在內生動態質押模型下，$L_{\text{slash}} = \lambda \times R_{\text{round}}$ 而 G_{attack} 的上界約為 $C \times R_{\text{round}}$（即攻擊者壟斷全部驗證獎勵），由於 $\lambda \gg C$，此條件穩定成立且不受代幣市場價值波動的影響。以本研究的實驗參數進行具體的數值分析：委員會規模 $C = 7$，每位驗證者的單輪獎勵為 1.0 單位，初始質押為 100 單位。攻擊者即使成功壟斷全部驗證獎勵，單輪最大收益 G_{attack} 上界約為 7.0 單位。然而，一旦惡意行為被偵測，參與共謀的至少 5 個惡意委員會成員各自損失全額質押 100 單位，攻擊者陣營的總損失 $L_{\text{slash}} = 500$ 單位，懲罰力度約為潛在收益的 71 倍。

如此極端的風險收益不對稱結構，使得理性攻擊者即使考慮到可能低估被偵測機率的僥倖心理，只要 P_{caught} 超過 $G_{\text{attack}}/L_{\text{slash}} \approx 1.4\%$ 的極低門檻，攻擊的預期收益即轉為負值。而 AC-BlockDFL 的安全性定理保證了 P_{caught} 趨近於 1，遠遠超過這一最低門檻。從長期均衡的角度而言，AC-BlockDFL 的罰沒機制成功打破了第 **??** 章所描述的漸進式委員會佔領攻擊所依賴的正反饋循環。在沒有罰沒機制的系統中，攻擊者可透過操縱委員會來獲取不當獎勵，進而增加質押權重並逐步掌控系統。而在 AC-BlockDFL 中，任何作惡嘗試都會導致質押的大幅減少而非增加，遭受罰沒的惡意節點不僅損失了當下的質押資產，更喪失了透過未來輪次逐步恢復影響力的經濟基礎，這種永久性的治理排除效應從根本上切斷了惡性循環的可能性 [**chiu2018incentive**]。

4.5 效率與開銷分析

第 **??** 節的分析揭示了傳統委員會架構面臨的根本性困境：在 BlockDFL 等現有系統中，安全性的保障完全依賴於「委員會中誠實節點佔據多數」這一機率性條件，而要提高此條件成立的機率，唯一的途徑便是擴大委員會規模，這又直接推高了通訊成本。本節將從通訊、運算與儲存三個維度論證 AC-BlockDFL 如何透過將安全性保障從「門檻安全性」轉移至「經濟安全性」，在維持等效安全保證的前提下實現顯著的效率提升。

4.5.1 通訊複雜度對比分析

BlockDFL 的安全性論證建立在超幾何分佈的機率運算之上：若要將委員會被惡意控制的風險壓制在可接受水準之下，系統必須維持足夠大的委員會規模。以第 **??** 節的數值分析為例，在全網節點數 $N = 100$、惡意節點佔比 $f = 30\%$ 的威脅環境下，若將風險閾值設定為 $p < 0.01$，委員會規模至少需要達到 $C = 9$。這種設計的深層問題在於其「悲觀併發控制」的本質：BlockDFL 預設每一輪都可能遭受攻擊，因此必須在每一輪都部署足以抵禦攻擊的防禦資源。然而在實際運作中，攻擊者成功控制委員會的情況畢竟屬於少數輪次，絕大多數時候系統處於正常運作狀態，此時維持大型委員會所付出的通訊成本便成為一種恆常的「預防溢價」。

AC-BlockDFL 對效率問題的回應並非追求「更好的機率保證」，而是從根本上改變了安全性的實現方式。傳統的門檻安全性聚焦於「如何降低委員會被攻破的機率」，這種思路必然導向更大的委員會規模。AC-BlockDFL 則採取截然不同的策略：與其執著於將被攻破的機率壓制至趨近於零，不如確保即使委員會被攻破，攻擊者也無法從中獲取正向收益。在經濟安全性的框架下，委員會被攻破的機率不再是唯一的安全性指標，因為異步挑戰機制確保了任何惡意行為都將面臨全額質押金的罰沒。由此，委員會規模的選擇便不再完全受制於安全性的機率運算，系統得以在滿足基本安全閾值的前提下採用相對較小的委員會來獲取效率優勢。

表 **??** 呈現了兩種架構在「委員會被惡意控制的風險低於 1%」這一統一安全性基準下的通訊成本對比。BlockDFL 必須採用 $C = 9$ 的委員會規模，每輪通訊複雜度固定為 $O(81)$；AC-BlockDFL 透過經濟安全性的補充保障，得以採用 $C = 7$ 的委員會規模，常態通訊複雜度降至 $O(49)$，實現了約 39.5% 的通訊成本削減。

從系統運作的動態視角來看，AC-BlockDFL 的通訊成本呈現條件式的特徵。在正常運作下，系統僅需支付 $O(C^2) = O(49)$ 的委員會共識成本；唯有當挑戰被觸發並進入全網仲裁時，才會產生額外的 $O(N^2)$ 通訊開銷。由於經濟懲罰機制有效消除了理性攻擊者的作惡誘因，挑戰觸發的機率 p 在長期均衡中將趨近於零，據此系統的期望通

表 4.1: BlockDFL 與 AC-BlockDFL 在相同安全性水平下的效率對比 ($N = 100$, $f = 30\%$, $p_{\text{risk}} < 0.01$)

評估維度	**BlockDFL**	**AC-BlockDFL**	**差異分析**
安全性實現方式	門檻安全性	經濟安全性	機率保證 vs. 激勵相容
所需委員會規模	$C = 9$	$C = 7$	規模縮減 22.2%
常態通訊複雜度	$O(C^2) = O(81)$	$O(C^2) = O(49)$	通訊成本降低 39.5%
安全性維護模式	每輪固定開銷	條件式觸發開銷	預防性 vs. 響應性

訊複雜度可表示為：

$$E[\text{Comm}] = (1-p) \cdot O(C^2) + p \cdot (O(C^2) + O(N^2)) = O(C^2) + p \cdot O(N^2) \tag{4.5}$$

當 $p \to 0$ 時，期望複雜度近似於常態值 $O(C^2)$，這意味著全網仲裁的高昂成本僅作為威懾手段存在而實際上鮮少被觸發。這種「按需付費」的安全模式，相較於 BlockDFL 每輪都必須支付的固定「預防溢價」，在資源利用上更為經濟。

4.5.2 儲存開銷的權衡分析

傳統的鏈上儲存方案要求每輪訓練中所有聚合提案的完整模型參數都被永久記錄於區塊鏈帳本之中，隨著訓練輪次的累積，帳本的體積將持續膨脹，對節點的儲存資源構成沉重負擔。AC-BlockDFL 透過將模型參數的沉重負載遷移至 IPFS 並搭配嚴格的生命週期管理策略，顯著緩解了這一問題。在此設計下，區塊鏈帳本僅記錄固定長度的 CID 雜湊值與相關元資料，鏈上儲存複雜度從 $O(\text{ModelSize})$ 降至 $O(\text{HashSize})$，對於典型的卷積神經網路模型而言，這意味著數個數量級的儲存節省。

IPFS 上的臨時儲存成本是此設計為獲得審計能力所付出的必要代價。在異步審計窗口的存續期間，參與節點需要釘選相關的提案資料以確保挑戰者能夠存取驗證所需的完整輸入，這會佔用一定的本地儲存空間。然而，由於審計窗口具有明確的時間上限（即生存時間 TTL），一旦窗口關閉且未發生挑戰，節點即可解除釘選以釋放空間。因此，每個節點在任一時刻所需維護的 IPFS 儲存量僅與當前處於審計窗口內的少數輪

次相關，而非與整個訓練歷史成正比。這種設計在審計所需的資料可用性與長期儲存成本之間建立了合理的平衡，使得 AC-BlockDFL 在引入異步審計能力的同時，不會對節點的儲存資源造成不可承受的負擔。

4.5.3 效率提升的本質：架構層面的解耦創新

綜合上述通訊與儲存兩個維度的分析，AC-BlockDFL 相對於 BlockDFL 的效率優勢並非源自共識協議本身的改進，而是源自架構層面的根本性創新，即將安全性與委員會規模之間的強耦合關係予以弱化。在 BlockDFL 的設計中，委員會規模是安全性的唯一保障手段，追求更高的安全性必然要求更大的委員會，而更大的委員會必然帶來更高的通訊成本。AC-BlockDFL 透過引入異步挑戰機制與經濟懲罰協議，為安全性開闢了獨立於委員會規模的第二條保障路徑，從而打破了這種強耦合。

這種解耦的實踐意義在於，系統設計者得以根據效率需求選擇較小的委員會規模，而無需過度顧慮安全性的機率運算。雖然 $C = 7$ 的委員會在純機率意義上的安全性略低於 $C = 9$，但經濟懲罰機制提供的額外威懾力足以彌補這一差距：攻擊者或許更容易獲得控制委員會的機會，但每一次攻擊嘗試都面臨著災難性的經濟後果，這種威懾足以使理性攻擊者放棄攻擊意圖。最終，系統在實際運作中達成了一種新的均衡，較小的委員會提供了效率優勢，而幾乎不會發生的攻擊確保了這種效率優勢不會被全網仲裁的開銷所侵蝕。

4.6 本章小結

本章提出的審計驅動型委員會 BlockDFL 代表了區塊鏈聯邦學習系統設計理念的一次重要轉變。AC-BlockDFL 建立在 BlockDFL 委員會模型之上，完整保留了其經過驗證的訓練流程與角色定義，同時透過三個層面的架構創新實現了從「門檻安全性」向「經濟安全性」的典範轉移。在角色擴展方面，挑戰者角色的設計遵循開放准入原則，任何持有足夠質押的節點均可擔任，將監督權力從少數委員會成員民主化到整個網路。在儲存架構方面，IPFS 整合與基於生命週期管理的釘選策略，將鏈上儲存複雜度從

$O(\text{ModelSize})$ 降至 $O(\text{HashSize})$，在確保審計期間資料可用性的同時有效控制了長期儲存成本。在質押定價方面，內生動態質押模型消除了對外部預言機的依賴，將罰沒金額錨定於系統內部經濟指標 R_{round} 的倍數，確保風險收益的不對稱結構在任何代幣市場環境下都穩定成立。

AC-BlockDFL 的安全性保障由三個形式化定理構成完整的邏輯鏈。定理 **??** 證明了在 1-of-N 誠實假設下任何偏離正確 Krum 結果的委員會決策必然被偵測，定理 **??** 進一步證明了在全網三分之二誠實假設下被偵測的惡意行為必然受到罰沒制裁，而定理 **??** 則量化了攻擊者完全規避懲罰所需的經濟成本下界。三者的結合揭示了 AC-BlockDFL 的核心安全特性：系統的安全性實質上由全網規模 N 決定而非委員會規模 C。通訊複雜度分析確認了在相同安全性要求下，AC-BlockDFL 得以採用 $C = 7$ 的委員會規模，相較於 BlockDFL 所需的 $C = 9$ 實現了約 39.5% 的通訊成本削減。博弈論分析則確保了這種架構在實踐中能夠長期穩定運作，在本研究的實驗配置下懲罰力度約為潛在收益的 71 倍，使得誠實行為成為所有理性參與者的最優策略，從根本上打破了漸進式委員會佔領攻擊所依賴的正反饋循環。下一章將透過多維度的模擬實驗驗證這些理論主張的有效性。

第五章　實驗評估

本章透過系統性的實驗設計，驗證審計驅動型委員會 BlockDFL 在應對漸進式委員會佔領攻擊時的防禦效能。有別於傳統聯邦學習安全性研究將模型準確率視為首要評估指標的慣例，本研究的核心關注點在於經濟安全性機制能否有效遏止理性攻擊者的惡意行為，進而維護系統的長期治理穩定性。這種評估視角的轉移根植於第 **??** 章所確立的設計哲學：當防禦機制的目標從「防止攻擊發生」轉向「確保攻擊無利可圖」時，衡量防禦效能的指標也應當相應地從模型品質轉向攻擊者的經濟決策空間。

為提供最嚴格的效能驗證，本章實驗採用「最壞情況分析」的設計哲學，假設攻擊者完全不顧經濟理性而執意發動所有可能的攻擊，藉此檢驗防禦機制在極端條件下的偵測與懲罰能力。這種實驗設計蘊含著一條重要的推論鏈：若機制在最壞情況下仍能確保每一次攻擊都被偵測並遭受懲罰，則理性攻擊者在事前評估預期收益時必然得出負值結論，從而自發地選擇不發動攻擊，系統便能在實務運作中自然趨向穩定均衡。換言之，本章所呈現的攻擊事件統計並非系統正常運作下預期會遭遇的攻擊頻率，而是防禦機制所能承受的最嚴苛壓力測試。

5.1　實驗設置

5.1.1　實驗配置概述

本研究選用 MNIST 手寫數字資料集作為聯邦學習任務的測試平台，採用包含兩個卷積層與兩個全連接層的標準卷積神經網路作為訓練模型。訓練資料以獨立同分布（IID）的方式均勻分配給所有參與的客戶端節點，此項設計選擇並非出於簡化實驗的考量，而是源於本研究防禦機制的本質特性。如第 **??** 章所闡述，審計驅動型委員會 BlockDFL 的核心防禦運作於共識層而非資料層，其功能在於透過經濟懲罰威懾惡意的委員會行為。攻擊能否成功取決於委員會組成與共識決策過程中的投票結果，而這些因素與底層訓練資料的統計分布特性相互獨立。無論資料呈現獨立同分布或高度異質

的 Non-IID 特性，委員會成員是否選擇誠實投票這一決策，在邏輯上並不受資料分布的影響，因此 IID 設定足以驗證共識層防禦機制的有效性。

為確保評估結果的公平性與可比較性，本實驗將審計驅動型委員會 BlockDFL 與原始 BlockDFL 進行系統性對照比較，兩種架構均採用相同的委員會規模 $C = 7$。攻擊場景嚴格遵循第 **??** 章所定義的漸進式委員會佔領攻擊模型，攻擊者採取理性的兩階段策略：在潛伏階段完全模仿誠實行為以累積權益，當成功獲得委員會超過三分之二席位時進入佔領階段，根據其對系統各組件的控制範圍選擇戰略性餓死或全棧投毒策略。表 **??** 彙整了本研究實驗所採用的完整系統參數配置。

表 5.1: 實驗參數配置

參數名稱	**設定值**
訓練輪數	$R = 300$（基礎實驗）/ $R = 2000$（長期實驗）
驗證者池規模	$N = 100$
委員會大小	$C = 7$
惡意節點數量	$M = 30$（初始權益佔比 30%）
每輪獎勵	驗證者 1.0，聚合者 1.0，更新提供者 0.05
罰沒規則	挑戰成功時全額沒收惡意節點質押

惡意節點初始佔比設定為 30% 代表了相當嚴峻的威脅情境，此比例已接近大多數拜占庭容錯系統所能容忍的理論上限。根據第 **??** 節的超幾何分布分析，在此條件下惡意節點於單輪中獲得委員會超過三分之二席位的機率約為 2.4%。這一機率乍看之下並不顯著，但考量到聯邦學習訓練通常需要經歷數百甚至數千輪迭代，累積下來足以產生數十次攻擊機會窗口，為驗證防禦機制在長期時間跨度中的持續效能提供了充分且嚴格的測試場景。

5.2 實驗結果與分析

本節依循「機制驗證、長期生存、服務品質」的層層遞進邏輯，從三個逐步擴大的觀察尺度呈現實驗結果。分析首先聚焦於 300 輪基礎實驗中防禦機制的即時響應特性，確認經濟懲罰機制在微觀層面確實能夠靈敏地偵測並懲罰異常行為；繼而將觀察視野

延伸至 2000 輪的長期模擬，驗證短期有效的機制是否足以引導系統趨向長期穩定的治理均衡；最終從系統服務品質的角度檢驗上述安全性保障是否以犧牲聯邦學習的核心效能為代價。這種從微觀機制到宏觀治理再到全局效能的遞進結構，旨在完整呈現經濟安全性機制在不同時間尺度與觀察維度上的防禦效能。

5.2.1 微觀機制的即時驗證

驗證經濟懲罰機制有效性的第一步，在於確認其能否在攻擊發生的當下產生即時且明確的響應。300 輪基礎實驗為此提供了理想的觀察窗口，因為在這一相對短暫的時間跨度內，每一次罰沒事件的觸發條件、執行過程與權益衝擊都能被精確地追蹤與解讀，而不至於被長期累積的統計噪音所模糊。

5.2.1.1 雙軌分歧：兩種架構的早期權益軌跡

權益比值作為衡量系統治理結構健康程度的核心指標，定義為惡意節點平均權益除以誠實節點平均權益，其數值變化直接反映了第 **??** 章所定義的「領先者優勢」是否正在被建立或瓦解。當此比值持續高於 1.0 時，意味著惡意節點正在累積治理層面的結構性優勢，其在委員會選舉中的入選機率將隨之提升，系統面臨逐步被滲透的風險；反之，當此比值被壓制至 1.0 以下時，則表明經濟激勵結構已成功將惡意節點邊緣化，使其在後續輪次中愈加難以組織有效的委員會佔領攻擊。基於此指標，本節透過 300 輪基礎實驗追蹤兩種架構下權益分布的演化軌跡，並結合具體的攻擊事件與罰沒資料，揭示經濟安全性機制在實驗早期即展現出的防禦效能。

如圖 **??** 所示，兩種架構的權益軌跡在實驗初期便呈現出截然相反的演化方向。在缺乏挑戰機制的 BlockDFL 架構中，300 輪實驗期間共發生 10 次委員會佔領攻擊事件，其中 4 次屬於攻擊者僅控制驗證委員會但未掌握聚合者角色的戰略性餓死攻擊，另外 6 次則是攻擊者同時控制委員會與聚合者的全棧投毒攻擊。這些攻擊事件平均約每 30 輪發生一次，而由於 BlockDFL 缺乏任何事後追責機制，全部 10 次攻擊均未受到經濟制裁，攻擊者得以在每次成功佔領中不受阻礙地獲取不當獎勵。這一資料為第 **??** 節的理

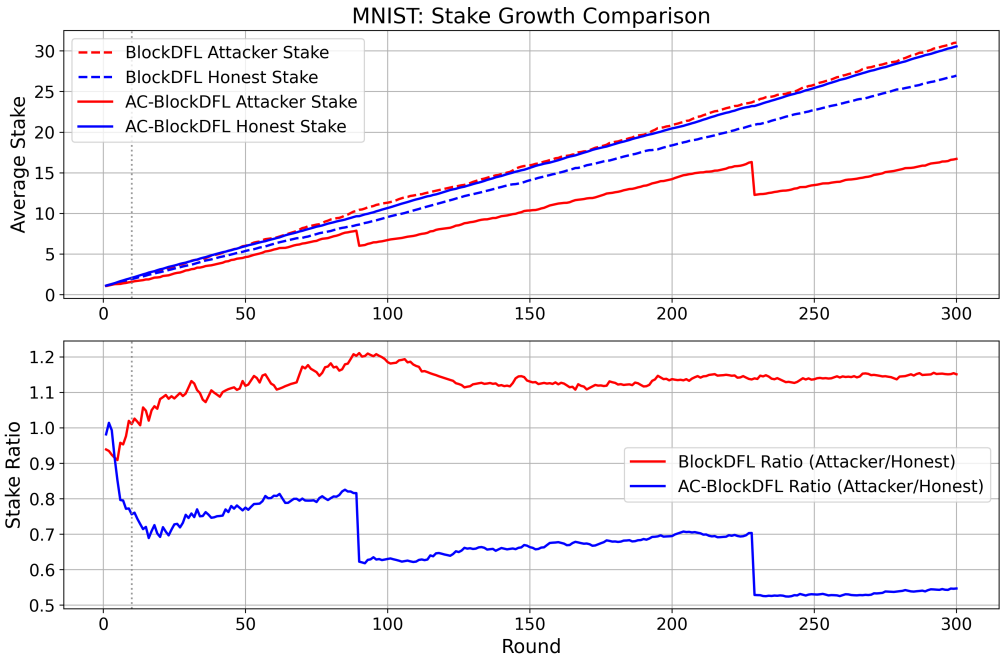

圖 5.1: BlockDFL 與 AC-BlockDFL 權益演化對比（300 輪基礎實驗）

論分析提供了直接的實證支持：獎勵機制所內建的正反饋特性確實使惡意節點的權益比值從初始的 1.0 開始穩定攀升，呈現出「權益優勢帶來更多獎勵，更多獎勵鞏固權益優勢」的累積效應。

5.2.1.2 罰沒機制的運作實證：兩次懲罰事件的詳細剖析

審計驅動型委員會 BlockDFL 在相同的 300 輪觀察期間內僅記錄到 2 次攻擊事件，分別發生在第 90 輪與第 229 輪，且兩次均為全棧投毒攻擊。相較於 BlockDFL 的 10 次未受懲罰的攻擊，這一資料差異的背後蘊含著罰沒機制對攻擊者決策空間的深層重塑。以下將逐一剖析這兩次懲罰事件的具體過程與經濟效果，藉此展示「偵測、舉證、懲罰」閉環在實務場景中的完整運作邏輯。

第一次攻擊發生在第 90 輪，此時惡意節點經過前 89 輪的誠實參與已累積了一定程度的權益優勢，其平均權益比值攀升至 1.25，意味著惡意節點的平均權益已高出誠實節點 25%。在權益加權的隨機選舉機制下，這種優勢轉化為更高的委員會入選機率，最終在第 90 輪使 5 名惡意節點同時被選入規模為 7 的驗證委員會，超過了三分之二的

控制閾值。由於攻擊者在該輪同時掌控了聚合者角色，遂發動全棧投毒攻擊，將惡意的模型更新寫入區塊鏈。然而，這一攻擊行為隨即被挑戰者偵測：挑戰者從 IPFS 下載該輪所有聚合提案的完整內容，在本地重新執行 Krum 演算法後發現委員會選定的結果與正確的 Krum 輸出存在明確偏離，遂提交挑戰交易並觸發全網仲裁。仲裁確認惡意行為後，智能合約自動執行罰沒操作，將參與共謀的 5 名惡意委員會成員的全額質押予以沒收。這次罰沒的經濟衝擊極為顯著：惡意節點的平均權益比值從罰沒前的 1.25 驟降至 0.62，在單一事件中便從「領先誠實節點 25%」逆轉為「僅及誠實節點 62%」，權益結構發生了根本性的翻轉。

這一數值變化的幅度值得從機制設計的角度進行深入解讀。5 名參與共謀的惡意委員會成員被全額沒收其自訓練開始以來所累積的全部權益，這意味著他們在前 89 輪誠實參與中透過正常獎勵機制所建立的經濟基礎被一次性清零。相較於該輪攻擊可能獲得的經濟收益上界，即當輪區塊獎勵的總和約 7.0 單位，罰沒所造成的損失遠超過其預期獲利，精確印證了第 **??** 節博弈論分析所推導的理論預測。更為關鍵的是，這 5 名遭受罰沒的惡意節點並非僅僅損失了當期的經濟利益，而是喪失了在後續所有輪次中參與委員會選舉的權益基礎。由於 BlockDFL 的角色分配機制以權益為權重進行隨機選舉，權益歸零的節點實質上被永久排除在高獎勵角色之外，這種「永久性治理排除」效應遠超單次經濟懲罰的意義，它從系統的參與者結構層面削弱了攻擊者的長期作戰能力。

第二次攻擊發生在第 229 輪，距離第一次罰沒已相隔 139 輪。這一時間間隔本身即反映了第一次罰沒對攻擊能力的有效壓制：在權益比值降至 0.62 的條件下，惡意節點需要經歷顯著更長的等待期，才能在隨機選舉中再度湊齊超過三分之二的委員會席位。值得注意的是，在第 90 輪至第 229 輪的 139 輪間隔中，惡意節點的平均權益比值從罰沒後的 0.62 緩慢回升至 0.70，這種回升源於存活的惡意節點（即未參與第 90 輪共謀而未被罰沒的節點）持續透過誠實參與獲取的正常獎勵。然而，0.70 的回升水平仍顯著低於第一次攻擊前的 1.25，這表明單次罰沒事件對惡意群體的整體權益基數造成了難以完全恢復的結構性損傷。第 229 輪的攻擊同樣由 5 名惡意節點在委員會中形成多數後發動，挑戰者再度成功偵測並觸發罰沒，5 名共謀成員的全額質押被沒收，惡意節點的平

均權益比值從 0.70 進一步下降至 0.52。

將兩次罰沒事件的資料進行對比，可以清晰地觀察到一個遞減的攻擊效力模式。第一次罰沒造成的權益比值降幅為 0.63（從 1.25 降至 0.62），第二次罰沒的降幅則縮小為 0.18（從 0.70 降至 0.52），這種降幅的收窄並非意味著罰沒機制的效力在衰減，而是反映了一個更為深層的動態：隨著惡意群體的權益基數在反覆罰沒中持續萎縮，每次被選入委員會的惡意節點所持有的質押金額也相應減少，因此單次罰沒所能沒收的絕對金額自然降低。然而，從攻擊者的視角來看，這種「可罰沒資產的縮減」並不構成任何安慰，因為與之同步縮減的還有其組織後續攻擊的能力：第二次罰沒後 0.52 的權益比值意味著惡意節點的平均權益僅略高於誠實節點的一半，在此條件下要透過隨機選舉同時將 5 名惡意節點送入規模為 7 的委員會，其機率已被壓縮至極低水平。事實上，從第 229 輪罰沒直至 300 輪實驗結束的 71 輪觀察期內，再未發生任何攻擊事件，這一觀測結果為上述分析提供了直接的實證支持。

5.2.1.3 模型收斂品質與訓練穩定性

從模型收斂品質的角度觀察，300 輪實驗結束時 BlockDFL 的模型準確率為 98.26%，而審計驅動型委員會 BlockDFL 達到 98.63%，兩者之間 0.37 個百分點的差距看似微小，但其背後隱含著截然不同的訓練穩定性特徵。BlockDFL 的 6 次全棧投毒攻擊意味著全域模型在 300 輪訓練過程中遭受了 6 次直接的惡意梯度注入，每一次注入都會導致模型準確率的急劇下降與後續數輪的修復過程，這種反覆的「下降與回升」不僅消耗了寶貴的訓練輪次，更使得收斂曲線呈現出顯著的波動特徵。相比之下，審計驅動型委員會 BlockDFL 僅經歷 2 次全棧投毒，模型所遭受的干擾次數減少了三分之二，訓練過程的連續性得到更好的保障，這正是其最終準確率略高的直接原因。

然而更為關鍵的觀察在於，兩種架構最終都能達到 98% 以上的準確率水平，這一事實恰好印證了第 **??** 章所闡述的核心設計理念：聯邦學習的迭代訓練特性賦予了系統內在的自癒能力，偶發性的模型偏差能夠被後續正常訓練自然修復，因此單純的模型品質指標並不足以全面評估系統安全性。BlockDFL 雖然在 300 輪後仍達到 98.26% 的準確率，但其間經歷的 10 次未受懲罰的攻擊已使惡意節點建立起穩固的權益優勢，這種

治理層面的失衡在短期內尚未對模型品質造成不可逆的損害，但隨著訓練輪次的延伸，攻擊頻率的持續攀升終將使自癒機制難以承受。這一觀察再次突顯了本研究聚焦於經濟安全性而非模型準確率的研究取向：真正需要防禦的並非偶發的模型偏差，而是攻擊者透過權益累積逐步奪取系統治理權的長期戰略性威脅。

5.2.1.4 從短期響應到長期均衡：延伸實驗的必要性

綜合 300 輪基礎實驗的各項觀測資料，兩種架構在權益分布演化上的對比結果清晰地揭示了經濟安全性機制的效能差異。在審計驅動型委員會 BlockDFL 中，罰沒機制展現出完整且有效的防禦閉環：2 次攻擊全部被成功偵測並執行罰沒，偵測率達 100%，惡意節點的平均權益比值從首次攻擊前的 1.25 經由兩次階梯式下降被壓制至 0.55，實驗結束時惡意節點的平均權益僅為誠實節點的約一半，沒有任何惡意行為能夠逃脫經濟制裁。與此形成鮮明對比的是，缺乏事後追責機制的 BlockDFL 中，惡意節點在 10 次未受懲罰的攻擊中持續鞏固權益優勢，其平均權益比值在 300 輪後攀升至 1.15，意味著攻擊者的經濟實力已超越誠實節點約 15%，在委員會選舉中享有更高的入選機率，為後續更頻繁的攻擊創造了結構性條件。兩條權益演化軌跡在方向上的根本分歧，清楚呈現了經濟懲罰機制對治理結構的重塑效應：一方將惡意節點向邊緣化方向驅趕，另一方則任由攻擊者逐步鞏固優勢地位。然而，300 輪的觀察期對於驗證經濟安全性機制的長期效能而言仍顯不足，這是因為 PCCA 攻擊的本質在於其漸進性與累積性，攻擊者能否透過長期的耐心等待逐步恢復被削減的權益，最終突破罰沒機制的持續壓制，是一個只有在更長時間尺度上才能獲得回答的問題。基於此考量，下一節將實驗觀察期延伸至 2000 輪，以完整呈現經濟懲罰機制作為「漸進式淨化」工具的長期運作特性，並驗證系統最終是否能夠收斂至一個誠實節點主導的穩定均衡狀態。

5.2.2 宏觀治理的長期均衡

短期實驗確認了經濟懲罰機制在微觀層面的即時有效性，但一個更為關鍵的問題隨之浮現：這種有效性能否在長期時間跨度中持續發揮作用，並最終將系統引導至一

個攻擊自然消亡的穩定均衡？2000 輪長期模擬實驗正是為回答這一問題而設計，其結果揭示了經濟懲罰機制作為「漸進式淨化」工具的完整動態過程，構成本章實驗分析的核心發現。

5.2.2.1 漸進式淨化：從短期懲罰到長期治理逆轉

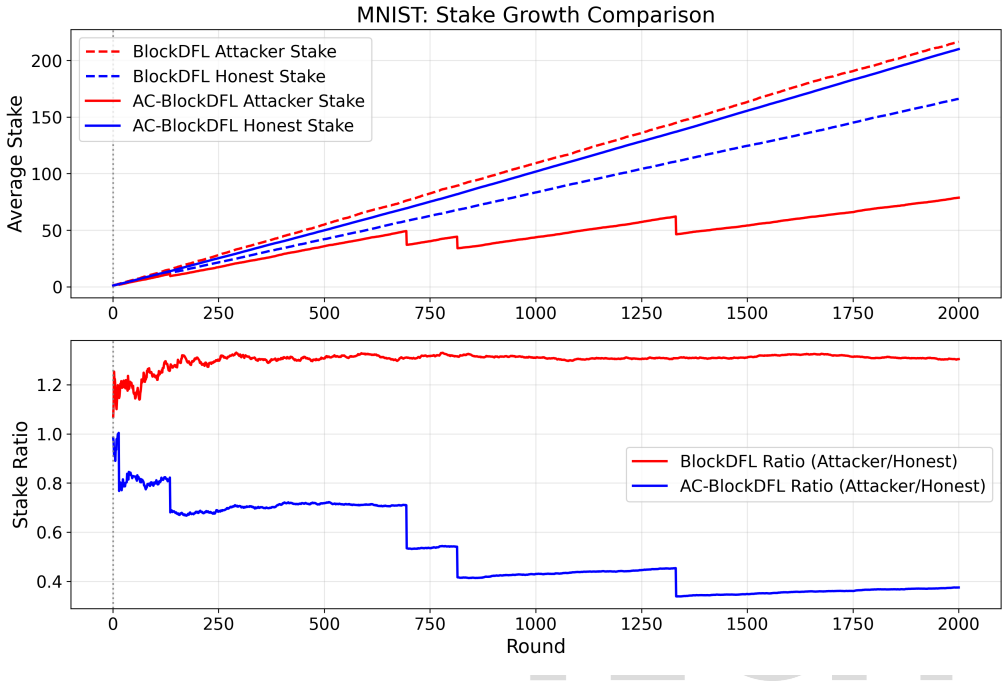

圖 5.2: 2000 輪長期模擬下的權益動態演化

如圖 **??** 所示，將觀察期延伸至 2000 輪後，兩種架構在治理結構層面的本質差異得到了充分的展現。BlockDFL 中惡意節點的權益比值在經歷約 300 輪的初期波動後穩定收斂至 1.3 左右，並在整個實驗期間持續維持在此水平。1.3 這一看似溫和的數值背後隱藏著深刻的治理危機：從超幾何分布的角度分析，權益比值為 1.3 意味著惡意節點在委員會選舉中享有顯著高於其初始比例的入選機率，這種機率優勢在 2000 輪的時間跨度中累積轉化為持續不斷的攻擊能力。更值得關注的是，由於 BlockDFL 缺乏任何事後追責機制，攻擊者從每一次成功佔領中獲取的不當利益反而進一步鞏固了其權益優勢，實證確認了第 **??** 節所預測的正反饋循環與常態化治理失衡。

AC-BlockDFL 的長期軌跡則呈現一幅截然不同的圖景。惡意節點的權益比值經歷

五次明確的階梯式下降，分別發生在第 15、136、695、815 與 1332 輪，從初始的 1.0 逐步被壓制至最終的 0.37。這一最終數值意味著實驗結束時惡意節點的平均權益僅為誠實節點的三分之一強，相較於 BlockDFL 中惡意節點維持 1.3 倍優勢的情況，兩種架構之間的權益比值差距達到 $1.3/0.37 \approx 3.5$ 倍。這種量級的差異並非漸進式的微幅改善，而是代表了系統治理結構的根本性逆轉：從攻擊者主導的失衡狀態轉變為誠實節點佔據明確優勢的健康格局。

5.2.2.2 攻擊事件：權益下降的因果證據

圖 **??** 中五次階梯式下降的背後，是五次被完整偵測並懲罰的攻擊事件。表 **??** 彙整了兩種架構在 2000 輪實驗中的攻擊事件統計，為理解權益為何下降提供了直接的因果證據。

表 5.2: 攻擊事件統計對比（2000 輪長期實驗）

評估指標	**BlockDFL**	**AC-BlockDFL**
攻擊總次數	107	5
戰略性餓死	18	2
全棧投毒	89	3
成功偵測並罰沒	0	5 (100%)
最終權益比值	1.30	0.37

BlockDFL 共記錄了 107 次委員會佔領攻擊事件，平均約每 19 輪發生一次，其中 18 次為戰略性餓死、89 次為全棧投毒，且全部 107 次攻擊未受任何形式的經濟制裁。AC-BlockDFL 則僅記錄 5 次攻擊事件，包含 2 次戰略性餓死與 3 次全棧投毒，而這 5 次攻擊全部被成功偵測並執行罰沒制裁，偵測率達到 100%。攻擊次數從 107 次降至 5 次，這種超過二十倍的數量級差異源於兩個相互強化的因果機制。第一個機制是罰沒事件對權益基數的直接削減：惡意節點在遭受全額質押罰沒後權益大幅縮水，其在後續輪次中被選入委員會的機率隨之降低，從根本上減少了攻擊機會窗口的出現頻率。第二個機制涉及權益分布變化對委員會控制門檻的影響：即使惡意節點在權益削減後仍被選入委員會，要在 7 名成員中同時佔據 5 名以上席位的難度也因權益比例的下降而

顯著提高。兩者共同作用的結果遠超任何單一機制的效果，這種雙重因果結構正是第 **??** 章博弈論分析中所預測的「永久性治理排除效應」的實證體現。

5.2.2.3 罰沒事件間隔的遞增趨勢：走向靜默

五次罰沒事件的時間間隔呈現出一種值得特別關注的遞增趨勢，此趨勢為判斷系統是否正在趨向長期均衡提供了關鍵線索。具體而言，第一次與第二次罰沒之間的間隔為 121 輪（第 15 輪至第 136 輪），第二次與第三次之間的間隔擴大至 559 輪（第 136 輪至第 695 輪），第三次與第四次的間隔為 120 輪（第 695 輪至第 815 輪），第四次與第五次的間隔再度擴大至 517 輪（第 815 輪至第 1332 輪），而第五次罰沒之後的 668 輪觀察期內則未再發生任何攻擊事件。

這種間隔遞增的整體趨勢並非偶然的統計波動，而是罰沒機制作用於權益分布後的必然數學結果。每一次罰沒事件都會削減惡意節點的權益基數，而根據超幾何分布的性質，惡意節點權益佔比的下降會直接降低其在後續委員會選舉中同時獲得五個以上席位的機率。以初始的 30% 權益佔比為參照，惡意節點獲得委員會控制權的單輪機率約為 2.4%；當權益比值被壓制至 0.37（對應約 20% 的權益佔比）時，此機率將降至不足 0.5%。機率的降低直接體現為攻擊機會窗口的稀疏化，進而表現為罰沒事件間隔的拉長。

更深層地看，這種動態揭示了一個自我強化的良性循環：罰沒削減權益，權益削減降低攻擊機率，攻擊機率降低減少罰沒的觸發頻率，而較低的觸發頻率又意味著系統在更長的時間段內以正常效率運作，印證了第 **??** 節關於「挑戰觸發機率 p 在長期均衡中趨近於零」的理論預測。第 1332 輪之後長達 668 輪的「靜默期」尤其具有說明力：在此期間惡意節點的權益已被削減至極低水平，其同時在委員會中獲得足夠席位以發動攻擊的可能性已變得微乎其微。從博弈論的視角來看，即使這些惡意節點在此後的某一輪中僥倖獲得委員會控制權，前四次罰沒的經驗已經清楚展示了攻擊的必然後果，攻擊事件的消失因此是一個結構性的、而非偶然性的結果。

5.2.3 安全性保障下的服務品質驗證

前兩節的分析已從微觀機制的即時響應與宏觀治理的長期均衡兩個維度，確認了 AC-BlockDFL 經濟安全性機制的有效性。一個隨之而來且同樣重要的問題是：這種安全性保障是否以犧牲聯邦學習的核心效能為代價？本節從系統可用性與模型收斂品質兩個互補的面向展開分析，其目的並非將模型準確率視為獨立的評估維度，而是作為一項合理性驗證，確認防禦機制在有效維護治理安全的同時，並未破壞系統作為機器學習基礎設施的實用價值。

5.2.3.1 系統可用性：從攻擊頻率到服務連續性

為量化攻擊事件對系統連續運作能力的實質衝擊，本研究定義「最低不可用率」作為評估指標，衡量系統因遭受全棧投毒攻擊而處於效能顯著下降狀態的時間比例。之所以聚焦於全棧投毒而非戰略性餓死，是因為前者直接注入惡意模型更新並導致全域模型準確率的急劇下降，其對系統功能的衝擊更為顯著且易於量化。根據實驗觀測，每次全棧投毒攻擊會導致模型準確率急劇下降，而聯邦學習的自我修復機制通常需要約 5 至 25 輪的正常訓練才能使效能恢復至攻擊前的水平。

採用最保守的 5 輪恢復期估計，BlockDFL 因 89 次全棧投毒攻擊累積產生至少 $89 \times 5 = 445$ 輪的效能下降期間，對應約 $445/2000 = 22.3\%$ 的最低不可用率。這意味著在整個 2000 輪的訓練過程中，系統有超過五分之一的時間處於模型品質受損的狀態，對於依賴模型輸出進行實時決策的應用場景（如自動駕駛或醫療診斷）而言，如此高的不可用率顯然難以接受。反觀 AC-BlockDFL，憑藉僅 3 次全棧投毒的記錄，最低不可用率被有效控制在 $3 \times 5/2000 = 0.75\%$ 以下，相較於 BlockDFL 實現了超過 96% 的改善幅度。值得注意的是，這一改善並非源自對攻擊行為本身的更好抵禦（兩種架構在面對全棧投毒時的單次受損程度相當），而是完全歸因於經濟懲罰機制對攻擊頻率的有效壓制。

5.2.3.2 模型收斂品質：訓練穩定性而非終點準確率

在確認系統可用性得到了實質性保障後，本節進一步檢驗模型訓練過程本身的品質特徵，探討防禦機制如何從根本上影響機器學習任務的執行效率。由於 AC-BlockDFL 在前 300 輪訓練內便已透過連續的罰沒事件大幅削弱了惡意節點的攻擊能力，這段集中交戰期間的模型表現格外值得深入剖析。它揭示了防禦機制如何在抵禦攻擊的同時，維護訓練過程的連續性與參數最佳化的穩定性，而非僅僅追求最終的一個數值指標。圖 **??** 呈現了兩種架構在 300 輪基礎實驗中的準確率演化軌跡，從中可以清晰辨識出截然不同的訓練動態特徵，這反映出底層治理結構對上層模型品質的深層形塑作用。

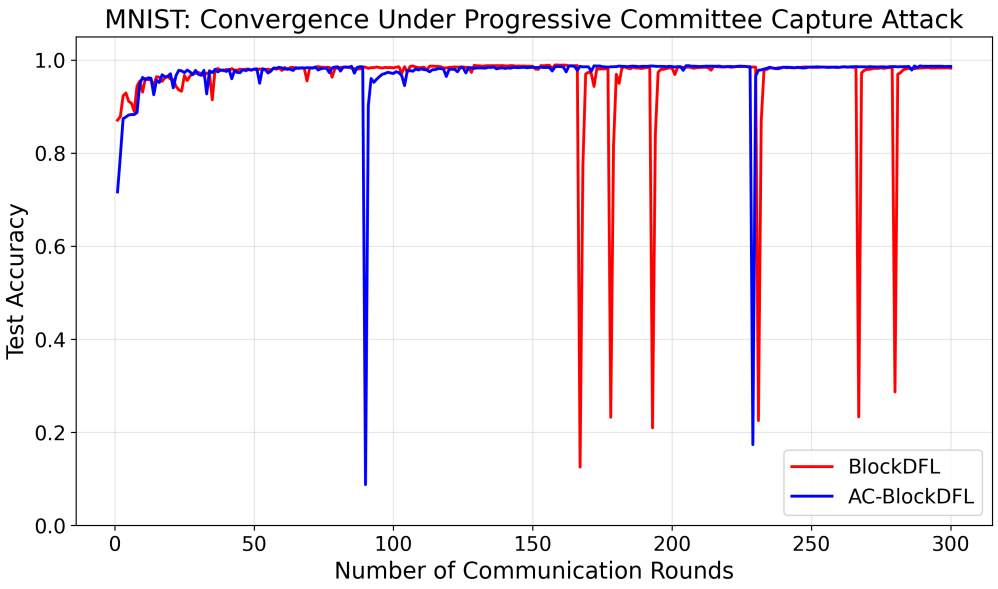

圖 5.3: 模型準確率收斂曲線比較

BlockDFL 的準確率曲線呈現出顯著的鋸齒狀波動，其中每一次突發性的效能下降都精確對應一次全棧投毒攻擊的瞬時衝擊。在 300 輪的實驗期間內，BlockDFL 共遭受 6 次破壞性的全棧投毒攻擊，而值得特別注意的是，系統最早遭受的威脅並非表現為準確率的下降。第 69 輪發生的首次攻擊採取了一種更為隱蔽的「戰略性餓死策略」，惡意委員會透過共謀排斥誠實委員的提案選擇，轉而批准能夠將獎勵分配給最多惡意更新

提供者的聚合結果。這種攻擊的目標並非直接摧毀模型品質，而是操縱經濟激勵的流向，藉此在不被察覺的情況下加速權益累積。從圖 ?? 的準確率監測角度觀察，這類戰略性餓死攻擊在模型效能指標上幾乎不留痕跡。傳統的準確率監控機制完全無法識別此類治理層面的攻擊，這也印證了第 ?? 章所強調的核心論點：僅依賴模型品質指標進行安全性評估存在根本性的盲區。

隨著 BlockDFL 中惡意節點透過戰略性餓死策略逐步建立起權益優勢，攻擊發生的頻率呈現出明顯的加速趨勢。實驗觀測顯示，惡意勢力平均約每 19 輪便能發動一次成功的委員會佔領攻擊，這種頻率遞增現象揭示了一種漸進式委員會佔領攻擊的正反饋邏輯。早期的戰略性餓死攻擊為惡意節點積累了足夠的權益基數，而提升的權益又轉化為後續輪次中獲得委員會控制權的高機率，使得攻擊窗口出現得愈發頻繁。當後續攻擊從隱蔽的戰略性餓死策略升級為破壞性的全棧投毒時，每一次成功的佔領都使模型在準確率曲線上留下深刻的凹陷，導致模型不得不被迫在偏離狀態與正常軌跡之間反覆擺盪。儘管聯邦學習的自癒特性使系統能在攻擊後逐步恢復，但這種持續的「偏離與修正」循環實質上消耗了大量本應用於模型最佳化的有效週期，大幅降低了運算資源的利用率。

相較之下，AC-BlockDFL 的準確率曲線展現出截然不同的動態穩定性特徵。在 300 輪實驗期間內，AC-BlockDFL 僅遭受 2 次全棧投毒攻擊，其中第 90 輪的極端案例提供了觀察防禦機制韌性的重要指標。在該輪次中，惡意節點執行了顯著的模型翻轉攻擊，將模型準確率從正常水準驟降至 9.5%，這一數值已接近 MNIST 十分類任務中的隨機猜測基準線。然而，系統展現了極強的自我修復能力，在經歷約 20 輪的正常訓練後成功恢復至攻擊前的水準。更有意義的發現是，隨著訓練的深入推進，模型對同類攻擊的抵抗力與恢復速度呈現持續改善的趨勢。至訓練後期，即使面臨同等強度的擾動，恢復所需的輪次已從初期的 20 輪大幅縮短至僅約 5 輪，這歸因於模型在長期訓練後建立的穩健參數結構提高了系統的容錯上限。

從最終收斂品質的角度審視，BlockDFL 在 300 輪訓練結束時達到 98.26% 的準確率，而 AC-BlockDFL 則達到 98.63%，兩者之間僅有 0.37 個百分點的微小差距。這一結果進一步印證了第 ?? 章第 ?? 節所制定的「不回滾策略」在實務上的合理性：聯邦學習

的迭代本質賦予了系統天然的自癒力，使偶發性的模型偏差能在後續正常訓練中被逐步消化，因此系統並不需要承擔回滾操作所帶來的巨大協調成本。然而，終點準確率的相近並不意味著兩種架構在訓練品質上等價。AC-BlockDFL 憑藉罰沒機制將全棧攻擊壓縮至僅 2 次，為模型訓練營造了近乎無干擾的連續最佳化環境，使運算資源得以更高效地轉化為模型效能的增益，而非被浪費在修復攻擊損害上。這對於運算資源受限的邊緣部署場景而言，具有尤為突出的實務意義。

5.3 本章小結

本章透過「機制驗證、長期生存、服務品質」的三階段遞進分析，從逐步擴大的觀察尺度驗證了審計驅動型委員會 BlockDFL 的防禦效能。300 輪基礎實驗首先確認了經濟懲罰機制在微觀層面的即時響應能力：每一次惡意委員會決策都被挑戰者靈敏偵測，並透過全網仲裁觸發罰沒制裁，權益軌跡的階梯式下降為機制的靈敏度與精準度提供了直觀的視覺證據。2000 輪長期模擬則進一步揭示了這種短期有效性如何轉化為長期的治理均衡：五次罰沒事件將惡意節點的權益比值從 1.0 逐步壓制至 0.37，罰沒間隔的遞增趨勢與最終長達 668 輪的靜默期，共同印證了系統確實被引導至攻擊自然消亡的穩定狀態。100% 的攻擊偵測率為第 **??** 節的博弈論分析提供了實證基礎：理性攻擊者若能預見攻擊必然被偵測並遭受損失，將選擇不發動攻擊以保全質押資產，系統因此能在實務運作中自然趨向穩定均衡，維持第 **??** 節所分析的 $O(C^2)$ 效率水平。

上述安全性保障的達成並非以犧牲系統效能為代價。服務品質分析顯示，AC-BlockDFL 將最低不可用率從 BlockDFL 的 22.3% 壓制至 0.75% 以下，同時為模型訓練提供了更為平穩且不受干擾的最佳化環境。這些實驗結果整體構成了一條完整的推論鏈：最壞情況測試證明所有攻擊皆被偵測，理性攻擊者因此預見懲罰而選擇不攻擊，系統在無攻擊的均衡狀態下同時維持了高效率與高品質的服務水準。

第六章　結論與未來展望

6.1　研究總結

區塊鏈聯邦學習的委員會架構在追求執行效率的同時，隱含著對「誠實多數假設」的過度依賴。本研究針對此安全性缺口進行系統性分析，從威脅識別、形式化建模到防禦機制設計，建構了完整的理論與實踐框架。我們識別並定義了「漸進式委員會佔領攻擊」，揭示理性攻擊者如何透過策略性權益累積逐步滲透委員會治理結構，從根本上顛覆去中心化系統的安全假設。

為彌補此缺口，本論文提出審計驅動型委員會 BlockDFL，其核心設計哲學在於將安全性保障與委員會規模解耦。透過異步審計機制與內部罰沒協議，本架構實現從「門檻安全性」向「經濟安全性」的典範轉移：與其執著於將被攻破機率壓至趨近於零，不如確保即使委員會被攻破，攻擊者也無法獲取正向收益。這種視角轉換使系統在面對理性對手時，仍能維持運作活性與模型聚合正確性，同時享有小規模委員會的效率優勢。

6.2　研究貢獻

本研究的首要貢獻在於對漸進式委員會佔領攻擊進行形式化威脅建模與實證驗證。我們定義了此攻擊的兩階段演化模型，刻畫攻擊者如何在潛伏階段偽裝誠實以累積權益，並在獲得控制權後切換至佔領階段執行惡意策略。長期模擬實驗證實了傳統委員會架構確實存在權益固化與治理失效風險，為防禦機制的必要性提供論據基礎。

第二項貢獻體現在經濟懲罰機制對攻擊者誘因結構的重塑。實驗結果顯示，罰沒機制成功打破惡意節點的權益累積正反饋循環，實現「漸進式淨化」效果。罰沒機制不僅懲罰個別惡意行為，更將攻擊失敗後果轉化為永久性治理排除，內部化作惡的外部性成本，迫使理性節點趨向誠實策略。

第三項貢獻在於證明「事前預防」轉向「事後追責」的架構創新能有效打破安全性

與通訊開銷的強耦合。透過引入經濟安全性作為獨立保障路徑，本架構在維持等效安全性的前提下顯著降低通訊成本，對資源受限的邊緣運算場景具有重要實務價值。

6.3 未來展望

本研究所提出的架構在應對理性攻擊者時展現優越防禦能力，然而仍存在若干值得探索的方向。首先，本研究的防禦策略建立在聯邦學習固有自癒能力之上，採用「僅懲罰不回滾」原則。當攻擊者目標從理性獲利轉為純粹破壞時，自癒機制的有效性邊界將成為關鍵問題，未來可探討如何設計高效的模型回溯復原機制。

其次，本研究驗證了罰沒機制在特定威脅環境下的淨化效果，但在惡意節點絕對數量更大的場景中，淨化所需輪次將延長。未來可探討如何透過動態調整委員會規模或設計更積極的挑戰觸發策略來最佳化淨化效率。

最後，實際部署環境往往具有高度異質性與動態性，包括低軌衛星網路的通訊窗口限制、工業物聯網的資源差異等。未來可探討如何建構自適應委員會機制，根據網路威脅監控資料與節點資源狀態動態調整配置，同時確保不會產生安全缺口或成為新的攻擊向量。