

Multidimensional Contract Design for In-Network Aggregation in Hierarchical Federated Learning

Rania Hidoussi, Gael S. Mubibya and Elmahdi Driouch

Université du Québec à Montréal (UQAM), Montréal, QC, Canada

hidoussi.rania@courrier.uqam.ca, mubibya.gael@courrier.uqam.ca, driouch.elmahdi@uqam.ca

Abstract—Integrating In-Network Aggregation (INA) into Hierarchical Federated Learning (HFL) is a promising approach to improve scalability and reduce communication overhead. However, efficiently managing edge nodes equipped with programmable switches (ENPS) remains challenging due to limited memory and radio resources, while selfish ENPS may refuse to participate without adequate incentives. This challenge is further aggravated by information asymmetry, as the server cannot observe the private operational costs of ENPS. To address this issue, we propose a multidimensional contract-theoretic incentive mechanism that jointly considers latency, model quality, and reward allocation. We develop a multidimensional ENPS cost model that preserves the single-crossing property, enabling a tractable contract design under information asymmetry. Simulations demonstrate that the proposed mechanism effectively incentivizes ENPS participation and significantly improves system utility under diverse network settings.

Index Terms—Hierarchical FL, in-network aggregation, multidimensional contract theory, incentive mechanism.

I. INTRODUCTION

While federated learning (FL) enables distributed training, communication bottlenecks and device heterogeneity limit its scalability over wireless networks [1]. Hierarchical FL (HFL) mitigates these issues through edge-based aggregation, reducing server communication and improving scalability in dense IoT networks [2]. Further performance gains can be achieved through in-network aggregation (INA), an application of in-network computing (INC) that leverages programmable switches as line-rate aggregators [3]. By performing aggregation within the network, INA reduces training latency, increases throughput, and enhances privacy through stateless processing [4]. However, deploying INA raises several challenges. Limited programmable switch memory and pipeline constraints restrict large-scale model aggregation and often require parameter quantization, degrading learning performance. In addition, shared switch resources motivate efficient resource management and incentive mechanisms [5].

While recent research works [3, 4] demonstrate the technical feasibility of INA, they largely overlook the problem of incentivizing edge nodes equipped with programmable switches (ENPS). In practical deployments, ENPS and the server are often operated by different administrative or commercial entities, where each party aims to maximize its utility. As a result, ENPS may be reluctant to allocate scarce resources, such as switch memory and wireless bandwidth, without appropriate compensation. This challenge is further complicated by information asymmetry, since the operational

costs and resource availability of ENPS are typically private information not observable by the server.

Contract theory provides a solid framework for designing incentive mechanisms under such information asymmetry in FL. Prior work has explored contract-based formulations for resource allocation and participation incentives from multiple perspectives. For example, [6] proposed a hierarchical incentive framework combining contract theory and coalitional game theory to encourage the provision of high-quality training data. The InFEDGE framework [7] leveraged blockchain technology to enable transparent and automated reward distribution, while [8] combined contract theory and Stackelberg games to promote truthful participation and improve system utility. Other studies have investigated blockchain-assisted incentive mechanisms for privacy-sensitive applications [9] and multidimensional contract designs for heterogeneous FL systems with straggler mitigation [10].

Despite these advances, existing approaches are not directly applicable to INA-enabled HFL systems. They typically assume a single-dimensional or partially observable cost structure, whereas ENPS exhibit multidimensional private information jointly characterized by switch memory usage and wireless resource availability. To address this gap, we develop a multidimensional contract-theoretic framework for heterogeneous ENPS under information asymmetry. The proposed framework jointly captures memory and radio resource costs while enabling tractable contract design through the preservation of the single-crossing property. The main contributions of this paper are as follows:

- We introduce an HFL system model enhanced with INA, where ENPS perform aggregation under coupled communication–computation resource constraints.
- We design a dynamic contract-theoretic incentive mechanism under information asymmetry, jointly optimizing latency, model quality, and resource-aware incentives.
- We develop a multidimensional ENPS cost model capturing memory, transmission power, and device association costs, establish its single-crossing property, and derive a tractable contract optimization framework.
- We validate the framework via simulations, showing gains in system utility and resource efficiency.

The rest of the paper is organized as follows. The system model and problem formulation are presented in Section II. Section III describes the contract theory design. Section IV

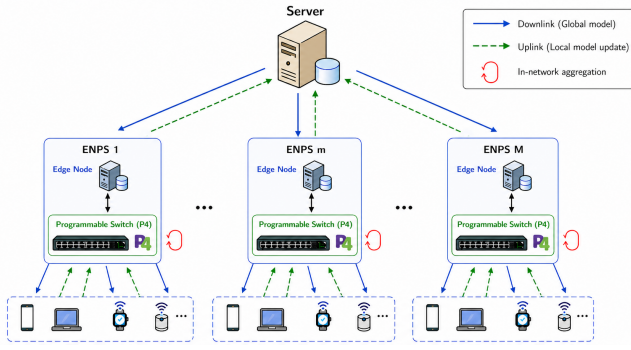


Fig. 1: System model

reports simulation results, and Section V concludes the paper.

II. SYSTEM MODEL AND PROBLEM FORMULATION

We consider an HFL system composed of a server, M ENPS, and N devices. Each ENPS $m \in \mathcal{M}$ acts as an intermediate aggregation node and is associated with a subset $\mathcal{N}_m$ of N_m devices. Fig. 1 illustrates the system model.

A. Hierarchical Federated Learning Model

The server broadcasts the global model θ^r to M ENPS during round $r \in \mathcal{R}$, each serving its subset of devices. Device $n \in \mathcal{N}$ owns a local dataset $\mathcal{D}_n$ of size D_n , i.e., $\{(\mathbf{x}_n^{(1)}, y_n^{(1)}), \dots, (\mathbf{x}_n^{(D_n)}, y_n^{(D_n)})\}$, where $\mathbf{x}_n^{(i)}$ and $y_n^{(i)}$ denote the i -th data sample and its corresponding label, respectively. Then, device n updates its local model by minimizing the local loss

$$F_n(\theta_n^r) = \frac{1}{D_n} \sum_{i \in \mathcal{D}_n} l(\theta_n^r, \mathbf{x}_n^{(i)}, y_n^{(i)}), \quad (1)$$

where $l(\cdot)$ denotes the sample-wise loss. After local training, selected devices upload their updates to the associated ENPS for intermediate aggregation:

$$\hat{F}_m(\theta_m^r) = \frac{1}{\hat{D}_m} \sum_{n \in \mathcal{N}_m} \sum_{i \in \mathcal{D}_n} l(\theta_m^r, \mathbf{x}_n^{(i)}, y_n^{(i)}), \quad (2)$$

where $\hat{D}_m = \sum_{n \in \mathcal{N}_m} D_n$. The ENPS then transmit their aggregated models to the server, which computes the global objective

$$F(\theta^r) = \frac{1}{D} \sum_{m \in \mathcal{M}} \hat{D}_m \hat{F}_m(\theta_m^r), \quad (3)$$

where $D = \sum_{m \in \mathcal{M}} \hat{D}_m$. The objective of HFL is to learn the optimal global model $\theta^* = \arg \min_{\theta} F(\theta)$.

B. Communication and Computation Model

Since the achievable transmission rates vary across communication rounds, the round index r is included only when necessary for clarity.

The server broadcasts the global model to ENPS. The server-to-ENPS transmission delay is

$$t_{s \rightarrow m}^D = \frac{\theta^{size}}{\min_{m \in \mathcal{M}} R_{s \rightarrow m}}, \quad (4)$$

where θ^{size} is the model size and $R_{s \rightarrow m}$ the achievable downlink rate between the server and ENPS m . Each ENPS

forwards the received global model to its associated devices. The corresponding transmission delay is

$$t_{m \rightarrow n}^D = \frac{\theta^{size}}{\min_{n \in \mathcal{N}_m} R_{m \rightarrow n}}. \quad (5)$$

where $R_{m \rightarrow n}$ is the achievable ENPS-to-device downlink rate. After receiving the global model, each device performs e local training iterations. The local computation delay of device n is

$$t_n^{cmp} = e \frac{C_n D_n}{f_n}, \quad (6)$$

where f_n is the CPU frequency of device n , and C_n denotes the number of CPU cycles required per data sample.

For uplink transmission, the local model is partitioned into $c = \lceil \theta^{size}/s_p \rceil$ chunks due to the limited memory capacity of the programmable switch, where each chunk has a size of s_p bits [11]. Since only a limited number of aggregation slots are available, ENPS m can process at most $S_m(r)$ chunks simultaneously during round r , where $S_m(r)$ denotes the number of available aggregation slots. Chunks exceeding this capacity must wait until previously processed chunks are acknowledged and the corresponding slots are released. Thus, the uplink delay from device n to ENPS m is expressed as

$$t_{n \rightarrow m}^U(r) = \frac{\theta^{size}}{R_{n \rightarrow m}} + \left(\left\lceil \frac{c}{S_m(r)} \right\rceil - 1 \right) t^{RTT}, \quad (7)$$

where t^{RTT} represents the round-trip time required to acknowledge processed chunks and release switch resources. The uplink transmission rate is

$$R_{n \rightarrow m} = b_{n \rightarrow m} \log_2 \left(1 + \frac{G_{n,m} P_n}{b_{n \rightarrow m} N_0} \right). \quad (8)$$

where $G_{n,m}$ is the channel gain between ENPS m and device n , $b_{n \rightarrow m}$ denotes the uplink bandwidth allocated to device n for transmitting to ENPS m , P_n is the transmission power of device n , and N_0 represents the power spectral density of the additive white Gaussian noise (AWGN). After intermediate aggregation, each ENPS uploads the aggregated model to the server using orthogonal frequency-division multiple access (OFDMA), with the bandwidth allocated $b_{m \rightarrow s}$ to each ENPS. The achievable uplink rate is

$$R_{m \rightarrow s}(r) = b_{m \rightarrow s} \log_2 \left(1 + \frac{G_{s,m} P_m(r)}{b_{m \rightarrow s} N_0} \right), \quad (9)$$

where $G_{s,m}$ is the channel gain between the server and ENPS m , $P_m(r)$ is the uplink transmission power of ENPS m at round r . The ENPS-to-server transmission delay is

$$t_{m \rightarrow s}^U(r) = \frac{\theta^{size}}{R_{m \rightarrow s}(r)}. \quad (10)$$

The total delay at ENPS m during round r is

$$T_m(r) = t_{s \rightarrow m}^D + t_{m \rightarrow n}^D + \max_{n \in \mathcal{N}_m} (t_n^{cmp} + t_{n \rightarrow m}^U(r)) + t_{m \rightarrow s}^U(r), \quad (11)$$

C. Problem Formulation

Our objective is to jointly optimize the ENPS memory allocation and transmission power to minimize the cumulative

latency.

$$\mathbf{P1} : \min_{\{S_m(r), P_m(r)\}} \sum_{r \in \mathcal{R}} \max_{m \in \mathcal{M}} T_m(r) \quad (12)$$

subject to:

$$\max_{m \in \mathcal{M}} T_m(r) \leq T_{max}, \quad \forall r \in \mathcal{R}, \quad (13)$$

$$S_m(r) s_p \leq H_{max}, \quad \forall m \in \mathcal{M}, \forall r \in \mathcal{R}, \quad (14)$$

$$P_m(r) \leq P_{max}, \quad \forall m \in \mathcal{M}, \forall r \in \mathcal{R}, \quad (15)$$

Constraint (13) ensures that the latency of each HFL round does not exceed the maximum tolerable delay T_{max} . Constraints (14) and (15) limit the aggregation memory and uplink transmission power of each ENPS, where H_{max} and P_{max} denote the maximum programmable switch memory capacity and transmission power, respectively.

Since ENPS incur heterogeneous aggregation costs and act as self-interested entities, the server cannot directly enforce the resource allocation required to solve **P1**. Moreover, these costs are private information unavailable to the server, resulting in information asymmetry. Therefore, an incentive mechanism is required to encourage voluntary ENPS participation while achieving the objective of **P1**.

III. CONTRACT THEORY DESIGN

A. Server's Contract

At each training round r , the server dynamically offers a menu of contracts, each represented by $\phi_i = (T_i, A_i, R_i)$, where T_i , A_i , and R_i denote the latency requirement, intermediate model accuracy requirement, and reward for a type- i ENPS, respectively. Owing to information asymmetry, the server only knows the probability λ_i that an ENPS belongs to type i , with $\sum_{i=1}^I \lambda_i = 1$. For notational simplicity, the round index is omitted in the remainder of this section, although the contracts are updated at every training round.

B. ENPS Type

The private information of ENPS m is characterized by its marginal costs associated with memory usage ω_{mem}^i , transmission power ω_{pow}^i , and device association capability (i.e., the cost of accommodating cost associated devices while meeting the required model quality) ω_{assoc}^i , denoted by $\theta_i = (\omega_{\text{mem}}^i, \omega_{\text{pow}}^i, \omega_{\text{assoc}}^i)$. The ENPS types are ordered monotonically as $\omega_{\text{mem}}^1 \leq \dots \leq \omega_{\text{mem}}^I$, $\omega_{\text{pow}}^1 \leq \dots \leq \omega_{\text{pow}}^I$, $\omega_{\text{assoc}}^1 \leq \dots \leq \omega_{\text{assoc}}^I$.

C. ENPS Cost Model

The total cost incurred by a type- i ENPS consists of memory, transmission power, and device association capability.

1) *Memory Cost*: When an ENPS selects contract ϕ_i , the required number of aggregation slots is determined by the latency requirement through a deterministic policy $S_i = S(T_i)$, where $S(\cdot)$ is non-increasing. Thus, stricter latency requirements require more aggregation slots. A linear allocation policy is

$$S(T_i) = \min \{S_{\max}, S_{\min} + \beta(T_{\max} - T_i)\}, \quad (16)$$

where $S_{\min}$ and $S_{\max}$ denote the minimum and maximum numbers of aggregation slots, respectively, and $\beta > 0$ is a scaling parameter. The normalized memory cost is

$$E^{\text{mem}}(T_i|\theta_i) = \omega_{\text{mem}}^i \frac{S(T_i) s_p}{H_{\max}}. \quad (17)$$

2) *Transmission Power Cost*: The available time budget for ENPS-to-server transmission is

$$t^U(T_i) = T_i - \left(t_{s \rightarrow m}^D + t_{m \rightarrow n}^D + \max_{n \in \mathcal{N}_m} t_n^{\text{cmp}} + \max_{n \in \mathcal{N}_m} t_{n \rightarrow m}^U(S_i) \right). \quad (18)$$

To guarantee feasibility, we require $t^U(T_i) > 0$. The uplink transmission rate is

$$R_{m \rightarrow s} = b_{m \rightarrow s} \log_2 \left(1 + \frac{P(T_i) G_{s,m}}{b_{m \rightarrow s} N_0} \right). \quad (19)$$

The minimum transmission power required to send the intermediate model within the available time budget is

$$P(T_i) = \min \left\{ (2^{\gamma_i} - 1) \frac{b_{m \rightarrow s} N_0}{G_{s,m}}, P_{\max} \right\}, \quad (20)$$

where $\gamma_i = \theta^{\text{size}} / b_{m \rightarrow s} t^U(T_i)$. The corresponding transmission power cost is

$$E^{\text{pow}}(T_i|\theta_i) = \omega_{\text{pow}}^i \frac{P(T_i)}{P_{\max}}. \quad (21)$$

3) *Model Quality Cost*: The number of devices associated with a type- i ENPS depends on its marginal device association cost and is modeled as

$$N_i = \min \left\{ N_{\max}, \max \left\{ N_{\min}, N - \alpha_2 (\omega_{\text{assoc}}^i - \omega_{\text{min}}^{\text{assoc}}) \right\} \right\}. \quad (22)$$

where $N_{\min}$ and $N_{\max}$ denote the minimum and maximum numbers of devices that can be associated with an ENPS, respectively, $\alpha_2 > 0$ controls the sensitivity of the association load to the marginal association cost, and $\omega_{\text{min}}^{\text{assoc}} = \min_i \omega_{\text{assoc}}^i$. To capture the statistical richness of the associated client data, we define the normalized effective diversity using the Gini-Simpson index [12, 13] as

$$\tilde{D}_i = \frac{1 - \sum_{l=1}^L p_{i,l}^2}{1 - \frac{1}{L}}, \quad (23)$$

where $p_{i,l}$ denotes the fraction of samples belonging to class l . The maximum achievable intermediate model accuracy of a type- i ENPS is modeled as

$$\bar{A}_i = 1 - \exp \left(-\xi \frac{N_i \tilde{D}_i}{N} \right), \quad (24)$$

where $\xi > 0$ is a scaling parameter. Hence, $\bar{A}_i$ increases with both the number of associated devices and their data diversity. The corresponding operational cost is modeled as

$$E^{\text{assoc}}(A_i|\theta_i) = \omega_{\text{assoc}}^i \left(\alpha_N \frac{N_i}{N} + \alpha_A A_i \right), \quad (25)$$

where $\alpha_N > 0$ and $\alpha_A > 0$ weight the impacts of the number of associated devices and the required intermediate model accuracy, respectively.

4) *ENPS Total Cost*: The total cost incurred by a type- i ENPS selecting contract ϕ_i is

$$E(T_i, A_i|\theta_i) = E^{\text{mem}}(T_i|\theta_i) + E^{\text{pow}}(T_i|\theta_i) + E^{\text{assoc}}(A_i|\theta_i). \quad (26)$$

D. ENPS Utility

For a feasible contract ϕ_j , the utility of a type- i ENPS is defined as

$$\mathcal{U}_i(\phi_j) = R_j - E(T_j, A_j | \theta_i), \quad (27)$$

where R_j is the reward associated with contract ϕ_j , and $E(T_j, A_j | \theta_i)$ is the total cost incurred by a type- i ENPS.

E. Server Utility

The server aims to improve HFL performance while minimizing incentive payments. For a signed contract ϕ_i , the server utility is defined as

$$\mathcal{U}_S(\phi_i) = \omega_T \left(1 - \frac{T_i}{T_{\max}}\right) + \omega_Q A_i - R_i, \quad (28)$$

where $\omega_T > 0$ and $\omega_Q > 0$ are weighting coefficients controlling the importance of latency reduction and model quality, respectively. The first term rewards latency reduction, whereas the second one captures the model quality.

F. Optimal Contract Formulation

Under information asymmetry, the contract design must satisfy two constraints, to ensure ENPS participation and truthful contract selection:

1) *Individual Rationality (IR)*: Each type- i ENPS should obtain a non-negative utility when selecting the contract designed for its own type:

$$\mathcal{U}_i(\phi_i) = R_i - E(T_i, A_i | \theta_i) \geq 0, \quad \forall i \in \mathcal{I}. \quad (29)$$

2) *Incentive Compatibility (IC)*: Each type- i ENPS should maximize its utility by selecting the contract intended for its own type rather than any other contract:

$$\mathcal{U}_i(\phi_i) \geq \mathcal{U}_i(\phi_j), \quad \forall i, j \in \mathcal{I}, i \neq j, \quad (30)$$

$$R_i - E(T_i, A_i | \theta_i) \geq R_j - E(T_j, A_j | \theta_i), \quad \forall i, j \in \mathcal{I}, i \neq j. \quad (31)$$

The objective of the server is to maximize its expected utility while ensuring voluntary participation and truthful contract selection. The contract optimization problem is formulated as

$$\mathbf{P2}: \max_{\{T_i, A_i, R_i\}_{i=1}^I} \sum_{i=1}^I \lambda_i \left[\omega_T \left(1 - \frac{T_i}{T_{\max}}\right) + \omega_Q A_i - R_i \right] \quad (32)$$

subject to

Constraints (29), (31) and

$$T_i \leq T_{\max}, \quad \forall i \in \mathcal{I}, \quad (33)$$

$$A_i \leq \rho \bar{A}_i, \quad \forall i \in \mathcal{I}, \quad (34)$$

$$\sum_{i=1}^I \lambda_i R_i \leq R_{\max}. \quad (35)$$

Constraint (33) limits the maximum contract latency. Constraint (34) ensures that the required intermediate model accuracy does not exceed a fraction $\rho \in (0, 1]$ of the maximum achievable accuracy $\bar{A}_i$ of a type- i ENPS. Constraint (35) restricts the server reward budget.

Since the contract design problem involves I IR constraints and $I(I - 1)$ global IC constraints, solving **P2** directly

becomes computationally challenging as the number of ENPS types increases. To obtain a more tractable formulation, we replace the global IC constraints with local IC ones.

Assumption 1 (Feasible Intended Contracts): the contract designed for a type- i ENPS is assumed to be feasible for that type. $A_i \leq \rho \bar{A}_i, \forall i \in \mathcal{I}$, ensuring that the required model quality does not exceed the ENPS aggregation capability.

Proposition 1 (Single-Crossing): Under the ordered ENPS types, the proposed ENPS cost model satisfies the single-crossing property. Specifically, for any two feasible contracts $\phi_a = (T_a, A_a, R_a)$ and $\phi_b = (T_b, A_b, R_b)$ satisfying $T_a \leq T_b$ and $A_a \geq A_b$, we have

$$\Delta_i(a, b) \leq \Delta_j(a, b), \quad \forall i < j, \quad (36)$$

where $\Delta_i(a, b) = E(T_a, A_a | \theta_i) - E(T_b, A_b | \theta_i)$.

Proof: The total ENPS cost is given by (26). Using the proposed cost model, this can be written as $E(T, A | \theta_i) = \omega_{\text{mem}}^i g_{\text{mem}}(T) + \omega_{\text{pow}}^i g_{\text{pow}}(T) + \omega_{\text{assoc}}^i g_{\text{assoc}}(A)$, where $g_{\text{mem}}(T) = S(T)s_p/H_{\max}$, $g_{\text{pow}}(T) = P(T)/P_{\max}$, and $g_{\text{assoc}}(A) = \alpha_N(N_i/N) + \alpha_A A$. For two contracts satisfying $T_a \leq T_b$ and $A_a \geq A_b$, contract ϕ_a is stricter than ϕ_b . Hence, $g_{\text{mem}}(T_a) - g_{\text{mem}}(T_b) \geq 0$, $g_{\text{pow}}(T_a) - g_{\text{pow}}(T_b) \geq 0$, and $g_{\text{assoc}}(A_a) - g_{\text{assoc}}(A_b) \geq 0$. Thus, the incremental cost of choosing ϕ_a instead of ϕ_b for a type- i ENPS is $\Delta_i(a, b) = \omega_{\text{mem}}^i [g_{\text{mem}}(T_a) - g_{\text{mem}}(T_b)] + \omega_{\text{pow}}^i [g_{\text{pow}}(T_a) - g_{\text{pow}}(T_b)] + \omega_{\text{assoc}}^i [g_{\text{assoc}}(A_a) - g_{\text{assoc}}(A_b)]$. Using the ordered ENPS types and the non-negativity of all bracketed terms, the inequality in (36) follows. Therefore, despite the multidimensional cost structure, the proposed ENPS cost model preserves the single-crossing property. ■

Lemma 1 (Monotonicity): for any $i < j$, the optimal contract satisfies

$$T_i \leq T_j, \quad A_i \geq A_j, \quad R_i \geq R_j, \quad (37)$$

which establishes an ordered structure of the contract menu.

Proof: Under the ordered efficiency structure and the *proposition 1*, lower-index ENPS types incur lower costs for stricter contracts. Therefore, assigning stricter latency and accuracy criteria to higher-cost types would violate the IC constraints. Since stricter contracts require higher effort, rewards must also be monotonic to preserve truthful participation. ■

Lemma 2 (Reduce IR constraints): Under *proposition 1*, and *lemma 1*, if the IR constraint of the highest-cost ENPS type is satisfied, then the IR constraints of all remaining ENPS types are automatically satisfied. Therefore, the I IR constraints can be reduced to the following single constraint:

$$R_I - E(T_I, A_I | \theta_I) \geq 0. \quad (38)$$

Proof: For any type- i ENPS with $i < I$, the IC constraint implies $R_i - E(T_i, A_i | \theta_i) \geq R_I - E(T_I, A_I | \theta_i)$. By the single-crossing property in *Proposition 1*, $E(T_I, A_I | \theta_i) \leq E(T_I, A_I | \theta_I)$ for all $i < I$. Combining the above inequalities yields $R_i - E(T_i, A_i | \theta_i) \geq R_I - E(T_I, A_I | \theta_I)$. Therefore, if the highest-cost ENPS type satisfies (38), then $R_i - E(T_i, A_i | \theta_i) \geq 0$ for all $i \in \mathcal{I}$. Hence, all IR constraints are implied by (38). ■

Lemma 3 (Reduce IC constraints): Under *proposition*

1, the global IC constraints are equivalent to the following local downward IC (LDIC) and local upward IC (LUIC) constraints:

$$R_i - E(T_i, A_i | \theta_i) \geq R_{i+1} - E(T_{i+1}, A_{i+1} | \theta_i), \quad i = 1, \dots, I-1, \quad (39)$$

$$R_i - E(T_i, A_i | \theta_i) \geq R_{i-1} - E(T_{i-1}, A_{i-1} | \theta_i), \quad i = 2, \dots, I. \quad (40)$$

Proof: By Lemma 3, the contract menu is monotonic. Consider a downward deviation from ϕ_i to a non-adjacent contract ϕ_j with $j > i$. Summing the LDIC constraints (39) from $k = i$ to $j-1$ yields

$$R_i - R_j \geq \sum_{k=i}^{j-1} \left(E(T_k, A_k | \theta_k) - E(T_{k+1}, A_{k+1} | \theta_k) \right). \quad (41)$$

Using the single-crossing property in Proposition 1, each term can be lower bounded by its counterpart evaluated at type θ_i , giving

$$R_i - R_j \geq \sum_{k=i}^{j-1} \left(E(T_k, A_k | \theta_i) - E(T_{k+1}, A_{k+1} | \theta_i) \right). \quad (42)$$

The right-hand side telescopes to $E(T_i, A_i | \theta_i) - E(T_j, A_j | \theta_i)$, which establishes the global downward IC constraint. The proof for LUIC follows similarly from (40). Hence, the $I(I-1)$ global IC constraints are reduced to the $2(I-1)$ local IC constraints in (39) and (40). ■

Accordingly, rather than relying only on LDIC, we adopt a stronger formulation that jointly enforces both LDIC and LUIC [14]. The final optimization problem is formulated as

$$\mathbf{P3} : \max_{\{T_i, A_i, R_i\}_{i=1}^I} \sum_{i=1}^I \lambda_i \left[\omega_T \left(1 - \frac{T_i}{T_{\max}} \right) + \omega_Q A_i - R_i \right] \quad (43)$$

subject to

Constraints (33), (34), (35), (37), (38) (39), (40).

IV. SIMULATION RESULTS

A. Simulation Parameters

We consider M ENPS uniformly deployed on a circle of radius 1.5 km, with equally spaced angular positions. Each subset of devices is generated around its associated ENPS. The wireless channel gains $G_{s,m}$ and $G_{n,m}$ are modeled using large-scale fading following the path-loss model $PL_{\text{dB}} = 128.1 + 37.6 \log_{10}(d)$, where d denotes the distance between a device and an ENPS or between an ENPS and the server. Table I summarizes the main simulation parameters used. We use the MNIST dataset [15]. The machine learning model used is a CNN with 21,840 trainable parameters [2]. Local training is performed using SGD. The batch size, learning rate, and learning rate decay are set to 64, 0.01, and 0.995, respectively. **P3** is implemented in AMPL and solved using the Gurobi optimizer.

TABLE I: Simulation Parameters

Notation	Description	Value
M	Number of ENPS	10
N	Total number of devices	200
I	Number of ENPS types	5
θ^{size}	Model size	110 kb
$b_{n \rightarrow m}^U = b_{m \rightarrow s}^U$	Uplink bandwidth	15×10^3 Hz
N_0	Noise spectral density	-174 dBm/Hz
P_n	Client transmit power	23 dBm
$P_{\max}$	Maximum transmit power	40 dBm
f_n	CPU frequency	$[10^5, 3 \cdot 10^5]$
C_n	Number of CPU cycles	20
$T_{\max}$	Maximum round latency	7 sec
$S_{\max}$	ENPS aggregation memory	24 slots [4]
s_p	Chunk size	1024 bits
ω_{mem}	Memory usage cost	[0.30, 0.65]
ω_{pow}	Transmission power cost	[0.30, 0.50]
ω_{assoc}	Device association cost	[0.10, 0.90]

B. Simulation results

Fig. 2 shows the utility of different ENPS types across the contract items. Each ENPS type reaches its maximum utility at its intended contract item, while any deviation results in an utility loss, confirming the IC of the proposed contract design. Lower-cost ENPS types benefit from stricter contracts, whereas higher-cost ENPS types prefer less demanding contracts. The vertical arrows highlight the optimal contracts.

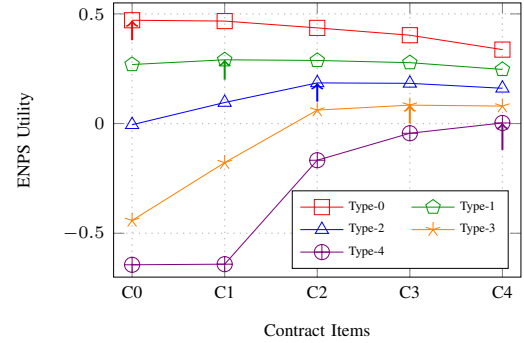


Fig. 2: Utility of different ENPS types across contract items.

Fig. 3 illustrates the actual server utility under different weight configurations (ω_T, ω_Q) . Prioritizing latency reduction, i.e., (1.6, 0.4), yields the highest utility, whereas emphasizing model quality with (0.4, 1.6) results in lower utility due to the higher cost of meeting stricter accuracy requirements. The balanced configuration (1, 1) provides an intermediate trade-off. Since the latency-oriented configuration achieves the highest server utility, the following results focus on this setting.

Fig. 4 shows the evolution of the average server utility for different numbers of ENPS. The utility increases over rounds and gradually saturates as the contract mechanism stabilizes. Moreover, increasing the number of ENPS improves the server utility since the workload is distributed among more aggregation entities, reducing the operational burden per ENPS and enabling more efficient contract satisfaction.

Finally, Fig. 5 shows the global accuracy versus cumulative time, capturing both communication and computation delays. The results reveal a trade-off with network density: although increasing the number of ENPS improves data diversity, the

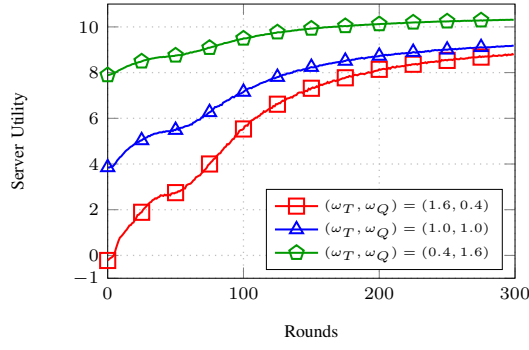


Fig. 3: Real server utility versus communication rounds under different weight configurations (ω_T, ω_Q)

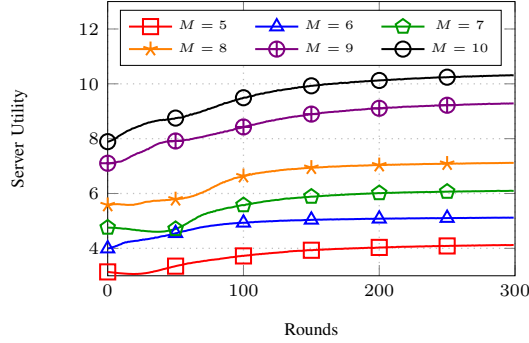


Fig. 4: Evolution of the average server utility over communication rounds for different numbers of ENPS.

best time efficiency is achieved at $M = 5$. Beyond this point, the overall round duration becomes dominated by the slowest ENPS, so additional aggregation delays outweigh the accuracy gains.

V. CONCLUSION

This paper investigated an HFL system enhanced with INA, where ENPS perform intermediate aggregation under information asymmetry. To motivate voluntary ENPS participation while minimizing training latency, we proposed a multidimensional contract-theoretic incentive mechanism that jointly optimizes latency, global model accuracy, and reward allocation. We further developed a multidimensional ENPS cost model capturing memory and radio resource costs, and proved that it preserves the single-crossing property, enabling a tractable contract design through monotonicity and the reduction of global IC constraints to LDIC and LUIC constraints. Simulation results demonstrated the effectiveness of the proposed mechanism in improving system utility and global model accuracy under various contract configurations and network settings. As future work, we will investigate a two-level incentive mechanism that jointly incentivizes both ENPS and devices.

REFERENCES

[1] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A Joint Learning and Communications Framework for Federated Learning Over Wireless Networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 269–283, 2021.

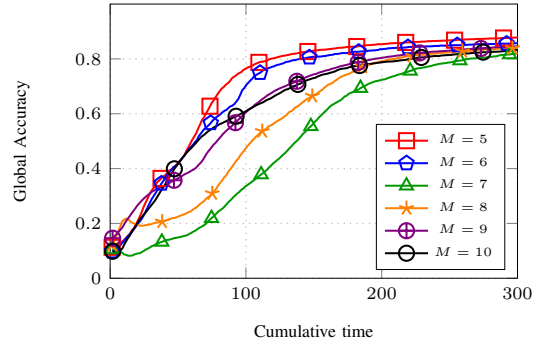


Fig. 5: Evolution of the global accuracy over cumulative training time for different numbers of ENPS.

[2] L. Liu, J. Zhang, S. Song, and K. B. Letaief, "Client-Edge-Cloud Hierarchical Federated Learning," in *IEEE ICC*, 2020, pp. 1–6.

[3] A. Sacco, A. Angi, G. Marchetto, and F. Esposito, "P4FL: An Architecture for Federating Learning With In-Network Processing," *IEEE Access*, vol. 11, pp. 103 650–103 658, 2023.

[4] J. Xia, G. Cheng, W. Wu, L. Luo, and D. Guo, "Exploring Communication-Efficient Federated Learning via Stateless in-Network Aggregation," *IEEE Trans. on Mobile Comput.*, vol. 24, no. 8, pp. 7423–7439, 2025.

[5] J. Xia, W. Wu, L. Luo, G. Cheng, D. Guo, and Q. Nian, "Accelerating and Securing Federated Learning with Stateless In-Network Aggregation at the Edge," in *IEEE ICDCS*, 2024, pp. 692–702.

[6] W. Y. B. Lim, Z. Xiong, C. Miao, D. Niyato, Q. Yang, C. Leung, and H. V. Poor, "Hierarchical Incentive Mechanism Design for Federated Machine Learning in Mobile Networks," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 9575–9588, 2020.

[7] X. Wang, Y. Zhao, C. Qiu, Z. Liu, J. Nie, and V. C. M. Leung, "InFEDge: A Blockchain-Based Incentive Mechanism in Hierarchical Federated Learning for End-Edge-Cloud Communications," *IEEE J. on Sel. Areas in Commun.*, vol. 40, no. 12, pp. 3325–3342, 2022.

[8] J. Huang, B. Ma, Y. Wu, Y. Chen, and X. Shen, "A Hierarchical Incentive Mechanism for Federated Learning," *IEEE Trans. on Mobile Comput.*, vol. 23, no. 12, pp. 12 731–12 747, 2024.

[9] J. Kang, J. Wen, D. Ye, B. Lai, T. Wu, Z. Xiong, J. Nie, D. Niyato, Y. Zhang, and S. Xie, "Blockchain-Empowered Federated Learning for Healthcare Metaverses: User-Centric Incentive Mechanism With Optimal Data Freshness," *IEEE Trans. on Cogn. Commun. and Netw.*, vol. 10, no. 1, pp. 348–362, 2024.

[10] X. Liu, J. Liu, X. Wei, and Y. Wang, "Incentive Mechanism for Blockchain-Enabled Coded Federated Learning in Edge Clouds," in *IEEE 31st ICPADS*, 2025, pp. 1–8.

[11] A. Sapio, M. Canini, C.-Y. Ho, J. Nelson, P. Kalnis, C. Kim, A. Krishnamurthy, M. Moshref, D. Ports, and P. Richtarik, "Scaling Distributed Machine Learning with In-Network Aggregation," in *18th USENIX Symp. Netw. Syst. Des. Implement. (NSDI)*. USENIX Association, Apr. 2021, pp. 785–808.

[12] E. H. Simpson, "Measurement of Diversity," *Nature*, vol. 163, p. 688, 1949.

[13] S. H. Hurlbert, "The Nonconcept of Species Diversity: A Critique and Alternative Parameters," *Ecology*, vol. 52, no. 4, pp. 577–586, 1971.

[14] P. Bolton and M. Dewatripont, *Contract theory*. MIT Press, 2005.

[15] L. Deng, "The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 141–142, 2012.