Towards Inference Delivery Networks: Distributing Machine Learning with Optimality Guarantees

Tareq Si Salem\textsuperscript{1}, Gabriele Castellano\textsuperscript{1,2}, Giovanni Neglia\textsuperscript{1}, Fabio Pianese\textsuperscript{2}, and Andrea Araldo\textsuperscript{3}

\textsuperscript{1}Inria, Université Côte d’Azur, France
\textsuperscript{2}Nokia Bell Labs, France
\textsuperscript{3}Télécom SudParis - Institut Polytechnique de Paris, France
Inference tasks are generated by users and executed through pretrained inference models.
Inference Delivery Networks

- Simpler models available locally have low accuracy
- Complex models at the cloud may not meet latency constraints
Integrate ML inference in the continuum between end-devices and the cloud.
Contributions

- Present inference delivery networks
- Propose INFIDA, a distributed online allocation algorithm for IDNs with strong guarantees even w.o. a prior on the request process
- Evaluate INFIDA experimentally with greedy heuristics
System Overview

Routing path from $p_1$

- Compute node
- Model load

Cloud-to-device continuum

Task
• We represent the inference delivery network (IDN) as a weighted graph $G(\mathcal{V}, \mathcal{E})$

• $\mathcal{N} = \{1, 2, \ldots, N\}$ is the set of tasks the system can serve
• Each node \( v \in \mathcal{V} \) has an allocation budget \( b_v^\mathcal{V} \in \mathbb{R}_+ \), and \( s_m^v \in \mathbb{R}_+ \) is the size of model \( m \in \mathcal{M} \).

• The budget constraints: \( \sum_{m \in \mathcal{M}} x_{m}^{v} s_{m}^{v} \leq b_v^\mathcal{V}, \forall v \in \mathcal{V} \)
We assume that every node has a predefined routing path towards a suitable repository node for each task $i \in \mathcal{N}$.

A request is determined by the pair $(i, p)$.
When serving request \( \rho = (i, p) \in \mathcal{R} \) on node \( p_j \) using model \( m \), the system experiences a cost \( C_{p,j}^{p} \in \mathbb{R}^+ \).

Our theoretical results hold under this very general cost model, but to be concrete we refer to the following simpler model:

\[
C_{p,j}^{p} = \sum_{j'=1}^{j-1} w_{p_{j'}, p_{j'+1}} + d_{m}^{p_j} + \alpha (1 - a_m),
\]

\( \text{round-trip latency} \quad \text{inference delay} \quad \text{trade-off parameter} \quad \text{inaccuracy} \)
During a slot $t$ the system receives a batch of requests

$$r_t = \left[ r^t_\rho \right]_{\rho \in \mathcal{R}} \in (\mathbb{N} \cup \{0\})^{\mathcal{R}}$$

Each model has an available capacity $l^t_{\rho,m} \in \mathbb{N} \cup \{0\}$
For a model $m$ allocated at $v$ having the $k$-th smallest service cost, we denote by:

$$\gamma^k_\rho = C^v_{p,m}$$, \quad $$\lambda^k_\rho(l_t) = l^t,v_{\rho,m}$$, \quad $$z^k_\rho(l_t, x) = x^v_m l^t,v_{\rho,m}$$.
The aggregate cost incurred by the system at time slot $t$ is

$$C(r_t, l_t, x) = \sum_{\rho \in \mathcal{R}} \sum_{k=1}^{K_{\rho}} \gamma^k_{\rho} \cdot \min \left\{ r^t - \sum_{k'=1}^{k-1} z^k_{\rho}(l_t, x), z^k_{\rho}(l_t, x) \right\}$$

(a)

$$\cdot 1 \left\{ \sum_{k'=1}^{k-1} z^k_{\rho}(l_t, x) < r^t_{\rho} \right\}.$$  

(b)

- (a) is the number of requests served by the $k$-th ranked model
- (b) is zero when all the requests can be served before the $k$-th ranked model

Equivalently, our objective is to maximize the allocation gain defined as $G(r_t, l_t, x) = C(r_t, l_t, \omega) - C(r_t, l_t, x)$. 

13
The allocation gain has the following equivalent expression:

$$G(r_t, l_t, x) = \sum_{\rho \in \mathcal{R}} \sum_{k=1}^{K_\rho - 1} \left( (\gamma_{\rho+1}^k - \gamma_\rho^k) \right) \left( Z_\rho^k(r_t, l_t, x) - Z_\rho^k(r_t, l_t, \omega) \right),$$

where $Z_\rho^k(r_t, l_t, x) \triangleq \min \left\{ r_\rho^t, \sum_{k'=1}^k z_\rho^{k'}(l_t, x) \right\}$.

Physical allocations $\mathcal{X}$]

Virtual allocations $\mathcal{Y}$

$G(r_t, l_t, y)$ is concave over $y$.
• We consider a “pessimistic” scenario where both requests and available capacities are selected by an adversary
Subgradients computation

For every \( v \in \mathcal{V} \) the gain function has a subgradient \( g^v_t \) at point \( y^v_t \in \mathcal{Y}^v \) given by

\[
g^v_t = \left[ \sum_{\rho \in \mathcal{R}} j^t, v_{\rho, m} \left( \gamma^t_{\rho}(y_t) - C^v_{p, m} \right) 1\{\kappa^t_{\rho}(v, m) < K^*_\rho(y_t)\} \right]_{m \in \mathcal{M}},
\]

where \( K^*_\rho(y_t) = \min \left\{ k \in [K^t_{\rho} - 1] : \sum_{k' = 1}^{k} z^k_{\rho}(l_t, y_t) \geq r^t_{\rho} \right\} \).
INFIDA distributed allocation

1: procedure INFIDA($y_1^\nu = \arg \min_{y^\nu \in Y^\nu \cap D^\nu} \Phi^\nu(y^\nu) \quad \eta^\nu \in \mathbb{R}_+$)

2: for $t = 1, 2, \ldots, T$ do

3: $x_t^\nu \leftarrow \text{DepRound}(y_t^\nu)$ ▷ Sample a physical allocation

4: Compute $g_t^\nu \in \partial_y G(r_t, l_t, y_t)$

5: $\hat{y}_t^\nu \leftarrow \nabla \Phi^\nu(y_t^\nu)$ ▷ Map state to the dual space

6: $\hat{h}_{t+1}^\nu \leftarrow \hat{y}_t^\nu + \eta^\nu g_t^\nu$ ▷ Take gradient step in the dual space

7: $h_{t+1}^\nu \leftarrow (\nabla \Phi^\nu)^{-1}(\hat{h}_{t+1}^\nu)$ ▷ Map dual state back to the primal space

8: $y_{t+1}^\nu \leftarrow P_{Y^\nu \cap D^\nu}^\nu(h_{t+1}^\nu)$ ▷ Project new state onto the feasible region

9: end for

10: end procedure
Theoretical Guarantees

We provide the optimality guarantees of INFIDA in terms of the $\psi$-regret ($\psi = 1 - \frac{1}{e}$).

$\psi$-Regret

$$\psi \underbrace{\text{Total gain of the static optimum} - \text{Expected total gain of INFIDA}}_{\text{discount}}$$

$$\leq \psi \frac{RL_{\text{max}} \Delta C}{s_{\text{min}}} \sqrt{2s_{\text{max}} |\mathcal{V}| |\mathcal{M}| \sum_{v \in \mathcal{V}} \min\{b^v, \|s^v\|_1\} \log \left( \frac{\|s^v\|_1}{\min\{b^v, \|s^v\|_1\}} \right) t}$$

$$= \mathcal{O} \left( \sqrt{T} \right).$$

- The average $\psi$-regret goes to 0 when $T$ is large enough; we perform as well as a $\psi$-approximation of the static optimum that knows the sequence of the models’ capacities and requests in hindsight!
- $\psi$ is the best approximation bound achievable for the problem, assuming $P \neq NP$. 

Experiments - Requests Traces

(a) Fixed Popularity Profile  (b) Sliding Popularity Profile

Performance Metric. The performance of a policy is evaluated in terms of the time averaged gain normalized to the number of requests per second (NTAG).
Experiments - Trade-off between Latency and Accuracy

Figure 2: Fractional allocation decisions $y_m^\nu$ of INFIDA on the various tiers of Network Topology I under Fixed Popularity Profile.

(a) $\alpha = 3$

(b) $\alpha = 4$

(c) $\alpha = 5$
Figure 3: Average latency (dashed line) and inaccuracy (solid line) costs experienced with INFIDA for different values of $\alpha$ under Network Topology I and Fixed Popularity Profile.
Figure 4: NTAG of the different policies under Sliding Popularity Profile and network topologies: (a) Network Topology I, and (b) Network Topology II.
Figure 5: Average Latency vs. Average Inaccuracy obtained for different values of $\alpha \in \{0.5, 1, 2, 3, 4, 5, 6\}$ under Fixed Popularity Profile and Network Topology II.
Experiments - Update Costs

Figure 6: (a) Models Updates (MU) and (b) NTAG of Greedy and INFIDA for different values of refresh period $B \in \{4, 8, 16\}$, and for a dynamic refresh period. The experiment is run under Network Topology I and Sliding Popularity Profile.
Experiments - Scalability on Requests Load

Figure 7: NTAG of the different policies for different request rates under Network Topology I.

(a) Fixed Popularity Profile    (b) Sliding Popularity Profile
Thank you for your attention!