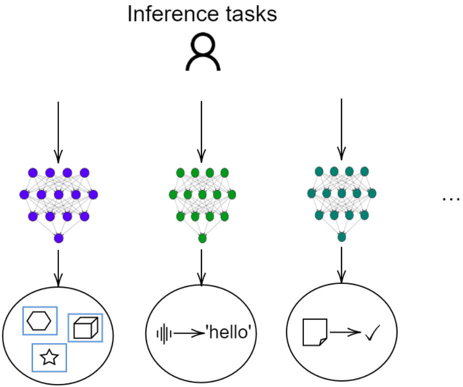
Towards Inference Delivery Networks: Distributing Machine Learning with Optimality Guarantees

Tareq Si Salem[1], Gabriele Castellano[1,2], Giovanni Neglia[1], Fabio Pianese[2], and Andrea Araldo[3]

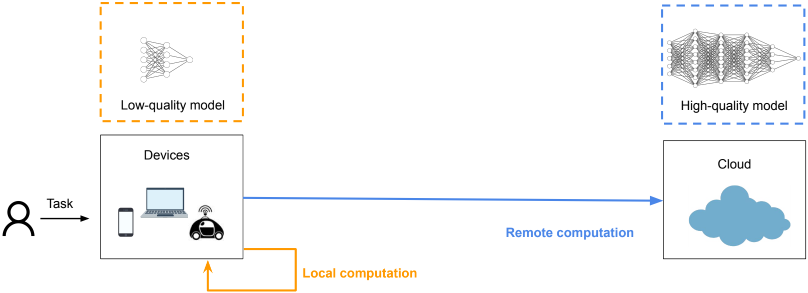
[1]Inria, Université Côte d'Azur, France
[2]Nokia Bell Labs, France
[3]Télécom SudParis - Institut Polytechnique de Paris, France

Machine Learning Tasks



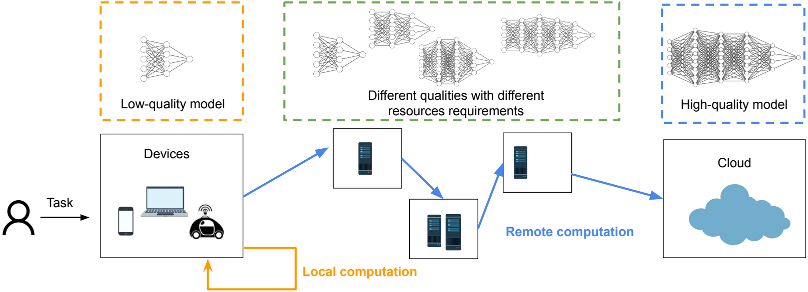
Inference tasks are generated by users and executed through pretrained inference models.

Inference Delivery Networks



- Simpler models available locally have low accuracy
- Complex models at the cloud may not meet latency constraints

Inference Delivery Networks

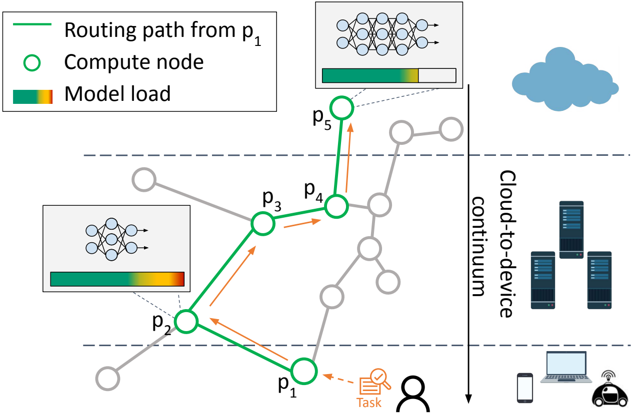


Integrate ML inference in the continuum between end-devices and the cloud.

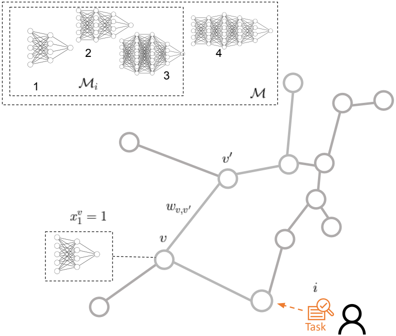
Contributions

- 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

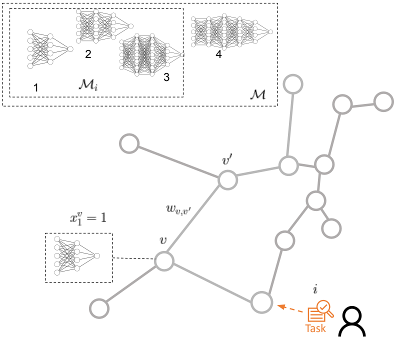


System Description - Compute Nodes and Models (1/2)



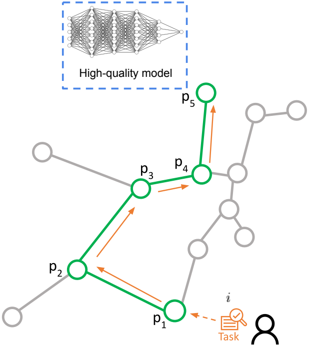
- We represent the inference delivery network (IDN) as a weighted graph $G(\mathcal{V}, \mathcal{E})$
- $\mathcal{N}=\{1, 2, \ldots, N\}$ is the set of tasks the system can serve

System Description - Compute Nodes and Models (2/2)



- Each node $v \in \mathcal{V}$ has an *allocation budget* $b^v \in \mathbb{R}_+$, and $s_m^v \in \mathbb{R}_+$ is the size of model $m \in \mathcal{M}$
- The budget constraints: $\sum_{m \in \mathcal{M}} x_m^v s_m^v \leq b^v, \forall v \in V$

System Description - Inference Requests



- 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 $(i, \boldsymbol{p})$

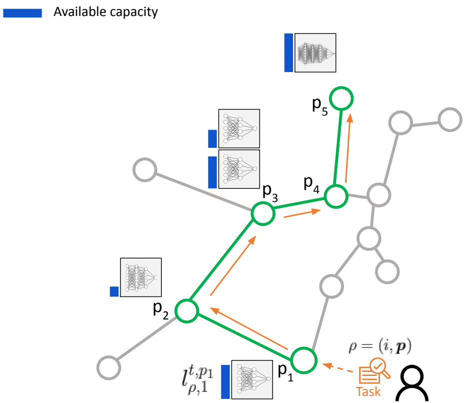
System Description - Cost Model

When serving request $\rho{=}(i,\boldsymbol{p}) \in \mathcal{R}$ on node p_j using model m, the system experiences a cost $C^{p_j}_{\boldsymbol{p},m} \in \mathbb{R}_+$.

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

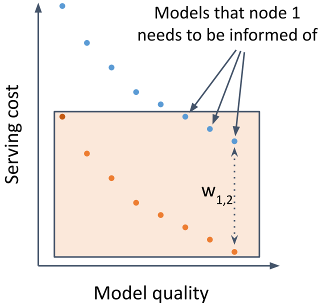
$$C^{p_j}_{\boldsymbol{p},m} = \underbrace{\sum_{j'=1}^{j-1} w_{p_{j'},p_{j'+1}}}_{\text{round-trip latency}} + \underbrace{d^{p_j}_m}_{\text{inference delay}} + \underbrace{\alpha}_{\text{trade-off parameter}} \underbrace{(1{-}a_m)}_{\text{inaccuracy}}.$$

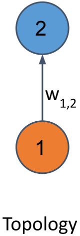
System Description - Request Load and Serving Capacity



- During a slot t the system receives a batch of requests $\boldsymbol{r}_t = [r_\rho^t]_{\rho \in \mathcal{R}} \in (\mathbb{N} \cup \{0\})^{\mathcal{R}}$
- Each model has an *available capacity* $l_{\rho,m}^{t,v} \in \mathbb{N} \cup \{0\}$

System Description - Serving Model (1/3)





For a model m allocated at v having the k-th smallest service cost, we denote by:

$$\underbrace{\gamma_\rho^k = C^v_{\boldsymbol{p},m}}_{\text{model service cost}}, \quad \underbrace{\lambda_\rho^k(\boldsymbol{l}_t) = l^{t,v}_{\rho,m}}_{\text{potential available capacity}}, \quad \underbrace{z_\rho^k(\boldsymbol{l}_t, \boldsymbol{x}) = x^v_m l^{t,v}_{\rho,m}}_{\text{effective available capacity}}.$$

System Description - Serving Model (2/3)

The aggregate cost incurred by the system at time slot t is

$$C(\boldsymbol{r}_t, \boldsymbol{l}_t, \boldsymbol{x}) = \sum_{\rho \in \mathcal{R}} \sum_{k=1}^{K_\rho} \gamma_\rho^k \cdot \underbrace{\min\left\{ r_\rho^t - \sum_{k'=1}^{k-1} z_\rho^{k'}(\boldsymbol{l}_t, \boldsymbol{x}),\, z_\rho^k(\boldsymbol{l}_t, \boldsymbol{x}) \right\}}_{\textbf{(a)}}$$

$$\cdot \underbrace{\mathbb{1}_{\left\{\sum_{k'=1}^{k-1} z_\rho^{k'}(\boldsymbol{l}_t, \boldsymbol{x}) < r_\rho^t\right\}}}_{\textbf{(b)}} .$$

- **(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

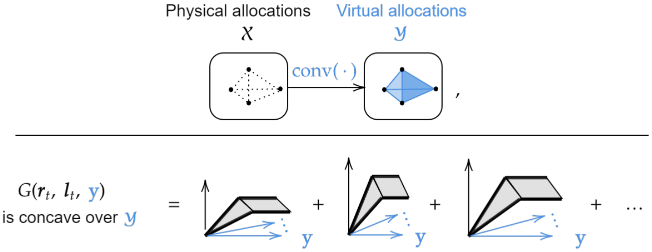
Equivalently, our objective is to maximize the *allocation gain* defined as $G(\boldsymbol{r}_t, \boldsymbol{l}_t, \boldsymbol{x}) = C(\boldsymbol{r}_t, \boldsymbol{l}_t, \boldsymbol{\omega}) - C(\boldsymbol{r}_t, \boldsymbol{l}_t, \boldsymbol{x})$.

System Description - Serving Model (3/3)

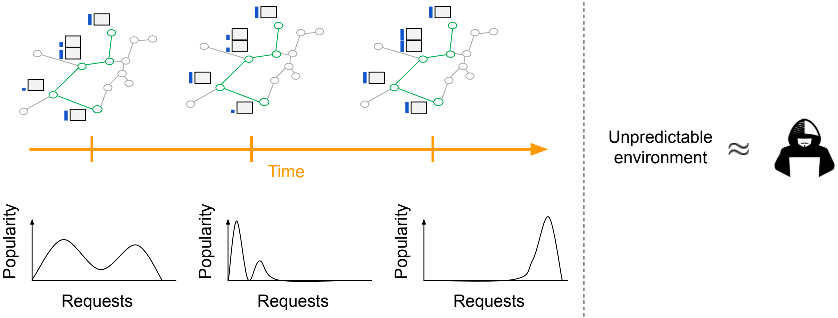
The allocation gain has the following equivalent expression:

$$G(\boldsymbol{r}_t, \boldsymbol{l}_t, \boldsymbol{x}) = \sum_{\rho \in \mathcal{R}} \sum_{k=1}^{K_\rho - 1} \underbrace{\left(\gamma_\rho^{k+1} - \gamma_\rho^{k}\right)}_{\text{cost saving}} \underbrace{\left(Z_\rho^k(\boldsymbol{r}_t, \boldsymbol{l}_t, \boldsymbol{x}) - Z_\rho^k(\boldsymbol{r}_t, \boldsymbol{l}_t, \boldsymbol{\omega})\right)}_{\text{additional requests}},$$

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



System Description - Adversarial Setting



- 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 $\boldsymbol{g}_t^v$ at point $\boldsymbol{y}_t^v \in \mathcal{Y}^v$ given by

$$\boldsymbol{g}_t^v = \left[\sum_{\rho \in \mathcal{R}} l_{\rho,m}^{t,v} \left(\gamma_\rho^{K_\rho^*(\boldsymbol{y}_t)} - C_{\boldsymbol{p},m}^v \right) \mathbb{1}_{\{\kappa_\rho(v,m) < K_\rho^*(\boldsymbol{y}_t)\}} \right]_{m \in \mathcal{M}},$$

where $K_\rho^*(\boldsymbol{y}_t) = \min \left\{ k \in [K_\rho - 1] : \sum_{k'=1}^{k} z_\rho^{k'}(\boldsymbol{l}_t, \boldsymbol{y}_t) \geq r_\rho^t \right\}$.

INFIDA distributed allocation

procedure $\text{INFIDA}(\boldsymbol{y}_1^v = \underset{\boldsymbol{y}^v \in \mathcal{Y}^v \cap \mathcal{D}^v}{\arg\min}\ \Phi^v(\boldsymbol{y}^v),\ \eta^v \in \mathbb{R}_+)$

  for $t = 1, 2, \dots, T$ **do**

    $\boldsymbol{x}_t^v \leftarrow \texttt{DepRound}(\boldsymbol{y}_t^v)$ ▷ Sample a physical allocation

    Compute $\boldsymbol{g}_t^v \in \partial_{\boldsymbol{y}^v} G(\boldsymbol{r}_t, \boldsymbol{l}_t, \boldsymbol{y}_t)$

    $\hat{\boldsymbol{y}}_t^v \leftarrow \nabla \Phi^v(\boldsymbol{y}_t^v)$ ▷ Map state to the dual space

    $\hat{\boldsymbol{h}}_{t+1}^v \leftarrow \hat{\boldsymbol{y}}_t^v + \eta^v \boldsymbol{g}_t^v$ ▷ Take gradient step in the dual space

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

    $\boldsymbol{y}_{t+1}^v \leftarrow \mathcal{P}_{\mathcal{Y}^v \cap \mathcal{D}^v}^{\Phi^v}(\boldsymbol{h}_{t+1}^v)$ ▷ Project new state onto the feasible region

  end for

end procedure

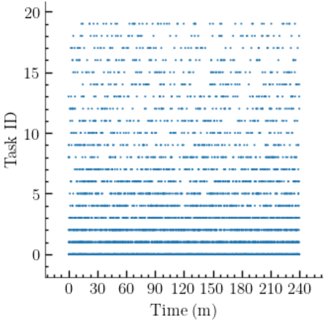
Theoretical Guarantees

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

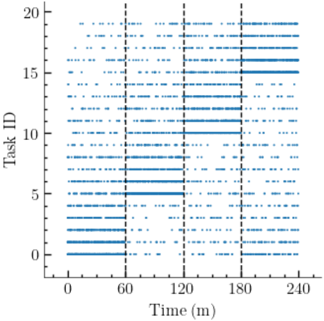
$$\begin{aligned}
&\psi\text{-Regret} \\
&= \underbrace{\psi}_{\text{discount}} \text{ Total gain of the static optimum} - \text{Expected total gain of } \texttt{INFIDA} \\
&\leq \psi \frac{R L_{\max} \Delta_{\mathcal{C}}}{s_{\min}} \sqrt{2 s_{\max} |\mathcal{V}||\mathcal{M}| \sum_{v \in \mathcal{V}} \min\{b^v, \|\boldsymbol{s}^v\|_1\} \log\left(\frac{\|\boldsymbol{s}^v\|_1}{\min\{b^v, \|\boldsymbol{s}^v\|_1\}}\right) T} \\
&= \mathcal{O}\left(\sqrt{T}\right).
\end{aligned}$$

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

Experiments - Requests Traces



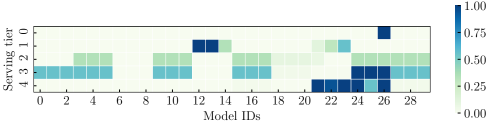
(a) Fixed Popularity Profile



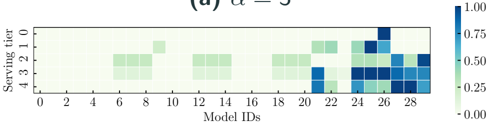
(b) Sliding Popularity Profile

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).

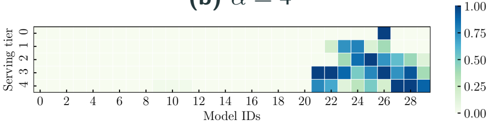
Experiments - Trade-off between Latency and Accuracy



(a) $\alpha = 3$



(b) $\alpha = 4$



(c) $\alpha = 5$

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

Experiments - Trade-off between Latency and Accuracy (1/3)

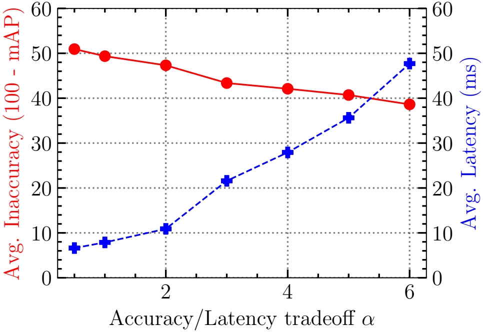


Figure 3: Average latency (dashed line) and inaccuracy (solid line) costs experienced with INFIDA for different values of α under *Network Topology I* and *Fixed Popularity Profile*.

Experiments - Trade-off between Latency and Accuracy (2/3)

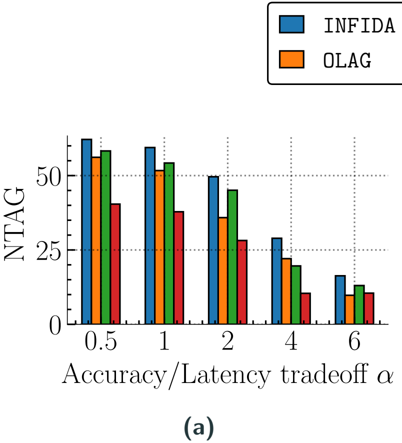


Figure 4: NTAG of the different policies under *Sliding Popularity Profile* and network topologies: (a) *Network Topology I*, and (b) *Network Topology II*.

Experiments - Trade-off between Latency and Accuracy (3/3)

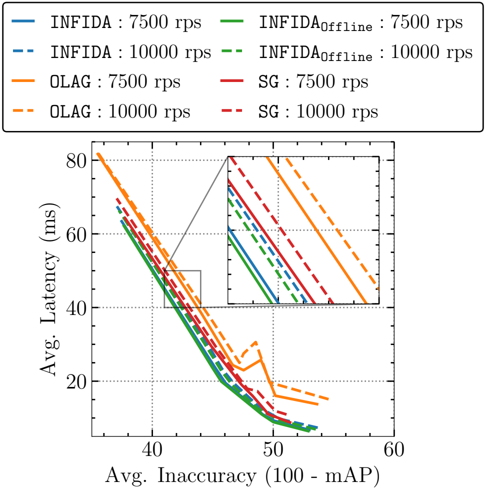
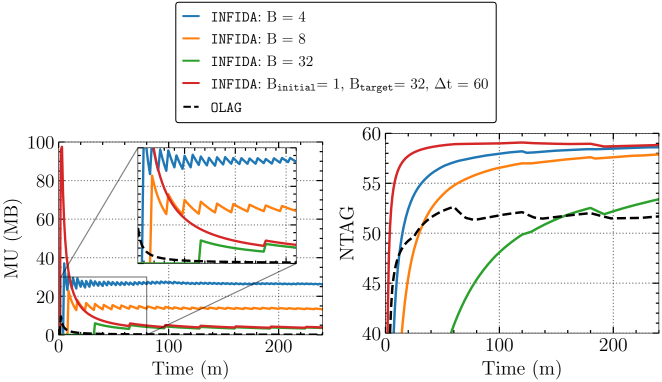


Figure 5: Average Latency vs. Average Inaccuracy obtained for different values of $\alpha \in \{0.5, 1, 2, 3, 4, 5, 6\}$ under *Fixed Popularity Profile* and *Network Topology II*.

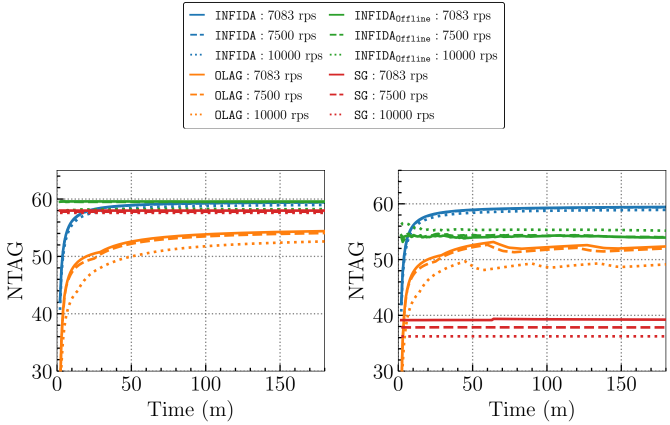
Experiments - Update Costs



(a) Models Updates (MU) **(b)** NTAG

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



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

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

Thank you for your attention!