Online Learning for Network Resource Allocation

Ph.D. Presentation

Tareq Si Salem

Inria center at Université Côte d'Azur

17 October 2022

Jury members:

Douglas LEITH	Professor, Trinity College Dublin, Ireland	Reviewer
Edmund YEH	Professor, Northeastern University, USA	Examiner
Giovanni NEGLIA	Research Director, Inria, France	Supervisor
György DÁN	Professor, KTH Royal Institute of Technology, Sweden	Examiner
Leandros TASSIULAS	Professor, Yale University, USA	Reviewer
Walid DABBOUS	Research Director, Inria, France	Examiner

Presentation Organization

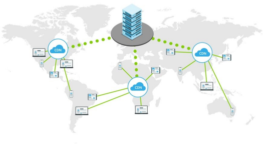
1. **Network Resource Allocation**
2. **Caching**
3. **Similarity Caching**
4. **Inference Delivery Networks**
5. **Fairness in Dynamic Resource Allocation**
6. **Concluding Remarks**

Presentation Organization

1. **Network Resource Allocation**
2. Caching
3. Similarity Caching
4. Inference Delivery Networks
5. Fairness in Dynamic Resource Allocation
6. Concluding Remarks

Network Resource Allocation

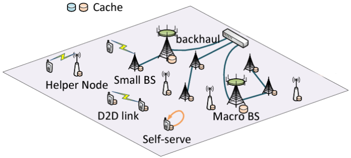
Network resource allocation is ubiquitous



(a) CDNs

(b) ICNs

(c) Cloud computing



(d) Edge/Wireless IoT

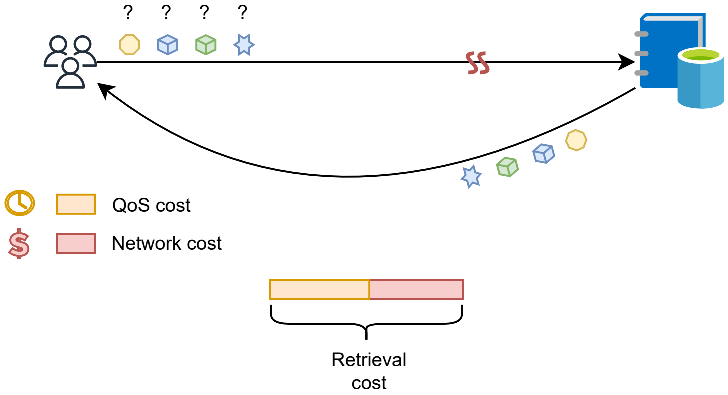
Goal

To provide *faster* service to demands generated by users (☺), or to *reduce* the computation or communication load on the system ($).

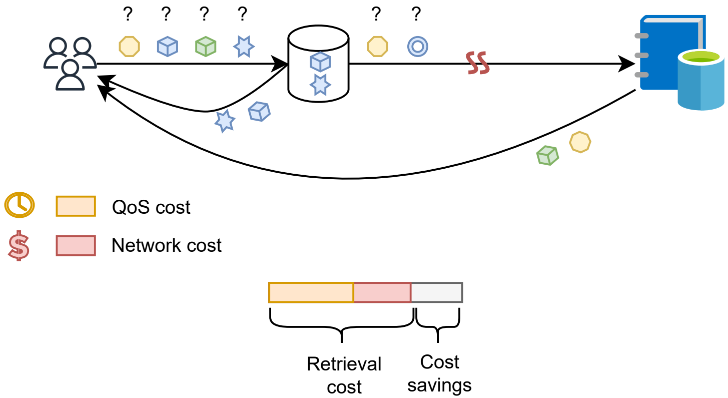
Presentation Organization

1. Network Resource Allocation
2. **Caching**
3. Similarity Caching
4. Inference Delivery Networks
5. Fairness in Dynamic Resource Allocation
6. Concluding Remarks

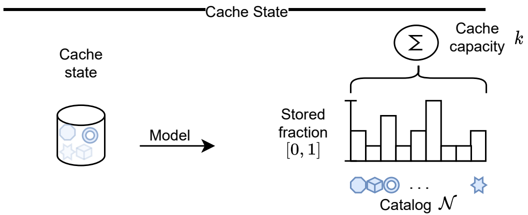
Caching



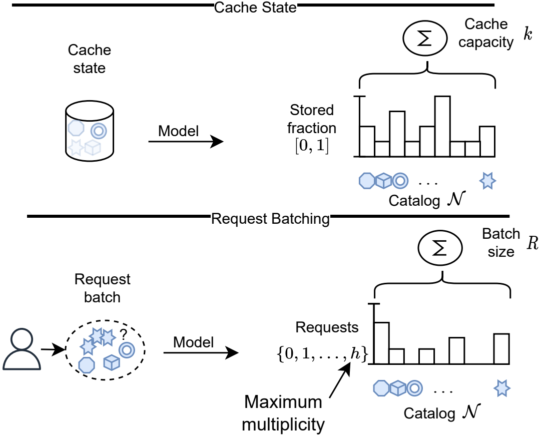
Caching



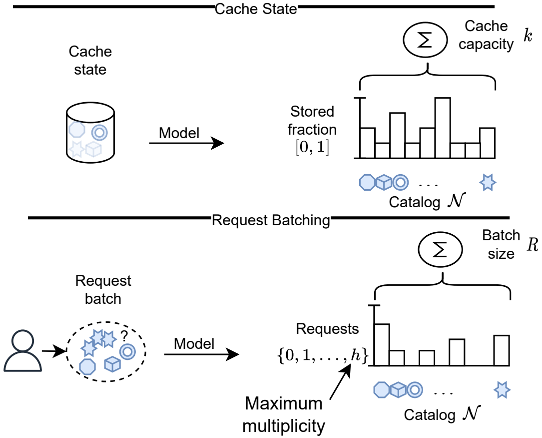
Caching – Model

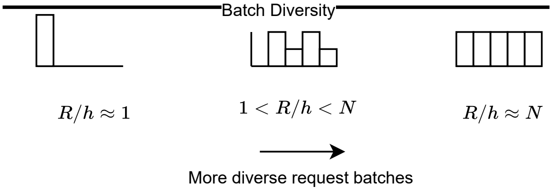


Caching – Model

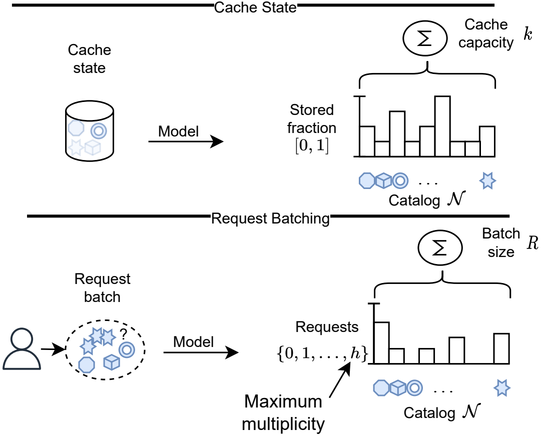


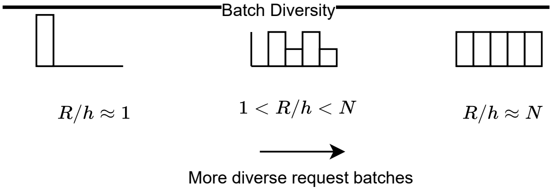
Caching – Model





Caching – Model

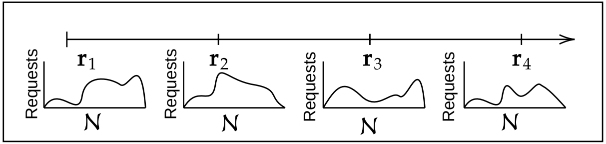




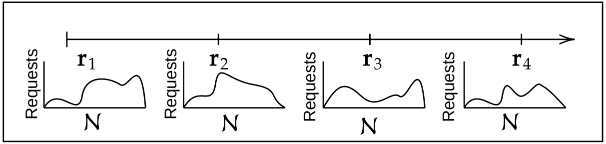
When a request batch $\mathbf{r}_t$ arrives, the cache incurs the following cost:

$$f_{\mathbf{r}_t}(\mathbf{x}_t) = \sum_{i=1}^{N} w_i r_{t,i} (1 - x_{t,i}).$$

Caching – Setting

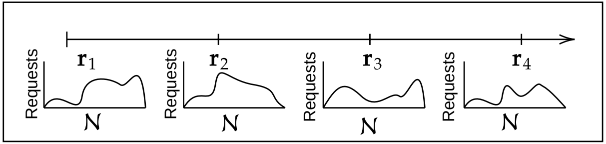


Caching – Setting

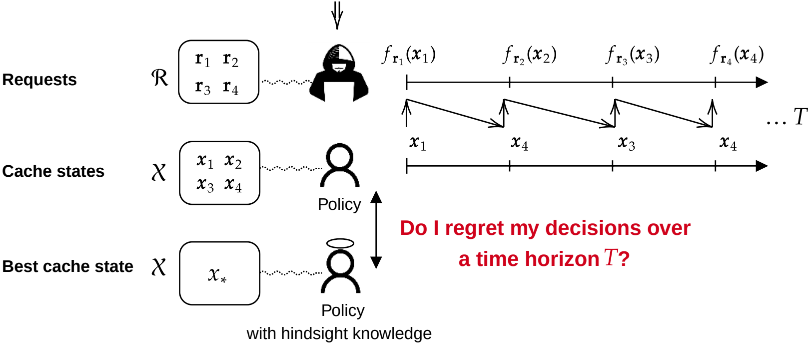


Noisy unpredictable environment can act as an **adversary** in the worst case scenario

Caching – Setting



Noisy unpredictable environment can act as an **adversary** in the worst case scenario



Caching – Performance Metric

Definition

The regret of a policy $\mathcal{A}$ is defined as

$$\mathrm{Regret}_T(\mathcal{A}) \triangleq \sup_{\{\boldsymbol{r}_t\}_{t=1}^T \in \mathcal{R}^T} \left\{ \sum_{t=1}^{T} f_{\boldsymbol{r}_t}(\boldsymbol{x}_t) - \min_{x \in \mathcal{X}} \sum_{t=1}^{T} f_{\boldsymbol{r}_t}(\boldsymbol{x}) \right\}.$$

When $\mathrm{Regret}_T(\mathcal{A})$ is sublinear in T, the policy $\mathcal{A}$ experiences no regret on average as $T \to \infty$.

Caching – Online Mirror Descent (OMD)

A mirror map $\Phi : \mathcal{D} \subset \mathbb{R}^{\mathcal{N}} \to \mathbb{R}$ defines a unique algorithm, e.g., $\Phi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_2^2$ defines OGD, and $\Phi(\boldsymbol{x}) = \sum_{i\in\mathcal{N}} x_i \log(x_i)$ (negative entropy) defines $\mathrm{OMD}_{\mathrm{NE}}$.

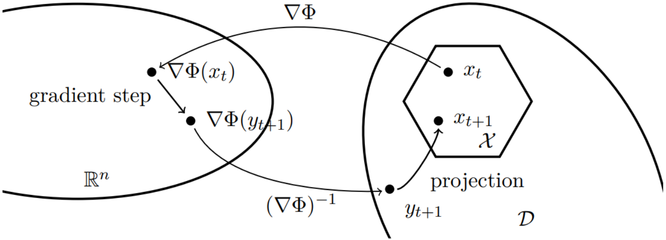


Figure: OMD update rule [Bub15].

Caching – Contributions

- We provide the first results to guide the selection of the best policy.

Theorem

OGD is optimal for $\frac{R}{h} \leq k$ (low diversity and large cache sizes). $\mathrm{OMD}_{\mathrm{NE}}$ is optimal for $\frac{R}{h} > 2\sqrt{Nk}$ (high diversity and small cache sizes).

Caching – Contributions

- We provide the first results to guide the selection of the best policy.

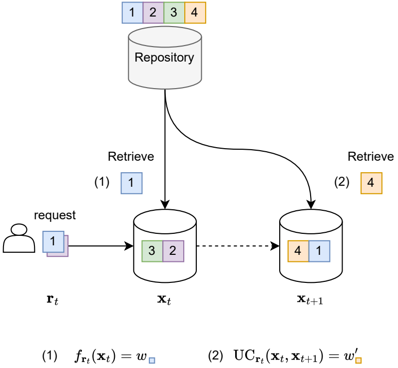
Theorem

OGD is optimal for $\frac{R}{h} \leq k$ (low diversity and large cache sizes). $\mathrm{OMD}_{\mathrm{NE}}$ is optimal for $\frac{R}{h} > 2\sqrt{Nk}$ (high diversity and small cache sizes).

- Highly efficient projection algorithm for $\mathrm{OMD}_{\mathrm{NE}}$ that yields a policy that has the lowest time-complexity per iteration among recent works [PDVI19, PS21, BBS20, MS21].

Caching – Update Costs

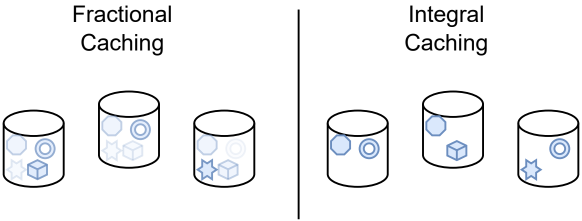
We define the update cost at time t as $\mathrm{UC}_{\boldsymbol{r}_t}(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}) \triangleq \sum_{i \notin \mathrm{supp}(\boldsymbol{r}_t)} w'_i \max\{0, x_{t+1,i} - x_{t,i}\}$.



Fractional Caching Incurs no Update Costs

We prove that any request batch $\boldsymbol{r}_t$, for $\mathrm{OMD}_{\mathrm{NE}}$ or OGD, it holds $\mathrm{UC}_{\boldsymbol{r}_t}(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}) = 0$.

Integral Caching – Necessity of Randomization



Proposition

Any deterministic policy restricted to select integral cache states in $\mathcal{Z} \triangleq \mathcal{X} \cap \{0,1\}^{\mathcal{N}}$ has linear regret, i.e.,

$$\text{Regret}_T(\mathcal{A}) \geq k\,(1 - k/N)\,T.$$

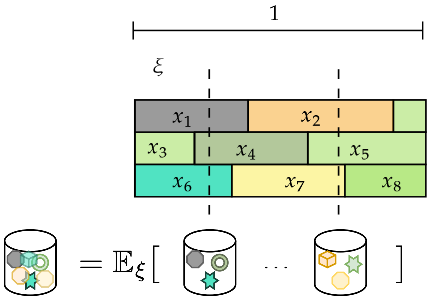
Randomized Integral Caching

We restrict ourselves to randomized rounding schemes that output $\boldsymbol{z}_t \in \mathcal{Z}$ such that $\mathbb{E}[\boldsymbol{z}_t] = \boldsymbol{x}_t$ through some rounding Ξ.

Remark

The expected regret $\mathrm{Regret}_T(\mathcal{A}, \Xi)$ is the same as the regret of $\mathcal{A}$.

A scheme that has this property is Madow's sampling [MM44]:



Randomized Integral Caching

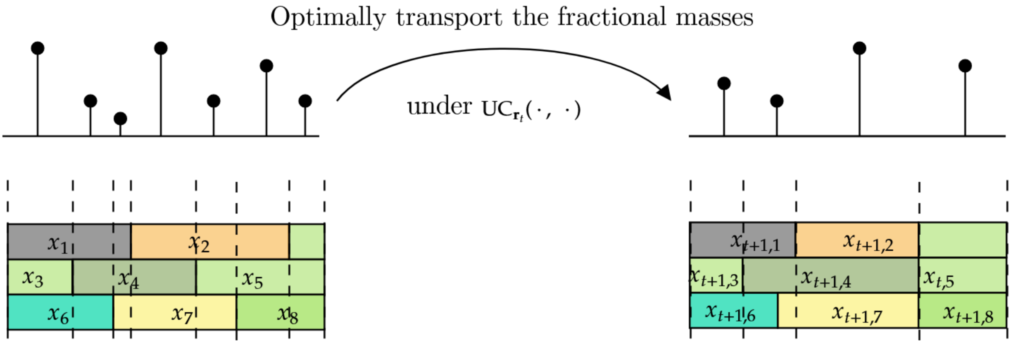
When considering the extended regret ($\text{E-Regret}_T(\mathcal{A}, \Xi)$) we lose immediately the regret guarantee:

Theorem

Any randomized caching policy constructed by an online policy $\mathcal{A}$ combined with online independent rounding as Ξ leads to $\Omega(T)$ $\text{E-Regret}(\mathcal{A}, \Xi)$.

Imposing dependence (coupling) between the two consecutive random states may significantly reduce the expected update cost.

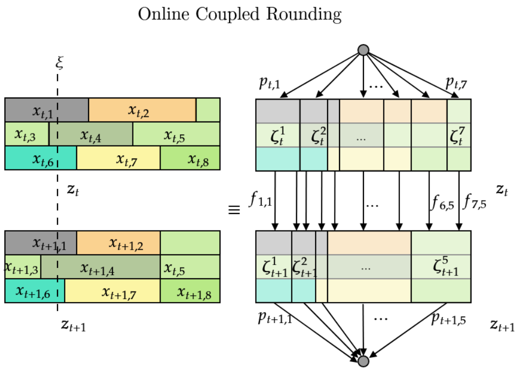
Randomized Integral Caching – Optimal Transport



Remark

We prove that this scheme selected as Ξ, coupled with a no-regret policy $\mathcal{A}$, has sublinear extended regret guarantee. However, it has a time-complexity $\mathcal{O}\left(N^3\right)$.

Randomized Integral Caching – Simpler Approach

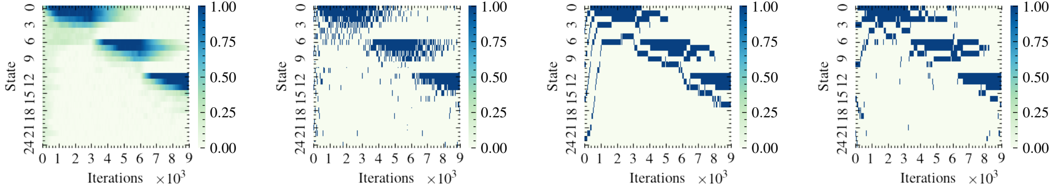


Theorem

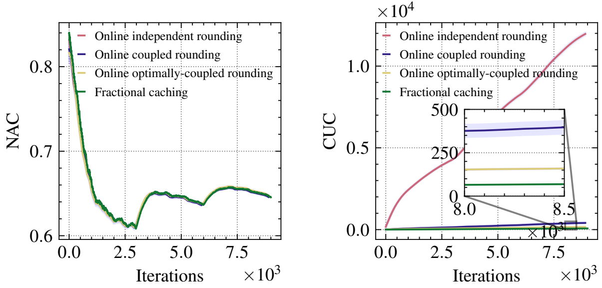
A no-regret policy $\mathcal{A}$ combined with online coupled rounding Ξ has $\mathcal{O}\left(\sqrt{T}\right)$ $\text{E-Regret}_T(\mathcal{A}, \Xi)$.

Online Coupled Rounding has linear time complexity.

Randomized Integral Caching – Summary



(a) Fractional cache states (b) Online independent rounding (c) Online coupled rounding (d) Online optimally-coupled rounding



(a) Normalized average cost (b) Cumulative update cost

Caching – Research Output

[C1] **No-Regret Caching via Online Mirror Descent**
IEEE ICC, 2021
T. Si Salem , G. Neglia, and S. Ioannidis

[C2] **Online Caching Networks with Adversarial Guarantees**
ACM SIGMETRICS, 2022
Y. Li**, T. Si Salem** et al.

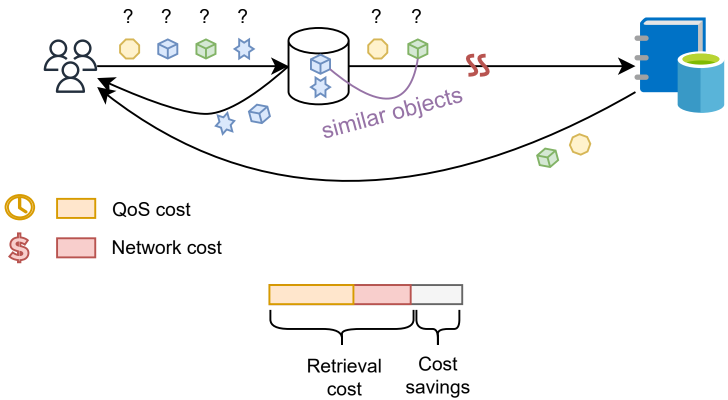
[J1] **Online Caching Networks with Adversarial Guarantees**
ACM POMACS, 2022
Y. Li**, T. Si Salem** et al.

[S1] **No-Regret Caching via Online Mirror Descent (extended)**
ACM ToMPECS (under review)
T. Si Salem , G. Neglia, and S. Ioannidis

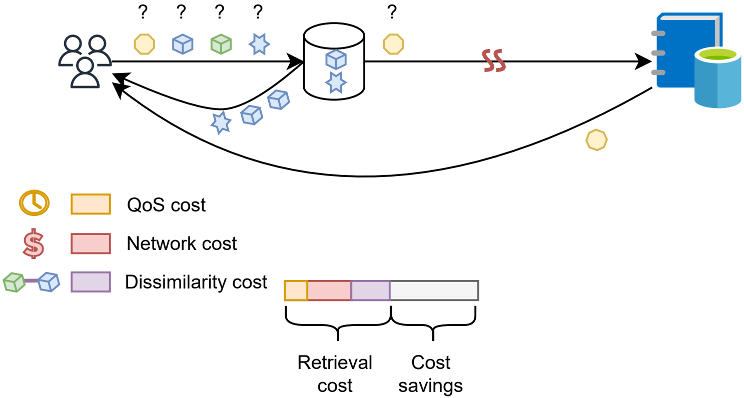
Presentation Organization

1. Network Resource Allocation
2. Caching
3. **Similarity Caching**
4. Inference Delivery Networks
5. Fairness in Dynamic Resource Allocation
6. Concluding Remarks

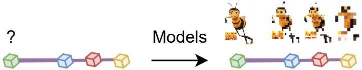
Similarity Caching



Similarity Caching

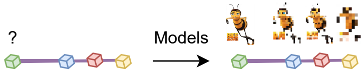


Similarity Caching – Motivation

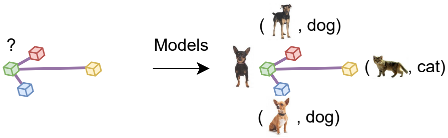


Video caching

Similarity Caching – Motivation

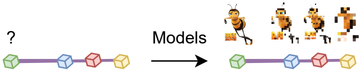


Video caching

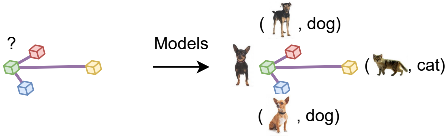


Classification results reuse \
Image retrieval

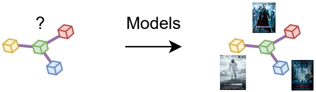
Similarity Caching – Motivation



Video caching

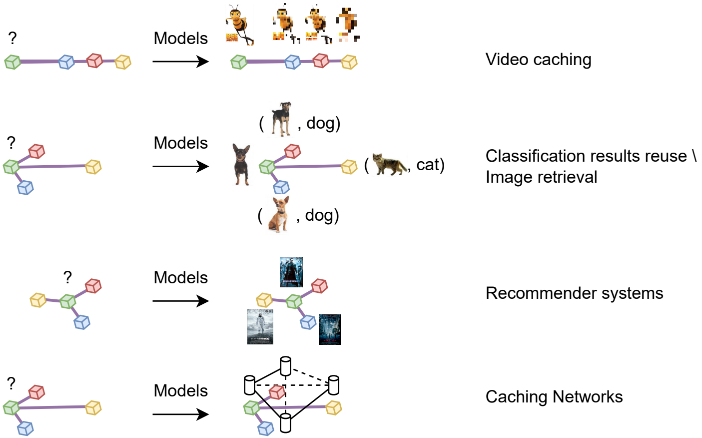


Classification results reuse \
Image retrieval



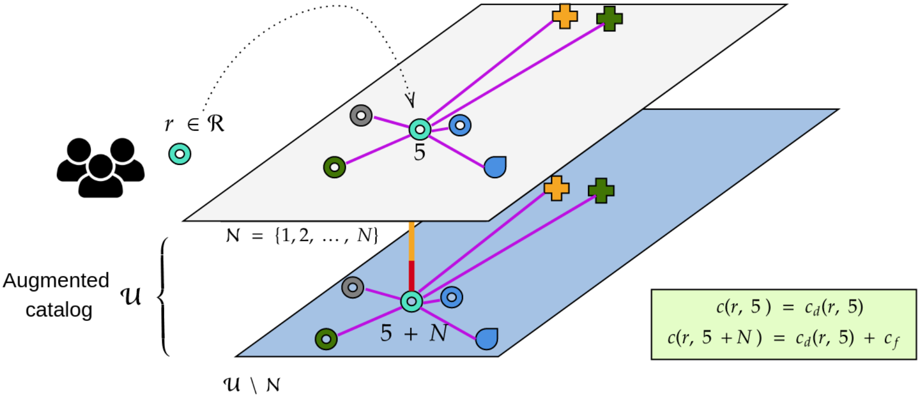
Recommender systems

Similarity Caching – Motivation



Content recommendation [SS16, SGSV18, CS20], content retrieval [FLO+08, PBC+09], Machine Learning serving [DGT+17, DGN17, CWZ+17, VGGK18, KBVA19].

Similarity Caching – Model



Similarity Caching – Caching Gain

When $r \in \mathcal{R}$ is received, a cache with allocation vector $\boldsymbol{x} \in \{0,1\}^{2N}$ incurs the cost

$$C(r, \boldsymbol{x}) = \sum_{i=1}^{2N} c(r, \pi_i^r) x_{\pi_i^r} \cdot \mathbb{1}\left(\sum_{j=1}^{i} x_{\pi_j^r} \leq k \right).$$

Objective

Our objective is to maximize the caching gain (cost savings) as the cache state $\boldsymbol{x}$ changes, given as

$$G(r, \boldsymbol{x}) \triangleq C(r, \text{empty cache}) - C(r, \boldsymbol{x}).$$

Similarity Caching – Performance Metric

Definition

The regret of the randomized policy $\mathcal{A}$ with the cache states $\{\boldsymbol{x}_t\}_{t=1}^{T}$ is given by

$$\psi\text{-Regret}(\mathcal{A}) = \sup_{\{r_1, r_2, \ldots, r_T\} \in \mathcal{R}^T} \left\{ \max_{x \in \mathcal{X}} \psi \sum_{t=1}^{T} G(r_t, \boldsymbol{x}) - \mathbb{E}\left[\sum_{t=1}^{T} G(r_t, \boldsymbol{x}_t) \right] \right\}.$$

The constant $\psi = 1 - 1/e$ is the best approximation ratio achievable in P to the NP-Hard static optimum

When ψ-Regret$(\mathcal{P})$ is sublinear in T, the policy experiences *no regret* on average as $T \to \infty$ w.r.t. the ψ-approximation of the offline problem in hindsight.

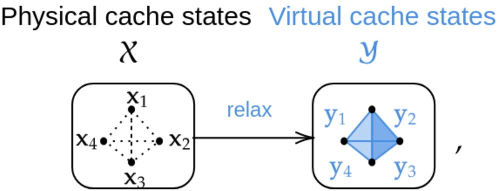
Similarity Caching – Exploiting OCO

Lemma

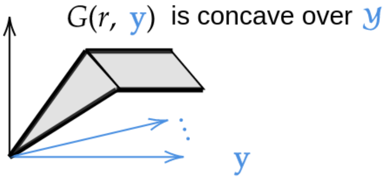
The caching gain can be expressed equivalently as

$$G(r,\boldsymbol{x}) = \sum_{i=1}^{K^r-1} \alpha_i^r \min\left\{k, \sum_{j=1}^{i} x_{\pi_j^r}\right\} + G_0,$$

where $\alpha_i^r \in \mathbb{R}_{\geq 0}$, $G_0 \in \mathbb{R}$, and $K^r \in \mathbb{N}$ are constants.



,



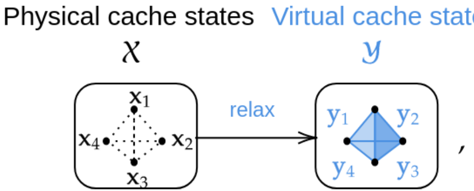
Similarity Caching – Exploiting OCO

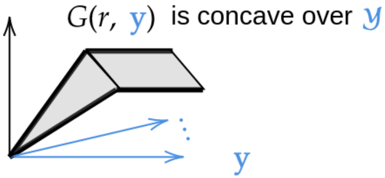
Lemma

The caching gain can be expressed equivalently as

$$G(r, \boldsymbol{x}) = \sum_{i=1}^{K^r-1} \alpha_i^r \min\left\{k, \sum_{j=1}^{i} x_{\pi_j^r}\right\} + G_0,$$

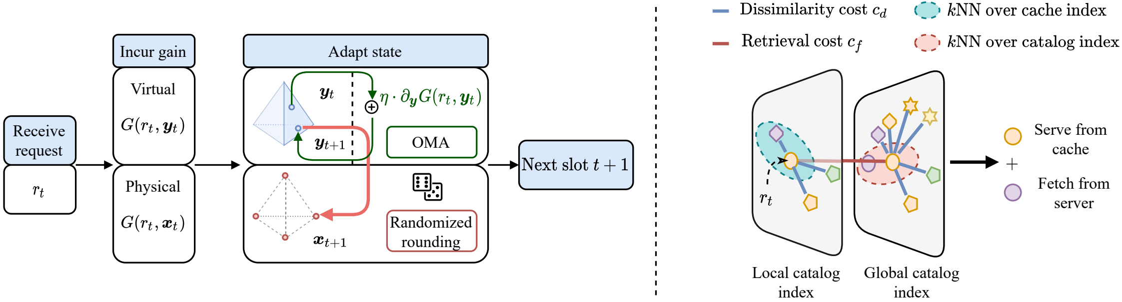
where $\alpha_i^r \in \mathbb{R}_{\geq 0}$, $G_0 \in \mathbb{R}$, and $K^r \in \mathbb{N}$ are constants.





The fractionally relaxed problem can be cast in the framework of OCO [Haz16] + Exploit the property $\mathbb{E}\left[G(r_t, \boldsymbol{x}_t)\right] \geq \psi G(r_t, \boldsymbol{y}_t)$.

Similarity Caching – AÇAI (Ascent Similarity Caching with Approximate Indexes) Policy



(a) (b)

Similarity Caching – AÇAI Policy Performance Guarantees

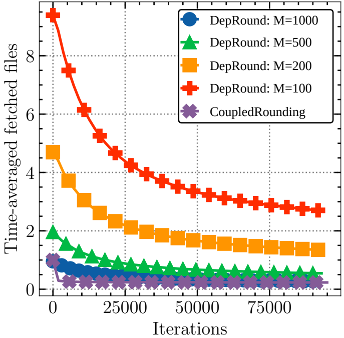
Theorem

AÇAI configured with a negentropy mirror map, learning rate η^*, and rounding scheme `ElasticCoupledRounding` or `DepRound` with a freezing period $M = \Theta\left(T^{\beta}\right)$ for $\beta \in [0, 1)$ satisfies

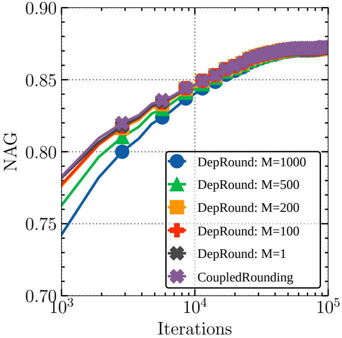
$$(1 - 1/e)\text{-Regret}_{\mathcal{X}}(\text{AÇAI}) = \mathcal{O}\left(T^{\frac{1+\beta}{2}}\right).$$

The parameter M reduces cache updates at the expense of reducing the cache reactivity. The update cost $\mathcal{C}_{\mathrm{UC},T}$ is given as $\mathcal{C}_{\mathrm{UC},T} = \mathcal{O}\left(T^{1-\beta}\right)$ for `DepRound` and $\mathcal{C}_{\mathrm{UC},T} = \mathcal{O}\left(\sqrt{T}\right)$ for `ElasticCoupledRounding`.

Similarity Caching – Service/Update Costs Tradeoff

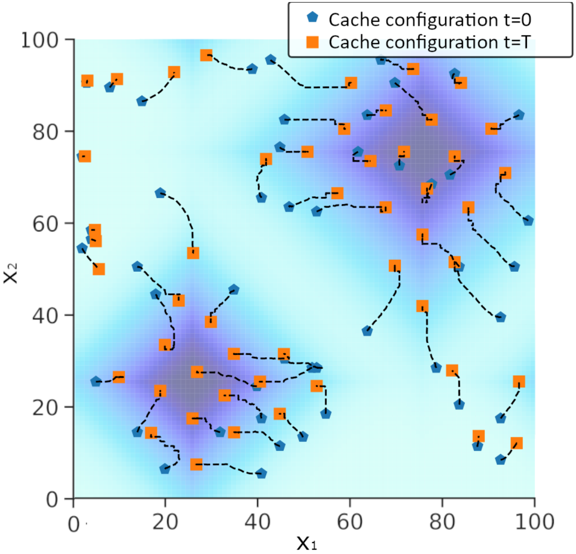


(a) Time-averaged fetched files



(b) Caching gain

Similarity Caching – A Heuristic under Continuous Catalogs



Caching Scheme

GRADES heuristic uses gradient descent to navigates the continuous space and find appropriate objects to store in the cache.

Similarity Caching – Research Output

GRADES: Gradient Descent for Similarity Caching
[C3] **IEEE INFOCOM, 2021**
A. Sabnis, **T. Si Salem** et al.

AÇAI: Ascent Similarity Caching with Approximate Indexes
[C4] **ITC 33, 2021**
T. Si Salem, G. Neglia, and Damiano Carra.

Best Paper Award

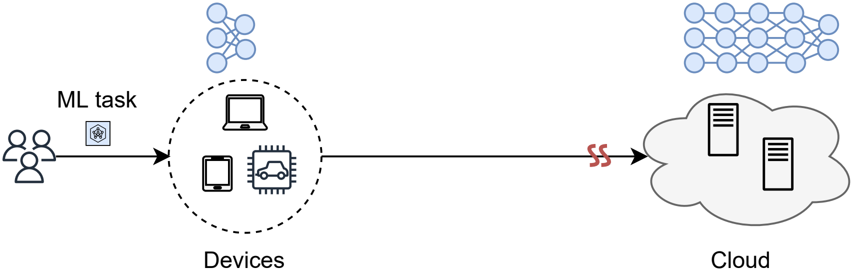
GRADES: Gradient Descent for Similarity Caching (extended)
[J2] **IEEE/ACM ToN, 2022**
A. Sabnis, **T. Si Salem** et al.

Ascent Similarity Caching with Approximate Indexes (extended)
[J3] **IEEE/ACM ToN, 2022**
T. Si Salem, G. Neglia, and Damiano Carra.

Presentation Organization

1. Network Resource Allocation
2. Caching
3. Similarity Caching
4. **Inference Delivery Networks**
5. Fairness in Dynamic Resource Allocation
6. Concluding Remarks

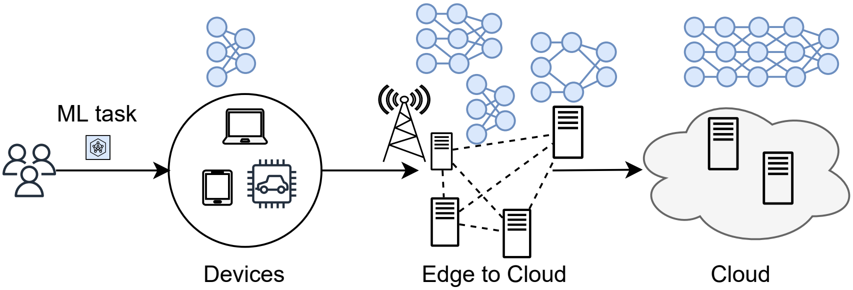
Inference Delivery Networks



Current ML deployment

Simpler models available locally have low accuracy. Complex models in the cloud may introduce high latency

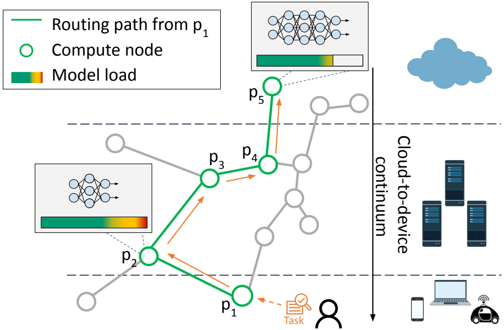
Inference Delivery Networks



IDNs

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

Inference Delivery Networks – Model



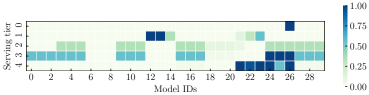
Differences with Vanilla Similarity Caching

- Models have serving capacity, and can saturate when their capacity is exceeded
- Distribute allocation decisions among computing nodes with limited information exchange

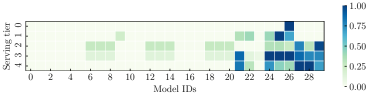
Inference Delivery Networks – Contributions

Contribution

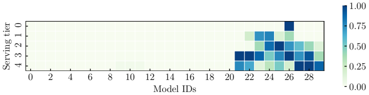
We propose a distributed online allocation algorithm for IDNs with a ψ-regret guarantee.



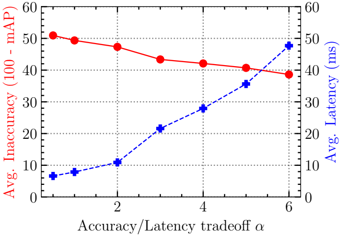
(a) $\alpha = 3$



(b) $\alpha = 4$



(c) $\alpha = 5$



Inference Delivery Networks – Research Output

[C5] **Towards Inference Delivery Networks: Distributing Machine Learning with Optimality Guarantees**

MedComNet, 2021
T. Si Salem et al.

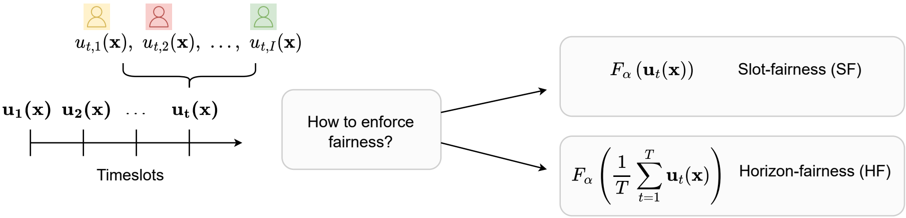
[S2] **Towards Inference Delivery Networks: Distributing Machine Learning with Optimality Guarantees**

IEEE/ACM ToN, 2022 (under review)
T. Si Salem et al.

Presentation Organization

1. Network Resource Allocation
2. Caching
3. Similarity Caching
4. Inference Delivery Networks
5. **Fairness in Dynamic Resource Allocation**
6. Concluding Remarks

Fairness in Dynamic Resource Allocation



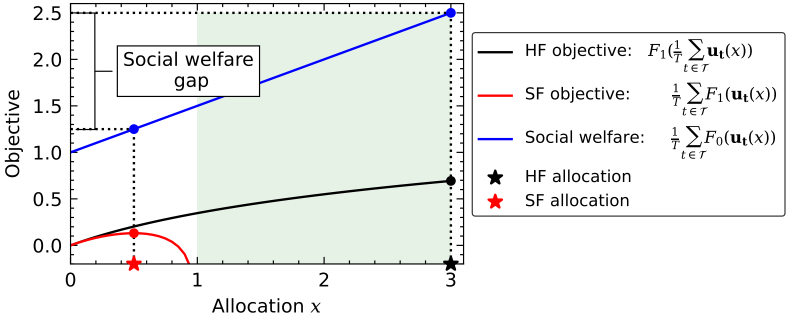
An α-fairness function $F_\alpha : \mathcal{U} \to \mathbb{R}$ is parameterized by the *inequality aversion parameter* $\alpha \in \mathbb{R}_{\geq 0}$, and it is given by

$$F_\alpha(\boldsymbol{u}) \triangleq \begin{cases} \sum_{i \in \mathcal{I}} \frac{u_i^{1-\alpha}-1}{1-\alpha}, & \text{for } \alpha \in \mathbb{R}_{\geq 0} \setminus \{1\}, \\ \sum_{i \in \mathcal{I}} \log(u_i), & \text{for } \alpha = 1, \end{cases}$$

Fairness in Dynamic Resource Allocation

Consider a system with two agents $\mathcal{I} = \{1, 2\}$, an allocation set $\mathcal{X} = [0, x_{\max}]$ with $x_{\max} > 1$, α-fairness criterion with $\alpha = 1$, even $T \in \mathbb{N}$, and the following sequence of utilities

$$\{\boldsymbol{u}_t(x)\}_{t=1}^{T} = \{(1 + x, 1 - x), (1 + x, 1 + x), \dots\}.$$



Price of Fairness under HF and SF objectives for $x_{\max} = 3$. The green shaded area provides the set of allocation unachievable by the SF objective but achievable by the HF objective.

Fairness in Dynamic Resource Allocation

We propose the *fairness regret* metric:

Definition

The long-term fairness regret of a policy $\mathcal{A}$ under α-fairness is defined as follows:

$$\mathfrak{R}_T\left(F_\alpha, \mathcal{A}\right) \triangleq \sup_{\{\boldsymbol{u}_t\}_{t=1}^T \in \mathcal{U}^T} \left\{ F_\alpha\left(\frac{1}{T}\sum_{t\in\mathcal{T}} \boldsymbol{u}_t(\boldsymbol{x}_\star)\right) - F_\alpha\left(\frac{1}{T}\sum_{t\in\mathcal{T}} \boldsymbol{u}_t(\boldsymbol{x}_t)\right)\right\}.$$

When $\lim_{T\to\infty} \mathfrak{R}_T\left(F_\alpha, \mathcal{A}\right) = 0$, policy $\mathcal{A}$ will attain the same fairness value as the static benchmark under any possible sequence of utility functions.

Fairness in Dynamic Resource Allocation

Impossibility Result

We prove that vanishing regret cannot be achieved in presence of an unrestricted adversary (as the one assumed in OCO).

We prove that *mild* restrictions on the adversary's capabilities make vanishing regret achievable. We provide an online policy that indeed guarantees vanishing regret under these restrictions.

Necessary Restrictions

These restrictions capture several practical utility patterns, such as non-stationary corruptions, ergodic and periodic inputs [LGK22, BLM22, ZLL⁺19, DAJJ12].

Fairness in Dynamic Resource Allocation - OHF Policy

Require:
$\mathcal{X},\ \alpha \in \mathbb{R}_{\geq 0},\ [u_{\star,\min}, u_{\star,\max}]$ ← No learning rate tuning

$\Theta \leftarrow \left[-1/u_{\star,\min}^{\alpha}, -1/u_{\star,\max}^{\alpha}\right]^{\mathcal{I}}$ ▷ Initialize the dual (conjugate) subspace

$\boldsymbol{x}_1 \in \mathcal{X};\ \boldsymbol{\theta}_1 \in \Theta;$ ▷ Initialize primal decision $\boldsymbol{x}_1$ and dual decision $\boldsymbol{\theta}_1$

for $t \in \mathcal{T}$ **do**

  Reveal $\Psi_{t,\alpha}(\boldsymbol{\theta}_t, \boldsymbol{x}_t) = (-F_\alpha)^\star(\boldsymbol{\theta}_t) - \boldsymbol{\theta}_t \cdot \boldsymbol{u}_t(\boldsymbol{x}_t)$ ▷ Incur reward $\Psi_{t,\alpha}(\boldsymbol{\theta}_t, \underline{\boldsymbol{x}_t})$ and loss $\Psi_{t,\alpha}(\underline{\boldsymbol{\theta}_t}, \boldsymbol{x}_t)$

  $\boldsymbol{g}_{\mathcal{X},t} \in \partial_{\boldsymbol{x}} \Psi_{t,\alpha}(\boldsymbol{\theta}_t, \boldsymbol{x}_t) = \sum_{i\in\mathcal{I}} \theta_{t,i} \partial_{\boldsymbol{x}} u_{t,i}$ ▷ Compute supergradient $\boldsymbol{g}_{\mathcal{X},t}$ at $\boldsymbol{x}_t$ of reward $\Psi_{t,\alpha}(\underline{\boldsymbol{\theta}_t}, \cdot)$

  $\boldsymbol{g}_{\Theta,t} = \nabla_{\boldsymbol{\theta}} \Psi_{t,\alpha}(\boldsymbol{\theta}_t, \boldsymbol{x}_t) = \left((-\theta_{t,i})^{-1/\alpha} - \boldsymbol{u}_t(\boldsymbol{x}_t)\right)_{i\in\mathcal{I}}$ ▷ Compute gradient $\boldsymbol{g}_{\Theta,t}$ at $\boldsymbol{\theta}_t$ of loss $\Psi_{t,\alpha}(\cdot, \boldsymbol{x}_t)$

  $\eta_{\mathcal{X},t} = \frac{\mathrm{diam}(\mathcal{X})}{\sqrt{\sum_{s=1}^{t} \left\lVert \boldsymbol{g}_{\mathcal{X},s} \right\rVert_2^2}};\ \eta_{\Theta,t} = \frac{\alpha u_{\min}^{-1-1/\alpha}}{t}$ ▷ Compute adaptive learning rates

  $\boldsymbol{x}_{t+1} = \Pi_{\mathcal{X}}\left(\boldsymbol{x}_t + \eta_{\mathcal{X},t} \boldsymbol{g}_{\mathcal{X},t}\right);\ \boldsymbol{\theta}_{t+1} = \Pi_{\Theta}\left(\boldsymbol{\theta}_t - \eta_{\Theta,t} \boldsymbol{g}_{\Theta,t}\right)$ ▷ Compute a new allocation through OGA and a new dual decision through OGD

end for

Fairness in Dynamic Resource Allocation - OHF Policy

Require:
$\mathcal{X}$, $\alpha \in \mathbb{R}_{\geq 0}$, $[u_{\star,\text{min}}, u_{\star,\text{max}}]$

$\Theta \leftarrow \left[-1/u^{\alpha}_{\star,\text{min}}, -1/u^{\alpha}_{\star,\text{max}}\right]^{\mathcal{I}}$ ▷ Initialize the dual (conjugate) subspace

$\boldsymbol{x}_1 \in \mathcal{X}$; $\boldsymbol{\theta}_1 \in \Theta$; ▷ Initialize primal decision $\boldsymbol{x}_1$ and dual decision $\boldsymbol{\theta}_1$

for $t \in \mathcal{T}$ **do**

Reveal $\Psi_{t,\alpha}(\boldsymbol{\theta}_t, \boldsymbol{x}_t) = (-F_\alpha)^\star(\boldsymbol{\theta}_t) - \boldsymbol{\theta}_t \cdot \boldsymbol{u}_t(\boldsymbol{x}_t)$ ▷ Incur reward $\Psi_{t,\alpha}(\boldsymbol{\theta}_t, \underline{\boldsymbol{x}_t})$ and loss $\Psi_{t,\alpha}(\underline{\boldsymbol{\theta}_t}, \boldsymbol{x}_t)$

$\boldsymbol{g}_{\mathcal{X},t} \in \partial_{\boldsymbol{x}} \Psi_{t,\alpha}(\boldsymbol{\theta}_t, \boldsymbol{x}_t) = \sum_{i \in \mathcal{I}} \theta_{t,i} \partial_{\boldsymbol{x}} u_{t,i}$ ▷ Compute supergradient $\boldsymbol{g}_{\mathcal{X},t}$ at $\boldsymbol{x}_t$ of reward $\Psi_{t,\alpha}(\boldsymbol{\theta}_t, \cdot)$

$\boldsymbol{g}_{\Theta,t} = \nabla_{\boldsymbol{\theta}} \Psi_{t,\alpha}(\boldsymbol{\theta}_t, \boldsymbol{x}_t) = \left((-\theta_{t,i})^{-1/\alpha} - \boldsymbol{u}_t(\boldsymbol{x}_t)\right)_{i \in \mathcal{I}}$ ▷ Compute gradient $\boldsymbol{g}_{\Theta,t}$ at $\boldsymbol{\theta}_t$ of loss $\Psi_{t,\alpha}(\cdot, \boldsymbol{x}_t)$

$\eta_{\mathcal{X},t} = \frac{\operatorname{diam}(\mathcal{X})}{\sqrt{\sum_{s=1}^{t} \left\|\boldsymbol{g}_{\mathcal{X},s}\right\|_2^2}}$; $\eta_{\Theta,t} = \frac{\alpha u_{\text{min}}^{-1-1/\alpha}}{t}$ ▷ Compute adaptive learning rates

$\boldsymbol{x}_{t+1} = \Pi_{\mathcal{X}}\left(\boldsymbol{x}_t + \eta_{\mathcal{X},t} \boldsymbol{g}_{\mathcal{X},t}\right)$; $\boldsymbol{\theta}_{t+1} = \Pi_{\Theta}\left(\boldsymbol{\theta}_t - \eta_{\Theta,t} \boldsymbol{g}_{\Theta,t}\right)$ ▷ Compute a new allocation through OGA and a new dual decision through OGD

end for

Formulate an Online Saddle Point problem

Fairness in Dynamic Resource Allocation - OHF Policy

Require:
$\mathcal{X}$, $\alpha \in \mathbb{R}_{\geq 0}$, $[u_{\star,\min}, u_{\star,\max}]$

$\Theta \leftarrow \left[-1/u_{\star,\min}^{\alpha}, -1/u_{\star,\max}^{\alpha}\right]^{\mathcal{I}}$ ▷ Initialize the dual (conjugate) subspace

$\mathbf{x}_1 \in \mathcal{X}$; $\boldsymbol{\theta}_1 \in \Theta$; ▷ Initialize primal decision $\mathbf{x}_1$ and dual decision $\boldsymbol{\theta}_1$

for $t \in \mathcal{T}$ **do**

  Reveal $\Psi_{t,\alpha}(\boldsymbol{\theta}_t, \mathbf{x}_t) = (-F_\alpha)^\star(\boldsymbol{\theta}_t) - \boldsymbol{\theta}_t \cdot \mathbf{u}_t(\mathbf{x}_t)$ ▷ Incur reward $\Psi_{t,\alpha}(\boldsymbol{\theta}_t, \underline{\mathbf{x}_t})$ and loss $\Psi_{t,\alpha}(\underline{\boldsymbol{\theta}_t}, \mathbf{x}_t)$

  $\mathbf{g}_{\mathcal{X},t} \in \partial_{\mathbf{x}} \Psi_{t,\alpha}(\boldsymbol{\theta}_t, \mathbf{x}_t) = \sum_{i \in \mathcal{I}} \theta_{t,i} \partial_{\mathbf{x}} u_{t,i}$ ▷ Compute supergradient $\mathbf{g}_{\mathcal{X},t}$ at $\mathbf{x}_t$ of reward $\Psi_{t,\alpha}(\boldsymbol{\theta}_t, \cdot)$

  $\mathbf{g}_{\Theta,t} = \nabla_{\boldsymbol{\theta}} \Psi_{t,\alpha}(\boldsymbol{\theta}_t, \mathbf{x}_t) = \left((-\theta_{t,i})^{-1/\alpha} - \mathbf{u}_t(\mathbf{x}_t)\right)_{i \in \mathcal{I}}$ ▷ Compute gradient $\mathbf{g}_{\Theta,t}$ at $\boldsymbol{\theta}_t$ of loss $\Psi_{t,\alpha}(\cdot, \mathbf{x}_t)$

  $\eta_{\mathcal{X},t} = \frac{\operatorname{diam}(\mathcal{X})}{\sqrt{\sum_{s=1}^{t} \left\|\mathbf{g}_{\mathcal{X},s}\right\|_2^2}}$; $\eta_{\Theta,t} = \frac{\alpha u_{\min}^{-1-1/\alpha}}{t}$ ▷ Compute adaptive learning rates

  $\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}\left(\mathbf{x}_t + \eta_{\mathcal{X},t} \mathbf{g}_{\mathcal{X},t}\right)$; $\boldsymbol{\theta}_{t+1} = \Pi_{\Theta}\left(\boldsymbol{\theta}_t - \eta_{\Theta,t} \mathbf{g}_{\Theta,t}\right)$ ▷ Compute a new allocation through OGA and a new dual decision through OGD

end for

Adapt allocation

Fairness in Dynamic Resource Allocation - OHF Policy

Require:

$\mathcal{X}$, $\alpha \in \mathbb{R}_{\geq 0}$, $[u_{\star,\text{min}}, u_{\star,\text{max}}]$

$\Theta \leftarrow \left[-1/u_{\star,\text{min}}^{\alpha}, -1/u_{\star,\text{max}}^{\alpha}\right]^{\mathcal{I}}$ ▷ Initialize the dual (conjugate) subspace

$\mathbf{x}_1 \in \mathcal{X}$; $\boldsymbol{\theta}_1 \in \Theta$; ▷ Initialize primal decision $\mathbf{x}_1$ and dual decision $\boldsymbol{\theta}_1$

for $t \in \mathcal{T}$ **do**

Reveal $\Psi_{t,\alpha}(\boldsymbol{\theta}_t, \mathbf{x}_t) = (-F_\alpha)^\star(\boldsymbol{\theta}_t) - \boldsymbol{\theta}_t \cdot \mathbf{u}_t(\mathbf{x}_t)$ ▷ Incur reward $\Psi_{t,\alpha}(\boldsymbol{\theta}_t, \underline{\mathbf{x}_t})$ and loss $\Psi_{t,\alpha}(\underline{\boldsymbol{\theta}_t}, \mathbf{x}_t)$

$\mathbf{g}_{\mathcal{X},t} \in \partial_{\mathbf{x}} \Psi_{t,\alpha}(\boldsymbol{\theta}_t, \mathbf{x}_t) = \sum_{i \in \mathcal{I}} \theta_{t,i} \partial_{\mathbf{x}} u_{t,i}$ ▷ Compute supergradient $\mathbf{g}_{\mathcal{X},t}$ at $\underline{\mathbf{x}_t}$ of reward $\Psi_{t,\alpha}(\underline{\boldsymbol{\theta}_t}, \cdot)$

$\mathbf{g}_{\Theta,t} = \nabla_{\boldsymbol{\theta}} \Psi_{t,\alpha}(\boldsymbol{\theta}_t, \mathbf{x}_t) = \left((-\theta_{t,i})^{-1/\alpha} - \mathbf{u}_t(\mathbf{x}_t)\right)_{i \in \mathcal{I}}$ ▷ Compute gradient $\mathbf{g}_{\Theta,t}$ at $\boldsymbol{\theta}_t$ of loss $\Psi_{t,\alpha}(\cdot, \mathbf{x}_t)$

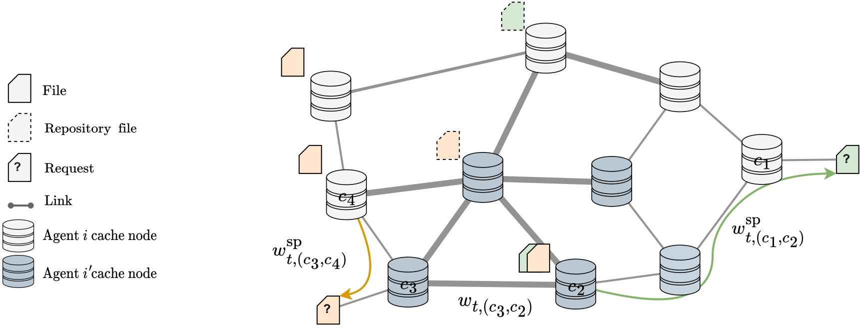
$\eta_{\mathcal{X},t} = \dfrac{\operatorname{diam}(\mathcal{X})}{\sqrt{\sum_{s=1}^{t} \|\mathbf{g}_{\mathcal{X},s}\|_2^2}}$; $\eta_{\Theta,t} = \dfrac{\alpha u_{\min}^{-1-1/\alpha}}{t}$ ▷ Compute adaptive learning rates

$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}(\mathbf{x}_t + \eta_{\mathcal{X},t}\mathbf{g}_{\mathcal{X},t})$; $\boldsymbol{\theta}_{t+1} = \Pi_{\Theta}(\boldsymbol{\theta}_t - \eta_{\Theta,t}\mathbf{g}_{\Theta,t})$ ▷ Compute a new allocation through OGA and a new dual decision through OGD

Correct fairness "weights"

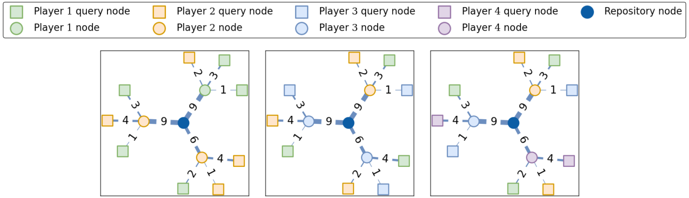
end for

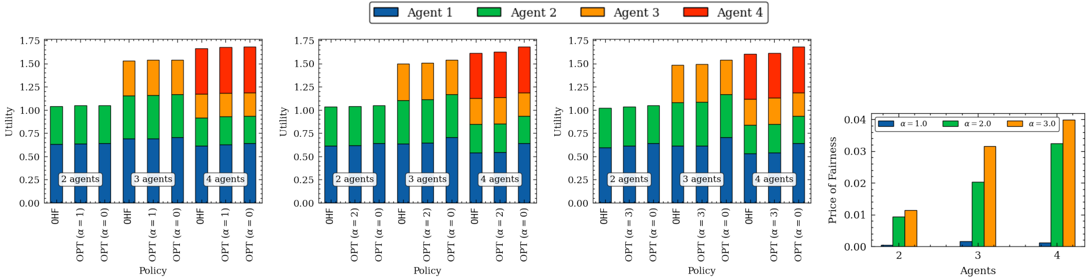
Fairness in Dynamic Resource Allocation



An application: a network comprised of a set of caching nodes $\mathcal{C}$. A request arrives at a cache node $c \in \mathcal{C}$, it can be partially served locally, and if needed, forwarded along the shortest retrieval path to another node to retrieve the remaining part of the file.

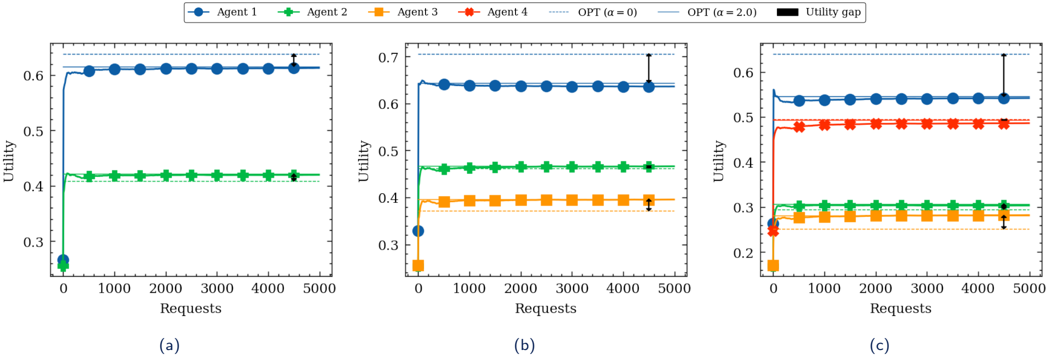
Fairness in Dynamic Resource Allocation – Some Results





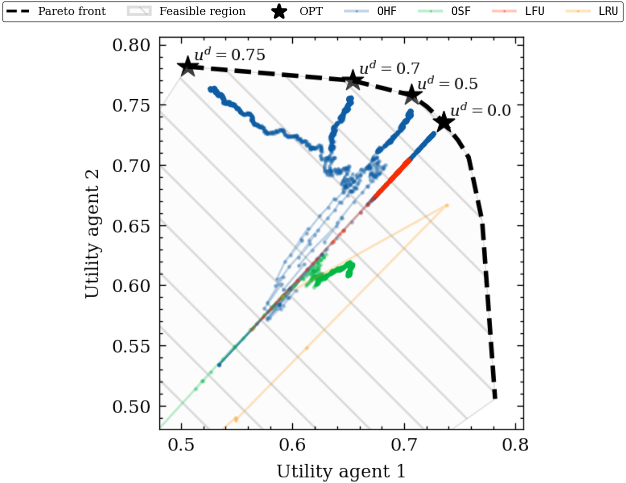
(a) $\alpha = 1$ (b) $\alpha = 2$ (c) $\alpha = 3$ (d) Price of Fairness

Fairness in Dynamic Resource Allocation – Some Results



The time-averaged utility across different agents obtained by OHF policy and OPT for $\alpha = 2$ under an increasing number of agents in $\{2, 3, 4\}$ and $\textsc{Tree}$ 1–3 network topology.

Fairness in Dynamic Resource Allocation – Some Results



Fairness in Dynamic Resource Allocation – Research Output

Enabling Long-term Fairness in Dynamic Resource Allocation
[C6]
ACM SIGMETRICS, 2023
T. Si Salem, G. Iosifidis, and G. Neglia

During 5 months visit at **TU Delft, Netherlands**

Enabling Long-term Fairness in Dynamic Resource Allocation
[J4]
ACM POMACS, 2023
T. Si Salem, G. Iosifidis, and G. Neglia.

Presentation Organization

1. Network Resource Allocation
2. Caching
3. Similarity Caching
4. Inference Delivery Networks
5. Fairness in Dynamic Resource Allocation
6. **Concluding Remarks**

Concluding Remarks

- We demonstrated the versatility of gradient algorithms on inherently combinatorial problems when paired with an opportune randomized rounding scheme.
- Our extensive experimental findings support the thesis that these algorithms are robust and can adapt to changing external system's parameters.
- We proposed a novel long-term online fairness framework for settings where the agents' utilities are subject to unknown, time-varying, and potentially adversarial perturbations.

Potential Future Work

- Investigate dimensionality reduction techniques to diminish the operational complexity of online learning algorithms.
- Bridge the horizon-fairness and slot-fairness criteria to target applications where the agents are interested in ensuring fairness within a target time window.
- Add support for coalition formation in our fairness framework.
- Consider a limited feedback scenario where only part of the utility is revealed to the agents.

Thank you for your attention.

References I

- Rajarshi Bhattacharjee, Subhankar Banerjee, and Abhishek Sinha, *Fundamental Limits on the Regret of Online Network-Caching*, Proceedings of the ACM on Measurement and Analysis of Computing Systems **4** (2020), no. 2.
- Santiago R Balseiro, Haihao Lu, and Vahab Mirrokni, *The Best of Many Worlds: Dual Mirror Descent for Online Allocation Problems*, Operations Research (2022).
- Sébastien Bubeck, *Convex Optimization: Algorithms and Complexity*, Foundations and Trends in Machine Learning **8** (2015), no. 3–4, 231–357.
- M. Costantini and T. Spyropoulos, *Impact of Popular Content Relational Structure on Joint Caching and Recommendation Policies*, 2020 18th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOPT), 2020, pp. 1–8.

References II

- Daniel Crankshaw, Xin Wang, Guilio Zhou, Michael J. Franklin, Joseph E. Gonzalez, and Ion Stoica, *Clipper: A Low-Latency Online Prediction Serving System*, 14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 17), 2017, pp. 613–627.
- John C Duchi, Alekh Agarwal, Mikael Johansson, and Michael I Jordan, *Ergodic Mirror Descent*, SIAM Journal on Optimization **22** (2012), no. 4, 1549–1578.
- Utsav Drolia, Katherine Guo, and Priya Narasimhan, *Precog: Prefetching for Image Recognition Applications at the Edge*, Proceedings of the Second ACM/IEEE Symposium on Edge Computing, SEC '17, 2017.
- Utsav Drolia, Katherine Guo, Jiaqi Tan, Rajeev Gandhi, and Priya Narasimhan, *Cachier: Edge-Caching for Recognition Applications*, 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS), 2017, pp. 276–286.

References III

- Fabrizio Falchi, Claudio Lucchese, Salvatore Orlando, Raffaele Perego, and Fausto Rabitti, *A Metric Cache for Similarity Search*, Proceedings of the 2008 ACM workshop on Large-Scale distributed systems for information retrieval, 2008, pp. 43–50.
- Elad Hazan, *Introduction to Online Convex Optimization*, Foundations and Trends® in Optimization **2** (2016), no. 3–4, 157–325.
- A. Kumar, A. Balasubramanian, S. Venkataraman, and A. Akella, *Accelerating deep learning inference via freezing*, 11th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 19), 2019.
- Luofeng Liao, Yuan Gao, and Christian Kroer, *Nonstationary Dual Averaging and Online Fair Allocation*, ArXiv e-prints (2022).
- William G Madow and Lillian H Madow, *On the Theory of Systematic Sampling*, Ann. Math. Statist. **15** (1944), no. 4, 1–24.

References IV

- Samrat Mukhopadhyay and Abhishek Sinha, *Online Caching with Optimal Switching Regret*, 2021 IEEE International Symposium on Information Theory (ISIT), 2021, pp. 1546–1551.
- Sandeep Pandey, Andrei Broder, Flavio Chierichetti, Vanja Josifovski, Ravi Kumar, and Sergei Vassilvitskii, *Nearest-Neighbor Caching for Content-Match Applications*, Proceedings of the 18th International Conference on World Wide Web, WWW '09, 2009, pp. 441–450.
- Georgios S Paschos, Apostolos Destounis, Luigi Vigneri, and George Iosifidis, *Learning to Cache With No Regrets*, IEEE INFOCOM 2019-IEEE Conference on Computer Communications, 2019, pp. 235–243.
- Debjit Paria and Abhishek Sinha, *LeadCache: Regret-Optimal Caching in Networks*, Advances in Neural Information Processing Systems **34** (2021), 4435–4447.

References V

- P. Sermpezis, T. Giannakas, T. Spyropoulos, and L. Vigneri, *Soft Cache Hits: Improving Performance Through Recommendation and Delivery of Related Content*, IEEE Journal on Selected Areas in Communications **36** (2018), no. 6, 1300–1313.
- Thrasyvoulos Spyropoulos and Pavlos Sermpezis, *Soft Cache Hits and the Impact of Alternative Content Recommendations on Mobile Edge Caching*, Proceedings of the Eleventh ACM Workshop on Challenged Networks, CHANTS '16, 2016, pp. 51–56.
- Srikumar Venugopal, Michele Gazzetti, Yiannis Gkoufas, and Kostas Katrinis, *Shadow Puppets: Cloud-level Accurate AI Inference at the Speed and Economy of Edge*, USENIX Workshop on Hot Topics in Edge Computing (HotEdge 18), July 2018.
- Yu-Hang Zhou, Chen Liang, Nan Li, Cheng Yang, Shenghuo Zhu, and Rong Jin, *Robust online matching with user arrival distribution drift*, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 459–466.