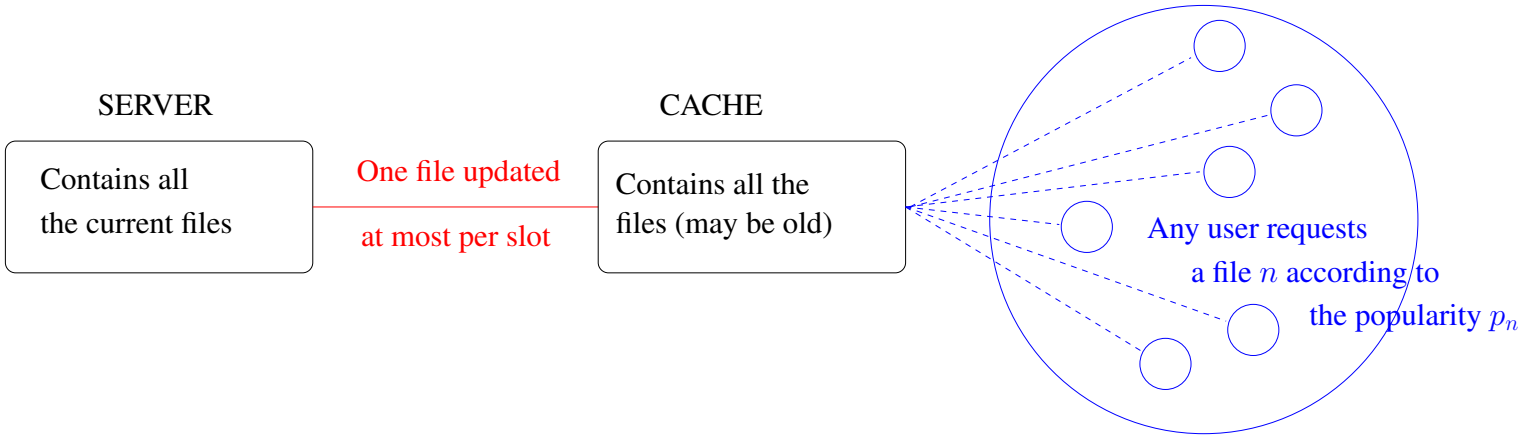
Age-Optimal Constrained Cache Updating

Philippe Ciblat

Joint work with Michèle Wigger, Roy Yates, and Aylin Yener

Fundings: ERC Starting Grant "CTOCom" and NSF award "CIF-1422988"

Problem statement



- When a user requests an item, the cache sends its local version.
- Issue : this version can be outdated since each item is time-varying, and the capacity-constrained server-cache link does not enable us to provide the latest version.

Question

- How should items be downloaded from the server in order to be as up-to-date as possible ?
- Equivalently, how should the server push updated versions to the cache s.t. users receive the most recent versions they request ?

Applications

- Data-Base Context
  - Content items = records in a database (server),
  - Cache is the local copy.
- Cloud RAN
  - Server = BBU in Cloud RAN or standard BTS.
  - Cache = small-cell base station (RRH in Cloud-RAN) delivering popular content to nearby mobile users.
  - Server-Cache link is rate-limited whereas cache-user links are short-range and ultra-high data rate.
- Satellite based Broadcasting System
  - Server = satellite broadcasting the same content updates to thousands of caches.
  - Cache = local storage in a TV/video news distribution

Mathematical model

- Set $\mathcal{N}$ of N timestamped items at the cache.
- Item n is requested from the cache with probability $p_n > 0$, independent of all other requests.

$$\mathbf{p} = [p_1, \cdots, p_N] \quad \text{(popularity vector)}$$

- $X_n(t)$ is the age of item n at time/slot t at the cache.
- If item n is updated from the server in slot t_0, it is available for downloading from the cache at slot t_0 too. And $X_n(t_0) = 1$.
- In slot t, a *single* file u_t may be updated from the server (u_t= 0 if no updated file). Let T be the number of slots.

$$\mathbf{u} = [u_1, \cdots, u_T] \quad \text{(update vector)}$$

- Cache is able to download K items from the server within T slots

$$\lambda = \lim_{T \to \infty} \frac{K}{T} \quad \text{(update rate)}$$

Optimization problem

The average age of a randomly requested item from the cache is

$$\overline{X}(\mathbf{u}) = \sum_{n=1}^{N} p_n \overline{X}_n(\mathbf{u}) \text{ with } \overline{X}_n(\mathbf{u}) = \frac{1}{T} \int_0^T X_n(t)\, dt$$

since a request for item n is uniformly-distributed over $[0, T]$.

Optimization problem : update scheduling

$$\overline{X}^*(K, T) = \min_{\mathbf{u}} \overline{X}(\mathbf{u})$$

s.t.

- $u_t \in \mathcal{N}$ for all t,
- $\sum_{t=1}^{T} \mathbf{1}\{u_t > 0\} = K$.

Additional implicit constraints :

C1 : Updates occur only in discrete time slots.

C2 : Updates are collision-free ; one item at most is updated in a slot.

Optimization problem

The average age of a randomly requested item from the cache is

$$\overline{X}(\mathbf{u}) = \sum_{n=1}^{N} p_n \overline{X}_n(\mathbf{u}) \text{ with } \overline{X}_n(\mathbf{u}) = \frac{1}{T} \int_0^T X_n(t)\, dt$$

since a request for item n is uniformly-distributed over $[0, T]$.

Optimization problem : update scheduling

$$\overline{X}^*(K, T) = \min_{\mathbf{u}} \overline{X}(\mathbf{u})$$

s.t.

- $u_t \in \mathcal{N}$ for all t,
- $\sum_{t=1}^{T} \mathbf{1}\{u_t > 0\} = K$.

Intractable combinatorial optimization problem

Additional implicit constraints :

C1 : Updates occur only in discrete time slots.

C2 : Updates are collision-free ; one item at most is updated in a slot.

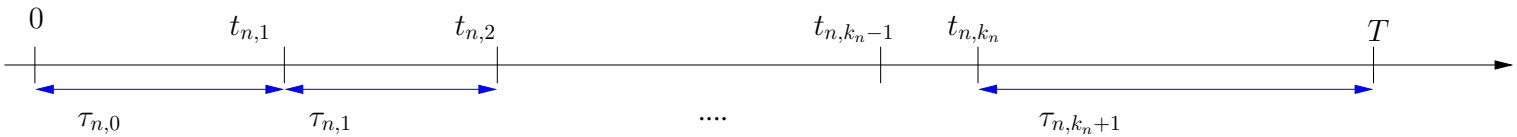
Problem Resolution : three steps

- Step 1 : we relax constraints C1 and C2 and rewrite the relaxed optimization problem in terms of the inter-update times $\{\tau_{n,i}\}$.
- Step 2 : we assume that k_n updates are allocated to item n during T slots, and find the inter-update times $\{\tau_{n,i}\}$ minimizing the age $\overline{X}_n(\mathbf{u})$.
- Step 3 : we optimize $k_1, \ldots, k_N$ in an asymptotic regime with large T ; integer optimization of k_n is replaced by continuous optimization of the asymptotic update rate $\lambda_n = \lim_{T\to\infty} k_n/T$, and

$$\lambda = \sum_{n=1}^{N} \lambda_n.$$

Step 1

Let $\tau_{n,i}$ the i-th inter update time and $t_{n,i}$ the update time ($u_{t_{n,i}} = n$).



Let $\boldsymbol{\tau}_n = [\tau_{n,1}, \cdots, \tau_{n,k_n+1}]$. We have

$$\overline{X}_n(\mathbf{u}) = \overline{X}_n(\boldsymbol{\tau}_n) = \frac{1}{2T} \sum_{i=1}^{k_n+1} \tau_{n,i}^2 + 1.$$

Relaxed optimization problem

$$\min_{\boldsymbol{\tau}_1,\ldots,\boldsymbol{\tau}_N} \sum_{n=1}^{N} p_n \overline{X}_n(\boldsymbol{\tau}_n)$$

s.t.

- $\tau_{n,i} \geq 0$, for all n, i,
- $\tau_{n,1} + \cdots + \tau_{n,k_n+1} = T$, for all n,
- $\sum_{n=1}^{N} k_n = K$ with $k_n \in \mathbb{N}$.

Step 2

Given k_n,

- the optimization problem becomes separable.
- the optimization problem for item n is convex with solution

$$\tau^*_{n,i} = \frac{T}{k_n + 1}, \qquad i = 1, \ldots, k_n + 1.$$

Identical deterministic inter-update times is an optimal policy

Next Optimization problem

$$\min_{k_1,\ldots,k_N} \sum_{n=1}^{N} p_n \overline{X}_n(\tau^\star_n)$$

or equivalently

$$\min_{k_1,\ldots,k_N} \sum_{n=1}^{N} \frac{p_n}{k_n + 1}$$

s.t.

- $\sum_{n=1}^{N} k_n = K$ with $k_n \in \mathbb{N}$.

Step 3

- Previous problem hard to solve as the integer assumption on k_n.
- To overcome this issue, **asymptotic approach**, i.e., large T.
- Replacing k_n with the update rate $\lambda_n \approx k_n/T \in \mathbb{R}^+$
- Actually,

$$\lambda_n = \lim_{T\to\infty} \frac{k_n}{T} \left(= \lim_{T\to\infty} \frac{k_n + 1}{T} \right).$$

New optimization problem

$$\min_{\lambda_1,\ldots,\lambda_N} \sum_{n=1}^{N} \frac{p_n}{\lambda_n}$$

s.t.

- $\lambda_n \geq 0$,
- $\lambda_1 + \cdots + \lambda_N = \lambda$.

Final results

Main result 1

Problem is convex and leads to

$$\lambda_n^* = \frac{\lambda\sqrt{p_n}}{\sum_{i=1}^{N}\sqrt{p_i}}$$

Update rate of item n follows a square-root law wrt. its popularity

Main result 2

The minimum asymptotic relaxed average age is

$$\overline{X}_{a.r.}^*(\mathbf{p},\lambda) = \frac{\Delta^*(\mathbf{p})}{\lambda} + 1,$$

with

$$\Delta^*(\mathbf{p}) = \frac{1}{2}\left(\sum_{i=1}^{N}\sqrt{p_i}\right)^2.$$

Practical protocol

- Issue 1 : updates occur in discrete-time.
  - Update item n with randomly quantized independent inter-update times $\mathcal{T}_{n,i}$ defined as follows
  - Let $\overline{\tau}_{n,i} = \lceil \tau^{\star}_{n,i} \rceil$ and $q_{n,i} = \overline{\tau}_{n,i} - \tau^{\star}_{n,i}$.

$$P_{\mathcal{T}_{n,i}}(z) = \begin{cases} q_{n,i} & z = \overline{\tau}_{n,i} - 1, \\ 1 - q_{n,i} & z = \overline{\tau}_{n,i}. \end{cases}$$

Results

$$\mathbb{E}[\mathcal{T}_{n,i}] = \tau^{\star}_{n,i}$$

An extra $\lambda/8$ of a time slot has to be added on the average age.

- Issue 2 : updates collisions have been ignored.

  If an update is scheduled in slot t having a backlog of b queued updates, then that update is postponed to slot $t + b$. But in slot $t + b$, the server sends the current version $t + b$.

Result

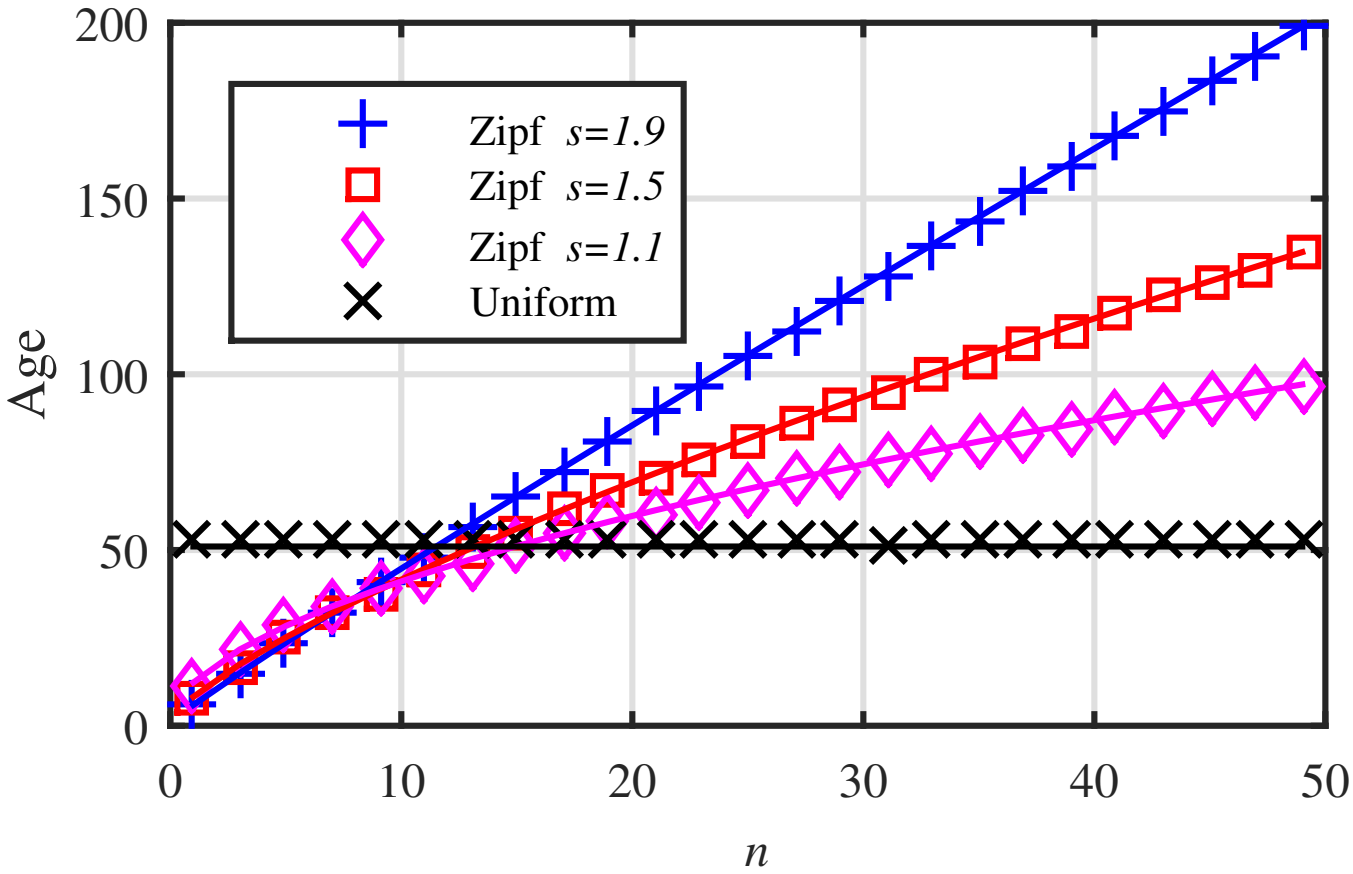
This adds some additional randomness to the inter-update times.

Simulations

Except otherwise stated,

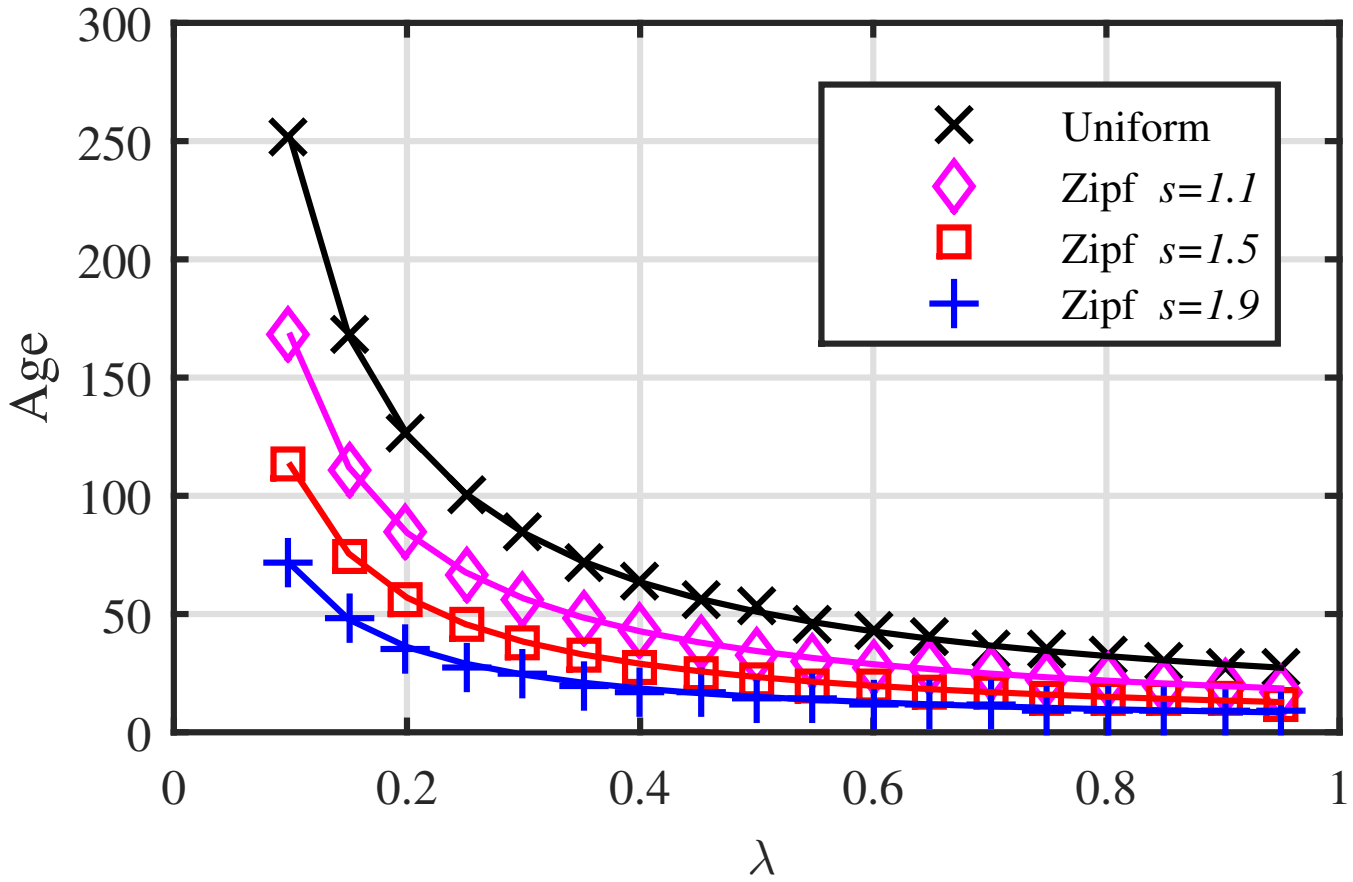
- $N = 50$ items,
- $\lambda = 0.5$,
- Solid lines represent the closed-form asymptotic relaxed expressions,
- Markers correspond to the evaluations through the practical protocol.

Average age for each item n



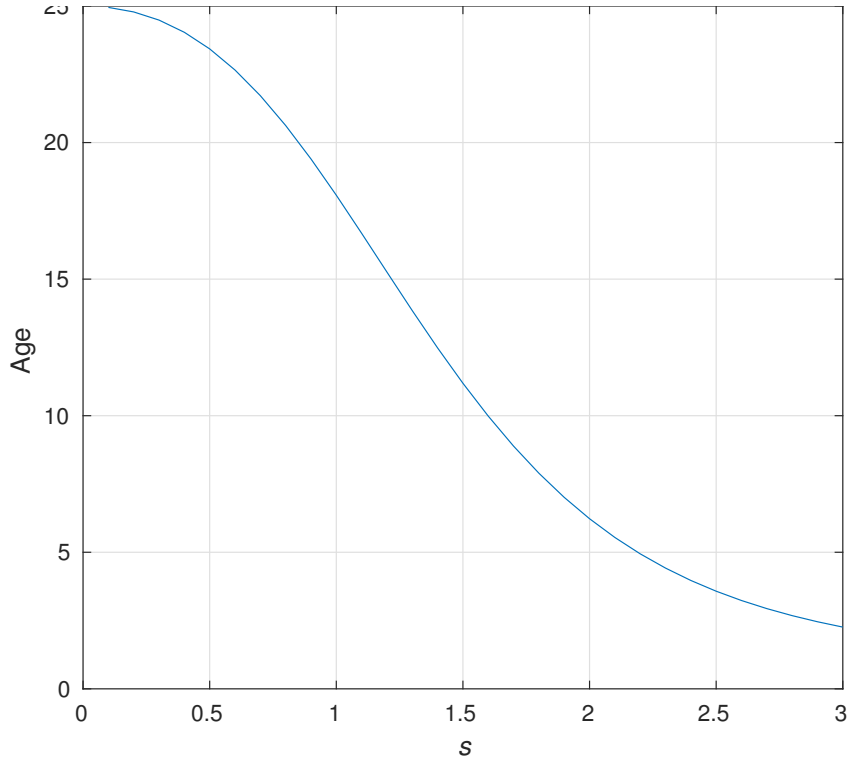
- Proposed policy reduces the average age of more popular items at the expense of less popular items,
- Effects of quantization and the scheduling queue are negligible for all items, independent of an item's popularity.

Average age vs. λ



- Once again, experimental protocol essentially matches the unconstained performance at all arrival rates λ,
- When Zipf parameter s increases, update rates optimization exploits the concentration in the popularities better.

Normalized average age $\Delta^*(\mathbf{p})$ vs. s



- When s incresaes, the average age decreases since the popularity concentrates on fewer but more popular items,
- When $s \to 0$ and Zipf approaches Uniform, the normalized average age approaches that of uniform popularity ($N/2$).

Conclusion and Perspectives

- Updating rates of the items in the local cache proportional to the square-root of items' popularity,
- Tractable solution to the optimal update schedule by using an asymptotic regime,
- Practical policy/protocol very close to the asymptotic one.
- Future research : item popularities evolving in time and/or are dependent on previous requests.