This work was supported in part by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 690893.

Proactive Wireless Content Caching

Deniz Gündüz
Imperial College London

11 January 2018
Communications & Signal Processing for 5G++ Workshop
Durham University

We gratefully acknowledge generous support received from the European Commission through BEACON, SCAVENGE and TactileNet projects.

www.imperial.ac.uk/ipc-lab twitter.com/Imperial_IPCL

erc European Research Council

HORIZON 2020

EPSRC Engineering and Physical Sciences Research Council

BRITISH COUNCIL

TOSHIBA Leading Innovation

Mobile Data Growth: Trends and Characteristics

- Video demand dominates traffic (78% by 2021)
- 75% of Facebook video browsing, 40% of Netflix downloads performed on smartphones
- We need a content aware network design

- Asymmetric resource usage
- Delay-tolerant, asynchronous access
- Most traffic due to a few viral/ popular video files
- Demand and access patterns highly predictable

Storage is relatively cheap, while bandwidth is extremely expensive!

- Video demand dominates traffic (78% by 2021)
- 75% of Facebook video browsing, 40% of Netflix downloads performed on smartphones
- We need a content aware network design
- Asymmetric resource usage
- Delay-tolerant, asynchronous access
- Most traffic due to a few viral/ popular video files
- Demand and access patterns highly predictable

Storage is relatively cheap, while bandwidth is extremely expensive!

Mobile Data Growth: Trends and Characteristics

- Video demand dominates traffic (78% by 2021)
- 75% of Facebook video browsing, 40% of Netflix downloads performed on smartphones
- We need a content aware network design

- Asymmetric resource usage
- Delay-tolerant, asynchronous access
- Most traffic due to a few viral/ popular video files
- Demand and access patterns highly predictable

Storage is relatively cheap, while bandwidth is extremely expensive!

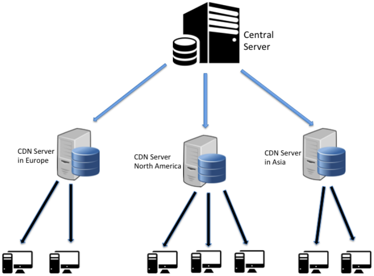
Mobile Data Growth: Trends and Characteristics

- Video demand dominates traffic (78% by 2021)
- 75% of Facebook video browsing, 40% of Netflix downloads performed on smartphones
- We need a content aware network design

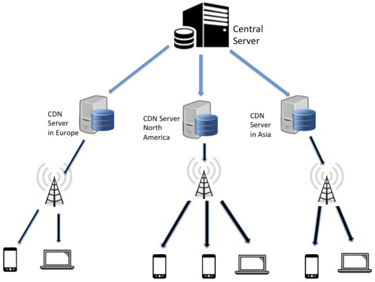
- Asymmetric resource usage
- Delay-tolerant, asynchronous access
- Most traffic due to a few viral/ popular video files
- Demand and access patterns highly predictable

Storage is relatively cheap, while bandwidth is extremely expensive!

Content Distribution Networks



- Content provider (e.g. Netflix, BBC, Facebook) contracts with a CDN (e.g. Akamai, LimeLight)
- Balance traffic, reduce latency, ...
- This is in the core network



- Content provider (e.g. Netflix, BBC, Facebook) contracts with a CDN (e.g. Akamai, LimeLight)
- Balance traffic, reduce latency, ...
- This is in the core network
- Bring content to the edge (e.g., Netflix Open Connect)

- Two-phase protocol:
  - **Placement phase**: off-peak hours, user demands unknown
  - **Delivery phase**: peak hours, demands revealed
- Library of N files, each consisting of F bits
- K users, each equipped with a cache of size M
- Each user requests one file
- **Error-free shared delivery link**: Satisfy all demands simultaneously

- **What is the minimum number of bits that must be delivered sufficient to satisfy all demand combinations?**
- **What is the trade-off between cache capacity and number of delivered bits?**

M. A. Maddah-Ali and U. Niesen, **Fundamental limits of caching**, IEEE Trans. Inform. Theory, vol. 60, no. 5, pp. 2856–2867, May 2014.

Coded Proactive Content Caching

- Two-phase protocol:
  - **Placement phase**: off-peak hours, user demands unknown
  - **Delivery phase**: peak hours, demands revealed
- Library of N files, each consisting of F bits
- K users, each equipped with a cache of size M
- Each user requests one file
- **Error-free shared delivery link**: Satisfy all demands simultaneously

- **What is the minimum number of bits that must be delivered sufficient to satisfy all demand combinations?**
- **What is the trade-off between cache capacity and number of delivered bits?**

M. A. Maddah-Ali and U. Niesen, **Fundamental limits of caching**, IEEE Trans. Inform. Theory, vol. 60, no. 5, pp. 2856–2867, May 2014.

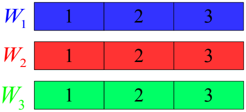
Coded Proactive Content Caching

- Two-phase protocol:
  - **Placement phase**: off-peak hours, user demands unknown
  - **Delivery phase**: peak hours, demands revealed
- Library of N files, each consisting of F bits
- K users, each equipped with a cache of size M
- Each user requests one file
- **Error-free shared delivery link**: Satisfy all demands simultaneously

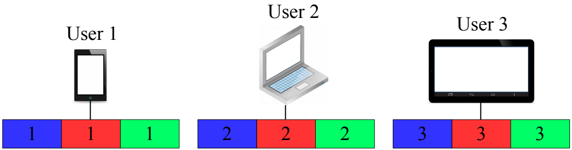
- **What is the minimum number of bits that must be delivered sufficient to satisfy all demand combinations?**
- **What is the trade-off between cache capacity and number of delivered bits?**

M. A. Maddah-Ali and U. Niesen, **Fundamental limits of caching**, IEEE Trans. Inform. Theory, vol. 60, no. 5, pp. 2856–2867, May 2014.

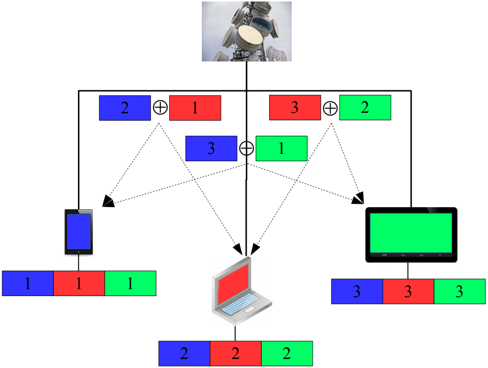
- $N = 3$ files
- $K = 3$ users
- Cache capacity: $M = 1$
- Split each file into 3 non-overlapping equal-size subfiles:



- Cache contents after placement phase:



- Delivery phase:



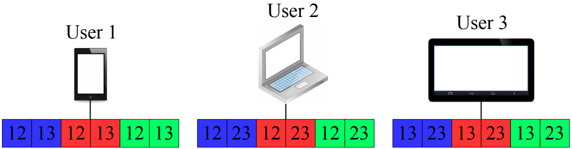
- Delivery rate: $R_{\mathrm{MAN}}(1) = 1$

Example 2

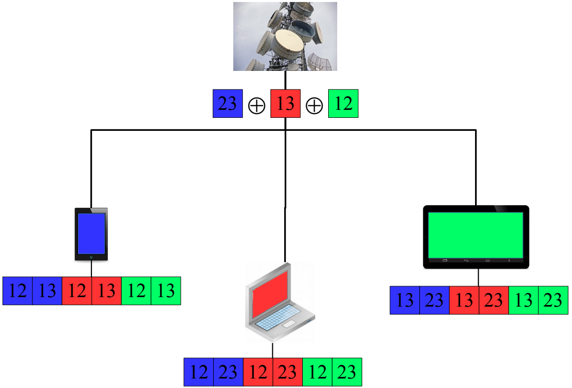
- $N = 3$ files
- $K = 3$ users
- Cache capacity: $M = 2$
- Split each file into 3 non-overlapping equal-size subfiles:

W_1	12	13	23
W_2	12	13	23
W_3	12	13	23

- Cache contents after placement phase:

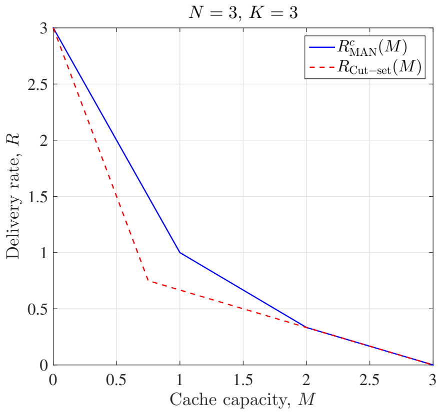


- Delivery phase:



- $R_{\text{MAN}}(2) = 1/3$

Delivery Rate-Cache Capacity Trade-off



- Many improvements and variations since then...

M. Mohammadi Amiri and D. Gündüz, Fundamental limits of caching: Improved delivery rate-cache capacity trade-off, IEEE Trans. on Communications, vol. 65, no. 2, pp. 806-815, Feb. 2017.

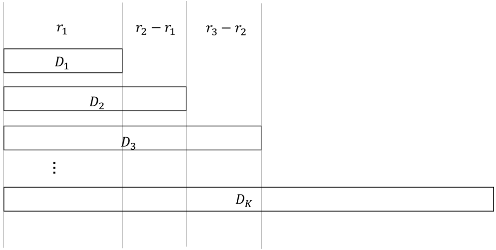
M. Mohammadi Amiri, Q. Yang and D. Gündüz, Decentralized coded caching with distinct cache capacities, IEEE Trans. on Communications, vol. 65, no. 11, pp. 4657 - 4669, Aug. 2017.

- Devices have different resolution/processing capabilities
- They may request the same file, but at different resolutions
- D_k: distortion requirement of user k. Without loss of generality, let

$$D_1 \geq D_2 \geq \cdots \geq D_K$$

- Devices have distinct cache capacities: M_k

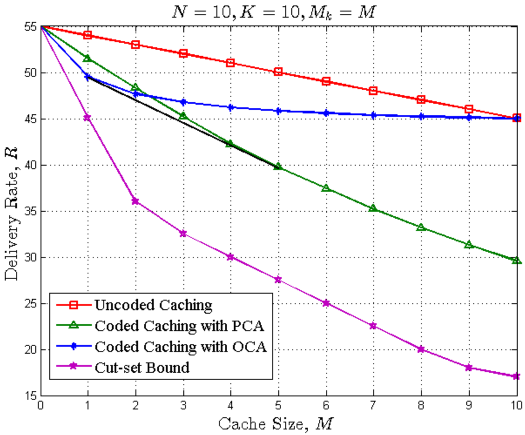
Q. Yang and D. Gündüz, **Coded caching and content delivery with heterogeneous distortion requirements**, revised, IEEE Trans. on Information Theory, 2016.



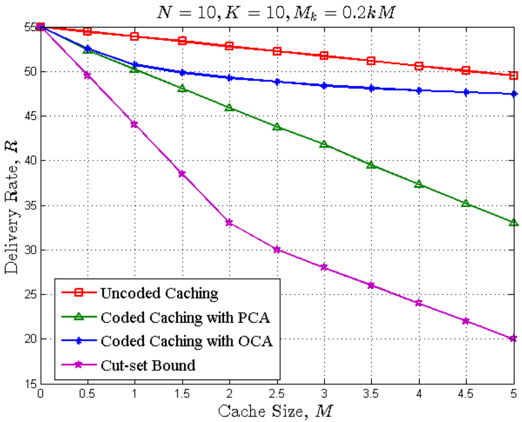
Compress video into multiple quality layers; e.g., **scalable video coding (SVC)** in H264/ MPEG

- First layer: r_1 bits/sample
- $k-$th layer: $r_k - r_{k-1}$ bits/sample
- User k wants $D_k \rightarrow$ needs first k layers

- $D_1 \geq D_2 \geq \cdots \geq D_{10}$: $r_k = k$, $k = 1, ..., 10$;
- Identical cache capacities, $M_k = M$.



- $D_1 \geq D_2 \geq \cdots \geq D_{10}$: $r_k = k$, $k = 1, ..., 10$;
- Heterogeneous cache capacities, $M_k = 0.2kM$.



Cache-Aided Delivery over the Wireless Edge

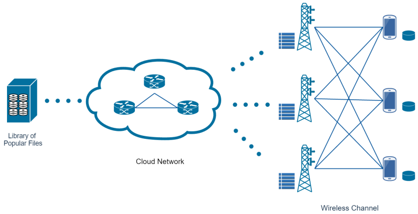
System overview

- $K_T \times K_R$ interference channel
- Transmitter cache: $M_T F$
- Receiver cache $M_R F$

Sum Degrees-of-Freedom

$$\mathrm{DoF}(M_T, M_R) = \liminf_{P \to \infty} \frac{C(M_T, M_R, P)}{\log(P)}.$$

- Decentralized caching at user terminals (RXs)



Novel scheme combining:

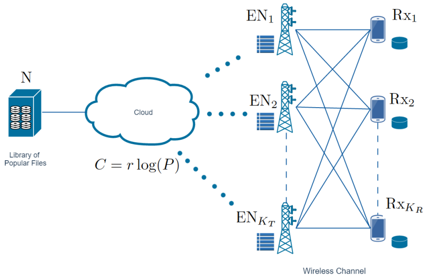
- **Zero-forcing**
- **Interference cancellation**
- **Interference alignment**

System overview

- Fronthaul connections to base stations
- Uncached contents can be delivered from the cloud server

Normalized Delivery Time

$$\delta(M_T, M_R) = \lim_{P\to\infty} \lim_{F\to\infty} \frac{T_F + T_E}{F/\log(P)}.$$

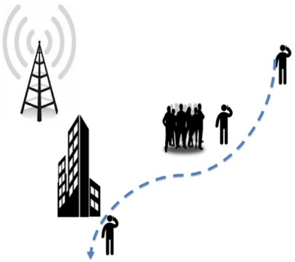


- Orthogonal backhaul links
- Fronthaul capacity r unknown during placement
- Serial/ pipelined fronthaul delivery
- Hard-transfer fronthauling
- Joint edge and cloud delivery

A. Sengupta, R. Tandon, and O. Simeone, **Cloud and cache-aided wireless networks: Fundamental latency trade-offs,**, IEEE Trans. on Information Theory, Nov. 2017.

J. Pujol-Roig, F. Tosato, and D. Gündüz, **Storage-latency trade-off in cache-aided fog radio access networks**, to appear in IEEE Int'l Conf. on Communications, Kansas City, MI, May. 2018.

Proactive Caching for Resource Optimization

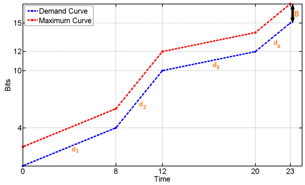


- Channel and network conditions vary over time
- State of the art: Reactive content delivery
- User behaviour (demands and mobility) are highly predictable
- **Contents can be pushed in advance when channel is good.**

A. C. Gungor and D. Gündüz, **Proactive wireless caching at mobile user devices for energy efficiency**, Int'l Symp. on Wireless Comm. Systems (ISWCS), 2015.

M. Gregori, J. Gomez-Vilardebo, J. Matamoros, and D. Gündüz, **Wireless content caching for small cell and D2D networks**, IEEE Journal on Selected Areas in Communications, May 2016.

- Demands known/ predicted in advance
- Finite capacity cache at user terminal
- System model:
  - Duration of time slot i: τ_i
  - User demand rate: d_i
  - Channel state: h_i
  - Cache capacity: B
  - Rate-power function: $r(t) = \log(1 + h(t)p(t))$



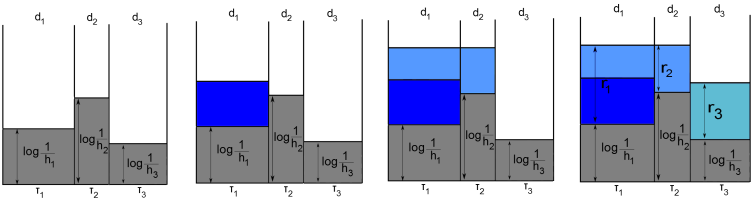
- Objective: Minimize energy consumption over N timeslots:

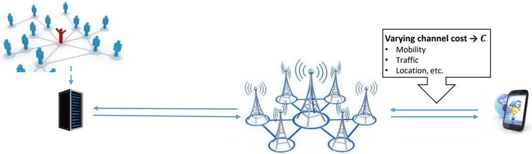
$$\min_{r_i \geq 0} \sum_{i=1}^{N} \tau_i \frac{e^{r_i} - 1}{h_i}$$

$$\text{s.t.} \sum_{i=1}^{n} \tau_i (d_i - r_i) \leq 0, \text{for } n = 1, \ldots, N,$$

$$\sum_{i=1}^{n} \tau_i (r_i - d_i) - B \leq 0, \text{for } n = 1, \ldots, N.$$

- Download demands over a longer period, and in better channel conditions
- Each file can be downloaded only in advance, not later than when it is requested
- Proactive caching amount is limited by cache memory

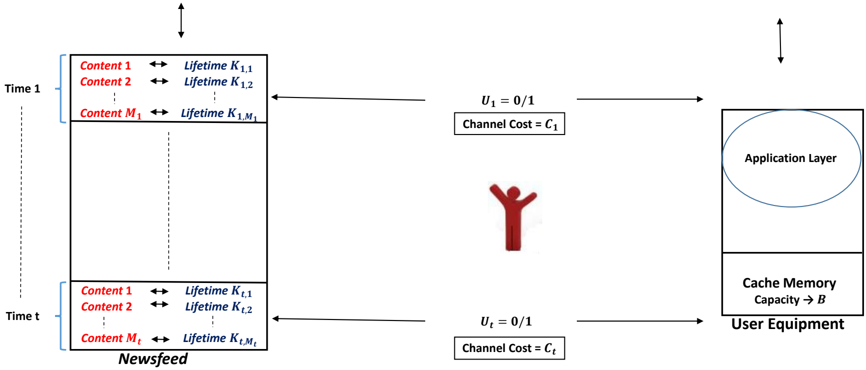




- Contents generated randomly, with random lifetime
- User accesses at random time instants to download all relevant contents (e.g., online social network)
- Cost = Channel cost of download × downloaded data
- **Goal: Minimize long-term average cost**
- Proactively cache content at favourable channel conditions

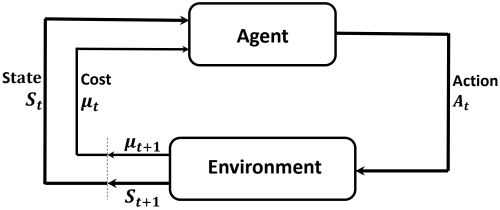
S. Somuyiwa, A. Gyorgy and D. Gündüz, **Improved policy representation and policy search for proactive content caching in wireless networks**, 2017 WiOpt.

S. Somuyiwa, A. Gyorgy and D. Gündüz, **Energy-efficient wireless content delivery with proactive caching,** 2nd Content Caching and Delivery in Wireless Networks Workshop.



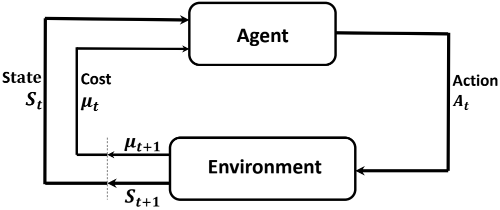
System State:

- Relevant contents outside cache ⇒ $\mathcal{O}_t$.
- Contents inside cache ⇒ $\mathcal{I}_t$ ($|\mathcal{I}_t| \leq B$).
- Elapsed time since last user access ⇒ E_t.
- Energy cost of downloading a content ⇒ C_t ($0 < C_t \leq C_{max}$): i.i.d. over time.



Markov decision process with side information (MDP-SI).

- State ($s \in \mathcal{S}$):
  - Controllable state: $(\mathcal{O}_t, \mathcal{I}_t, E_t)$.
  - Uncontrollable ~~state~~: $C_t \Rightarrow$ side information
- Action ($a \in \mathcal{A}_s$): $A_t = (A_t^{(1)}, A_t^{(2)})$.
- Transition probability: $P(S_{t+1}|S_t, A_t)$.
- Cost function: $\mu(S_t, A_t) = C_t \cdot |A_t^{(1)}|$.
- Objective function: $\rho = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \mu(S_t, A_t)\right]$.



Markov decision process with side information (MDP-SI).

- State ($s \in \mathcal{S}$):
  - Controllable state: $(\mathcal{O}_t, \mathcal{I}_t, E_t)$.
  - Uncontrollable ~~state~~: $C_t \Rightarrow$ side information
- Action ($a \in \mathcal{A}_s$): $A_t = (A_t^{(1)}, A_t^{(2)})$.
- Transition probability: $P(S_{t+1}|S_t, A_t)$.
- Cost function: $\mu(S_t, A_t) = C_t \cdot |A_t^{(1)}|$.
- Objective function: $\rho = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \mu(S_t, A_t)\right]$.

For any state $s = (\mathcal{O}, \mathcal{I}, E) \in \mathcal{S}$, the optimal policy $\pi^*(s)$ has a threshold structure with respect to cost C.

- Let
  - $l_1 \leq \cdots \leq l_B$:contents in the cache ($\mathcal{I}$).
  - $L_1 \geq \cdots \geq L_B$:B contents out of cache ($\mathcal{O}$) with highest lifetimes.
- $\exists B' \leq B$ and corresponding threshold values:

$$\mathcal{T}(a_{B'}) \leq \mathcal{T}(a_{B'-1}) \leq \cdots \leq \mathcal{T}(a_1) \leq C_{max},$$

and the optimal policy performs simple actions $a_i = (l_i | L_i)$, if $C \leq \mathcal{T}(a_i)$ and $E > 0$.

For any state $s = (\mathcal{O}, \mathcal{I}, E) \in \mathcal{S}$, the optimal policy $\pi^*(s)$ has a threshold structure with respect to cost C.

- Let
  - $l_1 \leq \cdots \leq l_B$:contents in the cache ($\mathcal{I}$).
  - $L_1 \geq \cdots \geq L_B$:B contents out of cache ($\mathcal{O}$) with highest lifetimes.
- $\exists B' \leq B$ and corresponding threshold values:

$$\mathcal{T}(a_{B'}) \leq \mathcal{T}(a_{B'-1}) \leq \cdots \leq \mathcal{T}(a_1) \leq C_{max},$$

and the optimal policy performs simple actions $a_i = (l_i | L_i)$, if $C \leq \mathcal{T}(a_i)$ and $E > 0$.

- **Longest lifetime in–Shortest lifetime out**:
  - Swap largest $L \in \mathcal{O}$ with the smallest $l \in \mathcal{I}$, if $C_t \leq \mathcal{T}(a)_{a=(l|L)}$, until no more swaps can be performed.
  - Single threshold value for each pair $(l|L)$ of lifetimes.
  - Parametrized by threshold values: $\boldsymbol{\theta} = \mathcal{T}(l|L)$ for all $L > l$.

Threshold values obtained using linear function approximation (LFA) as

$$\mathcal{T}(a)_{a=(l|L)} = \sum_{i=0}^{K_{max}} \phi(i)\theta_i(l,L) = \Phi^\top \theta(l,L) \; ,$$

K_{max}: maximum lifetime
$\Phi_t = [\phi_t(0), \phi_t(1), \ldots, \phi_t(K_{max})]$: *frequency vector*

$$\phi(i) \triangleq \frac{\sum_{l \in \mathcal{C}} \mathbb{I}_{\{l=i\}}}{B}, \quad \text{for} \quad i = 0, 1, \cdots, K_{max},$$

$\theta_i(l,L)$: coefficients to be optimized for each simple action.

- A model free policy search technique using stochastic gradient descent.

Policy Gradient Algorithm

- generate "samples" with $P(s'|s,a)$ and the probability density function $f_C(c)$
  - Results in *trajectory* $\tau_{\pi_{\boldsymbol{\theta}}} = (S_1, C_1, A_1), \ldots, , (S_T, C_T, A_T)$ i.e., $\tau_{\pi_{\boldsymbol{\theta}},T} \sim P_{\boldsymbol{\theta},T}(\tau_{\pi_{\boldsymbol{\theta}}}) = P(\tau_{\pi_{\boldsymbol{\theta}},T}|\boldsymbol{\theta})$.
- Evaluate average sample cost $J_{\pi_{\boldsymbol{\theta}}} = \frac{1}{T}\sum_{t=1}^{T}\mu(S_t, A_t)$
- Update $\boldsymbol{\theta}$ in the direction that decreases $\rho^{\pi_{\boldsymbol{\theta}}} = \mathbb{E}[J_{\pi_{\boldsymbol{\theta}}}]$:

$$\boldsymbol{\theta}_{j+1} = \boldsymbol{\theta}_j - \lambda \nabla_{\boldsymbol{\theta}} \rho^{\pi_{\boldsymbol{\theta}}},$$

where $\lambda > 0$ is the step size, j is the current iteration step and

$$\nabla_{\boldsymbol{\theta}} \rho^{\pi_{\boldsymbol{\theta}}} = \int_{\tau} \nabla_{\boldsymbol{\theta}} P_{\boldsymbol{\theta}}(\tau_{\pi_{\boldsymbol{\theta}}}) J_{\pi_{\boldsymbol{\theta}}} d\tau \ .$$

- Unlimited cache capacity (LB-UC)
  - Decouples actions for contents, $A_t^{(2)} = \emptyset, \ \forall t$
  - Threshold $\mathcal{T}_L$: Content with lifetime L is downloaded if $C \leq \mathcal{T}_L$.

$$0 \leq \mathcal{T}_1 \leq \cdots \leq \mathcal{T}_{K_{max}} \leq C_{max}$$

  - Threshold obtained using value iteration algorithm (VIA)

- Non-causal knowledge of user access times (LB-NCK)
  - For any time-to-user access t', contents are downloaded if $C_t \leq \mathcal{T}_{t'}$.

$$0 \leq \mathcal{T}_{D_{max}} \leq \cdots \leq \mathcal{T}_1 \leq C_{max}$$

    where D_{max} is the bound on the user access interval.
  - Threshold values obtained using VIA.

- Unlimited cache capacity (LB-UC)
  - Decouples actions for contents, $A_t^{(2)} = \emptyset, \ \forall t$
  - Threshold $\mathcal{T}_L$: Content with lifetime L is downloaded if $C \leq \mathcal{T}_L$.

$$0 \leq \mathcal{T}_1 \leq \cdots \leq \mathcal{T}_{K_{max}} \leq C_{max}$$

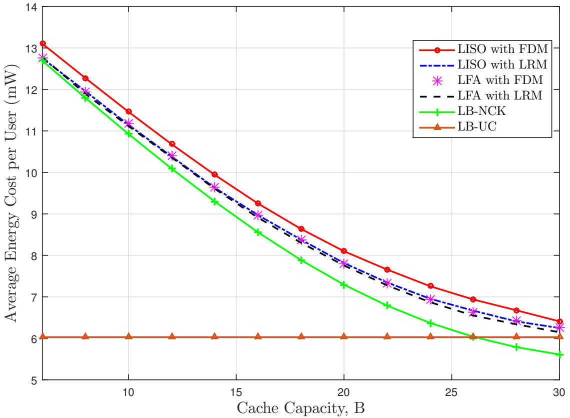
  - Threshold obtained using value iteration algorithm (VIA)

- Non-causal knowledge of user access times (LB-NCK)
  - For any time-to-user access t', contents are downloaded if $C_t \leq \mathcal{T}_{t'}$.

$$0 \leq \mathcal{T}_{D_{max}} \leq \cdots \leq \mathcal{T}_1 \leq C_{max}$$

    where D_{max} is the bound on the user access interval.
  - Threshold values obtained using VIA.

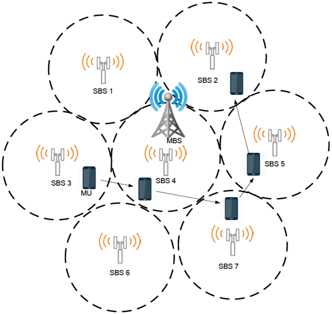
Average Energy Cost Vs Cache Capacity



Percentage Improvement over LISO with FDM:

▶ LFA with LRM → up to 5.6%. ▶ LFA with FDM → up to 4.4%.

▶ LISO with LRM → up to 4.2%.

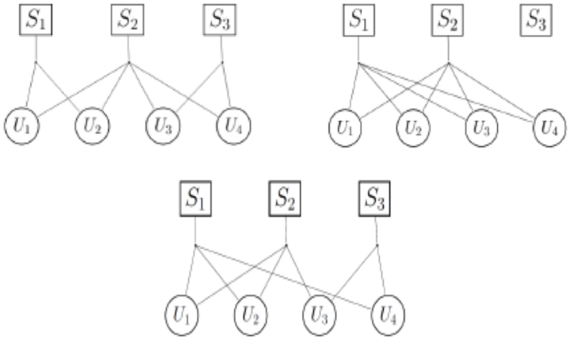


- Random mobility patterns
- Maximum distance separable (MDS) coded content storage
- How to allocate cached to contents with different popularities?

K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire. Femtocaching: Wireless content delivery through distributed caching helpers. IEEE Trans. Inf. Theory, Dec. 2013.

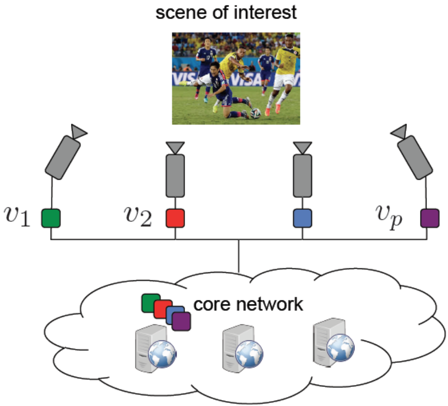
M. Ozfatura and D. Gündüz, Mobility and popularity aware coded small-cell caching, IEEE Communication Letters, 2017.

Multi-Server System with Random Topology



- Each user connects to ρ out of P servers
- Each server can cache N/ρ files
- Both coded caching and MDS coded storage need to be utilised

N. Mital, D. Gündüz and C. Ling, Coded caching in a multi-server system with distributed storage, to appear in Int'l Wireless Communications and Networking Conference, Barcelona, Spain, Apr. 2018.



- Interactive multiview streaming
- How to optimally cache and deliver multiview video content to improve the free viewpoint streaming experience?

E. Bourtsoulatze and D. Gündüz, Cache-aided interactive multiview video streaming in small cell networks, submitted for publication.

Thank You for Your Attention!