Caching Policies for Delay Minimization in Small Cell Networks with Joint Transmissions

Guilherme Iecker Ricardo
(*guilherme.ricardo@eurecom.fr*)

Giovanni Neglia
Thrasyvoulos Spyropoulos

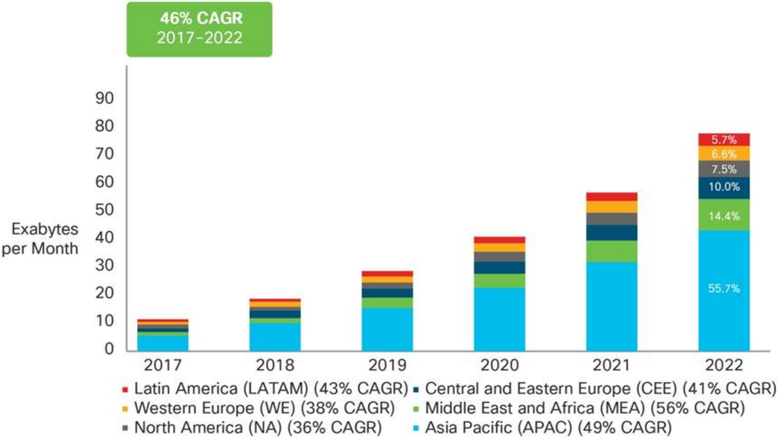
EURECOM
Sophia Antipolis

Inria
INVENTEURS DU MONDE NUMÉRIQUE
Member of UNIVERSITÉ CÔTE D'AZUR

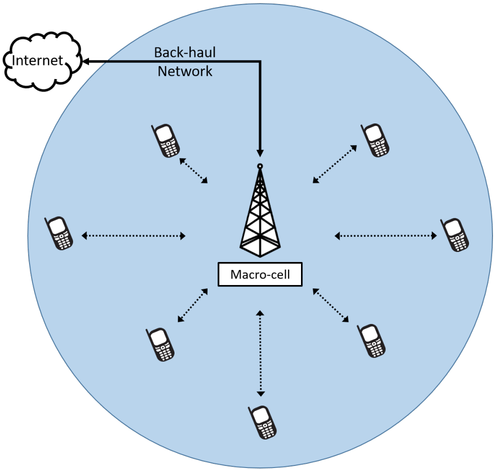
NEO Internal Meeting – May 4, 2020

Outline

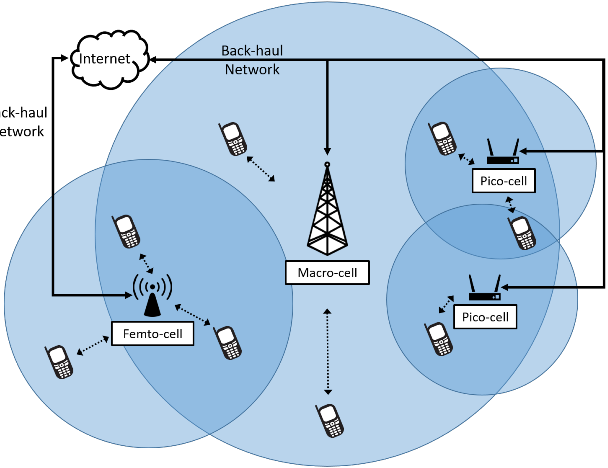
Motivation – Increase of Mobile Traffic (CISCO [1])



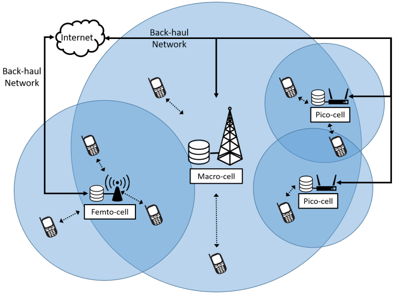
Motivation – Heterogeneous Networks (Bhushan et al. [2])



Motivation – Heterogeneous Networks (Bhushan et al. [2])

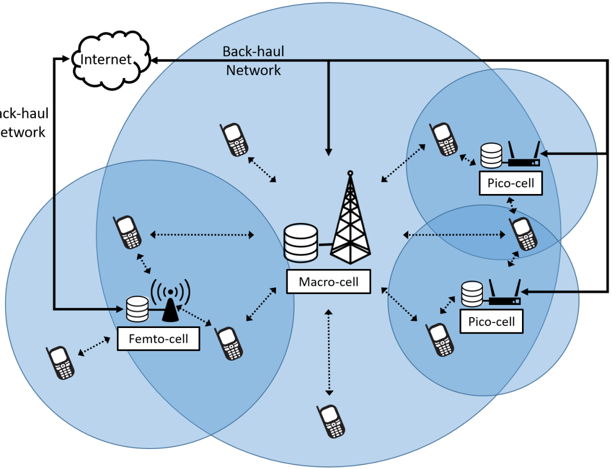


Problem: How to Improve QoE? (1) Edge-Caching!



- Static Solution:
  - FemtoCaching Greedy Algorithm (Shanmugam et al. [3])
  - Dynamic Programming (Ayenew et al. [4])
- Online Policies:
  - LRU-All and LRU-One (Giovanidis and Avranas [5])
  - qLRU-Lazy and 2LRU-Lazy (Leonardi and Neglia [6])

Problem: How to Improve QoE? (2) CoMP Techniques



- Coordinated Multi-Point (CoMP) (Lee et al. [7])
- Static Solution: Delay Minimization (Tuholukova et al. [8])

Methodology

- Model network transmissions and caching operations
- Revisit Delay Minimization static solution
    - Optimization formulation – Properties and Solutions
    - Full-coverage scenario – Application and Results
- Propose novel online caching policies:
    - The qLRU-Δd Policy (algorithmic description, optimality sketch proof, and examples)
    - The 2LRU-Δd Policy (algorithmic description and examples)
- Perform experiments
    - Observe qLRU-Δd convergence
    - Evaluate performance through simulations
    - Evaluate performance under heterogeneous SNR scenario

Notation

H	Number of Helpers
U	Number of User Equipments (UEs)
$g_{h,u}$	SNR [dB] from helper h to UE u
d_B	Backhaul access delay [ms]
F	Catalog size [number of files]
p_f	Popularity of file f, $[p_f, f = 1, ..., F] \sim \text{Zipf}(\alpha)$
C	Cache capacity [number of files]
$X_f^{(h)}$	Indicator whether helper h caches file f
$\mathbf{X} \in \{0,1\}^{F \times H}$	Allocation matrix
I_u	Set of helpers covering UE u
$J_{u,f} \subseteq I_u$	Set of helpers covering UE u and caching file f

Table 1: Initial notation

The Total E2E Delay

The delay UE u experiences to get file f under allocation $\mathbf{X}$ is:

$$d(u,f,\mathbf{X}) = \begin{cases} d_B + \frac{M}{W\log_2(1+g_{h^*,u})}, & |J_{u,f}| = 0 \quad \text{(Miss)} \\ \frac{M}{W\log_2\left(1+\sum_{h\in J_{u,f}} g_{h,u}\right)}, & |J_{u,f}| > 0 \quad \text{(Hit)}, \end{cases}$$

where W is the channel bandwidth [Hz] and M is the file size [bits].

The Total E2E Delay

The delay UE u experiences to get file f under allocation $\mathbf{X}$ is:

$$d(u,f,\mathbf{X}) = \begin{cases} d_B + \frac{M}{W\log_2\left(1+g_{h^*,u}\right)}, & |J_{u,f}| = 0 \quad \text{(Miss)} \\ \frac{M}{W\log_2\left(1+\sum\limits_{h\in J_{u,f}} g_{h,u}\right)}, & |J_{u,f}| > 0 \quad \text{(Hit)}, \end{cases}$$

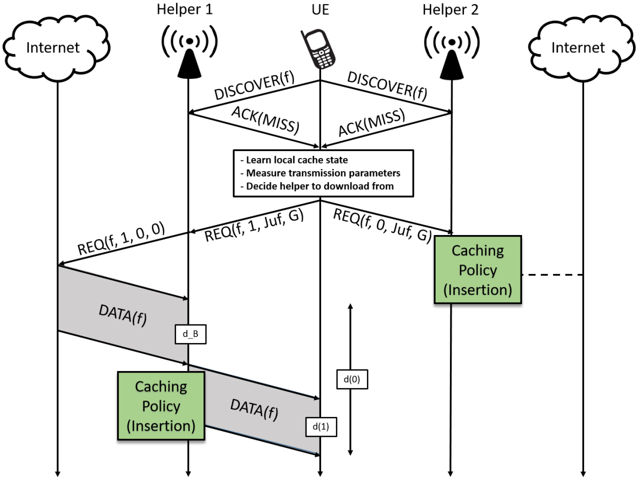
where W is the channel bandwidth [Hz] and M is the file size [bits].

For homogeneous SNRs (i.e., when $g_{h,u} = g$, for all pairs (u,h)):

$$d(k(u,f,\mathbf{X})) = \begin{cases} d_B + \frac{M}{W\log_2(1+g)}, & k(u,f,\mathbf{X}) = 0 \quad \text{(Miss)} \\ \frac{M}{W\log_2(1+k(u,f,\mathbf{X})g)}, & k(u,f,\mathbf{X}) > 0 \quad \text{(Hit)}, \end{cases}$$

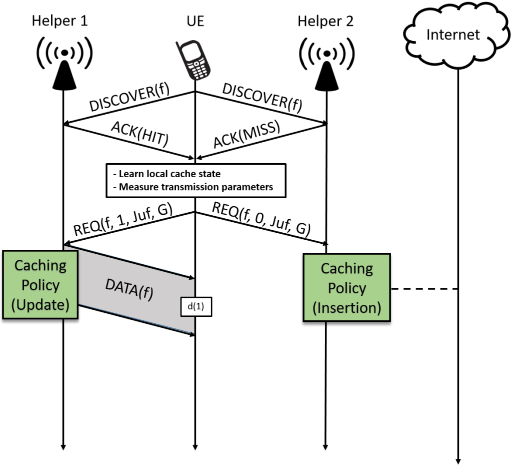
where $k(u,f,\mathbf{X}) \triangleq \sum_{h\in J_{u,f}} X_f^{(h)} = |J_{u,f}|$.

Network Operation: 2 Helpers Example

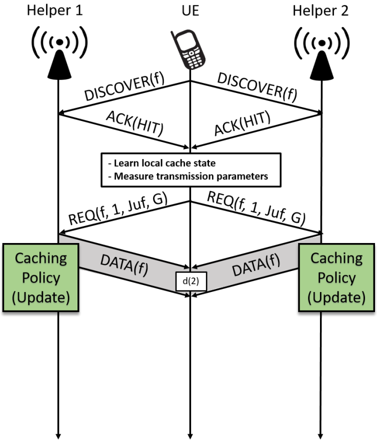


- Learn local cache state
- Measure transmission parameters
- Decide helper to download from

Network Operation: 2 Helpers Example



Network Operation: 2 Helpers Example



Optimization Formulation

Average Delay Minimization (ADMin) Problem

$$
\begin{aligned}
(ADMin) \quad & \text{minimize} \quad && \bar{d}(\mathbf{X}) = \sum_{f=1}^{F} \frac{1}{U} \sum_{u=1}^{U} p_f \cdot d(u, f, \mathbf{X}) \\
& \text{subject to} \quad && \sum_{f=1}^{F} X_f^{(h)} = C, \;\; h \in [H] \\
& && X_f^{(h)} \in \{0, 1\}, h \in [H], f \in [F]
\end{aligned}
$$

Properties and Solutions

Theorem

ADMin is a NP-Hard problem.

Theorem

If $d_B \geq d(1)$, in the homogeneous SNR case, or if $d_B \geq d(1)$, in the general case, then $\bar{d}(\mathbf{X})$ is submodular.

Proposition

Because

- *$\bar{d}(\mathbf{X})$ is a **monotone**, **submodular** set function, and*
- *ADMin Problem's constraints form a **partition matroid**,*

*then ADMin Problem can be **efficiently approximated by a Greedy algorithm** within a factor of 0.5 from the optimal solution.*

Full-Coverage Scenario

Proposition

In the full-coverage scenario, if $d(1) \leq d_B$, an allocation provided by the greedy algorithm is optimal.

Proof.

- 1 cache with capacity HC and up to H copies of the same file
- Delay function $d(k(u,f))$ replaced with constant d_k
- New variables: $\mathbf{Y} \in \{0,1\}^{F\times H}$;
- New objective: $F(\mathbf{Y}) = \sum_{f=1}^{F}\sum_{k=0}^{B} p_f d_k y_{f,k}$
- Knapsack constraints: $\sum_{f=1}^{F}\sum_{k=0}^{B} y_{f,k} = BC$

☐

Full-Coverage Scenario – Extreme Allocation Bounds

Proposition

In the full-coverage scenario, if $d(1) \leq d_B$, *full-diversity is an optimal allocation if, and only if,*

$$d_B \geq D_{FD} \triangleq \frac{p_1}{p_{HC}}(d(1) - d(2)),$$

and full-replication is an optimal allocation if and only if

$$d_B \leq D_{FR} \triangleq \frac{p_C}{p_{C+1}}(d(H-1) - d(H)).$$

Full-Coverage Scenario – Parametric Bounds

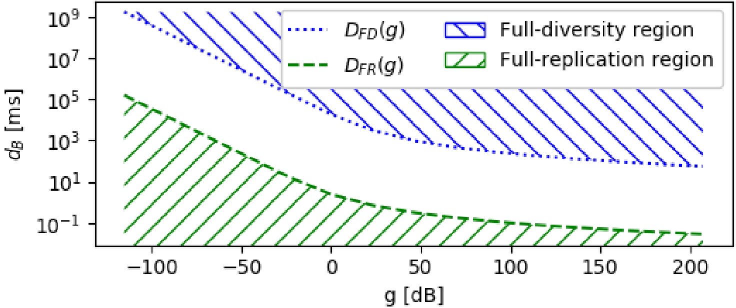


Figure 1: Boundaries of (d_B, g) for extreme allocations: $H = 10$, $\alpha = 1.5$, $F = 10^6$, and $C = 100$. The axis d_B is in log scale.

Extreme Allocation General Conditions

Corollary

Assuming homogeneous SNRs and for general network topologies, then:

- $d_B \geq D_{FD}$ *is a* ***necessary condition*** *for the full-diversity allocation to be locally optimal*
- $d_B \leq D_{FR}$ *is a* ***sufficient condition*** *for the full-replication allocation to be locally optimal.*

Online Caching Policies: Introduction to qLRU

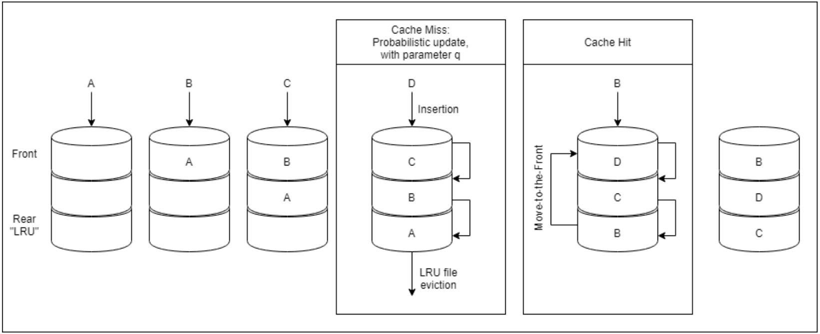


Figure 2: qLRU Policy

Online Caching Policies: Additional Notation

The Gain Function Δd

$$\Delta d_f^{(h)}(u, \mathbf{X}_f) \triangleq d(u, f, \mathbf{X}_f \ominus \mathbf{e}^{(h)}) - d(u, f, \mathbf{X}_f)$$

"Move-to-the-Front" Probability

$$\rho_f^{(h)}(u, \mathbf{X}_f) = \beta \cdot \Delta d_f^{(h)}(u, \mathbf{X}_f)$$

Insertion Probability

$$q_f^{(h)}(u, \mathbf{X}_f) = q \cdot \sigma_f^{(h)}(u, \mathbf{X}_f),$$

where

$$\sigma_f^{(h)}(u, \mathbf{X}_f) = \gamma \cdot \Delta d_f^{(h)}(u, \mathbf{X}_f \oplus \mathbf{e}^{(h)})$$

Online Caching Policies: qLRU-Δd Description

qLRU-Δd Policy

Upon a request (u, f), $\forall h \in I_u$:

- if $h \in J_{u,f}$, **move f to the front** of h's cache with probability $\rho_f^{(h)}(u, \mathbf{X}_f)$, and
- if $h \in I_u \backslash J_{u,f}$, evict the LRU file from and **insert f** to h's cache with probability $q_f^{(h)}(u, \mathbf{X}_f)$.

Online Caching Policies: qLRU-Δd Convergence

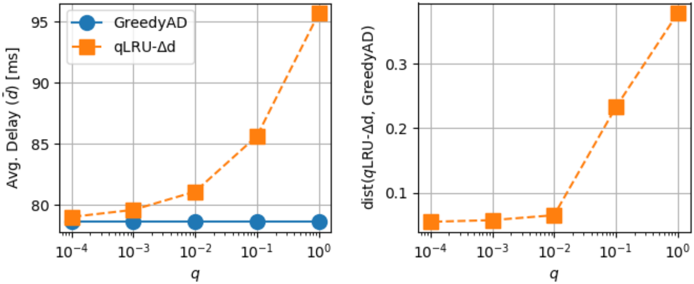


Figure 3: Convergence analysis: delay (left) and allocation (right) convergence with q, for $\alpha = 1.2$, $d_B = 100$ms, and $g = 10$dB.

Online Caching Policies: Introduction to 2LRU

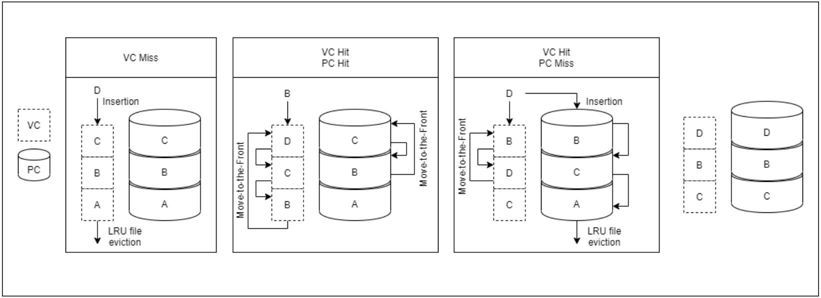


Figure 4: 2LRU Policy

Online Caching Policies: 2LRU-Δd Description

2LRU-Δd Policy

Upon a request (u, f), $\forall h \in I_u$:

- if $h \in \hat{J}_{u,f}$, **move f's ID to the front** of h's VC and,
  - if $h \in J_{u,f}$, **move f's ID to the front** of h's PC with prob. $\rho_f^{(h)}(u, \mathbf{X}_f)$, or
  - if $h \notin J_{u,f}$, evict LRU file from and insert f to h's PC.
- if $h \notin \hat{J}_{u,f}$, evict LRU file from and insert f to h's VC with prob. $\sigma_f^{(h)}(u, \mathbf{X}_f)$.

IRM Homogeneous SNR Results

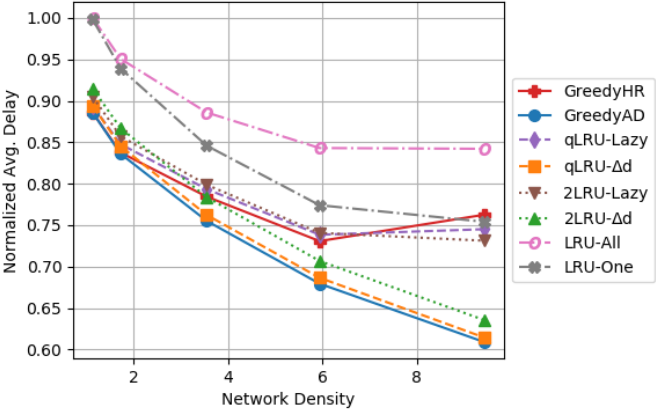


Figure 5: Performance analysis of various policies in a real topology with IRM request process ($\alpha = 1.2$)

Real Trace Homogeneous SNR Results

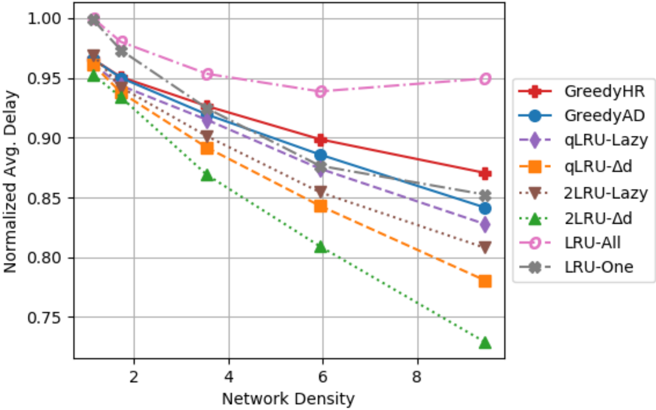


Figure 6: Performance analysis of various policies in a real topology with Akamai trace.

Real Trace Heterogeneous SNR Results

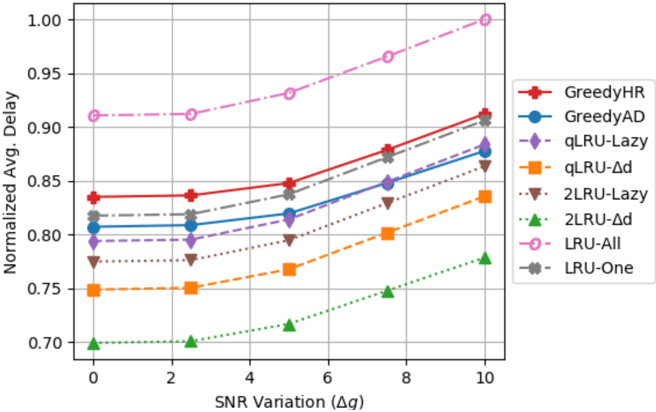


Figure 7: Heterogeneous SNRs: Berlin topology with density 9.4, $g_0 = 10.0$dB, and $d_B = 100.0$ms with Akamai trace.

Conclusion

- GreedyAD and GreedyHR can provide different allocations depending on network parameters
- qLRU-Δd provides locally optimal under IRM
- 2LRU-Δd provides good results for trace-based request processes
- The two policies are better than other policies from the literature
- The SNR variability affects proportionally all policies

Future Work

- Heterogeneous SNR Experiments
- Heterogeneous File size model
- Physical Interference model
- Joint optimization of caching and cell power/location

Caching Policies for Delay Minimization in Small Cell Networks with Joint Transmissions

Guilherme Iecker Ricardo
(*guilherme.ricardo@eurecom.fr*)

Giovanni Neglia
Thrasyvoulos Spyropoulos

EURECOM
Sophia Antipolis

Inria
INVENTEURS DU MONDE NUMÉRIQUE
Member of UNIVERSITÉ CÔTE D'AZUR

NEO Internal Meeting – May 4, 2020

References

[1] CISCO. *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021 White Paper*. Tech. rep. CISCO, Feb. 2017.

[2] N. Bhushan et al. "Network densification: the dominant theme for wireless evolution into 5G". In: *IEEE Communications Magazine* 52.2 (Feb. 2014), pp. 82–89. DOI: `10.1109/mcom.2014.6736747`.

[3] K. Shanmugam et al. "FemtoCaching: Wireless Content Delivery Through Distributed Caching Helpers". In: *IEEE Transactions on Information Theory* 59.12 (Dec. 2013), pp. 8402–8413. DOI: `10.1109/TIT.2013.2281606`.

References

[4] T. M. Ayenew et al. "A Novel Content Placement Strategy for Heterogeneous Cellular Networks with Small Cells". In: *IEEE Networking Letters* (2019), pp. 1–1. ISSN: 2576-3156. DOI: `10.1109/LNET.2019.2950990`.

[5] Anastasios Giovanidis and Apostolos Avranas. "Spatial Multi-LRU Caching for Wireless Networks with Coverage Overlaps". In: *SIGMETRICS Perform. Eval. Rev.* 44.1 (June 2016), pp. 403–405. ISSN: 0163-5999. DOI: `10.1145/2964791.2901483`.

[6] Emilio Leonardi and Giovanni Neglia. "Implicit Coordination of Caches in Small Cell Networks Under Unknown Popularity Profiles". In: *IEEE Journal on Selected Areas in Communications* 36.6 (June 2018), pp. 1276–1285. ISSN: 0733-8716. DOI: `10.1109/JSAC.2018.2844982`.

References

[7] Daewon Lee et al. "Coordinated multipoint transmission and reception in LTE-advanced: deployment scenarios and operational challenges". In: *IEEE Communications Magazine* 50.2 (Feb. 2012), pp. 148–155. DOI: `10.1109/mcom.2012.6146494`.

[8] A. Tuholukova, G. Neglia, and T. Spyropoulos. "Optimal cache allocation for femto helpers with joint transmission capabilities". In: *2017 IEEE International Conference on Communications (ICC)*. May 2017, pp. 1–7. DOI: `10.1109/ICC.2017.7996469`.