Invited Talks

The Role of Caching in 5G Wireless Networks

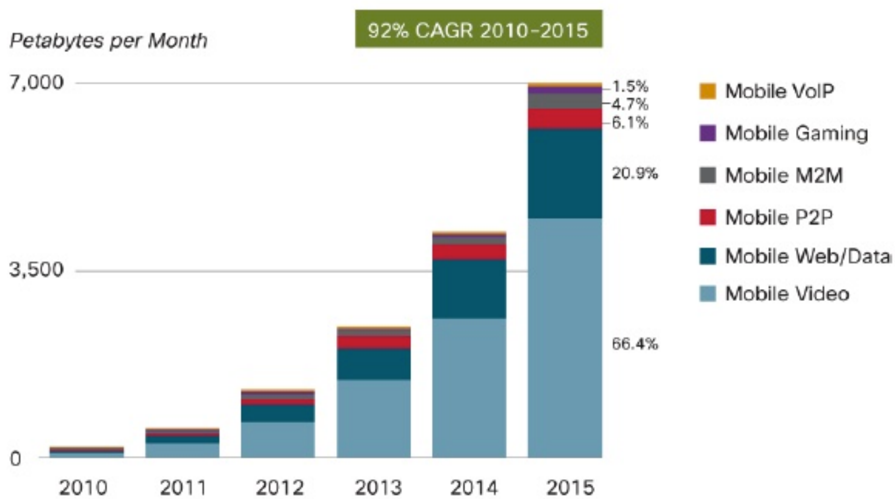
Giuseppe Caire

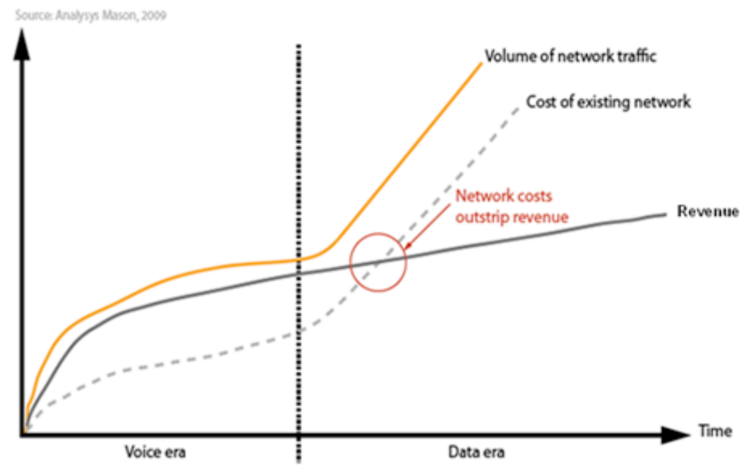
(joint work with Mingyue Ji, Dilip Benabothla, Michael J. Neely, Andreas F. Molisch)

University of Southern California, Viterbi School of Engineering, Los Angeles, CA

IEEE ICC – Budapest 06-10/13-2013

Wireless operators' nightmare



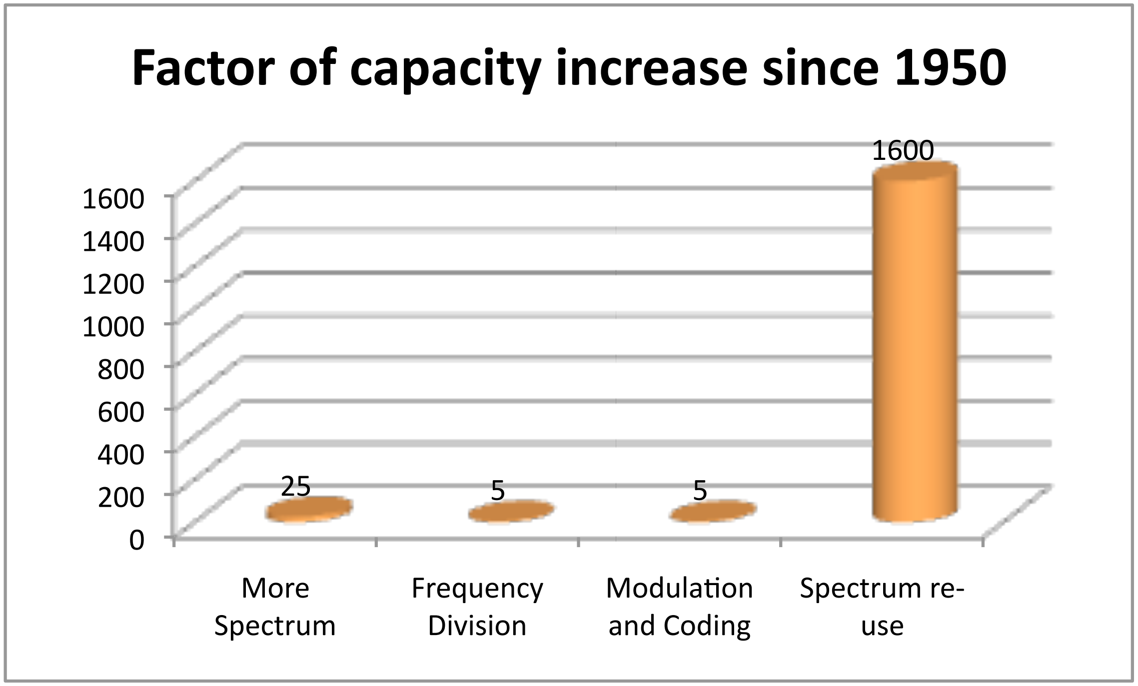


Mobile Broadband's Profitability Gap

- 100x Data traffic increase, due to the introduction of powerful multimedia capable user devices.
- Operating costs not matched by revenues.

A Clear Case for Denser Spatial Reuse

- If user-destination distance is $O(1/\sqrt{n})$, with transport capacity $O(\sqrt{n})$, we trivially achieve $O(1)$ throughput per user.



Small Cells: Challenges

- Handling mobility: we need (at least) two tiers, small cells to provide throughput, underneath macro-cells to provide coverage).
- Lack of carefully centralized planning $\Longrightarrow$ wild inter-tier and intra-tier interference scenarios, SoN.
- Open access versus closed access.... and other “femtocells” stories.
- Deployment of a high-capacity wired backbone (by far the most costly operation in terms of CapEX).

Small Cells: Challenges

- Handling mobility: we need (at least) two tiers, small cells to provide throughput, underneath macro-cells to provide coverage).
- Lack of carefully centralized planning $\Longrightarrow$ wild inter-tier and intra-tier interference scenarios, SoN.
- Open access versus closed access.... and other "femtocells" stories.
- **Deployment of a high-capacity wired backbone (by far the most costly operation in terms of CapEX).**

Video-Aware Wireless Networks

- Video is responsible for 66% of the traffic demand increase.
- Internet browsing for another 21%.
- On-demand video streaming and Internet browsing have important common features:
  1. Asynchronous content reuse (traffic generated by a few popular files, which are accessed in a totally asynchronous way).
  2. Highly predictable demand distribution (we can predict what, when and where will be requested).
  3. Delay tolerant, variable quality, ideally suited for best-effort (goodbye QoS, welcome QoE).

Video-Aware Wireless Networks

- Video is responsible for 66% of the traffic demand increase.
- Internet browsing for another 21%.
- On-demand video streaming and Internet browsing have important common features:
  1. **Asynchronous content reuse** (traffic generated by a few popular files, which are accessed in a totally asynchronous way).
  2. Highly predictable demand distribution (we can predict what, when and where will be requested).
  3. Delay tolerant, variable quality, ideally suited for best-effort (goodbye QoS, welcome QoE).

Video-Aware Wireless Networks

- Video is responsible for 66% of the traffic demand increase.
- Internet browsing for another 21%.
- On-demand video streaming and Internet browsing have important common features:
  1. Asynchronous content reuse (traffic generated by a few popular files, which are accessed in a totally asynchronous way).
  2. **Highly predictable demand distribution** (we can predict what, when and where will be requested).
  3. Delay tolerant, variable quality, ideally suited for best-effort (goodbye QoS, welcome QoE).

Video-Aware Wireless Networks

- Video is responsible for 66% of the traffic demand increase.
- Internet browsing for another 21%.
- On-demand video streaming and Internet browsing have important common features:
  1. Asynchronous content reuse (traffic generated by a few popular files, which are accessed in a totally asynchronous way).
  2. Highly predictable demand distribution (we can predict what, when and where will be requested).
  3. **Delay tolerant, variable quality, ideally suited for best-effort** (goodbye QoS, welcome QoE).

Video-Aware Wireless Networks

- Video is responsible for 66% of the traffic demand increase.

- Internet browsing for another 21%.

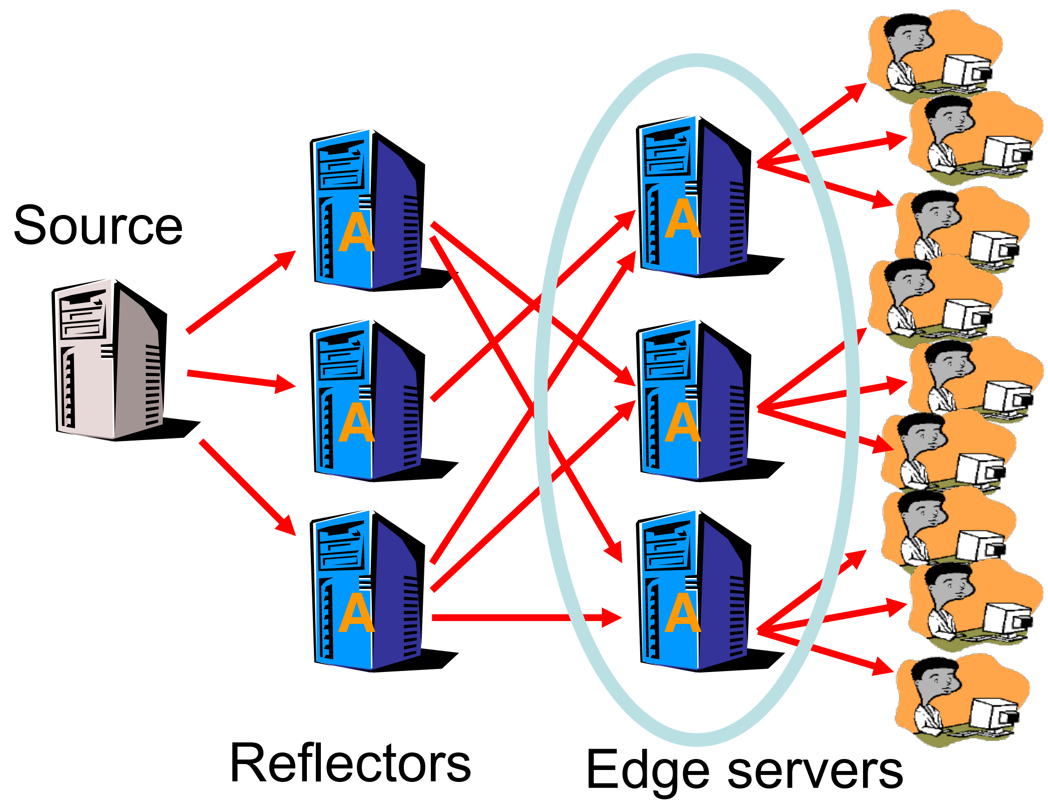
- On-demand video streaming and Internet browsing have important common features:

  1. Asynchronous content reuse (traffic generated by a few popular files, which are accessed in a totally asynchronous way).
  2. Highly predictable demand distribution (we can predict what, when and where will be requested).
  3. Delay tolerant, variable quality, ideally suited for best-effort (goodbye QoS, welcome QoE).

- VAWN Project:

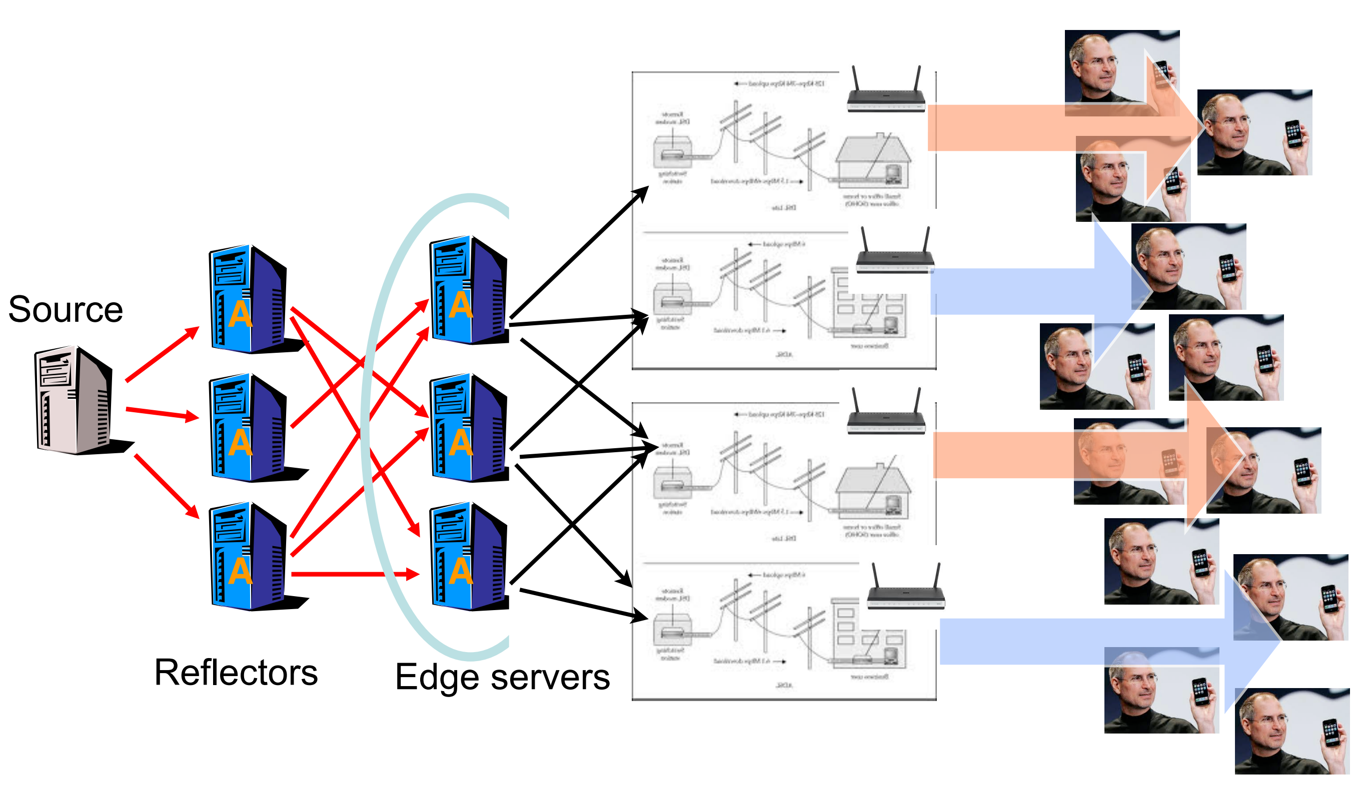
Well-Known Solution in Wired Networks: CDNs

- Caching is implemented in the core network (transparent to the wireless segment).



Why the Problem is Not (Yet) Solved?

- The wired backhaul to small cells is inexistent, weak or expensive.
- To a lesser extent: interference in the wireless segment.



Caching at the Wireless Edge

- If the CDN nodes are in the core-network, there is not enough bit-rate to the wireless edge (DSL, Cable ... not fast enough, US fiber-to-the home penetration .... scarce and costly ... ask Google Fiber!).

- Femto Caching: a radical view ... helper nodes everywhere with caches possibly refreshed by the LTE network at off-peak times.

- D2D Caching: an even more radical view ... cache directly in the user devices, and enable LTE-D2D.

- Caching wireless helpers: 10TB nodes $\times$ 100 nodes/km^2 $=$ 1000 TB/km^2 of distributed storage capacity.

- Near future user devices: 100GB of memory per device $\times$ 10000 people/km^2 $=$ 1000 TB/km^2 of distributed storage capacity.

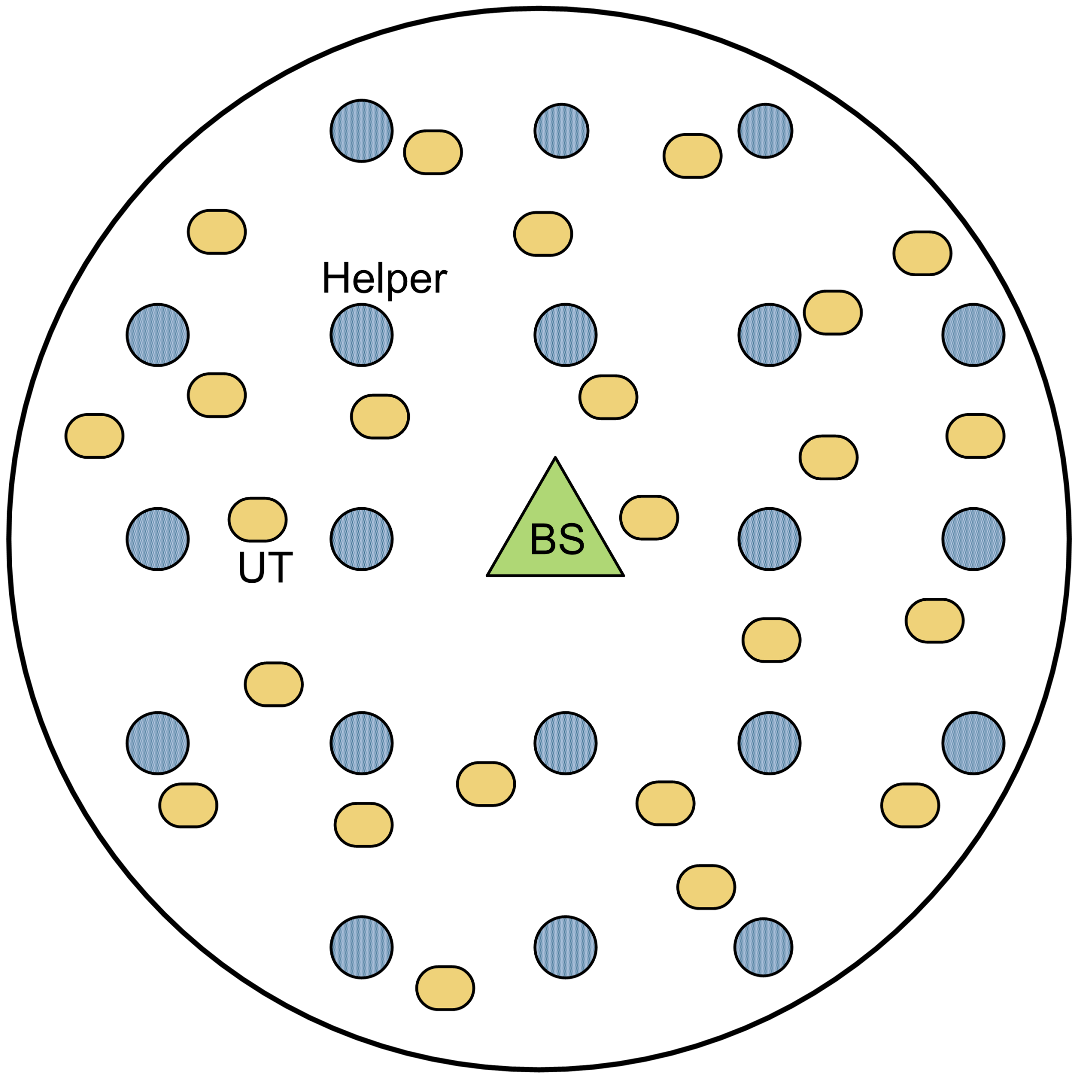
Research Problems

- What to cache, when and where: Predictive networks, using context side information (e.g., social networks).
- Efficient video-streaming in a wireless D2D network: video-quality aware admission control and scheduling.
- Efficient PHY/MAC: how to cope with interference in a dense self-organizing network (WiFi-offload, forthcoming Small-Cells Standards, LTE-D2D).
- Performance Analysis: throughput-outage tradeoff of caching networks.

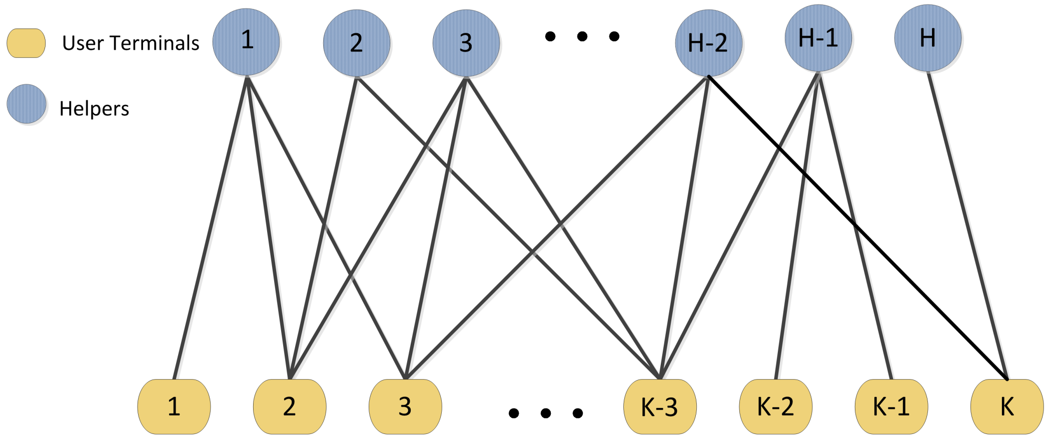
Cache Placement: use the LTE base station at off-peak times

LTE Multicast Stream
(Fountain-encoded)

A Cell with Caching Helpers

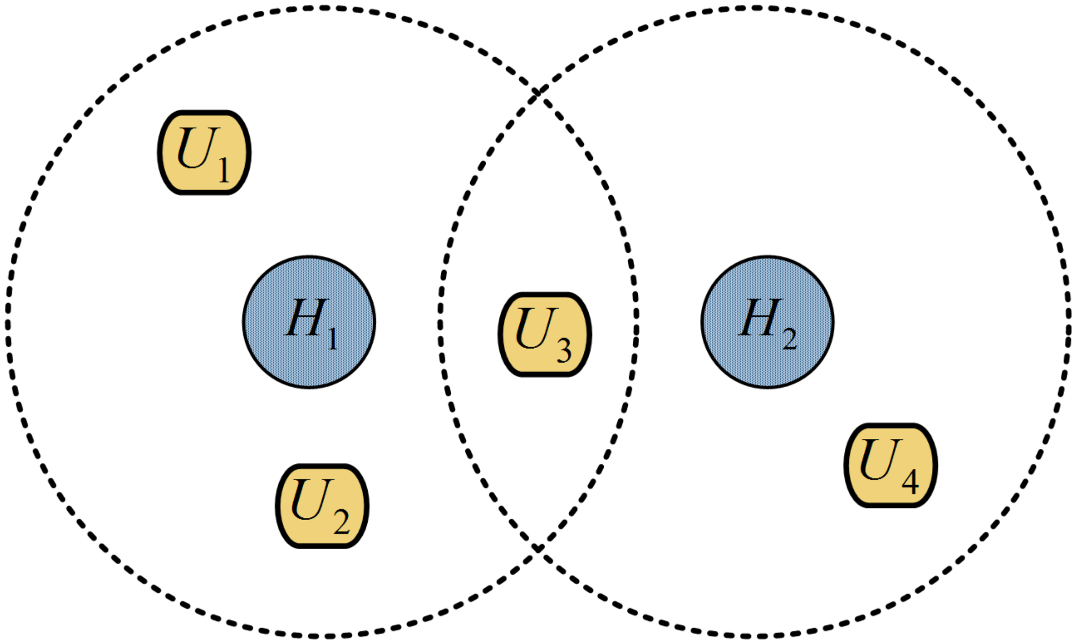


The Cache Placement Problem



- Helpers $\mathcal{H}$ of size $H + 1$, users $\mathcal{U}$ of size U and a library of files $\mathcal{F}$ of size F.
- Bipartite connectivity graph $\mathcal{G} = (\mathcal{H}, \mathcal{U}, \mathcal{E})$.
- Helper $h = 0$ (base station) is connected to all users.
- $\mathbf{\Omega} = [\omega_{h,u}]$ is the matrix of downloading times per information bit over each link.

The Problem is Far from Trivial



- Average downloading delay per information bit for user u:

$$\bar{D}_u = \sum_{j=1}^{|\mathcal{H}(u)|-1} \omega_{(j)_u,u} \sum_{f=1}^{F} \left[\prod_{i=1}^{j-1} (1 - x_{f,(i)_u}) \right] x_{f,(j)_u} P_r(f)$$
$$+ \omega_{0,u} \sum_{f=1}^{F} \left[\prod_{i=1}^{|\mathcal{H}(u)|-1} (1 - x_{f,(i)_u}) \right] P_r(f).$$

- In order to see this: $\left[\prod_{i=1}^{j-1}(1 - x_{f,(i)_u})\right] x_{f,(j)_u}$ is the indicator function of the condition that file f is in the cache of helper $(j)_u$ (the j-th lowest delay helper for user u), and it is not in any of the helpers with lower delay $(i)_u$, for $i = 1, \ldots, j-1$.

- Integer programming problem (combinatorial optimization):

$$\begin{aligned} \text{maximize} \quad & \sum_{u=1}^{U} \left(\omega_{0,u} - \bar{D}_u\right) \\ \text{subject to} \quad & \sum_{f=1}^{F} x_{f,h} \leq M, \quad \forall \;\; h, \\ & \mathbf{X} \in \{0,1\}^{F \times H}. \end{aligned}$$

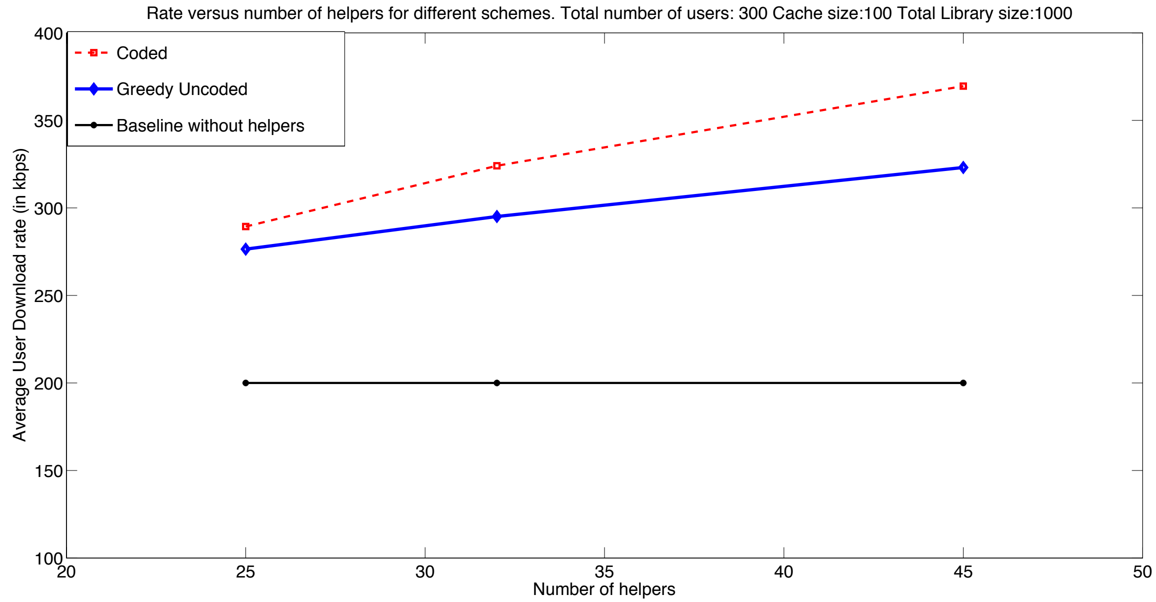
- We can show that the problem is NP-hard.
- Fortunately: we can formulate it as the maximization of a sub modular function subject to a matroid constraint (greedy is good!).
- Convex relaxation: we obtain an LP, with the meaning of intra-session fountain coding.
- Details in: FemtoCaching: Wireless Video Content Delivery through Distributed Caching Helpers, ArXiv Preprint, submitted to IEEE Trans. on Inform. Theory, (2011, revised 2013).

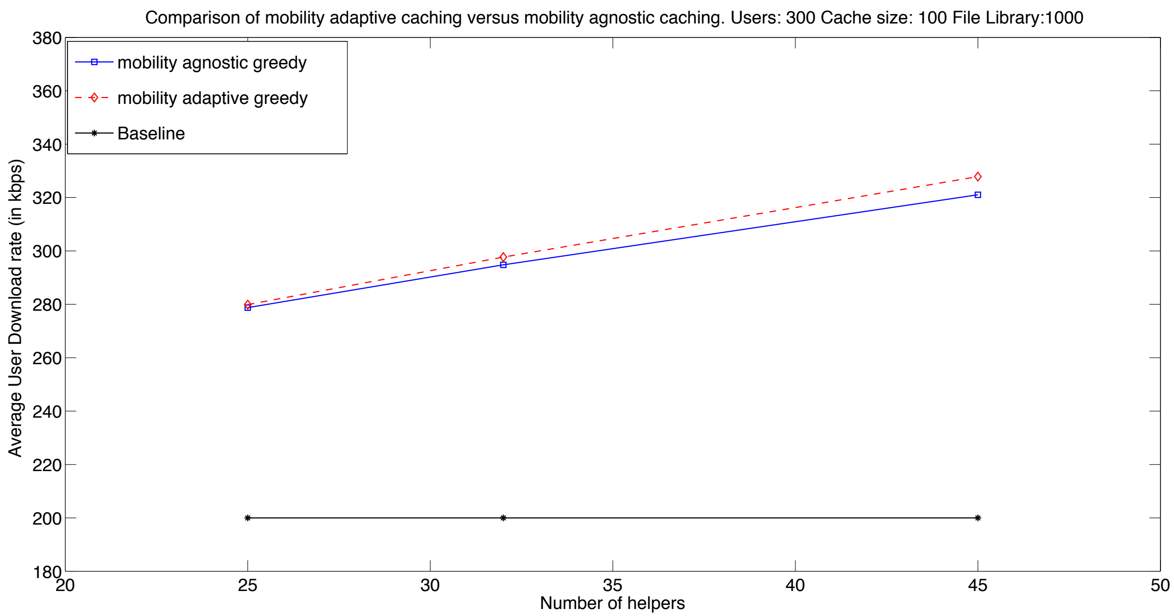
Numerical Results

- Cell of radius 350m.
- Helpers connectivity range 70m
- BS and helpers operates at 3 bit/s/Hz over 20 MHz of bandwidth:

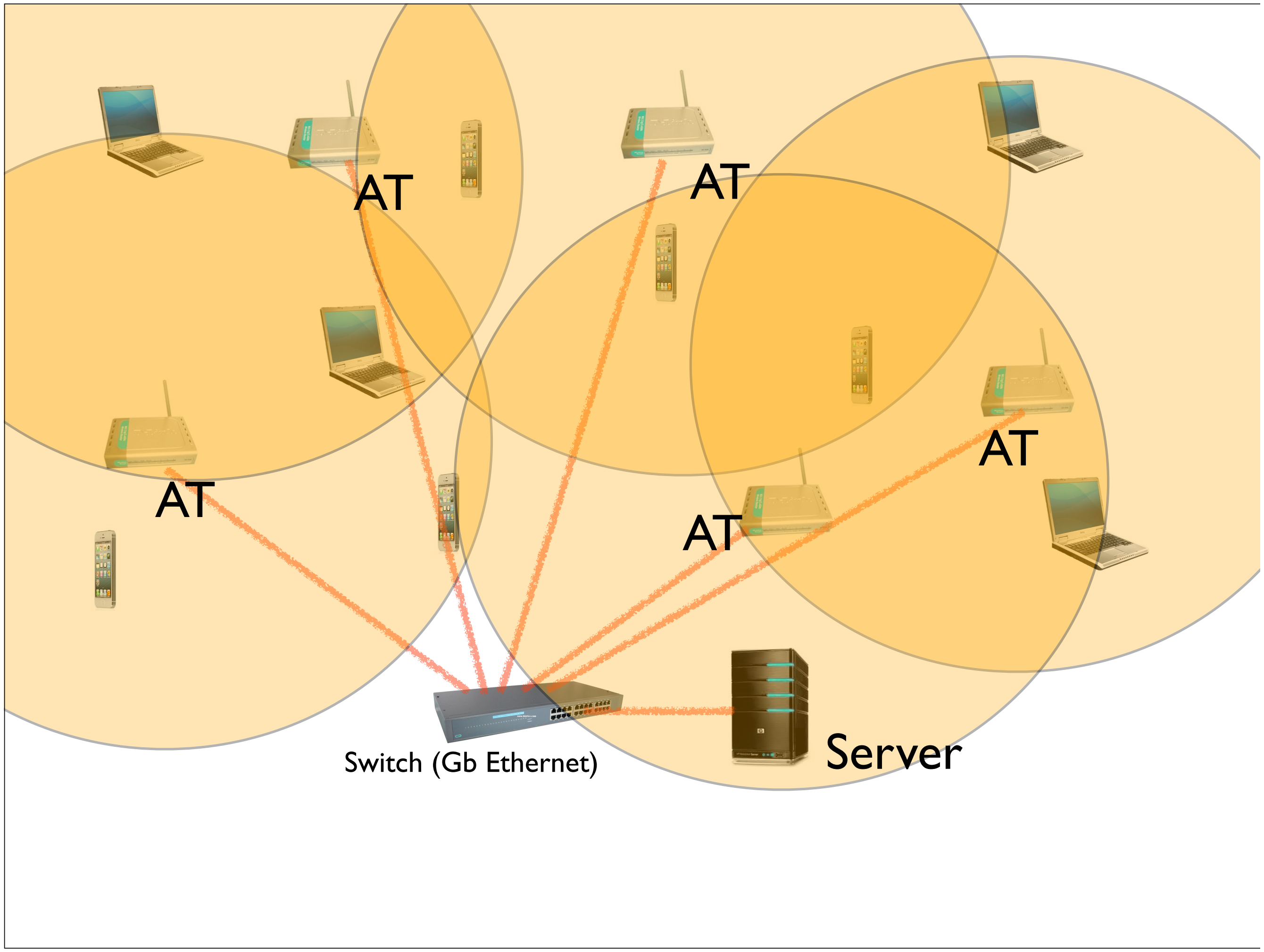
$$\text{Rate} = \frac{\text{Spectral Efficiency} \times \text{Bandwidth}}{\text{Number of connected users}}$$

- Helpers are placed on a regular grid over the cell area.
- $F = 1000$, $M = 100$, request distribution is Zipf with parameter 0.56.


Rate versus number of helpers for different schemes. Total number of users: 300 Cache size:100 Total Library size:1000
Coded
Greedy Uncoded
Baseline without helpers
Average User Download rate (in kbps)
Number of helpers
400
350
300
250
200
150
100
20
25
30
35
40
45
50


Comparison of mobility adaptive caching versus mobility agnostic caching. Users: 300 Cache size: 100 File Library:1000
mobility agnostic greedy
mobility adaptive greedy
Baseline
Average User Download rate (in kbps)
Number of helpers
380
360
340
320
300
280
260
240
220
200
180
20
25
30
35
40
45
50

Operating the Helpers Cooperatively: Distributed MU-MIMO



Impact of Caching on Massive MIMO: distributed implementation

- Consider a distributed implementation of massive MIMO, with conjugate beamforming:

$$\mathbf{y} = \mathbf{H}^{\mathsf{H}}\mathbf{x} + \mathbf{z}, \quad \mathbf{x} = \mathbf{H}\mathbf{d}$$

- Each Antenna Terminal (AT) i needs to estimate the i-th row of $\mathbf{H} \in \mathbb{C}^{M \times K}$ from the orthogonal uplink pilots, and produce the local data linear combination

$$x_i = \sum_{j=1}^{K} h_{i,j} d_j$$

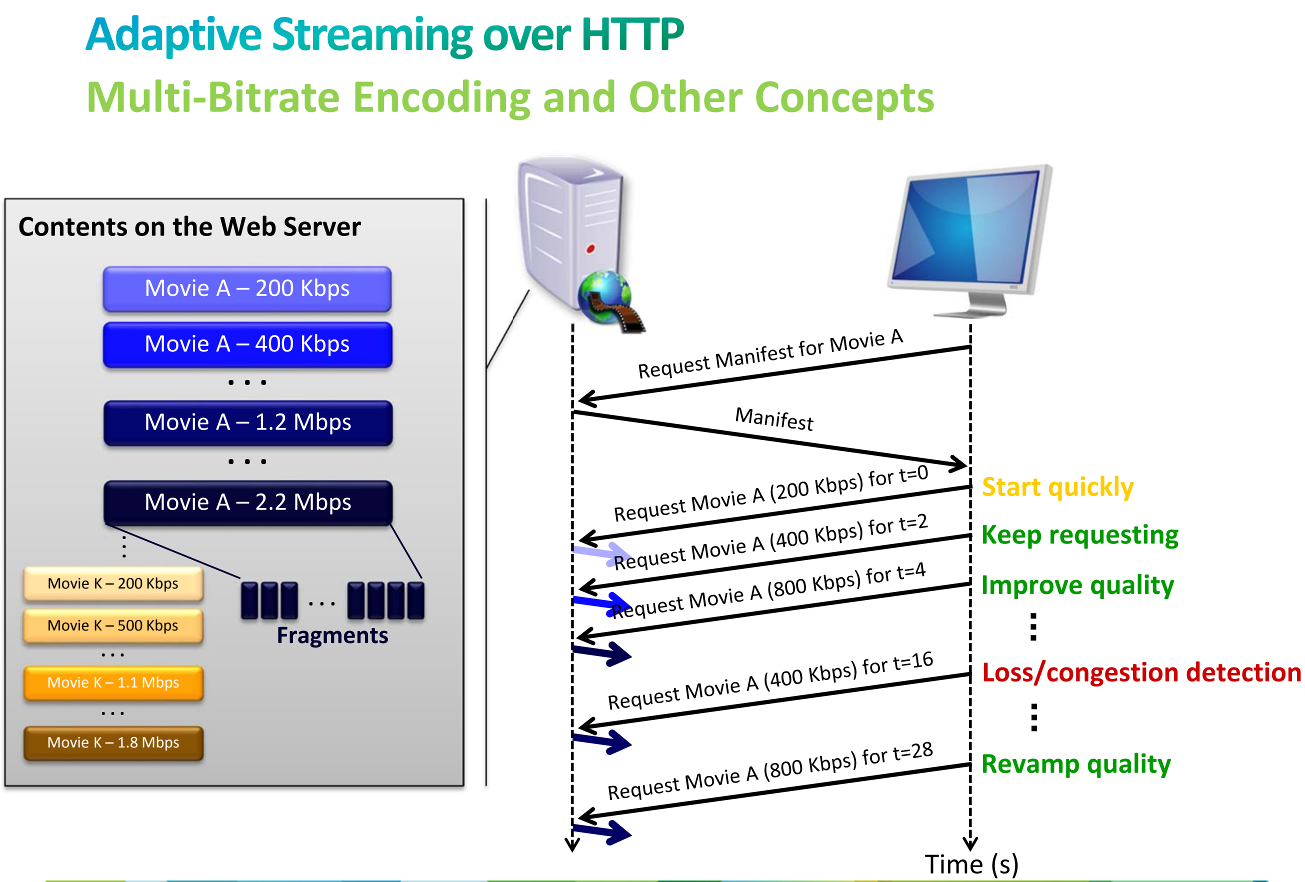
- In order to do so, the data for all K users must be delivered to each of the M ATs (big stress on the backhaul).

- If these data are cached in advance, we can operate the helper nodes cooperatively, without requiring a K-fold increase of the backhaul capacity.

Is the Average Downloading Delay Meaningful?

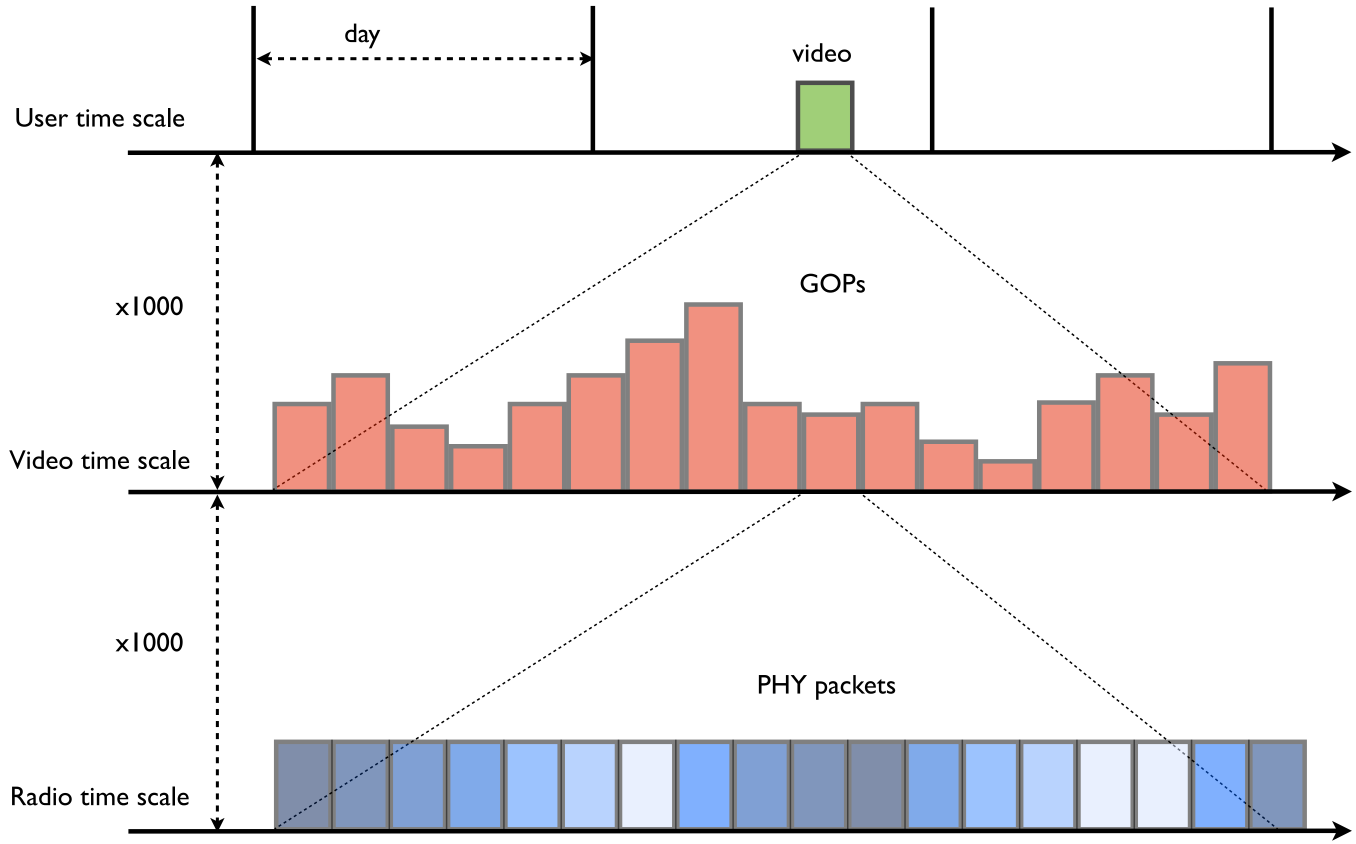
- Streaming is characterized by a small pre-buffering delay with respect to the total file playback time.
- Average downloading delay $\leq$ total video playback is a necessary condition for streaming without stall.
- Statistical fluctuations must be handled by scheduling, and are smoothed out by the playback buffer.
- Several common schemes: MicroSoft Smooth Streaming, Flash Dynamic Streaming, Apple HTTP Adaptive Bitrate Streaming.

DASH (Dynamic Adaptive Streaming over HTTP)

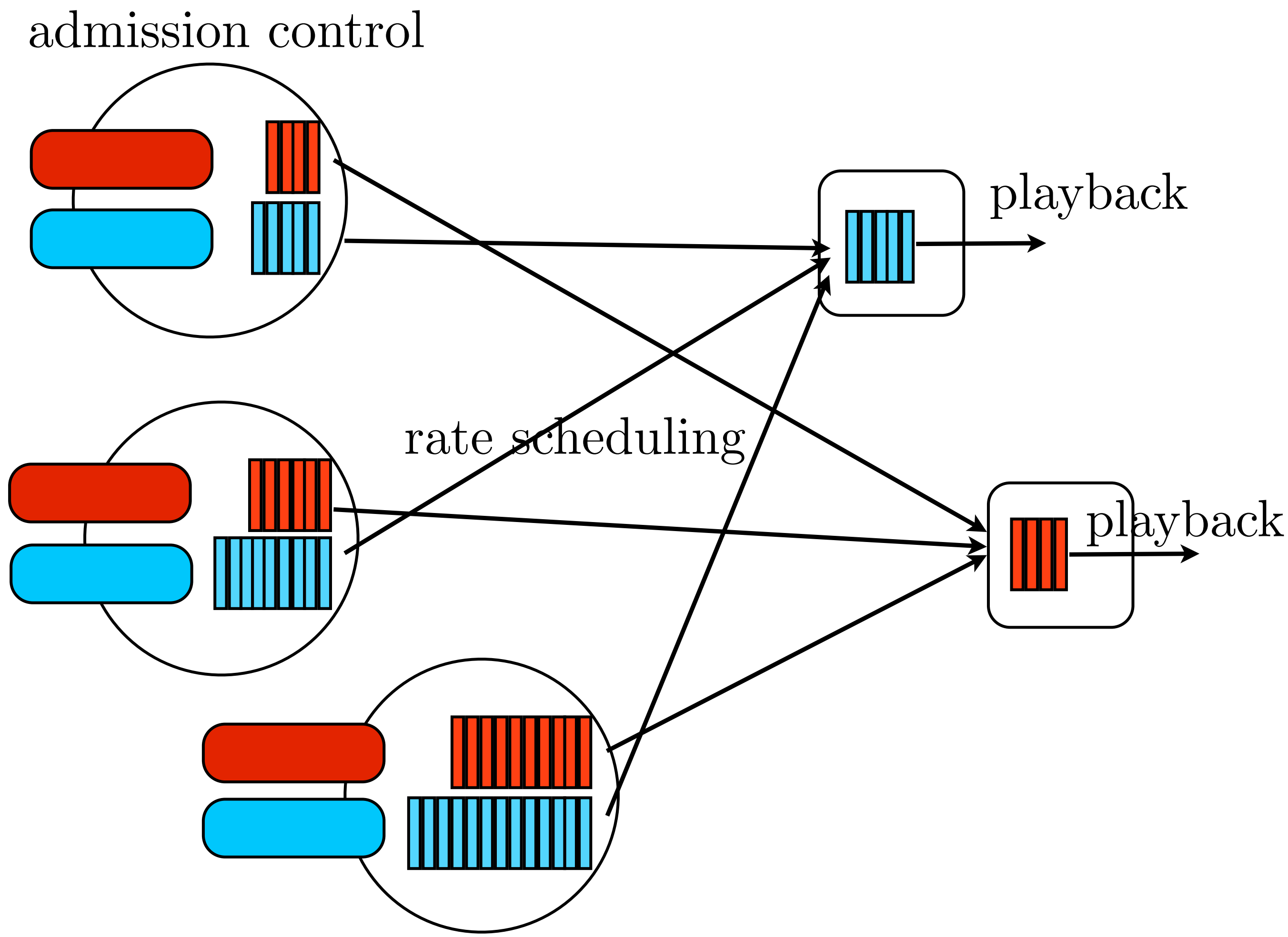


Time-Scale Decomposition

- We perform scheduling at the level of the video chunk (GOP).



Video-Aware Admission Controls and Scheduling



Dynamic Stochastic Optimization Problem

- The dynamics of the transmission queues at the helpers is given by:

$$Q_{hu}(t+1) = \max\{Q_{hu}(t) - n\mu_{hu}(t), 0\} + kR_{hu}(t), \quad \forall\, (h,u) \in \mathcal{E},$$

- Downlink rate region at each helper node:

$$\sum_{u \in \mathcal{N}(h)} \frac{\mu_{hu}(t)}{C_{hu}(t)} \leq 1, \quad \forall\, h \in \mathcal{H},$$

where

$$C_{hu}(t) = \mathbb{E}\left[\log\left(1 + \frac{P_h g_{hu}(t)|a_{hu}|^2}{1 + \sum_{h' \neq h} P_{h'} g_{h'u}(t)|a_{h'u}|^2}\right)\right].$$

(This corresponds to FDMA/TDMA orthogonal sharing of the downlink).

- Optimization Problem:

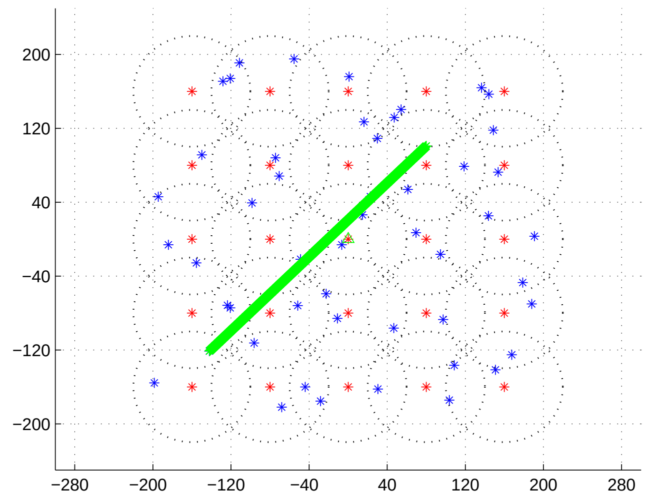
$$\begin{aligned} \text{maximize} \quad & \sum_{u \in \mathcal{U}} \phi_u(\overline{D}_u) \\ \text{subject to} \quad & \overline{Q}_{hu} < \infty \; \forall \; (h,u) \in \mathcal{E} \\ & \alpha(t) \in A_{\omega(t)} \; \forall \; t, \end{aligned}$$

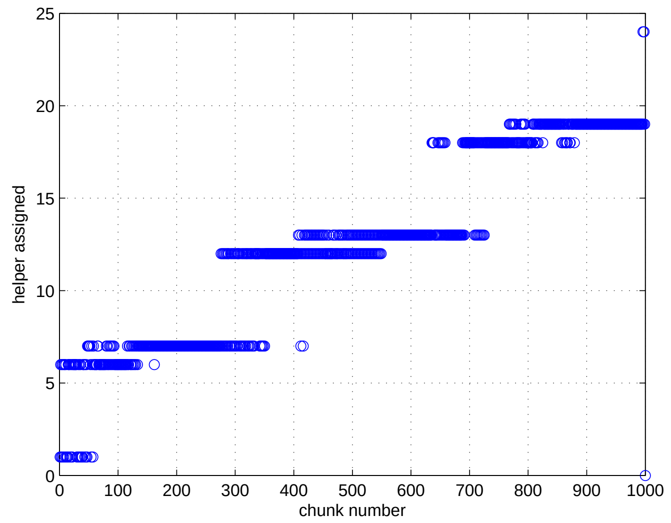
- We used the classical method of Liapunov Drift Plus Penalty (DPP).

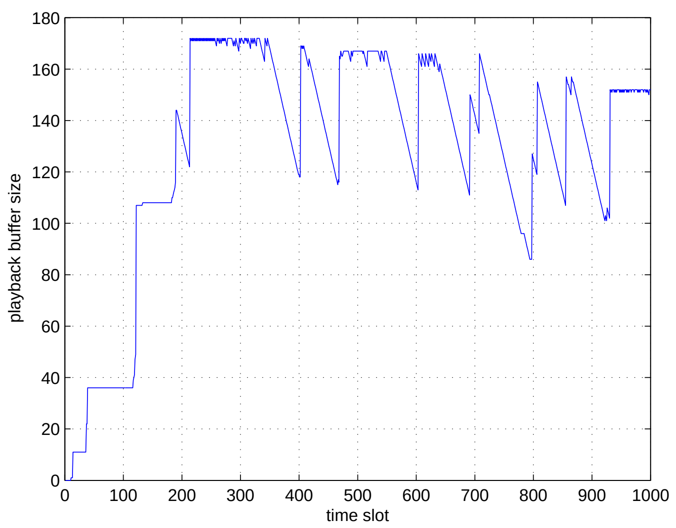
- The problem decomposes naturally into three decentralized subproblems: admission control, transmission scheduling, and greedy objective function maximization.

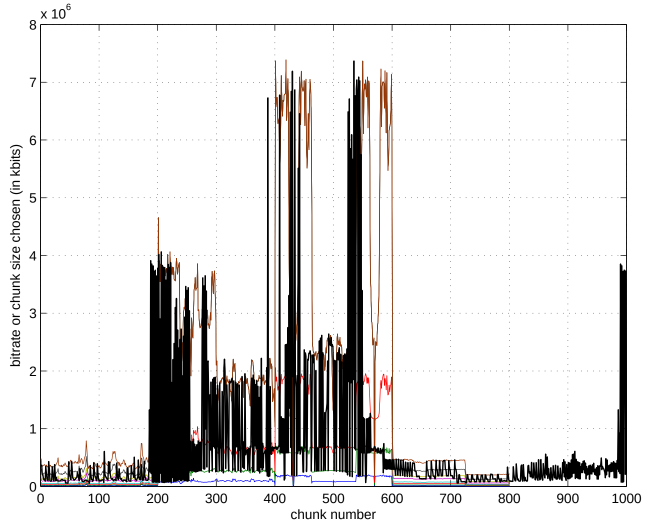
- Details in: Joint Transmission Scheduling and Congestion Control for Adaptive Video Streaming in Small-Cell Networks, ArXiv Preprint, submitted to IEEE Trans. on Commun., (2013).

Mobility experiment with VBR coded video



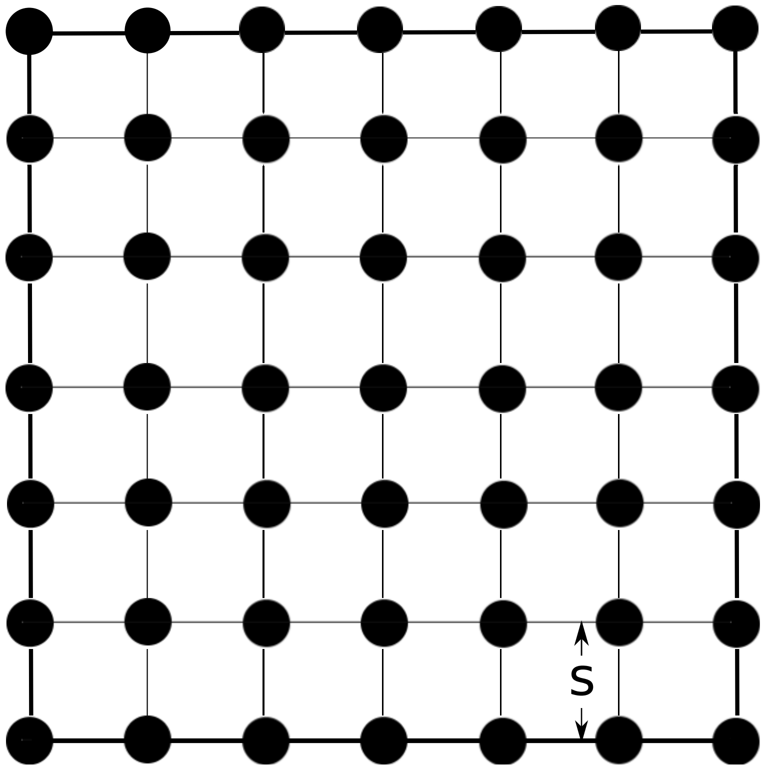


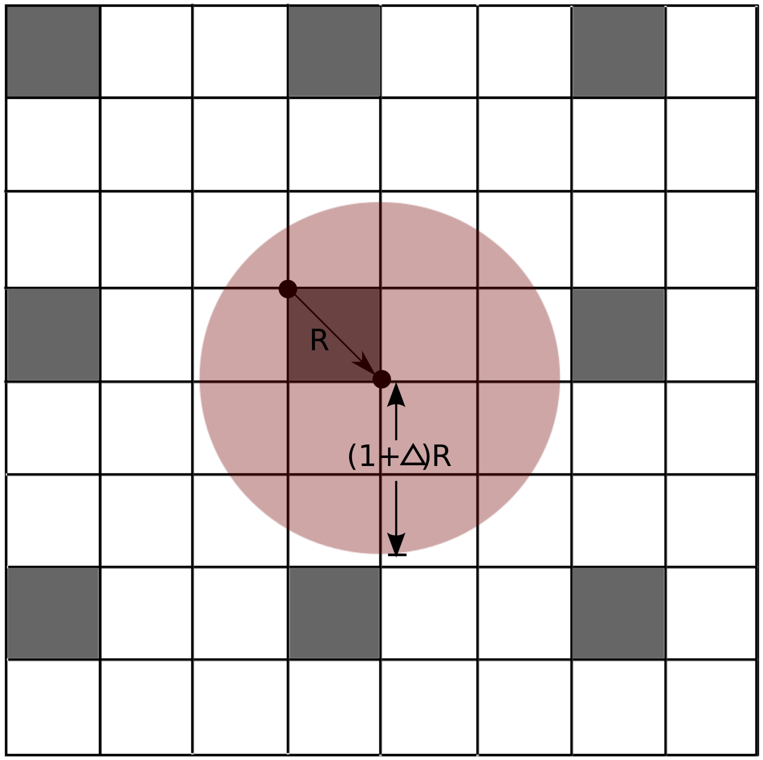




Throughput-Outage Tradeoff of One-Hop Caching Networks

- Dense network, distance $1/\sqrt{n}$, nodes on a grid, protocol model:





- Independent requests with a Zipf distribution $P_r(f) : f = 1, \dots, m$ with parameter $\gamma_r \in (0, 1)$.

- Interference avoidance transmission (independent set scheduling, by the protocol model).

- Random caching: each node cache at random, according to some probability distribution $P_c(f)$, up to M files.

- For a given set of scheduled links A, user u gets the rate

$$T_u = \sum_{v:(u,v)\in \mathsf{A}} c_{u,v} 1\{f_u \in \mathsf{G}(v)\}$$

- Minimum average per-user throughput:

$$\overline{T}_{\min} = \min_{u \in \mathcal{U}} \mathbb{E}[T_u]$$

- Number of users in outage:

$$N_o = \sum_{u \in \mathcal{U}} 1\{\mathbb{E}[T_u | \mathsf{f}, \mathsf{G}] = 0\}$$

- Outage probability:

$$p_o = \frac{1}{n}\mathbb{E}[N_o] = \frac{1}{n}\sum_{u \in \mathcal{U}} \mathbb{P}\left(\mathbb{E}[T_u | \mathsf{f}, \mathsf{G}] = 0\right).$$

- Throughput-Outage Tradeoff: the set of points $(T^*(p), p)$ solution of

$$\begin{array}{ll} \text{maximize} & \overline{T}_{\min} \\ \text{subject to} & p_o \leq p, \end{array}$$

(maximization with respect to the cache placement and transmission policies).

Tight Scaling Result

- In the regime $m, n \to \infty$, M finite, $\gamma_r < 1$, and $p \in (0,1)$ we have

$$T^*(p) = \Theta\left(\max\left\{\frac{M}{m}, \frac{1}{n}\right\}\right)$$

- Details in: Wireless Device-to-Device Caching Networks: Basic Principles and System Performance, ArXiv preprint and submitted to IEEE JSAC (2013).

  Optimal Throughput-Outage Trade-off in Wireless One-Hop Caching Networks, ArXiv preprint, to appear at IEEE ISIT (2013).

  Fundamental Limits of Distributed Caching in D2D Wireless Networks, ArXiv preprint, submitted to IEEE ITW (2013).

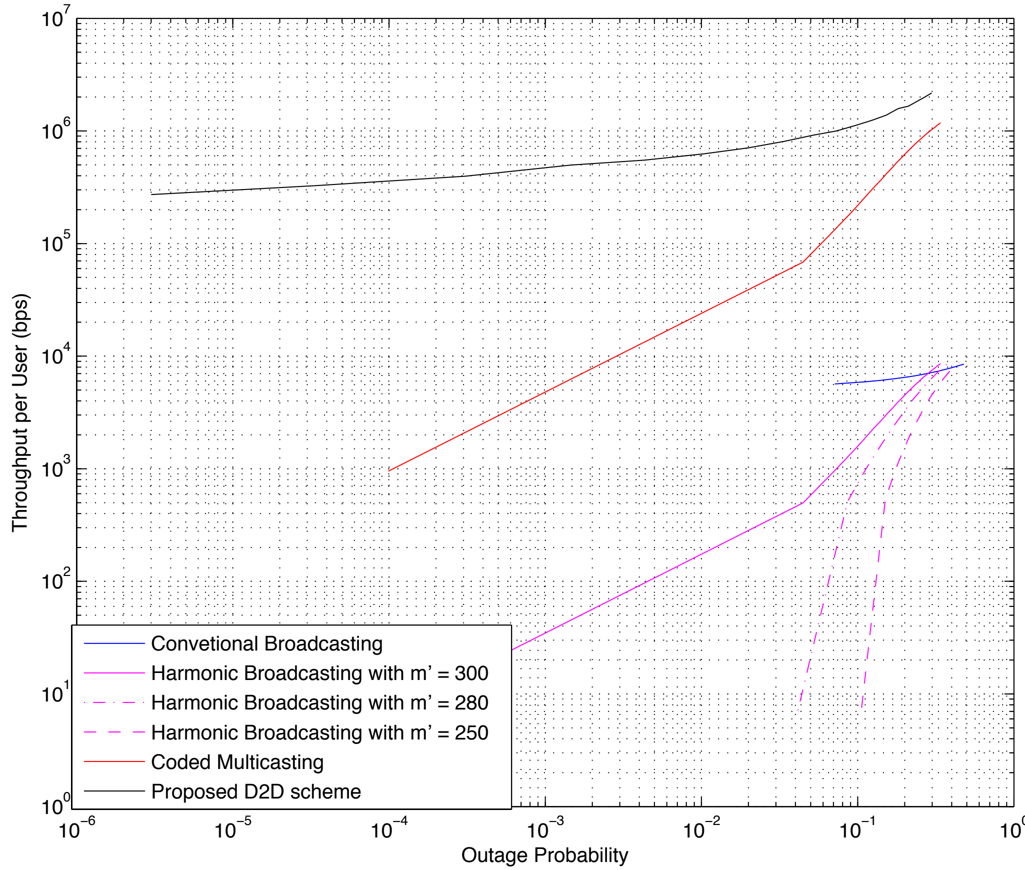
Competitor Schemes

- Conventional broadcasting (TCP connection for each individual streaming session), yields $\Theta\left(\frac{1}{n}\right)$.
- Harmonic broadcasting (UDP stream, from which all users grab what they need), yields $\Theta\left(\frac{1}{m \log L}\right)$.
- Coded multicasting (Maddah-Ali and Niesen, ArXiv 2012-2013) yields also

$$T_u = \Theta\left(\max\left\{\frac{M}{m}, \frac{1}{n}\right\}\right)$$

- Remarkably and surprisingly, coded multicasting from the base station and random caching with D2D spatial reuse achieve the **same order of throughput**. The difference is in the actual rates!!

Results (indoor outdoor campus scenario)

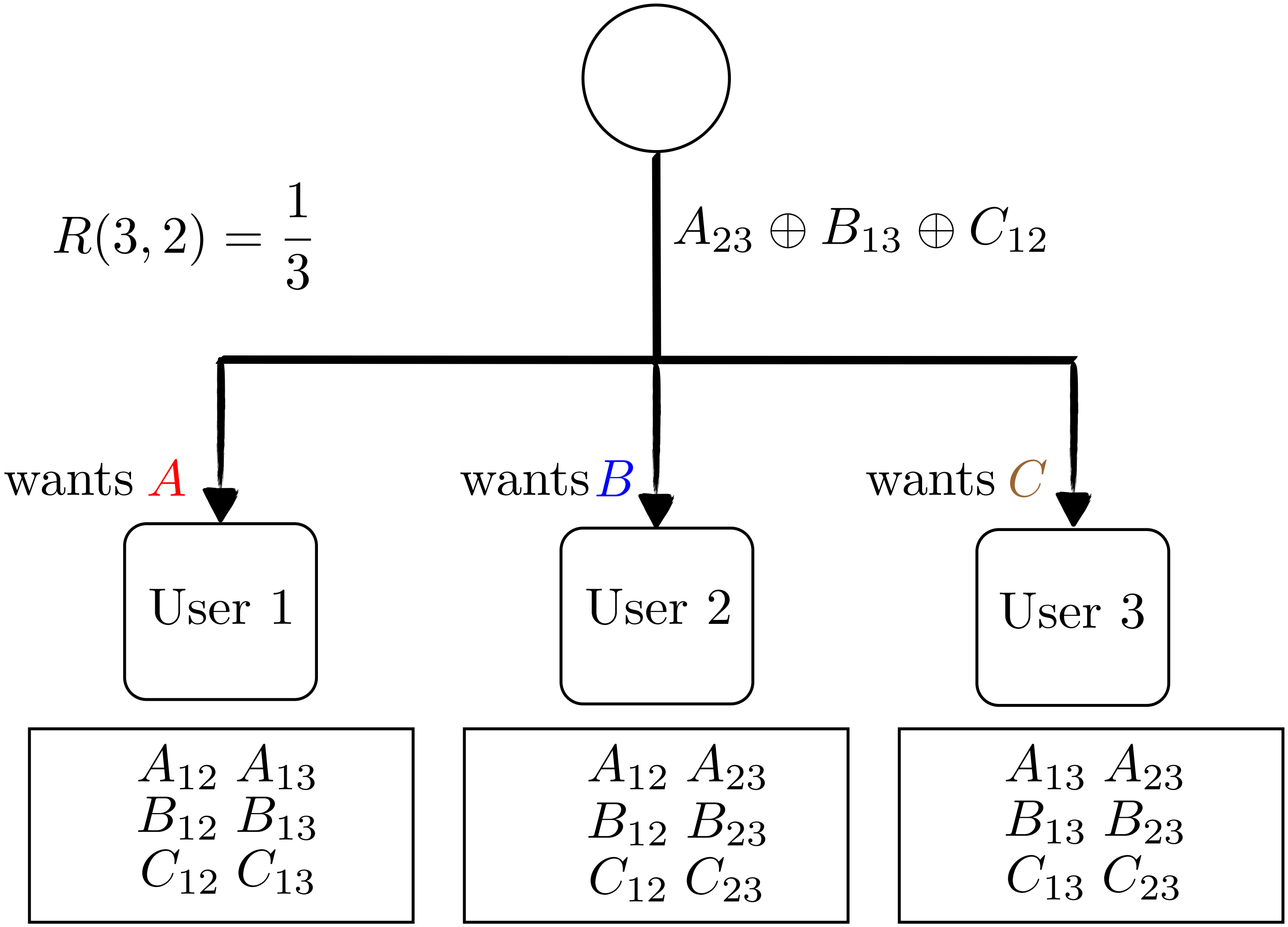


Simulation results for the throughput-outage trade-off for different schemes under the realistic indoor/outdoor propagation environment, $n = 10000$, $m = 300$, $M = 20$ and $\gamma_r = 0.4$.

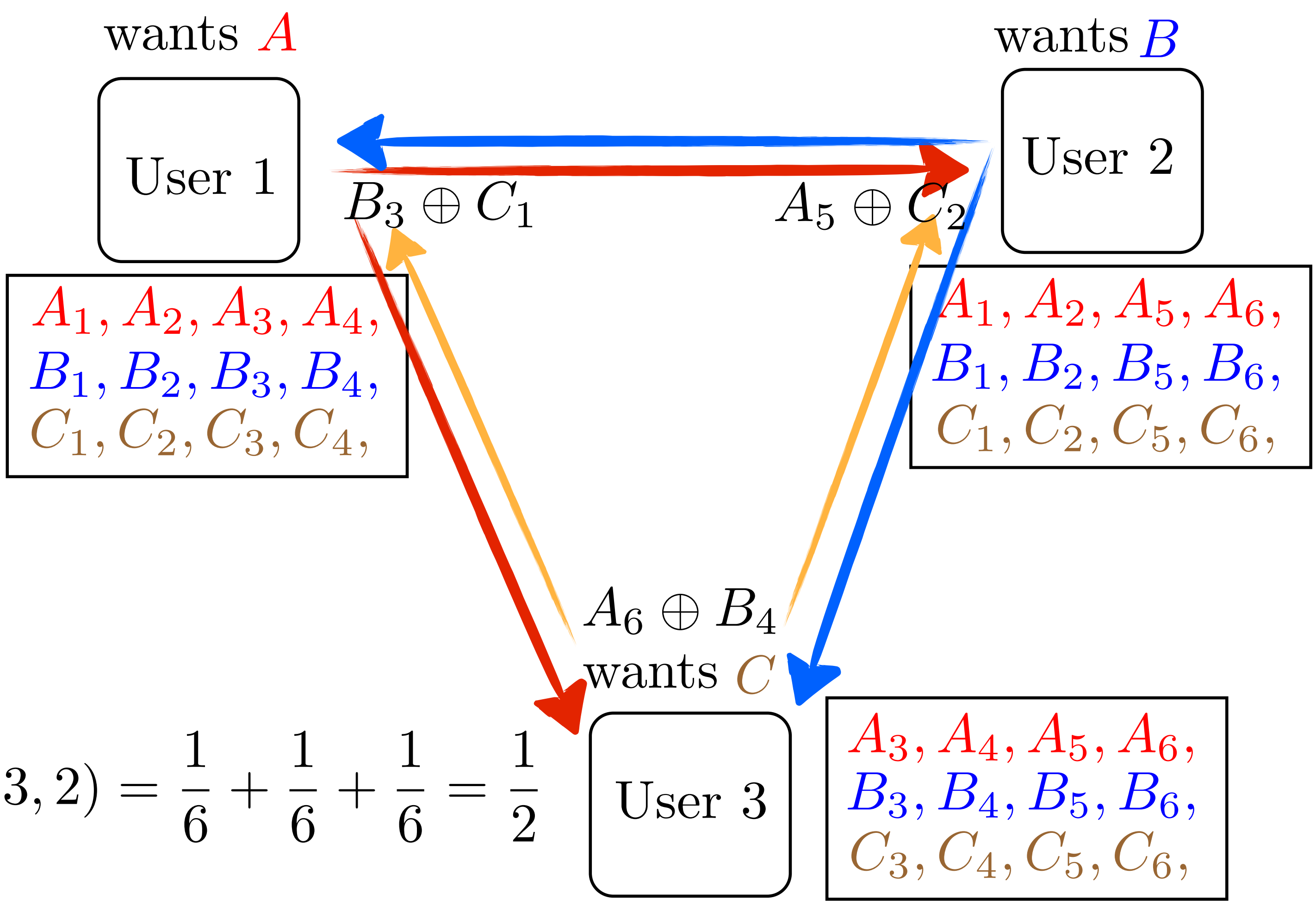
Can we combine coded multicasting and D2D reuse?

- A tempting idea: can we combine both gains?
- We have proposed a combinatorial (non-random) caching at the user (helper) nodes (ArXiv preprint).
- D2D network-coded delivery phase, tight result within a gap from information theoretic cut-set bound.
- Let's take a closer look at the Maddah-Ali and Niesen scheme.....

Coded Multicasting ($n = m = 3$, $M = 2$)



D2D Coded Delivery ($n = m = 3$, $M = 2$)



$$R(3,2) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

General Tight Results

- For the base-station coded multicasting scheme, the number of transmitted bits (normalized to the file size) is:

$$R(n, m, M) = n\left(1 - \frac{M}{m}\right)\frac{1}{1 + \frac{nM}{m}}$$

- For the D2D coded delivery scheme, the number of transmitted bits (normalized to the file size) is:

$$R(n, m, M) = n\left(1 - \frac{M}{m}\right)\frac{m}{nM}$$

- In the interesting regime $nM \gg m$ these quantities are almost identical.

- In both cases, the throughput behaves as:

$$T_u = \Theta\left(\max\left\{\frac{M}{m}, \frac{1}{n}\right\}\right)$$

- By clustering and replicating the scheme in space we loose the TDMA factor! **Coding and spatial reuse gains do not cumulate, at least in terms of scaling laws!**

Conclusions

- Exploiting the asynchronous content reuse of wireless data killer apps is key for achieving the required 100x.
- Caching at the wireless edge has a great potential, since it relaxes the constraints on the backhaul (expensive network component).
- We have proposed FemtoCaching (helper nodes), and D2D Caching network (caching at the user devices).
- We have developed optimal or near-optimal algorithms for cache placement, scheduling for adaptive video streaming, and D2D cluster-based interference avoidance link scheduling.
- Theoretical results and simulations show the effectiveness of the approach.
- Good news for LTE operators: new use of the macro-cellular base stations at off-peak times.

Thank You

Harmonic Broadcasting (example)