ICC 2017

Wireless Edge Caching: Promises and Recent Advances

Giuseppe Caire

Communications and Information Theory Group
Technische Universität Berlin

Alexander von Humboldt
Stiftung/Foundation

Paris, May 24, 2017

CommIT
Communications and Information Theory Chair

Motivation (1)

- Users like (Multi-Media) content.

Motivation (2)

- Typical Netflix movie: 2.5GB

Motivation (2)

- Typical Netflix movie: 2.5GB
- Typical streaming bit-rate:

$$R_b = \frac{2.5 \times 8 \times 10^9}{1.5 \times 3600} \approx 3.7 \text{ Mb/s}$$

Motivation (2)

- Typical Netflix movie: 2.5GB
- Typical streaming bit-rate:

$$R_b = \frac{2.5 \times 8 \times 10^9}{1.5 \times 3600} \approx 3.7 \text{ Mb/s}$$

- As a matter of fact, streaming at rates between 400 kb/s and 4Mb/s are “OK”

Motivation (2)

- Typical Netflix movie: 2.5GB
- Typical streaming bit-rate:

$$R_b = \frac{2.5 \times 8 \times 10^9}{1.5 \times 3600} \approx 3.7 \text{ Mb/s}$$

- As a matter of fact, streaming at rates between 400 kb/s and 4Mb/s are “OK”
- Focus of most “5G” development: Peak rates (multiple Gb/s)

Motivation (2)

- Typical Netflix movie: 2.5GB
- Typical streaming bit-rate:

$$R_b = \frac{2.5 \times 8 \times 10^9}{1.5 \times 3600} \approx 3.7 \text{ Mb/s}$$

- As a matter of fact, streaming at rates between 400 kb/s and 4Mb/s are “OK”
- Focus of most “5G” development: Peak rates (multiple Gb/s)
- Current typical LTE data plans: 3GB/month

Motivation (2)

- Typical Netflix movie: 2.5GB
- Typical streaming bit-rate:

$$R_b = \frac{2.5 \times 8 \times 10^9}{1.5 \times 3600} \approx 3.7 \text{ Mb/s}$$

- As a matter of fact, streaming at rates between 400 kb/s and 4Mb/s are “OK”
- Focus of most “5G” development: Peak rates (multiple Gb/s)
- Current typical LTE data plans: 3GB/month
- $\Rightarrow$ **There is a significant mismatch between user consumption demand and operator offers!**

Motivation (3)

Paradigm shift: from Gb/s to Tb/month

Motivation (3)

Paradigm shift: from Gb/s to Tb/month

- 1Tb/month means $R_b \approx 400$ kb/s for all users, 24/7.
- This must be stable (constant rate, low latency jitter, not really “low-delay”).
- Definitely not an easy task (impossible to do on LTE and most probably on 5G early deployment).
- **Especially difficult at 10EUR/month.**

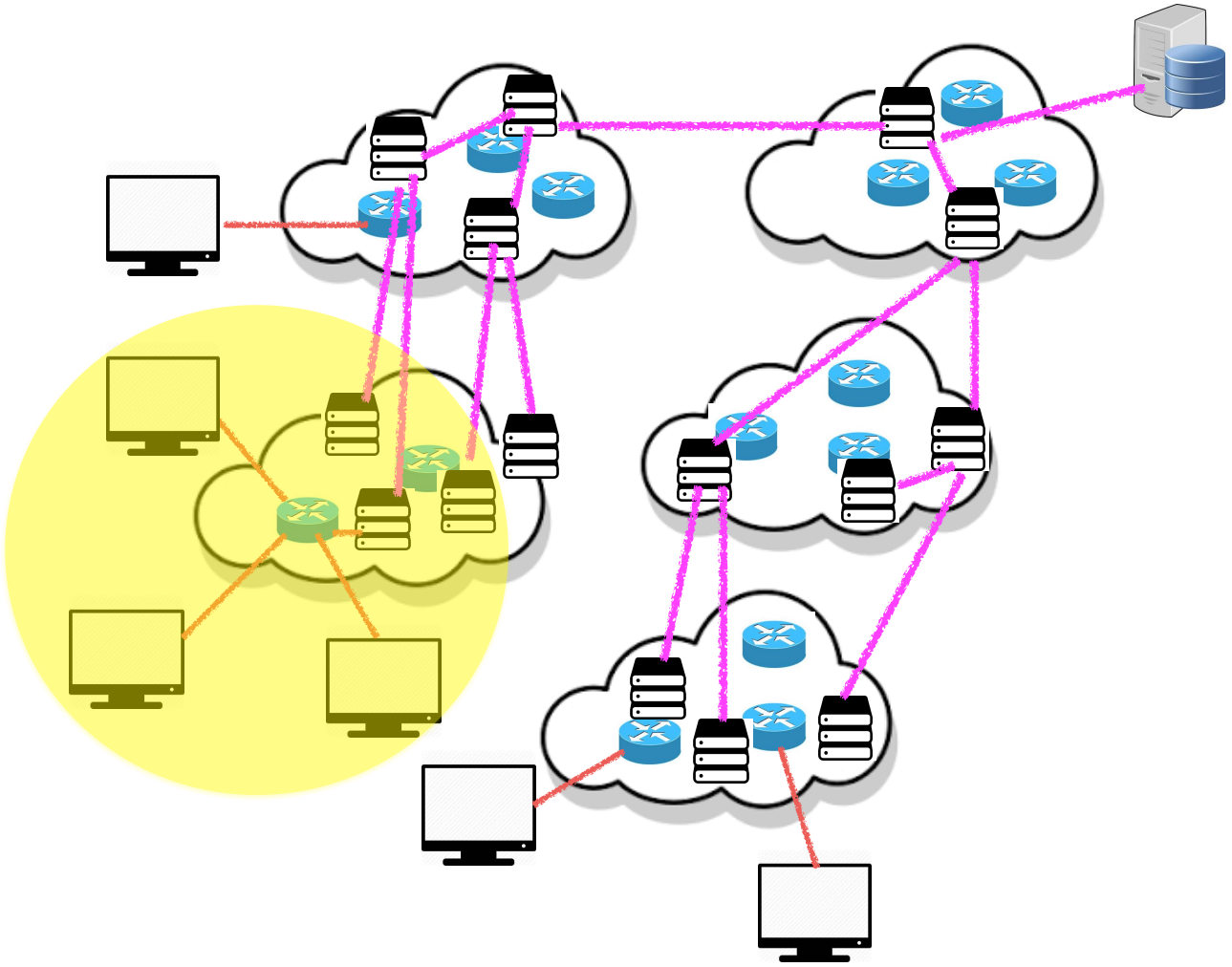
Motivation (4)

- Wifi is not a solution (yes .. I know .. it works at home ... but ...).
- Opportunity for “Internet companies” to reach their customers directly” (bypassing cable & cellular).
- Opportunity for “new comers” (startups) to get into a difficult and competitive market.

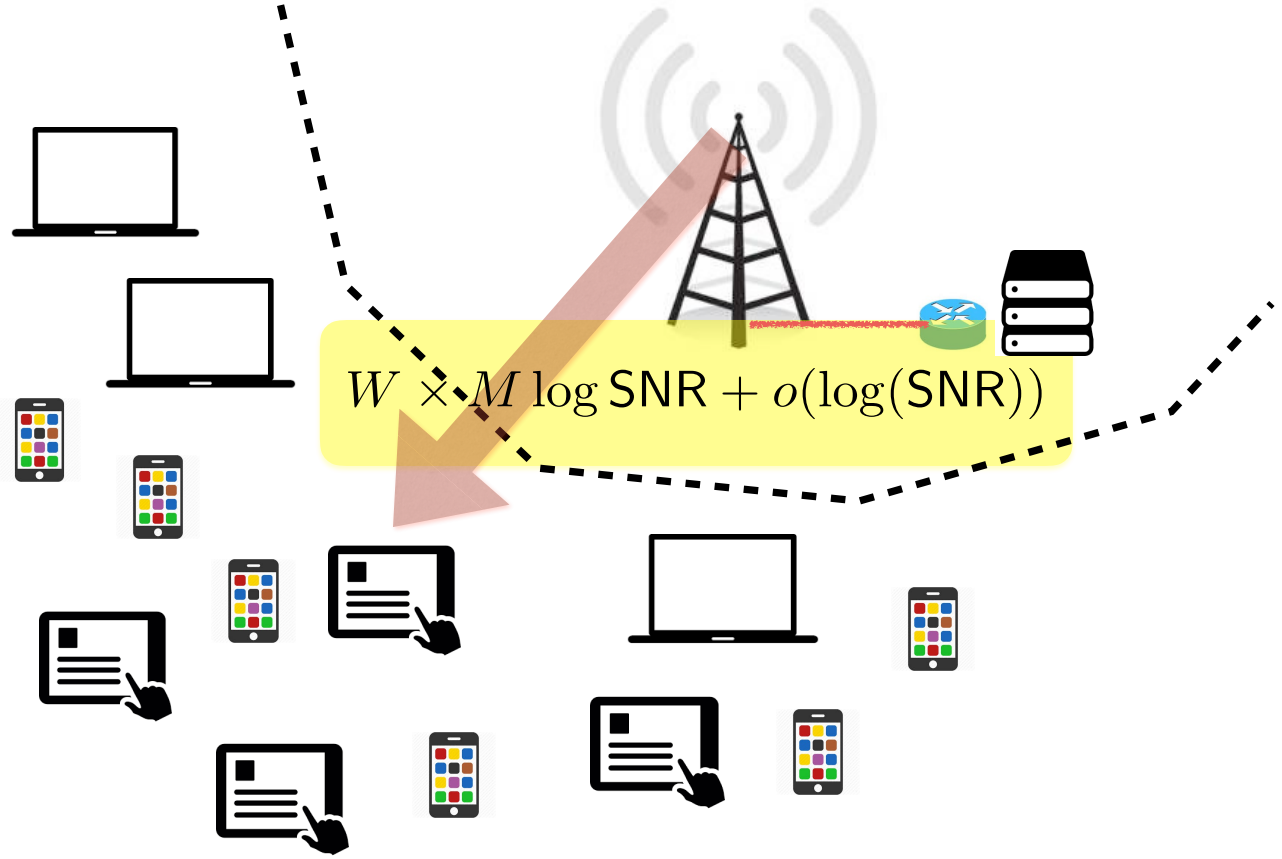
Caching: turning memory into bandwidth

- On-demand MM content has special features:
  1. **Asynchronous content reuse** (traffic generated by a few popular files, which are accessed in a totally asynchronous way).
  2. **Easy-to-learn/predict demand distribution** (Providers do this already!).
  3. **Easy-to-shape demand distribution** (restricted library refreshed at low rate ... Providers do this already).

Content Distribution over Internet (CDNs)



Wireless Edge Bottleneck



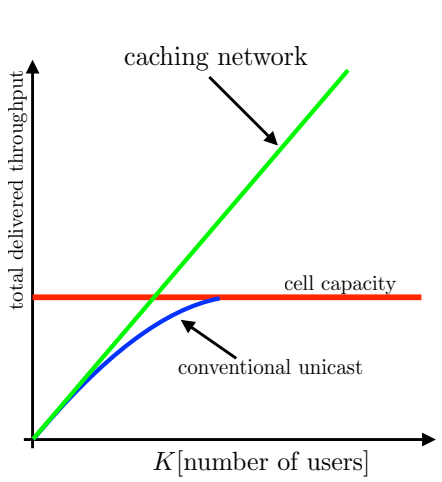
Can Caching help at the RAN Level?

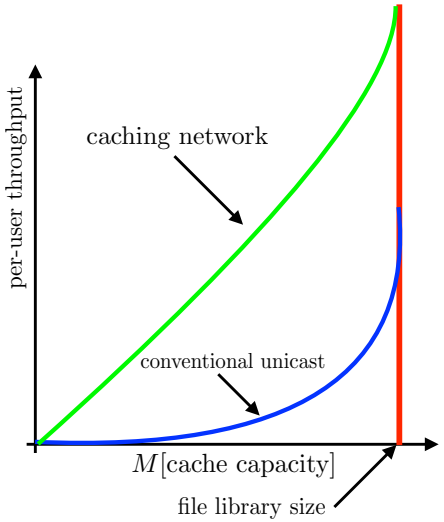
In this Talk

- Focus on the recent (2012 – 2017) set of results emerged in Information Theory / Wireless Comm.
- System assumptions: time-scale decomposition, “off-line caching”.
- Infrastructure-only approaches: the FemtoCaching and Cache-Aided PHY.
- Infrastructure-less approaches: leveraging spatial reuse.
- Coded caching: leveraging network coding .... best of both worlds?

In this Talk

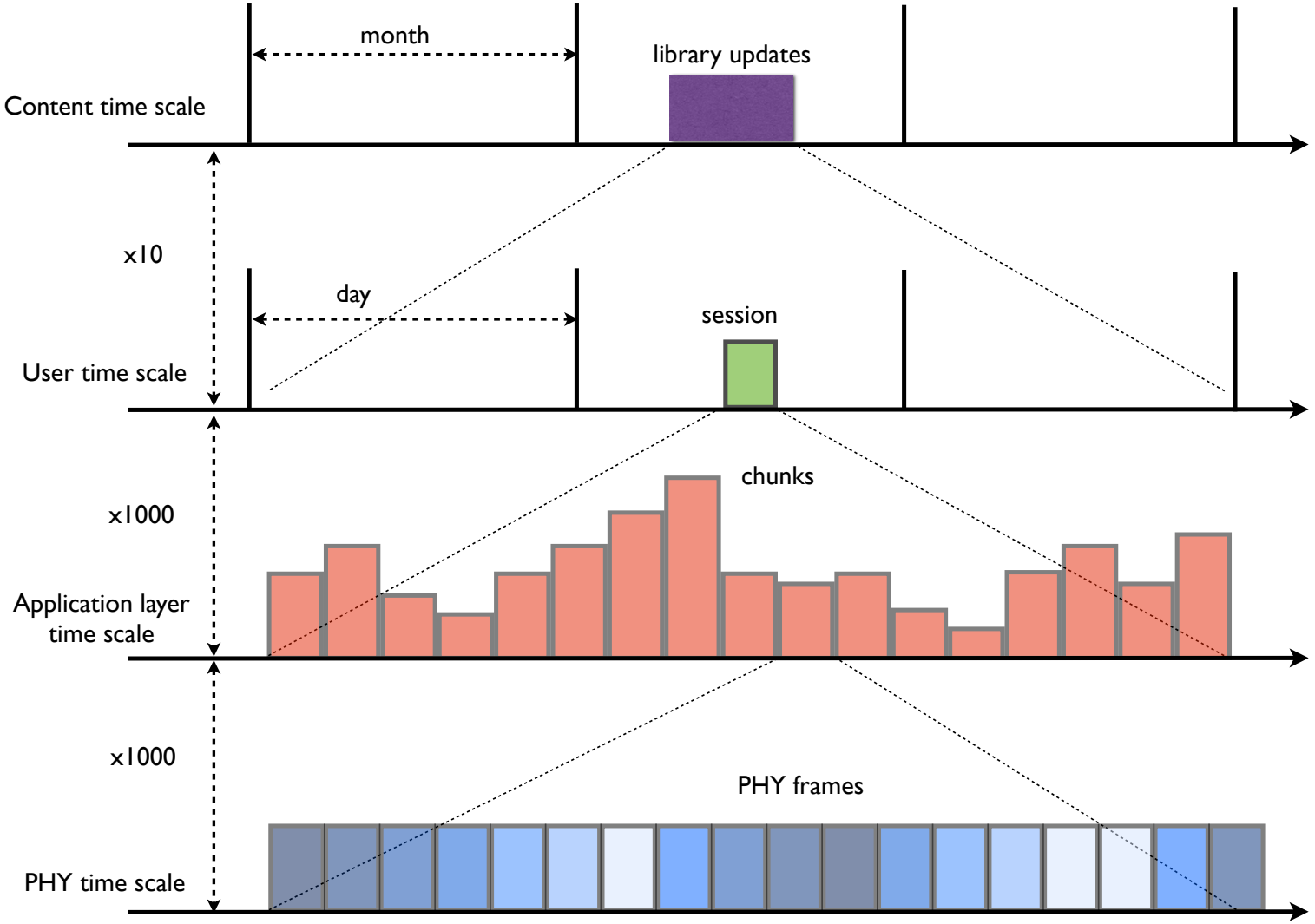
- Focus on the recent (2012 – 2017) set of results emerged in Information Theory / Wireless Comm.
- System assumptions: time-scale decomposition, “off-line caching”.
- Infrastructure-only approaches: the FemtoCaching and Cache-Aided PHY.
- Infrastructure-less approaches: leveraging spatial reuse.
- Coded caching: leveraging network coding .... best of both worlds?
- Holy grail: **achieving scalability**



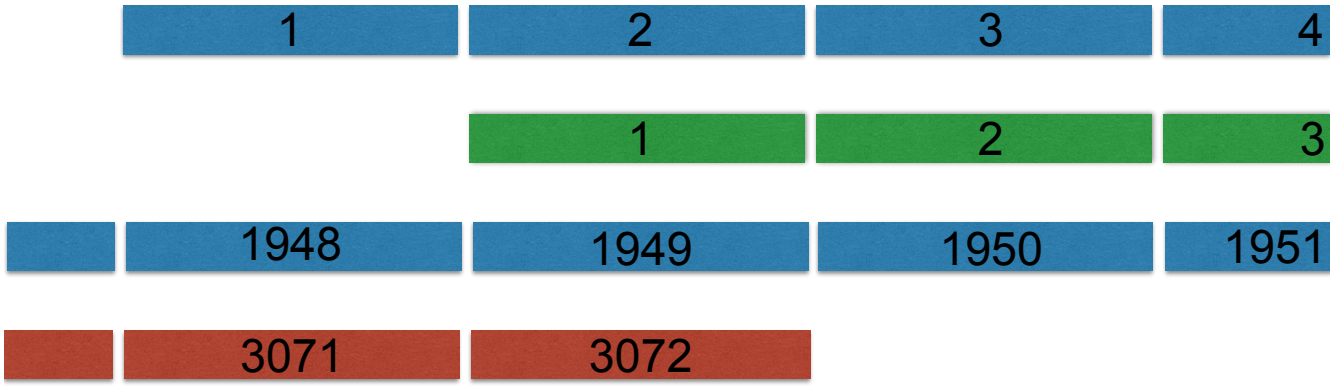


Assumptions

Timescale decomposition



Asynchronous Content Reuse



- Unlike live TV, the probability that two two users access the same content item at the same time is negligible.
- In contrast, multiple users may demand the same content item at different times.
- Mathematically: measure caches in "files", where each file is formed by multiple "content items", and let the number of content items per file be sufficiently large such that $N \geq K$.
- **This prevents from taking advantage of "naive multicasting" (number of content items less than number of users)**.

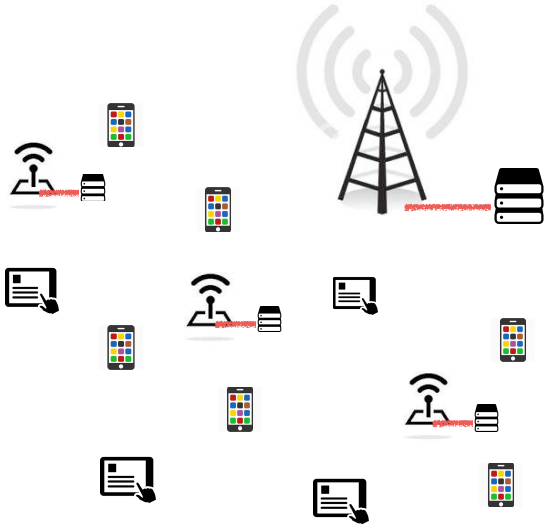
Non-Dynamic Content-Centric Networking

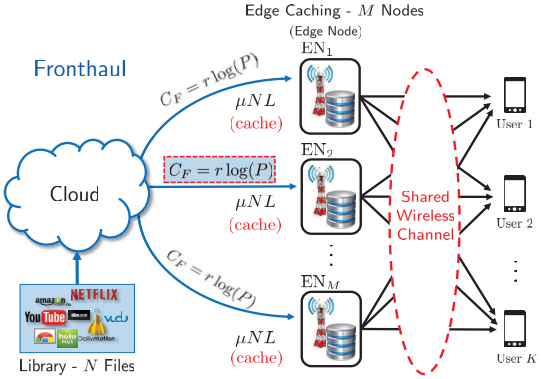
- Fixed library of content items (files).
- Content items (or functions thereof) are placed in the network in some caching nodes: off-line caching.
- Users access content by identity and not by location.
- Cache Placement: done a priori, knowing the network topology and possibly the demand distribution, but not the realization of the demand vector.
- Content Delivery: given a realization of the demand vector, the network satisfies the users' demands by sending (coded) messages.
- Goal: minimize load $\sim$ minimize latency of delivery $\sim$ maximize per-user "goodput".

Approaches

Tremendous Interest – Lots of Papers! (1)

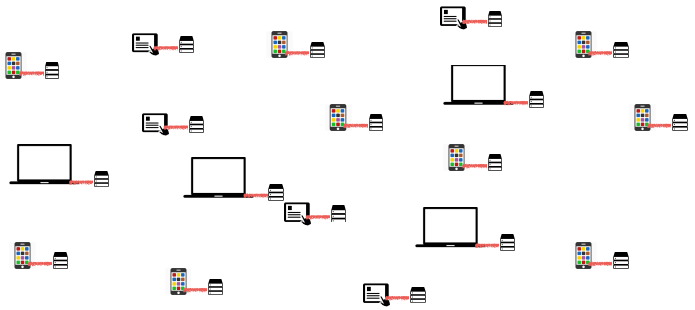
- FemtoCaching: Caching at the infrastructure side (SBS, Helpers)
  - Golrezaei, Shanmugam, Dimakis, Molisch, and GC, 2012, March. "Femtocaching: Wireless video content delivery through distributed caching helpers." In INFOCOM, 2012 Proceedings IEEE (pp. 1107-1115). IEEE. (see also: IEEE Trans. on IT, 2013)
  - Sengupta, Tandon, and Simeone, 2016. "Cloud and cache-aided wireless networks: Fundamental latency trade-offs." arXiv preprint arXiv:1605.01690.





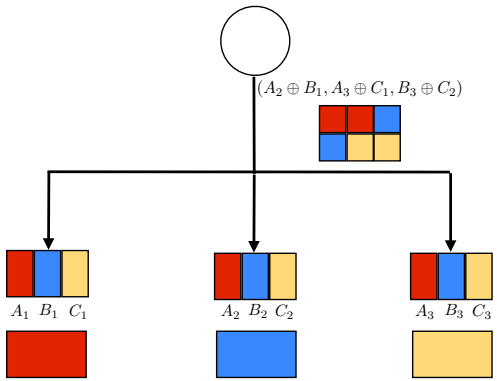
Tremendous Interest – Lots of Papers! (2)

- D2D Caching: Caching in the user devices, direct Device-to-Device transmission
  - Gitzenis, Paschos, and Tassiulas, 2013. “Asymptotic laws for joint content replication and delivery in wireless networks.” IEEE Trans. on IT, 59(5), pp.2760-2776.
  - Ji, GC and Molisch, 2015. “The throughput-outage tradeoff of wireless one-hop caching networks.” IEEE Trans. on IT, 61(12), pp.6833-6859.
  - Jeon, Hong, Ji, GC and Molisch, 2017. “Wireless multihop device-to-device caching networks.” IEEE Trans. on IT, 63(3), pp.1662-1676.
  - Liu, A., Lau, V. and GC, 2016. Cache-induced Hierarchical Cooperation in Wireless Device-to-Device Caching Networks. arXiv preprint arXiv:1612.07417.



Tremendous Interest – Lots of Papers! (3)

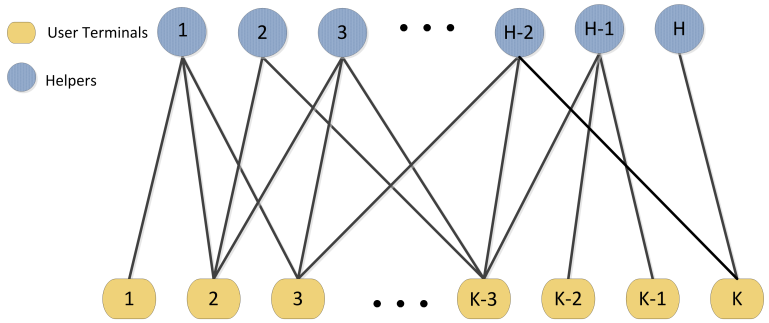
- Coded Caching: Caching in the user devices, delivery through network coding
  - Maddah-Ali, and Niesen, 2014. “Fundamental limits of caching.” IEEE Trans. on IT, 60(5), pp.2856-2867.
  - Maddah-Ali, and Niesen, 2015. “Decentralized coded caching attains order-optimal memory-rate tradeoff.” IEEE/ACM ToN, 23(4), pp.1029-1040.
  - Naderializadeh, Maddah-Ali, and Avestimehr, 2017. “On the Optimality of Separation between Caching and Delivery in General Cache Networks.” arXiv preprint arXiv:1701.05881.



FemtoCaching

Femtocaching (1)

- Network represented as a bipartite graph $\mathcal{G}(\mathcal{H}, \mathcal{U}, \mathcal{E})$, with a matrix of link downloading delays $\mathbf{\Omega} = [\omega_{\ell,k}]$ (s/bit).
- Library $\mathcal{W} = \{W_1, \ldots, W_N\}$ of N content items of equal size B bits.
- Infrastructure node 0 (Server/BS) has the whole library, RRH nodes $\ell \neq 0$ have cache memory size equivalent to M content items (i.e., MB bits).
- Known demand distribution $\{P_n : n = 1, \ldots, N\}$ such that $P_n = \mathbb{P}(d_k = n)$.
- Problem: Minimize average delivery (sum-)delay w.r.t. content placement.



Femtocaching (2)

- Let $\mathbf{X} = [x_{n,\ell}] \in \{0,1\}^{N \times H}$ is the (integer-valued) placement matrix. The Av. delay of user k is:

$$D_k = \sum_{\ell \in \mathcal{H}(k) \setminus \{0\}} \omega_{(\ell)_k,k} \sum_{n=1}^{N} \left[\prod_{i=1}^{\ell-1} (1 - x_{n,(i)_k}) \right] x_{n,(\ell)_k} P_n + \omega_{0,k} \sum_{n=1}^{N} \left[\prod_{i=1}^{|\mathcal{H}(k)|-1} (1 - x_{n,(i)_k}) \right] P_n$$

- Problem:

$$\text{maximize} \quad \sum_{k=1}^{K} (\omega_{0,k} - D_k)$$

$$\text{subject to} \quad \sum_{n=1}^{N} x_{n,\ell} \leq M \quad \forall \ \ell, \qquad \mathbf{X} \in \{0,1\}^{N \times H}$$

- Results: NP-Complete, but efficient approximation possible (monotone sub-modular function with matroid constraint).

Femtocaching (3)

- Intra-session coding: Using MDS coding, content items become "fluid": an item is delivered when enough coded symbols are received. Let $\mathbf{R} = \rho_{n,\ell}] \in \mathbb{R}_{+}^{N \times H}$ be the fractional placement matrix. The Av. delay of user k for file n is:

$$D_k^n = \omega_{(\ell)_k,k} - \sum_{i=1}^{\ell-1} \rho_{n,(i)_k} (\omega_{(\ell)_k,k} - \omega_{(i)_k,k}), \quad \text{if} \quad \sum_{i=1}^{\ell-1} \rho_{n,(i)_k} < 1 \leq \sum_{i=1}^{\ell} \rho_{n,(i)_k}$$

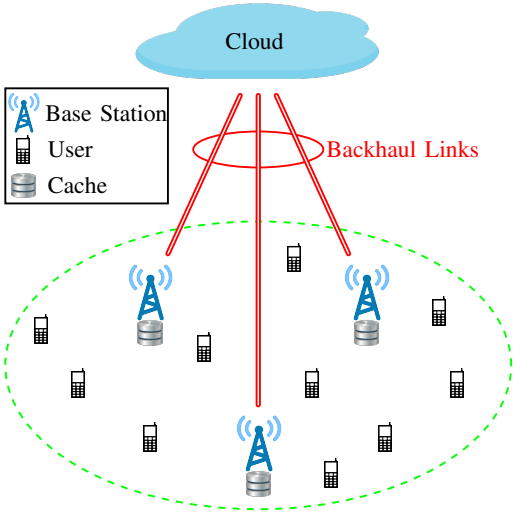
- Observation: D_k^n is the pointwise maximum of affine functions of $\mathbf{R}$ and therefore is convex.
- Convex Optimization Problem:

$$\text{minimize} \quad \sum_{k=1}^{K} \sum_{n=1}^{N} D_k^n P_n$$

$$\text{subject to} \quad \sum_{n=1}^{N} \rho_{n,\ell} \leq M \quad \forall \ \ell, \quad \mathbf{R} \in \mathbb{R}_{+}^{N \times H}$$

Joint Placement and PHY Optimization

- Optimization of sparse beamforming subject to individual SINR constraints:
  cost of backhaul C_{bh} decreases with sparsity
  cost of Tx power C_{tx} decreases with sparsity
- Minimize $C_{bh} + \lambda C_{tx}$, and obtain a backhaul/Tx power tradeoff Pareto boundary.

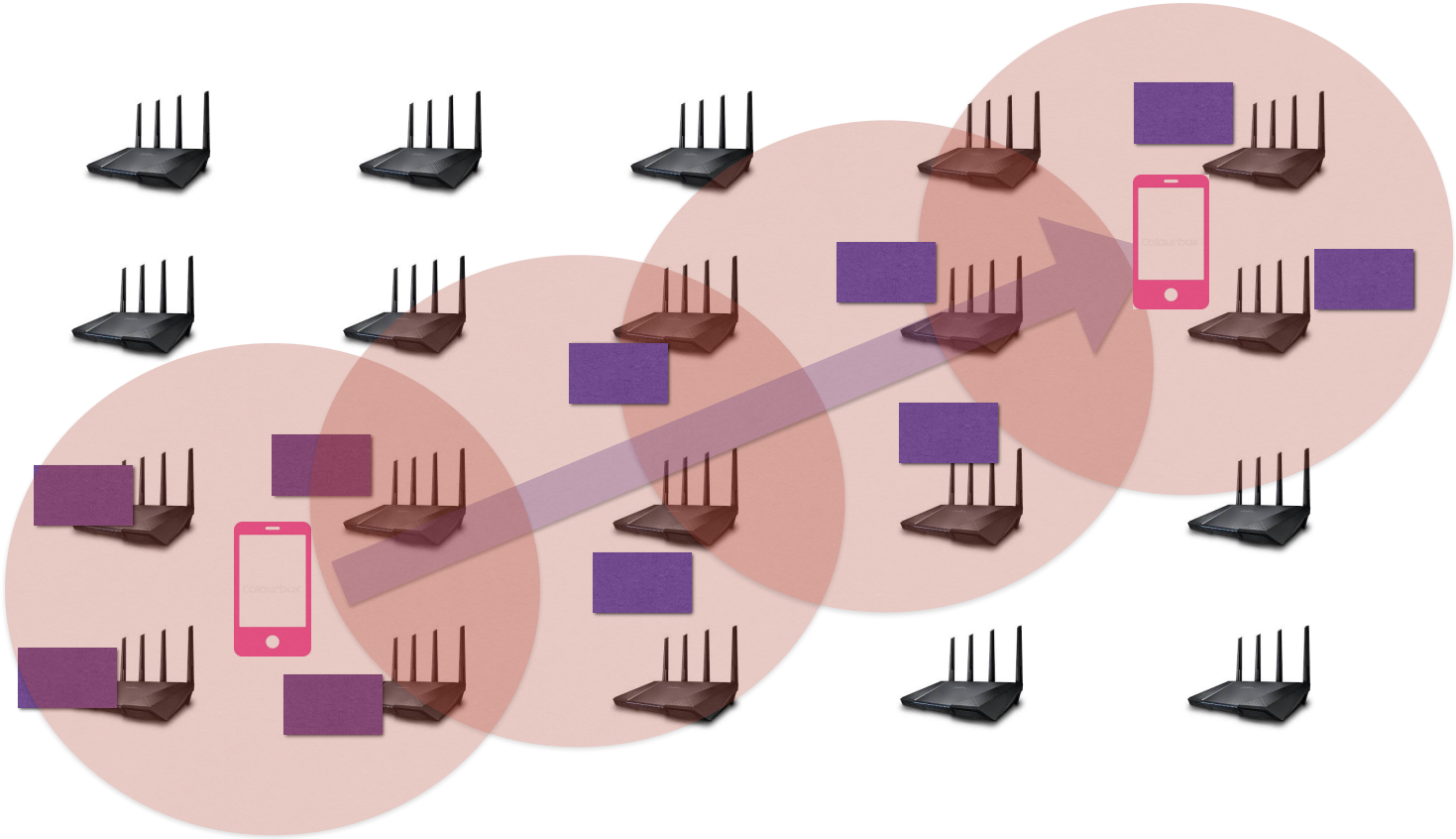


Ugur, Y., Awan, Z.H. and Sezgin, A., 2016, March. "Cloud radio access networks with coded caching." In Smart Antennas (WSA 2016); Proceedings of the 20th International ITG Workshop on (pp. 1-5). VDE.

Tao, M., Chen, E., Zhou, H. and Yu, W., 2016. "Content-centric sparse multicast beamforming for cache-enabled cloud RAN." IEEE Transactions on Wireless Communications, 15(9), pp. 6118-6131.

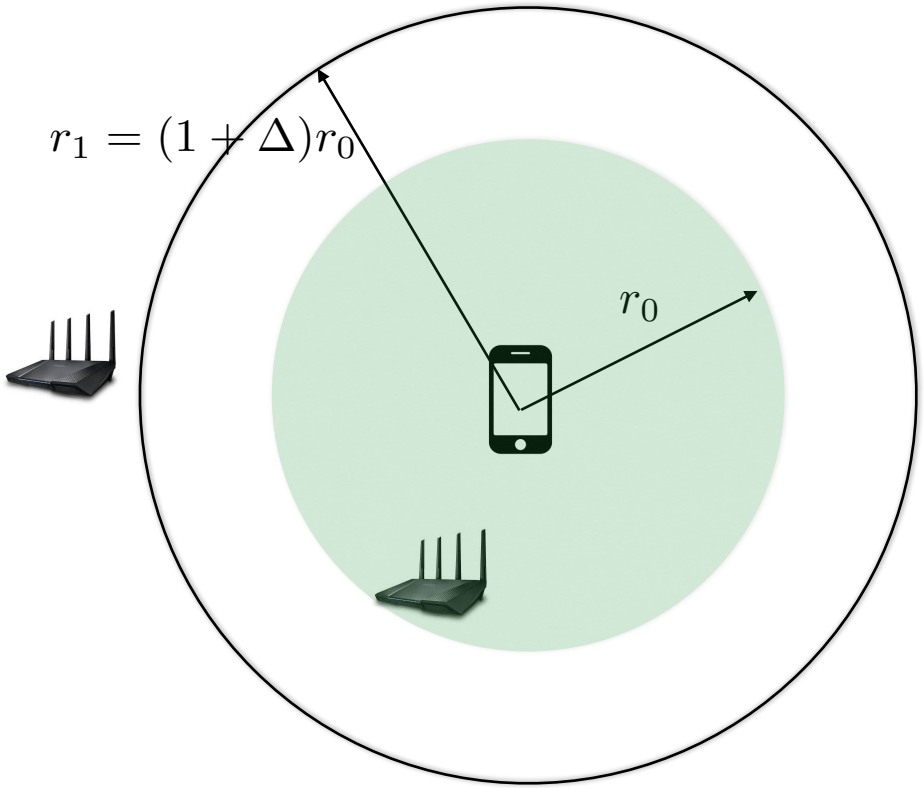
Implicit Sparse BF by Pilot Collision Control

- Bursalioglu, Wang, Papadopoulos, and GC, 2016, May. “RRH based massive MIMO with on the Fly pilot contamination control.” IEEE ICC (pp. 1-7).



Scaling Laws of FemtoCaching (1)

- Assume a Zipf “heavy tail” demand distribution $P_n = \frac{n^{-\tau}}{\sum_{m=1}^{N} m^{-\tau}}$ with $\tau < 1$.
- Assume the “protocol model” for successful transmission



Scaling Laws of FemtoCaching (2)

- Using the results from "one-hop" D2D wireless networks, it is immediate to conclude that if $H = O(K)$ and $N^{\frac{1-\tau}{2-\tau}} = o(H)$, the per-user delivery delay as $K \to \infty$ is

$$D = \Theta\left(1/\mu\right)$$

  where $\mu = M/N$ is the per-RRH fractional cache memory, with no backhaul or centralized server.

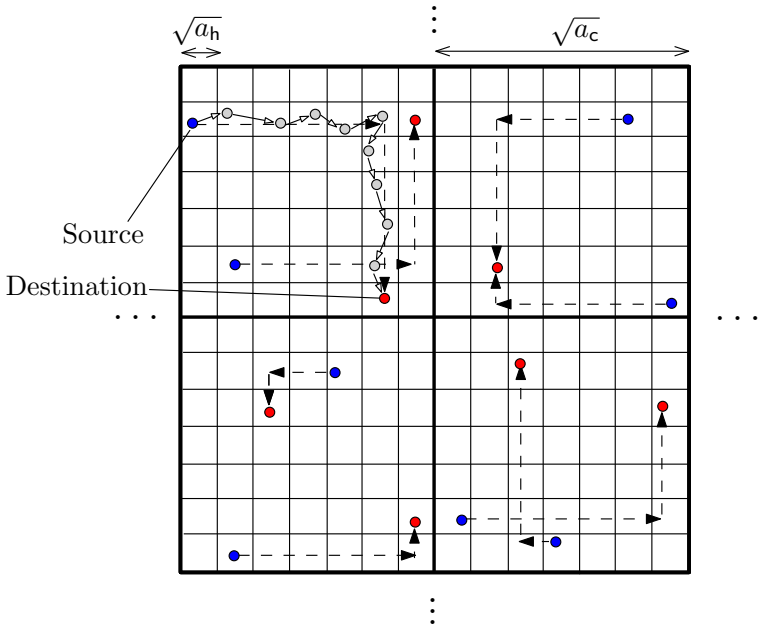
- Intuition:
  - Suppose $N = \nu MH$ for some $\nu \in (0, 1]$ ... $\Rightarrow$ the entire library can be cached in a cooperation cluster of size νH.
  - Divide the network in disjoint clusters each containing $\nu H = 1/\mu$ RRH and $\approx \nu H(K/H)$ users, and activate one user per cluster in round robin ... $\Rightarrow$ delivery delay $O(\nu K) = O(1/\mu)$.

D2D Caching

D2D Caching: Getting Rid of the Infrastructure

- With the advent of Device-to-Device communications (is it coming? LTE-D2D, 5G-D2D), and the ever-growing on-board memory of wireless devices, we can replace the RRH in a FemtoCaching network with the user devices (which becomes at the same time users and helpers).

- Key Idea: content replication and multi-hop.



D2D Caching: Scaling Laws under the Protocol Model

- For the protocol model, we have

$$D = \Theta(\sqrt{1/\mu})$$

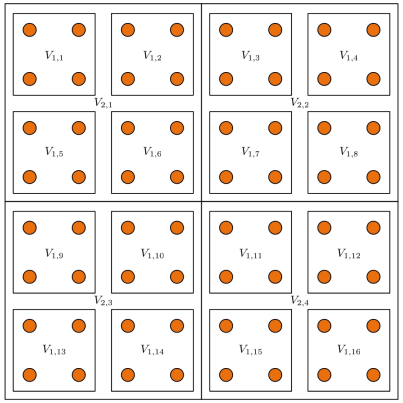
with is significantly better than the one-hop network (under the physical model restricting to multihop the same can be achieved).

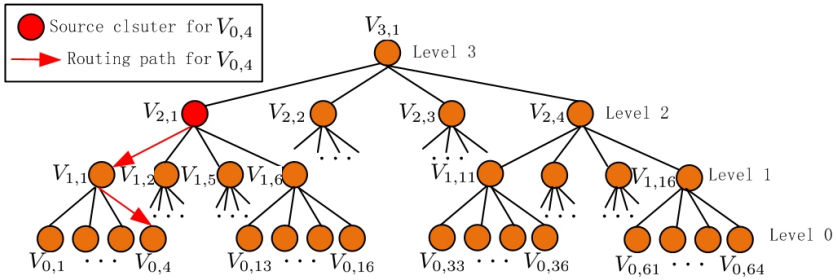
- Intuition:
  - Suppose $N = \nu MK$ for some $\nu \in (0, 1]$ ... $\Rightarrow$ the entire library can be cached in a cooperation cluster of size νK.
  - Divide the network in disjoint clusters of size $\nu K = 1/\mu$, and serve the users in each cluster by multihop.
  - Gupta and Kumar IT-2000 famous result: per-user throughput $\Theta(\frac{1}{\sqrt{\nu K}})$ ... $\Rightarrow \quad D = \Theta(\sqrt{1/\mu})$.

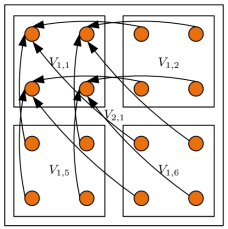
D2D Caching: Cache-Enabled Hierarchical Cooperation

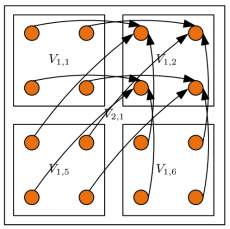
- The network is organized into an L-level hierarchy of nested partitions with groups size $1, 4, 4^2, 4^3, \ldots, 4^L = K$.
- Content items are broken into chunks. Content item at level ℓ is broken into 4^ℓ chunks, and replicated in each group of the partition at level ℓ.
- More probable files are allocated to lower levels (i.e., higher rate of replication in the network).
- Delivery phase: let user k demand content item n at hierarchy level ℓ. The group of 4^ℓ nodes containing user k "aggregates" the packets of the requested content item onto user k. Aggregation occurs in successive steps (in this case ℓ steps).
- Cache-enabled cooperative MIMO: just as in Ozgur, Leveque, and Tse, IT2010, the group of users sharing chunks of a given content item operate as a giant distributed antenna array, and create a point-to-point MIMO in a cooperative fashion.

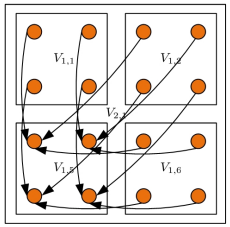
D2D Caching: Cache-Enabled Hierarchical Cooperation

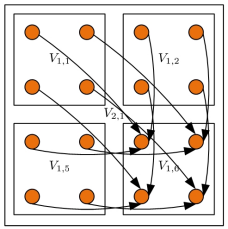


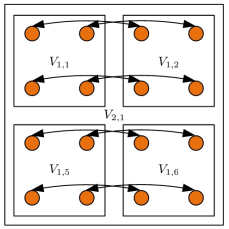


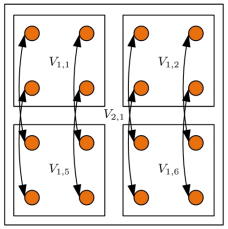

(a1) Concentration of content from parent node $V_{2,1}$ to child node $V_{1,1}$

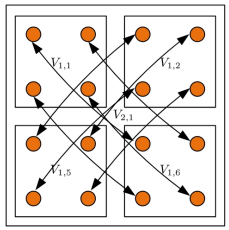

(a2) Concentration of content from parent node $V_{2,1}$ to child node $V_{1,2}$


(a3) Concentration of content from parent node $V_{2,1}$ to child node $V_{1,5}$


(a4) Concentration of content from parent node $V_{2,1}$ to child node $V_{1,6}$


(b1) First group of 4^2 uniform permutation traffic in $V_{2,1}$


(b2) Second group of 4^2 uniform permutation traffic in $V_{2,1}$


(b3) Third group of 4^2 uniform permutation traffic in $V_{2,1}$

D2D Caching: Cache-Enabled Hierarchical Cooperation

- Results for extended networks (constant user density, Area $= \Theta(K)$).

		Protocol model	Physical model
W/o cache	Scaling	$\Theta(K^{\frac{1}{2}})$	$\Theta(K^{\frac{\min\{3,\alpha\}}{2}-1})$
	Strategy	Multihop	Hierarchical coop.
With cache heavy tail $\{P_n\}$	Scaling	$\Theta(\mu^{-1/2})$	$\Theta(\mu^{1-\frac{\min\{3,\alpha\}}{2}})$
	Strategy	Random caching + Multihop	Cache-enabled H-Coop
With cache general $\{P_n\}$	Scaling	Unknown	Known for Zipf
	Strategy	Unknown	Known for Zipf

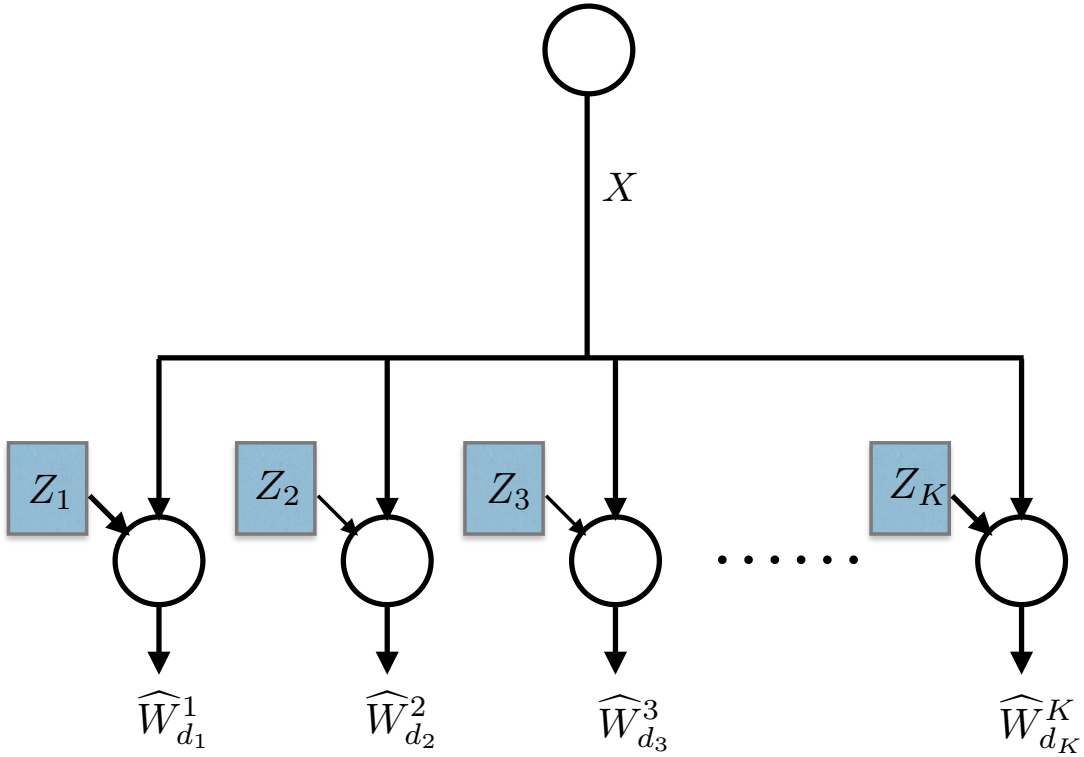
Coded Caching

Can we achieve scalability with finite infrastructure?

- FemtoCaching requires infrastructure nodes to grow linearly with the users.
- D2D Caching requires no infrastructure, but it is very hard to do (no good D2D standard in place, coordination across a large network).
- Can we achieve scalability with finite infrastructure and no D2D communication?
- **Yes we can!** Caching at the user nodes, and delivery by Coded Multicasting.
- (Network) Coding turns unicast traffic into multicast traffic.

The Single Shared Link Network

- Model proposed and analyzed by Maddah-Ali and Niesen in a series of papers.

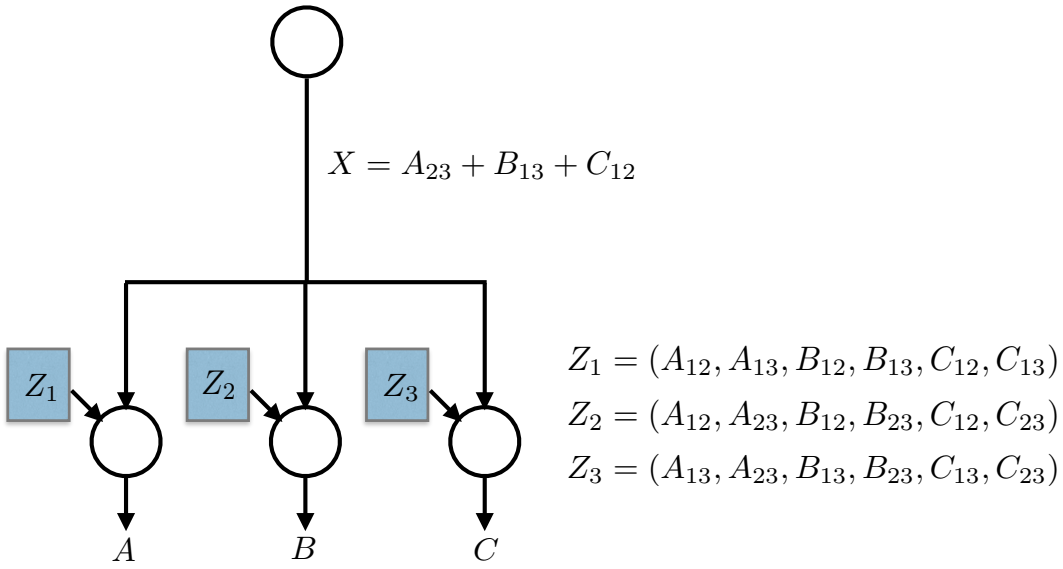


Example: $K = N = 3$, $M = 2$

- Files are divided into three sub-packets of size $1/3$ each:

$$A = (A_{12}, A_{13}, A_{23}), \quad B = (B_{12}, B_{13}, B_{23}), \quad C = (C_{12}, C_{13}, C_{23})$$

- For example, the demand vector $\mathbf{d} = (A, B, C)$ is satisfied by transmitting: $X = A_{23} + B_{13} + C_{12}$.
- Achieved rate $R = 1/3$.



Delivery Time of Coded Caching

- Centralized (generalizing the previous example):

$$D = \frac{K(1-\mu)}{1+\mu K}$$

- Decentralized (users cache random chunks independently):

$$D = \frac{K(1-\mu)(1-(1-\mu)^K)}{\mu K}$$

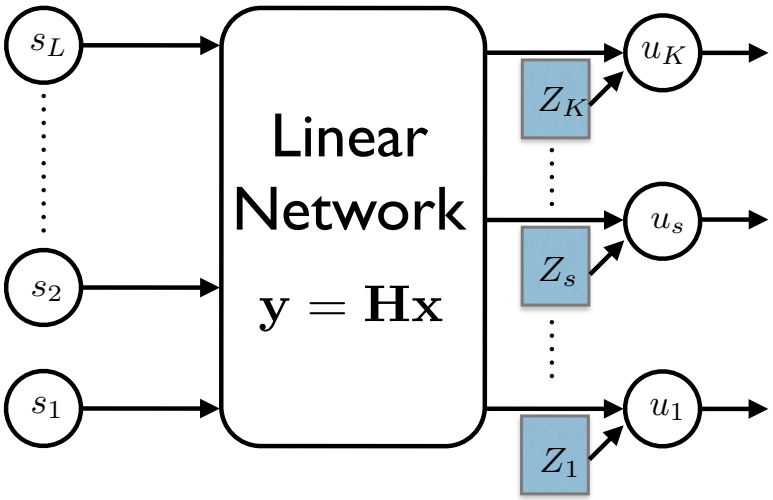
- In both cases, the scaling for large K is

$$D = \frac{1}{\mu} - 1 \approx \frac{1}{\mu}$$

- Pretty much the same scaling of the FemtoCaching network (or one-hop D2D network) $D = \Theta(1/\mu)$.

Coded Caching in More Practical Networks

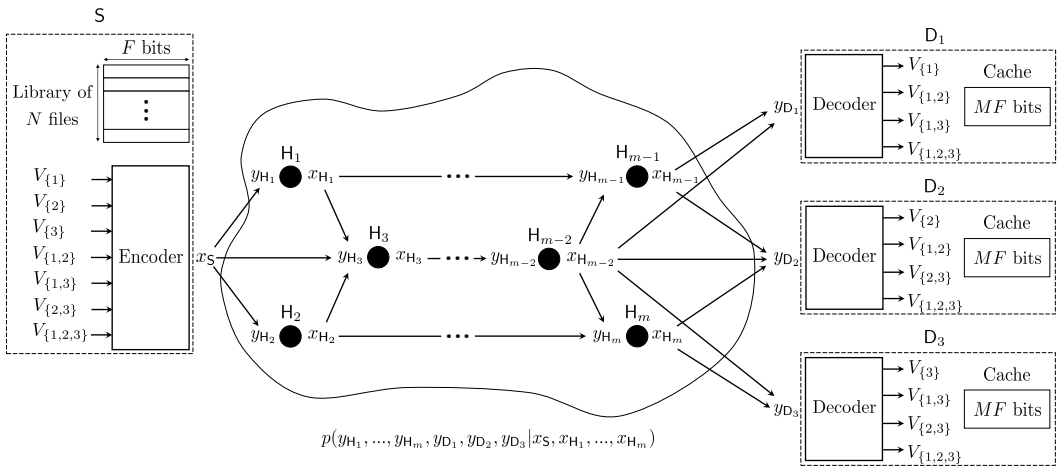
- For every network topology we have to solve a new problem (somehow difficult, boring, and impractical).
- Naturally matched to wireless broadcast networks (Evolved Multimedia Broadcast Multicast Services, eMBMS).
- Naturally matched to MU-MIMO (through the "multiserver" linear network extension of Shariatpanahi, Motahari, and Khalaj, IT2016).



Universality of the Shared Link Network

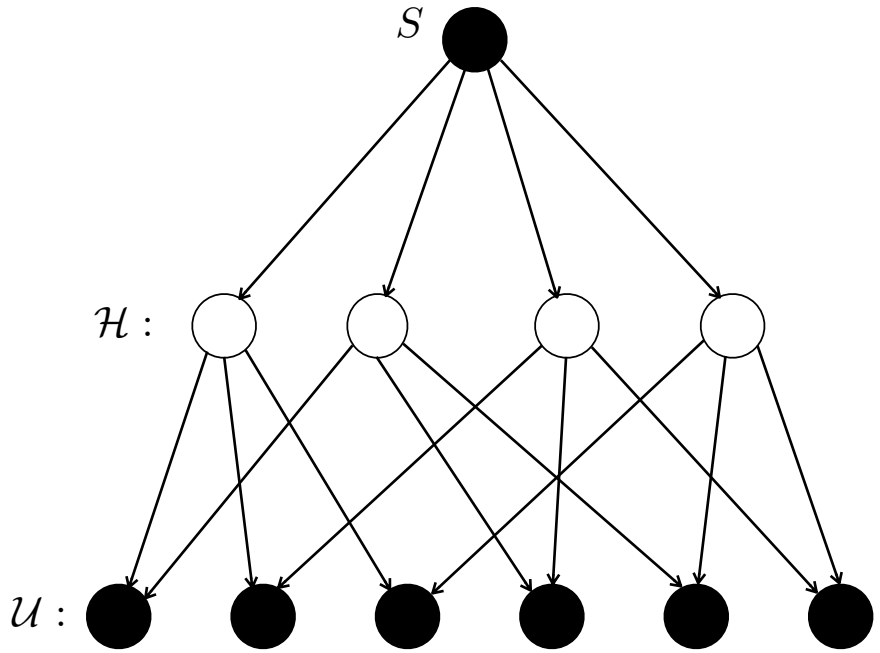
- Naderializadeh, Maddah-Ali, and Avestimehr, arXiv:1701.05881, 2017, have proved a very important universality result:

  APPROX. OPTIMALITY OF CACHING AND DELIVERY: Given a many-to-one memoryless network with one transmitter, arbitrary internal nodes (relays), and K users (destinations) with caches, the delivery time achieved by using the shared link network coded caching scheme followed by **optimal multicast** in the network is approximately optimal.



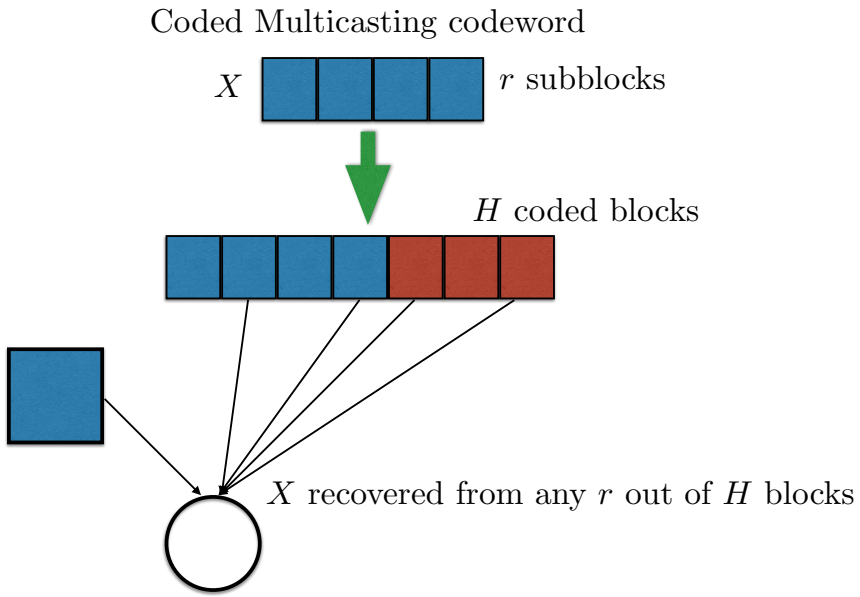
Example: Combination Networks

- H relays (helpers, mirrors, RRHs) with backhaul (H orthogonal links).
- Each user is connected to $r \leq H$ distinct relays.



A Simple Concatenated Scheme

- Outer Code: coded caching as for the shared link network.
- Inner Code: the multicast codeword is encoded by an (r, H) MDS code $\Rightarrow$ the whole multicast message is recovered from any r out of H MDS-coded blocks.
- Speedup factor equal to r (e.g., degree of parallelization, bandwidth aggregation).



Challenges with Coded Caching

- Packetization order: For the class of achievability scheme proposed so far, the file size B scales with N, M, K, B must scale super-exponentially fast with $t = \mu K$ in order to achieve the $\Theta(K)$ coded caching gain.

  Shanmugam, Tulino, Llorca, and Dimakis, "Finite-Length Analysis of Caching-Aided Coded Multicasting." IEEE Trans. on IT, 62(10), pp.5524-5537, 2016.

- Users with different quality requirements: getting closer to actual video streaming (DASH)

  Cacciapuoti, Caleffi, Ji, Llorca, and Tulino, "Speeding up future video distribution via channel-aware caching-aided coded multicast." IEEE JSAC, 34(8), pp.2207-2218, 2016.

  Yang, and Gündüz, "Centralized coded caching for heterogeneous lossy requests." ISIT 2016.

Conclusions

Concluding on a Positive Note

- Caching yields per-user throughput scalability for $\mu = M/N > 0$.
- Caching at the user devices is key to spare on infrastructure.
- Coded caching is much more general than what it looks like ... close to one-size fits all.
- Despite widespread skepticism and serious theoretical challenges .... **IT WORKS!**

CADAMI

(WT10 - Advanced Caching in Wireless Networks: Thursday May 24, Room 231M/232M, 14:00 – 17:30, Keynote by Michael Heindlmaier)

Thank You