Caching at the Edge: Distributed Phy-aware Caching Policies for 5G Cellular Networks

Supervisors: Thrasyvoulos Spyropoulos (EURECOM) and Giovanni Neglia (INRIA Maestro)

Labex themes: Data Centric Networking, Infrastructures: Heterogeneity and Efficiency

Summary

Driven by the exponential growth in wireless data traffic, operators are increasingly considering denser, heterogeneous network (HetNet) deployments, overlaying the macro-cellular network with many small cell base-stations (BS). However, CAPEX and OPEX considerations are making the operators turn to cheap, under-provisioned solutions for the backhaul support of small cells, making backhaul capacity a potential new bottleneck. To alleviate this bottleneck, recent work has proposed `caching-at-the-edge' methods that cache popular content closer to the user to reduce the number of duplicate backhaul transmissions.

While a number of recent works consider the problem of caching at BSs or even user devices, these works are not addressing the complex interplay between caching and advanced communication techniques considered for beyond 4G networks, such as Coordinated Multi-Point (CoMP) transmission. In CoMP multiple BSs transmit the same content concurrently, aiming at diversity or multiplexing gains on the radio interface, at the cost of a significant additional overhead on the backhaul to retrieve the content at all the BSs involved. Hence, for edge caching to facilitate CoMP, the *same* content must be cached at all cooperating BSs.

We propose to jointly consider and optimize cooperative caching at the edge with cooperative communication technologies. A key goal in this proposal is to perform a first-of-its-kind exploration of the elusive tradeoff between edge caching that reduces backhaul traffic, and caching that improves radio access performance. At the center of this tradeoff lies the question of how caching algorithms can adapt to accommodate potential Coordinate Multi-Point (CoMP) transmission opportunities: caching the same replicated content in multiple nearby caches can exploit coordinated transmissions, but might waste valuable storage space and lead to cache misses; yet storing different contents to improve cache hits might lead to missing CoMP opportunities, and degrade communication performance. To this end, we plan to investigate distributed implementations of our optimization solutions, in order to (a) deal with the high complexity of such cooperative caching problems, and (b) significantly reduce the amount of additional (signaling) information transmitted over the already congested backhaul links.

This research action is the opportunity to develop a new collaboration between two researchers in the Labex, but also to start working on a hot research topic, that is currently not represented in Sophia Antipolis' research panorama. The thesis will benefit from the co-advisers' complementary expertise in cellular networks and edge caching systems (Thrasyvoulos), and caching models and distributed optimization algorithms (Giovanni), as well as from many ongoing external collaborations.

Scientific context

The majority of research on dense small-cell setups emphasizes on self-organizing radio resource management, inter-cell and interference coordination, energy-efficiency and MIMO techniques. However, these works usually do not consider the backhaul network which, in order to support this cell densification, must have a (backhaul) capacity that scales at unrealistic rates. As a result, small cells are expected to be mostly connected by inexpensive wired or wireless links and thus will often be underprovisioned [SC-forum].

At the same time though, owing to the cheap storage/memory prices, storing content closer to the user has emerged as a promising solution in order to: (a) avoid congesting the capacity-limited backhaul links, and (b) reduce the longer end-to-end access time along the path to an Internet server. For example, in [Bocc2014], local content caching has been identified as one of the five most disruptive enablers for 5G networks, sparking a tremendous interest of academia and industry alike, as a cost-efficient way of tackling the data tsunami. Furthermore, studies based on real cellular traffic suggest substantial benefits of forward caching at different locations inside a cellular network [Erman2011].

While caching has been widely used in fixed line peer-to-peer systems and content distribution networks (CDNs) [Borst2010], the seminal work of [DimCa2013] was the first to propose caching content on a set of inexpensive edge nodes (called "helper") nodes, which can be, for example, femto-cells with limited backhaul capacity. Dense topologies are considered together with a set of contents of different popularity, and the optimal content placement problem is investigated assuming both coded and uncoded versions of the content. [MN14] takes an information theoretic approach and lays down some fundamental limits of caching in terms of local and global caching gains. More recently, [Tass2014] considers the problem of joint caching and routing in HetNets, as well as the tradeoff between storing different video versions or layers, and [NMB2015] investigates tradeoffs between edge caching and user request routing. [Mehdi2014] has further investigated the benefits of both spatial and social caching. Finally, due to the increasing storage capabilities of user equipment (e.g. smartphone, tablets, etc.), User Equipments (UEs) have also been considered as potential caching locations (see, e.g., [DimCa2012], [Hui2012], [Serm2014], [Serm2015]). However, these latter works using device-based storage and D2D focus on offloading the radio access, rather than the backhaul. Furthermore, all the aforementioned works do not consider advanced cooperative communication techniques.

Somewhat closer to our vision is the work of [Lau2014], where the authors use a coding scheme to introduce redundancy in caches, and create CoMP opportunities for cooperative transmissions. However, this work mostly focuses on PHY layer issues, and does not consider the important cost of cache misses.

Thesis' goal

Joint caching and transmission in 5G - new opportunities and challenges: While the above discussed works already demonstrate the potential benefits of caching in wireless networks, the problem becomes much more interesting in future 5G networks, where caching and transmission are intertwined, due to (a) a significant coverage overlap between small cells (and their local storage) in dense deployments, and (b) an increased connectivity over backhaul links (e.g. through the X2 LTE interface), to support BS cooperation. This new 5G-inspired setting, creates opportunities for both Cooperative Caching and Coordinated Multi-Point (CoMP) transmission, which we will jointly consider here. A UE has concurrent access to multiple nearby small cell caches (either directly or over one or more backhaul links) essentially seeing sets of larger, partially overlapping caches. Furthermore, CoMP transmission capitalizes on the co-placement of data to perform advanced interference cancellation and significantly increase the transmission rates in future cellular systems. Nevertheless, to fully exploit CoMP opportunities, the requested content must be available in the caches of all BSs involved in a coordinated transmission, otherwise the effective rate of one or more links might be severed by a congested backhaul link. This is a new challenge that novel caching techniques must account for. Furthermore, the already considerable complexity of cooperative caching schemes is aggravated by the further coupling of different caches through CoMP.

We envision a joint treatment of storage and transmission of contents in wireless networks, resulting in novel caching algorithms that deviate from the existing focus on content popularity, to include facilitation of multinode cooperative CoMP communication. As a key step toward designing such novel algorithms, we plan to model and analyze the interesting tradeoffs that arise between, on the one hand, storing popular contents in multiple small cells (to maximize CoMP opportunities, at the cost of taking up valuable storage capacity), and on the other hand, maximizing the number of unique contents stored at the edge, at the cost of fewer CoMP opportunities and potentially lower wireless transmission rates. As a second step, we will investigate efficient distributed caching algorithms that jointly consider backhaul and radio access performance, while coping with the computationally-complex nature of cooperative caching problems (cooperative caching is known to be of NP nature in many setups).

To the best of our knowledge, this effort will constitute the first work in the context of edge caching that jointly explores coordinated multi-point transmission and optimization complexity issues. To support this vision, the proposed doctoral thesis will address problems along the following two main directions.

PHY-aware cooperative caching: The majority of caching schemes for small cells or UEs focuses on content popularity, when deciding which contents to store. Consideration of the physical layer is usually limited to the supported rate between the BS and UE involved (see e.g., [Mehdi2014]). Yet, Cooperative Multi Point (CoMP) transmission techniques, envisioned in beyond 4G systems, *couples* the placement decision on the caches involved: caching policies should now take into account potential CoMP transmission opportunities, and the concept of `cache hit' is no longer binary, as the more the `cache hits' for the same content in nearby caches the better usually the performance.

Furthermore, the caching policy is affected by the type of cooperative transmission technique used. For example, cooperation can be used to increase diversity (and thus the communication rate) of a transmission to a single user, by coherently combining the transmission of the *same content* from multiple BS in parallel. Stringent synchronization constraints, and potentially congested backhaul links imply that the same content must be cached on all the BSs involved. Alternately, even if some co-located users request *different contents,* but each of them happens to be stored in nearby BSs that can coordinate, multi-user MIMO (MU-MIMO) techniques could still be used to transmit these different contents *concurrently* to each UE, without interference (e.g., by zero-forcing), thus creating multiplexing gains on the radio access. Mixed schemes can also use some of the available degrees of freedom for multiplexing gain and the rest for diversity.

In view of these, we will investigate appropriate models that try to predict CoMP opportunities and related gains on the radio access, based on the amount of coverage overlap between BSs (e.g., using known tools from stochastic geometry), and available CoMP techniques. Based on these predictions together with content popularity statistics, we will investigate optimal policies that jointly select the best combination of cooperative caching and transmission schemes in order to tradeoff radio access gains with cache misses and resulting rate/latency degradation on the backhaul.

Distributed implementation of caching policies: While our first research goal is to explore the theoretical trade-offs achievable by jointly exploiting CoMP and caching, the second goal is to design practical rules that can achieve the target operation point. In principle, one could collect all the necessary information from the BSs about the channel quality and users' content requests at a central node that could calculate the optimal configuration, according to the previous theoretical analysis.

However, such a solution would require a lot of backhaul signaling, further congesting the already capacity-limited backhaul links, and thus increasing the impact of a cache miss. While some link capacity can be available for signaling during off-peak hours (and in particular during the night), content popularity can change significantly over a few hours time horizon (this is particularly true for media contents generated on a social network). What is more, wireless channel characteristics and related coordinated transmission gains change even more rapidly. Hence, a centralized implementation bears the further risk of operating on the basis of stale information.

For all these reasons we plan to investigate distributed policies that can allow BSs to coordinate their caching policies with limited or almost zero signaling. We plan to build on the TTL-cache model that Giovanni and his co-authors have recently proposed [Chou14]. The family of TTL caches is very large, given that it well approximates the usual LRU, Random and FIFO policies. The advantage of this model is that it makes possible to calculate the main performance metrics for a network of TTL caches. Recently, Don Towsley (one of the authors of [Chou14]) has extended the original framework by showing that, by suitably selecting the local TTL value, the eviction policy at the cache can implicitly optimize a local utility function (this work is currently under submission). Our idea is then to explore the possibility to define the TTLs at each cache, so that the caches are jointly solving a global optimization problem.

Some *local* information exchange can still be required for a cache to take the right decision from the global point of view. E.g., nearby caches could exchange with each other (e.g. through the X2 interface) a summary of the contents they store by a Bloom filter. The fingerprint of cache A allows cache B to know if a given content is stored at A (even if at the risk of false positives) and then to take more informed local decisions. This approach is intrinsically tunable: the more the fingerprint's bits, the smaller the false positive probability but the higher the bandwidth requirements.

Furthermore, additional *local* feeback about channel quality is needed for CoMP operation. (although limited-feedback cooperation techniques are currently being explored). The capacity constraint on the backhaul links intrinsically couples the decision about the amounts of cache-related and channel-related feedback to provide, thus creating another interesting tradeoff that we plan to explore. In this direction, we plan to exploit the fact that the time scales at which cached content and channel performance change are different.

Supervisors

Thrasyvoulos Spyropoulos and Giovanni Neglia will share equally the supervision task. They have both submitted their application to be enrolled for the *Habilitation à diriger de recherches* at University of Nice. If the registration is granted, they plan to defend their HDR by Spring 2016. The student will be hired by Eurecom.

Thrasyvoulos has been recently working on small cell and device-based caching and communication frameworks for cellular networks [Serm2014][Serm2015]. He has also been working on a number of topics related to future small cell networks and HetNets, such as user association, energy management, and offloading [Sap2014][Sap2015], and has a solid background on modern cellular systems and communication theory. Giovanni has been recently working on distributed optimization in networks (e.g. [Alo10], [Mas11]), as well as on evaluating the performance of cache networks [Chou14], proposing a unified and tractable model for a large class of eviction policies like LRU, FIFO, Random, etc.

Their backgrounds complement each other in order to properly address the two research topics identified. Additionally, this project will create a completely new collaboration between two young researchers of two distinct institutions of the Labex UCN@Sophia (EURECOM and INRIA), along the lines of a very timely topic.

Finally, the hired student will benefit from our current collaborations with different researchers with expertise related to this project, such as Petros Elia from EURECOM who is working on information-theoretic and complexity aspects of caching, Navid Nikaein from EURECOM with whom we frequently consult about cellular technology issues and practical constraints in 5G, George Paschos from Huawei Labs - Paris and Don Towsley from University of Massachusetts – Amherst, with whom we are in contact on the topic of caching in networks, as well as Vaishnav Janardhan and Ming Feng from Akamai, San Francisco, who provided us some long traces with content requests at Akamai CDN caches.

References

[Alo10] S. Alouf, G. Neglia, I. Carreras, D. Miorandi, A. Fialho, Fitting genetic algorithms to distributed on-line evolution of network protocols, Elsevier Computer Networks, Volume 54, Issue 18, December 2010.

[Bocc2014] F. Boccardi, R. Heath, A. Lozano, T. Marzetta, and P. Popovski, “Five Disruptive Technology Directions for 5G,” IEEE Comm. Mag. SI on 5G Prospects and Challenges, Feb. 2014.

[Borst2010] S. Borst, V. Gupta, and A. Walid, “Distributed caching algorithms for content distribution networks,” in Proc. of IEEE INFOCOM, 2010.

[Chou14] N. Choungmo Fofack, P. Nain, G. Neglia, D. Towsley, “Performance evaluation of hierarchical TTL-based cache networks,” Elsevier Computer Networks, Volume 65, 2 June 2014.

[DimCa2012] N. Golrezaei, A.F. Molisch, A.G. Dimakis and G. Caire, “Device-to-Device Collaboration: A New Architecture for Wireless Video Distribution,” IEEE Communications Magazine, 2012.

[DimCa2013] K. Shanmugam, N. Golrezaei, A. Dimakis, A. Molisch and G. Caire, “FemtoCaching: Wireless video content delivery through distributed caching helpers,” IEEE Trans. on Information Theory, Dec. 2013.

[Erman2011] Erman, J.; Gerber, A.; Hajiaghayi, M.; Dan Pei; Sen, S.; Spatscheck, O., "To Cache or Not to Cache: The 3G Case," IEEE Internet Computing, IEEE , vol.15, no.2, pp.27,34, March-April 2011.

[Tass2014] K. Poularakis, G. Iosifidis, A. Argyriou, L. Tassiulas, ” Video Delivery over Heterogeneous Cellular Networks: Optimizing Cost and Performance ” in Proc. of IEEE INFOCOM 2014.

[Hui2012] B. Han, P. Hui, V. S. A. Kumar, M. V. Marathe, J. Shao, A. Srinivasan: Mobile Data Offloading through Opportunistic Communications and Social Participation. IEEE Trans. Mob. Comput. 11(5), 2012.

[Lau2014] A. Liu and V. K. N. Lau, “Cache-enabled opportunistic cooperative MIMO for video streaming in wireless systems,” IEEE Trans. Signal Processing, Jan. 2014.

[Mas11] R. Masiero, G. Neglia, “Distributed Sub-gradient Method for Delay Tolerant Networks,” in Proc. of 30th IEEE Conf. on Computer Communications (INFOCOM), Mini-Conference, 10-15 April 2011, Shangai, China.

[Mehdi2014] E. Baştuğ, M. Bennis and M. Debbah, "Social and Spatial Proactive Caching for Mobile Data Offloading," in Proc. IEEE ICC 2014, Sydney, Australia.

[MN:14] M. A. Maddah-Ali and U. Niesen, “Fundamental limits of caching,” IEEE Trans. on Information Theory*, May.* 2014.

[NMB2015] K.P. Naveen, L. Massoulie, E. Baccelli, A. C. Viana, and D. Towsley, “On the Interaction between Content Caching and Request Assignment in Cellular Cache Networks,” In Proc. of the 5th Workshop on All Things Cellular: Operations, Applications and Challenges (AllThingsCellular '15), 2015.

[Sap2014] N. Sapountzis, S. Sarantidis, T. Spyropoulos, N. Nikaein, U. Salim, “Reducing the energy consumption of small cell networks subject to QoE constraints,” in Proc. of IEEE GLOBECOM 2014.

[Sap2015] N. Sapountzis, T. Spyropoulos, N. Nikaein, and U. Salim, “An Analytical Framework for Optimal Downlink-Uplink User Association in HetNets with Traffic Differentiation,” in Proc. of IEEE GLOBECOM 2015.

[Serm2014] P. Sermpezis, T. Spyropoulos, “Not all content is created equal: effect of popularity and availability for content-centric opportunistic networking,” in Proc. of ACM MobiHoc 2014.

[Serm2015] Pavlos Sermpezis, Luigi Vigneri, Thrasyvoulos Spyropoulos, "Offloading on the Edge: Analysis and optimization of local data storage and offloading in HetNets", submitted to IEEE/ACM Transactions on Networking, September 2015.

[SC-forum] Small Cell Forum, “Backhaul technologies for small cells: Use Cases, Requirements, and Solutions,” White Paper, Feb. 2013.