Coded Caching: Basic Schemes, Fundamental Limits, and Practical Implementation

Giuseppe Caire

In the past few years, a new paradigm for content delivery exploiting proactive caching and network-coded delivery.
In conventional caching, some segments of the content files are cached at off-peak times, such that during the peak times of network utilization the system must only deliver the remaining segments of the requested content files. A typical example is the so-called prefix caching,
where the initial part of the most popular files is pre-placed into the user devices such that the streaming playback can start immediately, without an initial prefetching phase.
In contrast, the new paradigm commonly denoted as ``coded caching'' is able to exploit partial overlap of the caches contents and network coding, such that the (coded) multicast messages are simultaneously useful for a large number of users. The gain due to coded multicasting over conventional caching is very remarkable, and it is nearly proportional to the ration between the total cache memory in the system (across all users) and the size of the content library. Since the total memory in the system scales with the number of users, the gain grows with the number of users leading to ``scalable'' systems, where the per-user rate does not vanish when the number of users grow, despite the fact that the total downlink capacity is constant.

In this talk we review the basic (information theoretic and network-coding theoretic) schemes and results. Then, we discuss some practical (but important) limitations of the coding theoretic models, and discuss methods to overcome such limitations. In particular, we shall show that a combination of cache content replication and spatial reuse (either using multiple antennas or spatially distributed multiple cells) is able to achieve optimal throughout scaling in very practical scenarios of number of users, file sizes, asynchronous streaming sessions, and encrypted HTTP requests as in adaptive streaming.