CACHE NETWORKS: AN INFORMATION-THEORETIC VIEW

by Mohammad Ali Maddah-Ali and Urs Niesen

Caching is an essential technique to improve throughput and latency in a vast variety of applications such as virtual memory hierarchies in CPU design, web caching for content delivery networks (CDNs), and inquiry caching in domain name systems. Companies like Akamai, Facebook, Netflix, Google, etc. are heavily investing in their cache networks to increase system performance.

There is a rich and beautiful theory, developed mostly in the computer science community during the 80s and 90s, for systems with a single cache. However, when it comes to networks of caches, the existing theory falls short, and engineers instead rely on heuristics and the intuition gained from the analysis of single-cache systems.

Recent results suggest that information theory can in fact provide the theoretical underpinnings for the deployment and operation of cache networks. Applying information-theoretic tools for the analysis of cache networks reveals that the conventional way to operate these networks can be substantially suboptimal, and that new concepts such as coded multicasting are needed.

The goal of this tutorial is to discuss opportunities and challenges for cache networks with emphasis on the role of information theory in offering a fundamental view on this problem. We start by giving an overview of caching systems followed by a review of the theory of single-cache systems. We proceed with recent results on cache networks in a variety of scenarios. We will also present a video-streaming prototype demonstrating the application of these information-theoretic ideas in a practical setting. During the tutorial we will discuss various open problems motivated by real-life caching applications.

"ISPs are busily setting up caches and CDNs to scalably distribute video and audio. Caching is a necessary part of the solution, but there is no part of today's networking—from Information, Queuing, or Traffic Theory down to the Internet protocol specs—that tells us how to engineer and deploy it." — Van Jacobson

About the speakers: Mohammad Ali Maddah-Ali and Urs Niesen have been investigating cache networks from an information-theoretic perspective for the past four years. They have co-authored five journal papers and many conference papers on this topic, and co-filed several patents in this area. Their work led to the first characterization of the rate-memory trade-off for some basic cache network configurations and scenarios, including tree and hierarchical cache networks, non-uniform and asynchronous demands, online cache updating, and cache-aided video streaming. Their results particularly show that, by employing coded caching, the performance of conventional caching schemes can be improved by a factor on the order of the number of caches in the network. They have also developed a video streaming prototype, demonstrating these gains in a practical setting. Their work on coding for caching has received an IEEE ICC 2014 best paper award, and their video streaming prototype was selected as an official Bell Labs demo.