A new coded caching scheme to improve online video delivery

August 28 2018, by Ingrid Fadelli

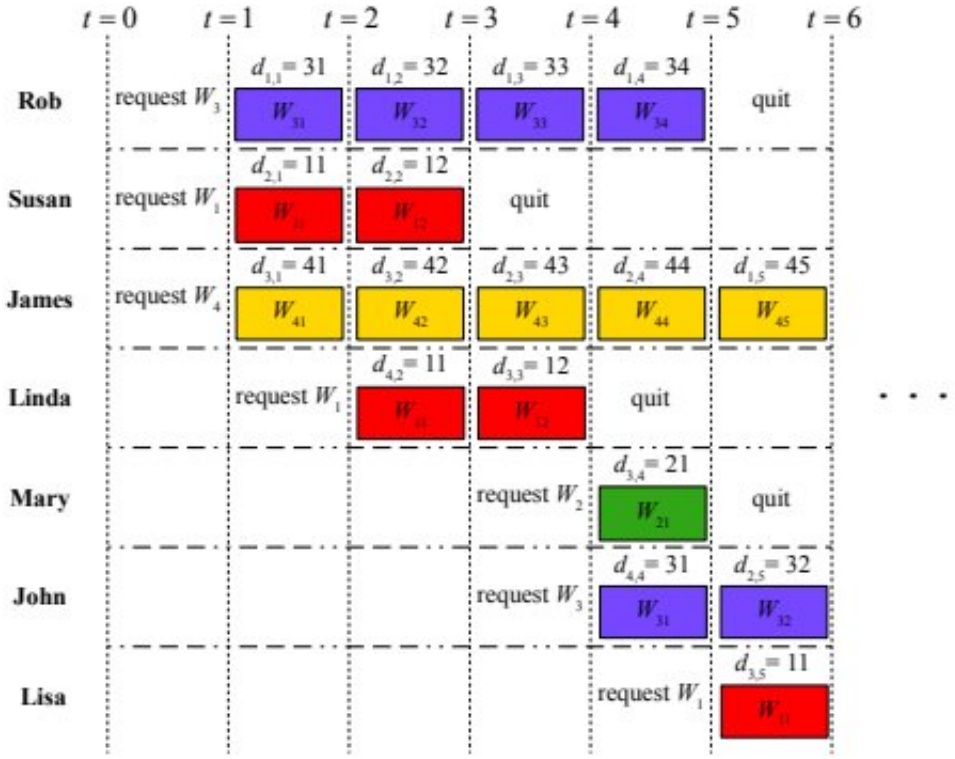


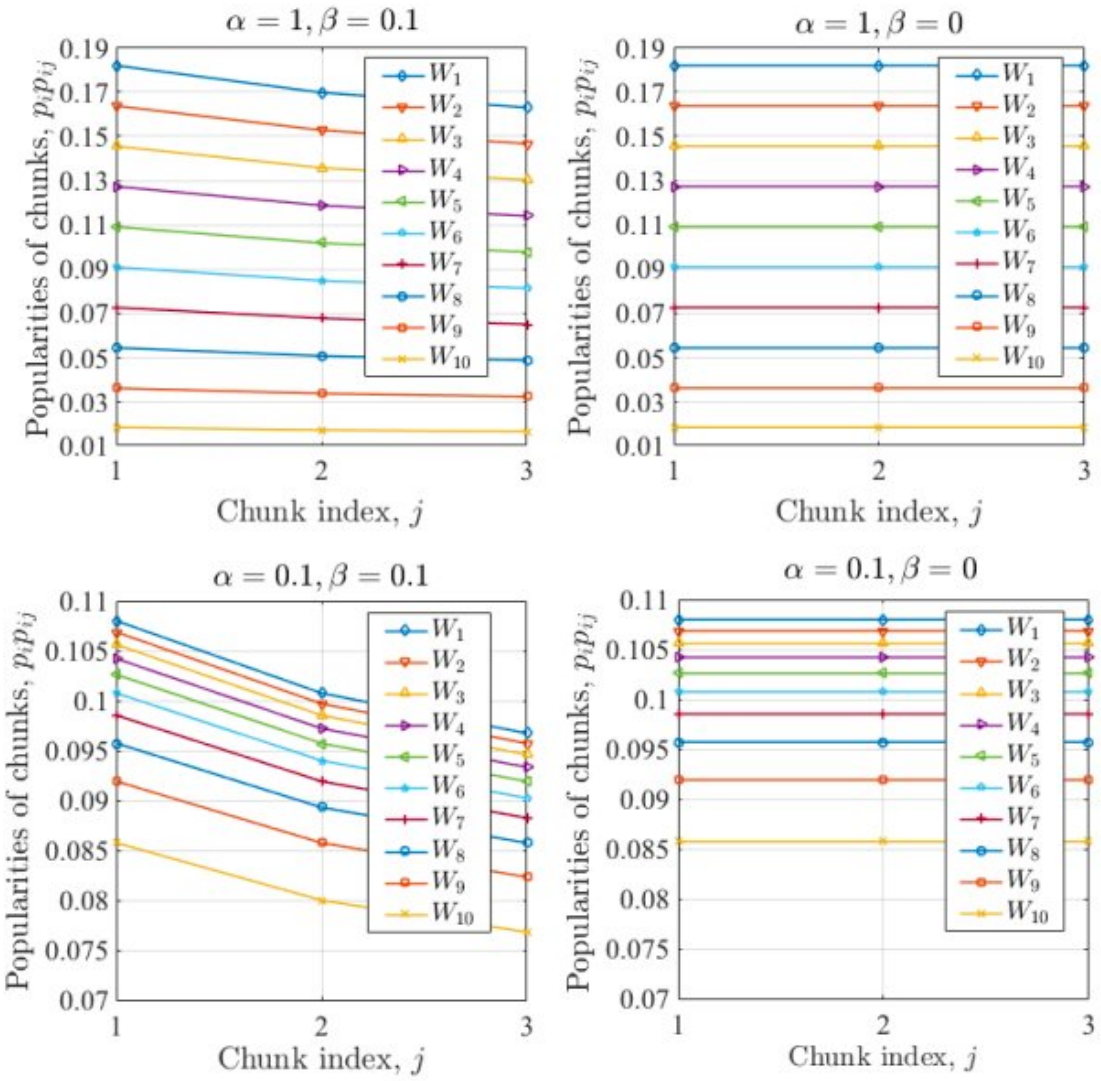
Illustration of the demand arrivals for an asynchronous caching system with N ≥ 4 files and Amax ≥ 3 for time slots t = 1 to 6 of the delivery phase. In the caching setting under consideration, we have a1 = 3, a2 = 1, a3 = 0, a4 = 2, a5 = 1 demands, and K(1) = 3, K(2) = 4, K(3) = 3, K(4) = 4, K(5) = 3 users served at each time slot. Credit: Yang, Amiri & Gündüz

Researchers at Imperial College London have developed a new method for coded caching that could improve the delivery of popular video content online. A research paper outlining their findings was pre-published on arXiv, outlining the technique and its performance in comparison to other caching schemes.

More and more people are streaming video content online, and some videos become particularly popular, dominating wireless data traffic. This has led to the development of proactive caching systems, which pre-fetch video content over off-peak traffic periods and store it at the network's edge or directly in users' devices. These systems can alleviate the traffic load and reduce latency on particularly popular video content.

Proactive caching has two phases: the placement phase, in which the system fills users' caches during off-peak traffic periods and the delivery phase, which takes place once users' demands are revealed (at times of high-peak traffic). Traditional un-coded caching schemes use orthogonal unicast transmissions, which entail a one-to-one association between the sender of the information and its destination, with every destination identifying a single receiver.

A new paradigm, called coded caching, exploits cache resources across a network, optimizing the placement and delivery phases by creating opportunities for multicasting transmission, which entails datagrams being routed simultaneously to many recipients in a single transmission. In their study, the researchers proposed a new strategy that addresses two limitations of existing coded caching systems.

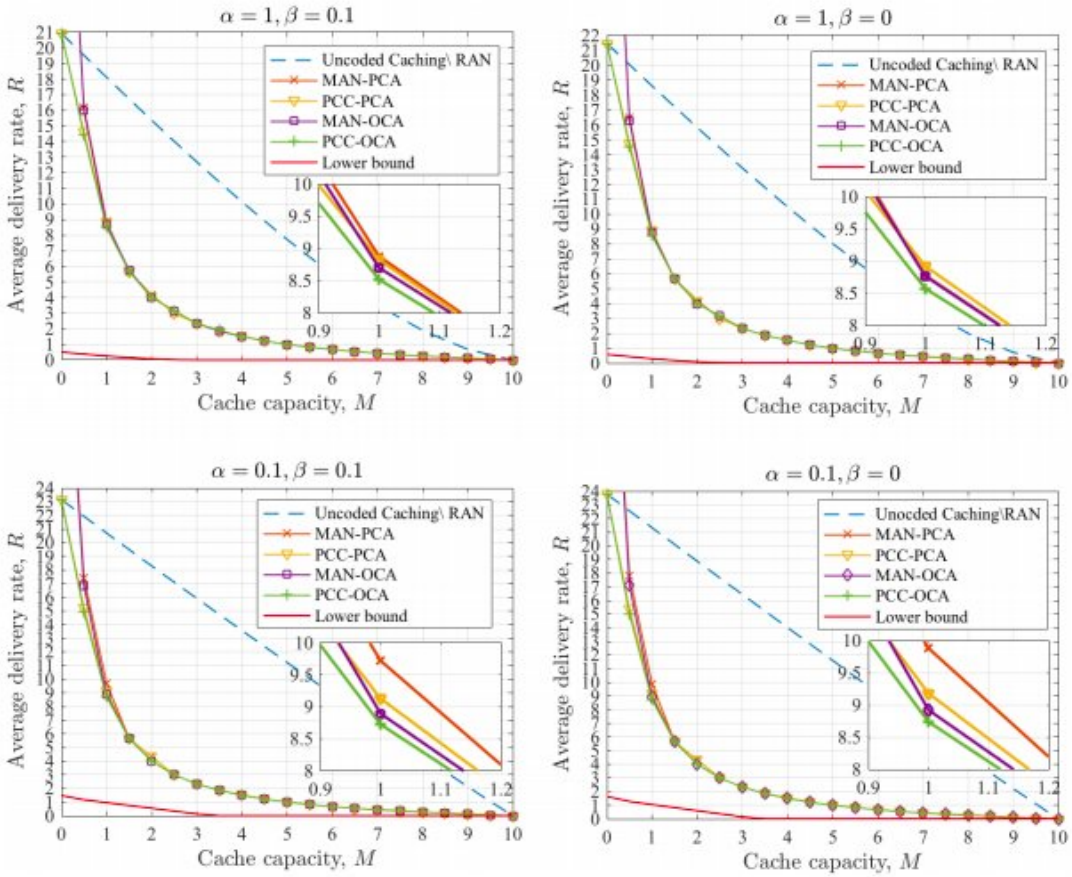


The popularity of video chunks Wij , i.e., pipij given different values of α and β. Credit: Yang, Amiri & Gündüz

So far, most studies developing methods for coded caching have primarily focused on static scenarios, in which a fixed number of users simultaneously place requests from a content library. The performance of these caching schemes is measured by the latency in satisfying the

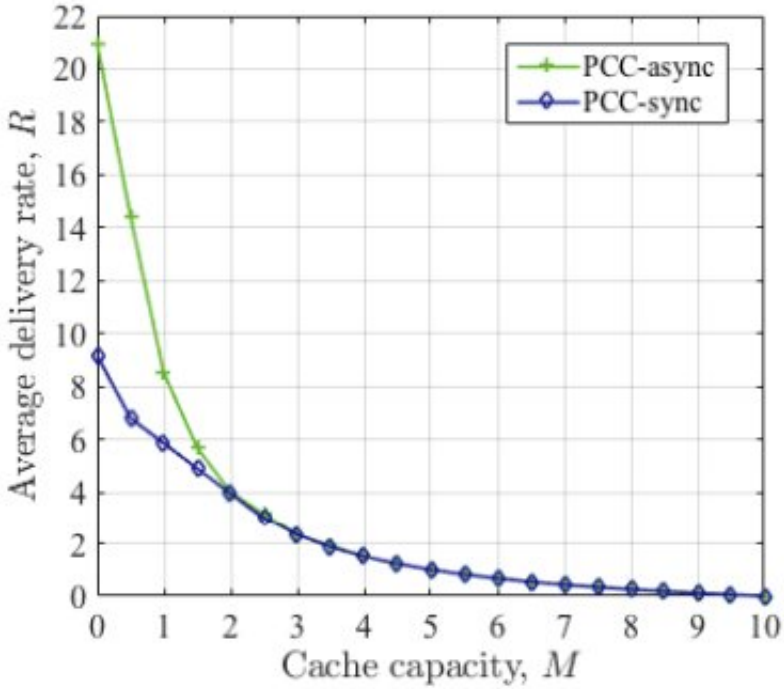
demands of all users. However, in reality, people in different places actually start watching a video online at different points in time, sometimes interrupting it before its end. This particular user behavior is represented by the audience retention rate, a measurement introduced by mainstream video platforms such as YouTube and Netfix, which defines the portion of a particular video that is watched by users, on average.

Audience retention rates can help streaming services to better understand and model the popularity of different sections of video content among users. In their study, the researchers found that partial caching, in which only most viewed portions of a video are cached, could help to achieve more efficient caching.



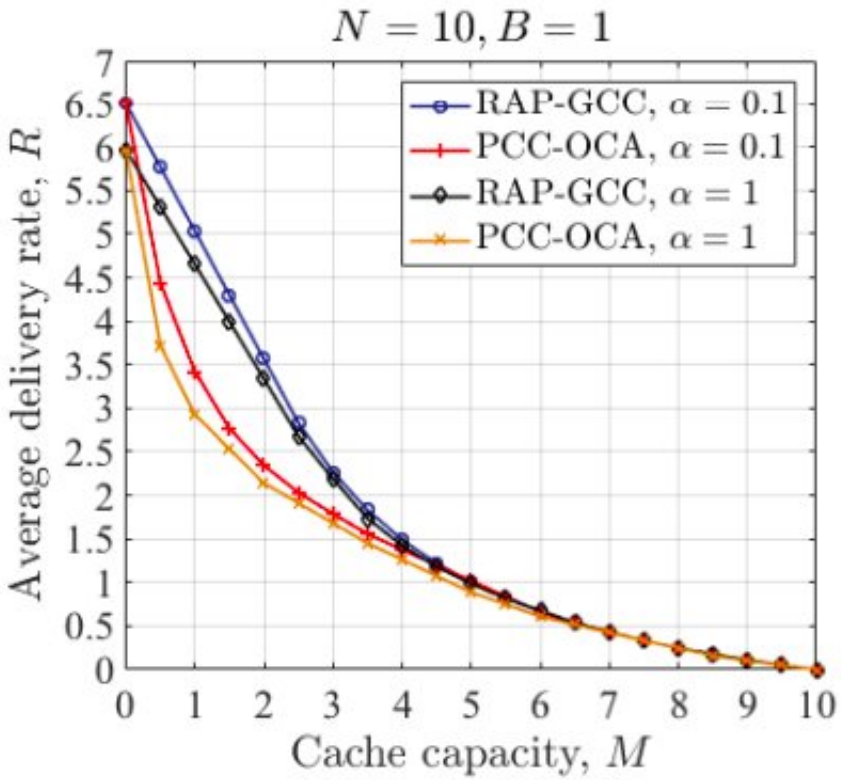
Comparison between PCC, MAN, uncoded caching and the lower bound given different values of α and β. Credit: Yang, Amiri & Gündüz

"We investigate coded caching of video files taking into account the audience retention rate for each video," the researchers explained in their paper. "We consider that each video file consists of equal-length chunks, and the audience retention rate of each chunk is the fraction of users watching this chunk among total views of the corresponding video."



Comparison between the asynchronous and synchronous demand arrival scenarios, α = 1 and β = 0.1. Credit: Yang, Amiri & Gündüz

Contrarily to prior literature on coded caching, in which users are assumed to reveal their demands simultaneously, the researchers propose a dynamic demand arrival model, called partial coded caching (PCC). This model is more realistic, in that it considers that users start and stop watching a video at different points in time. In addition, the researchers proposed two different cache allocation schemes, which allocate users' caches to different chunks of the video files in the library; called optimal cache allocation (OCA) and popularity based cache allocation (PCA).



Comparison between PCC with OCA and RAP-GCC with α = 0.1 and α = 1. Credit: Yang, Amiri & Gündüz

"The results showed a significant improvement with the proposed

scheme over uncoded caching in terms of the average delivery rate, or the extension of other known delivery methods to the asynchronous scenario," the researchers wrote in their paper.

In the future, this new partial coded caching scheme could help to tackle issues with low latency and improve video delivery of popular videos online at times of high data traffic. This could be very useful for popular streaming platforms, such as YouTube, Netflix, and Amazon Prime Video.

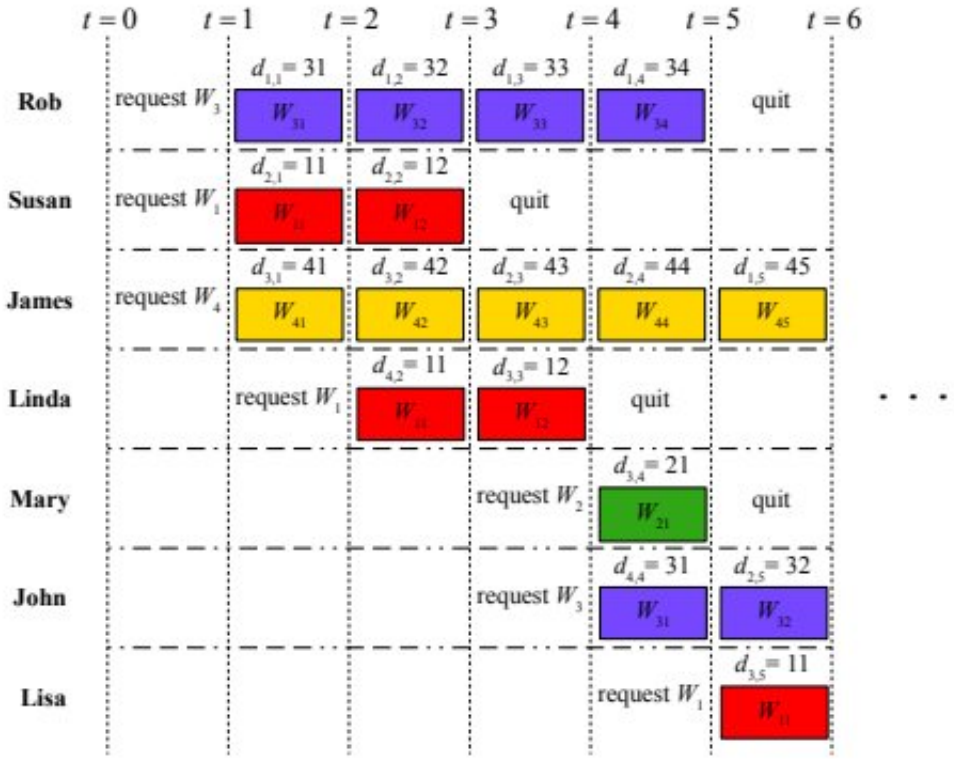


Illustration of the demand arrivals for an asynchronous caching system with N ≥ 4 files and Amax ≥ 3 for time slots t = 1 to 6 of the delivery phase. In the caching setting under consideration, we have a1 = 3, a2 = 1, a3 = 0, a4 = 2, a5 = 1 demands, and K(1) = 3, K(2) = 4, K(3) = 3, K(4) = 4, K(5) = 3 users served at each time slot. Credit: Yang, Amiri & Gündüz

More information: Audience-Retention-Rate-Aware Caching and Coded Video Delivery with Asynchronous Demands arXiv:1808.04835v1 [cs.IT]. arxiv.org/abs/1808.04835

Abstract

Most results on coded caching focus on a static scenario, in which a fixed number of users synchronously place their requests from a content library, and the performance is measured in terms of the latency in satisfying all of these demands. In practice, however, users start watching an online video content asynchronously over time, and often abort watching a video before it is completed. The latter behaviour is captured by the notion of audience retention rate, which measures the portion of a video content watched on average. In order to bring coded caching one step closer to practice, asynchronous user demands are considered in this paper, by allowing user demands to arrive randomly over time, and both the popularity of video files, and the audience retention rates are taken into account. A decentralized partial coded caching (PCC) scheme is proposed, together with two cache allocation schemes; namely the optimal cache allocation (OCA) and the popularity-based cache allocation (PCA), which allocate users' caches among different chunks of the video files in the library. Numerical results validate that the proposed PCC scheme, either with OCA or PCA, outperforms conventional uncoded caching as well as the state-of-the-art decentralized caching schemes, which consider only the file popularities, and are designed for synchronous demand arrivals. An information-theoretical lower bound on the average delivery rate is also presented.