The growth of Internet traffic fueled by online video, new mobile devices, and ever-expanding web services presents a challenge to content delivery networks (CDNs) delivering content via their server infrastructure and internet service providers (ISPs) carrying the traffic on the core network links.

Towards the goal of providing high-performance content delivery in a cost-effective manner, this thesis studies resource management strategies at the CDN-level and at the ISP-level.

We study two approaches by which ISPs can increase effective network capacity:

(1) Route optimization: While route optimization is a well-known technique, we present a re-evaluation of several techniques based on user-perceived metrics taking into account the location diversity in the Internet, aspects largely ignored in prior work.

(2) Integrating CDN functionality: By offering CDN services in their network, an ISP can reduce network costs due to content caching and generate additional revenue. We evaluate several schemes for managing content delivery and route optimization in network CDNs.

In the context of CDNs, we study two approaches that respectively enable distributing large files with minimal server bandwidth and allow dynamic content to be served with small latency and a small update cost.

(1) Hybrid CDN: A hybrid-CDN design reduces bandwidth costs to distribute large files by leveraging clients’ upload bandwidth, which supplements the server bandwidth contributed by CDN. We study how a hybrid CDN should manage its sever bandwidth across content being delivered to optimize performance/cost objectives.

(2) Dynamic content: We design locality and load-aware replication strategies for dynamic content, that reduce the cost of keeping dynamic content updated by restricting the amount of replication, and yet achieves small response latencies by replicating content in a locality-aware manner.

In this process, we consider two types of strategies: (1) planned strategies, which assume that content workloads in future will resemble those in the past, and allocate resources globally using an offline analysis of past workloads, and (2) unplanned strategies, which manage resources either using a fixed policy or using an online algorithm based on local information.

Potentially, a planned strategy can outperform an unplanned approach due to a global optimization of resources and a long-term knowledge of workloads. However, we find that simple unplanned approaches perform well in many cases. In some cases, a planned approach indeed gives better performance than an unplanned approach, but a planned approach is not always effective due to unpredictability of the workload and because it is infeasible to estimate the optimal solution or even its approximation.