Telecommunications Policy xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Telecommunications Policy

journal homepage: www.elsevier.com/locate/telpol

The growing complexity of content delivery networks: Challenges and implications for the Internet ecosystem

Volker Stocker[a,*], Georgios Smaragdakis[b,c,1], William Lehr[c,2], Steven Bauer[c]

[a] University of Freiburg, Abteilung für Netzökonomie, Wettbewerbsökonomie und Verkehrswissenschaft, Platz der Alten Synagoge, 79085 Freiburg im Breisgau, Germany
[b] Technical University of Berlin, Germany
[c] Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory (CSAIL), Cambridge, MA, United States

ARTICLE INFO

Keywords:
CDN
Interconnection
Peering
QoS
QoE
Pricing
Internet evolution

ABSTRACT

Since the commercialization of the Internet, content and related applications, including video streaming, news, advertisements, and social interaction have moved online. It is broadly recognized that the rise of all of these different types of content (static and dynamic, and increasingly multimedia) has been one of the main forces behind the phenomenal growth of the Internet, and its emergence as essential infrastructure for how individuals across the globe gain access to the content sources they want. To accelerate the delivery of diverse content in the Internet and to provide commercial-grade performance for video delivery and the Web, Content Delivery Networks (CDNs) were introduced. This paper describes the current CDN ecosystem and the forces that have driven its evolution. We outline the different CDN architectures and consider their relative strengths and weaknesses. Our analysis highlights the role of location, the growing complexity of the CDN ecosystem, and their relationship to and implications for interconnection markets.

1. Introduction

Content Delivery Networks (CDNs) emerged as overlay networks on the Internet in order to provide better support for delivering commercial content than was available using basic, "best-effort" Internet packet transport services. As the volume, complexity, and heterogeneity of Internet traffic has grown and evolved, so too have the Internet Service Providers (ISPs) and CDNs that provide the services used to deliver this traffic. The importance of CDNs within the Internet ecosystem has grown significantly over time – recent reports expect that CDNs will soon be handling over half of the global traffic on the Internet (cf. Cisco, 2016, p. 18).

In this paper, we describe the current CDN ecosystem and the forces that have driven its evolution. Section 2 provides an overview of the basic operation of CDNs and recounts how they have evolved since their market inception in the 1990s. Section 3 then presents a typology that sets forth the multiple types of CDNs and their relative strengths and weaknesses. Section 4 focuses on how CDN choices of where to locate their servers impact their operations, and Section 5 offers our speculations on where we see the CDN ecosystem going. Section 6 concludes and offers some thoughts and implications for the future Internet ecosystem.

* Corresponding author.
E-mail addresses: volker.stocker@vwl.uni-freiburg.de (V. Stocker), gsmaragd@csail.mit.edu (G. Smaragdakis), wlehr@mit.edu (W. Lehr), bauer@mit.edu (S. Bauer).

[1] Author was supported by the ERC Starting Grant ResolutioNet (679158).
[2] Author would like to acknowledge support from NSF Awards 1413973, 1547265 and the MIT Communications Futures Program (http://cfp.mit.edu).

http://dx.doi.org/10.1016/j.telpol.2017.02.004
Received 26 November 2016; Received in revised form 6 February 2017; Accepted 12 February 2017

2. On the evolution of content delivery networks

Although the Internet's basic infrastructure has scaled remarkably well, its end-to-end, “best effort” design was premised on a communication paradigm based on passive traffic management. The basic protocols like the Transmission Control Protocol (TCP) that manage packet transmission on an end-to-end basis seek to provide decentralized congestion management and fair resource allocations across competing flows; however, these protocols fail to support the predictable, end-to-end Quality of Experience (QoE)[3] that is increasingly being demanded by commercial content and application providers.[4] Moreover, networks offering only a single service class are not well-suited for supporting the heterogeneous needs of traffic types with very different Quality of Service (QoS) requirements.

While attempts have been made to expand the set of Internet protocols to enable better support for heterogeneous QoS requirements,[5] modifying the basic Internet infrastructure is difficult because it requires coordinating the widespread adoption of new protocols. In the Internet, control is decentralized and dispersed among many different and often competing entities. Although individual networks may deploy enhanced capabilities for active traffic management of on-net traffic, widespread inter-provider deployments are not generally available (cf. e.g., Stocker, 2015). CDNs evolved to address the need for better support for differentiated content distribution without requiring modifications to the basic Internet architecture.

2.1. A primer on CDN service provision

CDNs employ a scalable architecture of cache servers which are strategically distributed across the Internet and constitute an overlay “on top” of the Internet's basic packet transport infrastructure.[6] A typical CDN maintains a hierarchy of servers, with back-end servers ensuring the efficient intra-CDN distribution of content, and front-end servers at the edges used for handling user-server communications. These servers allow replicated copies of content to be stored at multiple locations across the Internet.

CDN providers employ complex software to match incoming requests for content to the “best” server for meeting each end-user request.[7] Requests flow from end-users to the selected front-end server, which delivers a copy of the requested content to the end-user, if a copy is already available on the server. If not, the front-end server passes the request further up the CDN-server hierarchy until a copy is located, which may entail pulling a copy from the content provider's origin server if a closer copy is not available (cf. Nygren et al., 2010). In this way, content is replicated and distributed across the CDN's footprint of servers. Deciding what content to store in which servers and for how long to retain copies, and how to best manage requests for serving content is complicated and depends on the nature of the content (i.e., static versus dynamic),[8] the preferences of the content provider, end-user demand for the content (i.e., content popularity), what else is going on in the Internet (e.g., congested links), and the capabilities of the CDN provider. Further complications arise if origin servers are located at different content providers (e.g., as might be the case for advertisements and background text), or specialized security or stringent QoS requirements necessitate specialized routing treatment by the CDN provider.

Replicating content in multiple locations and jointly managing multiple servers in real-time, allows CDN providers to better balance server loads, which enhances the utilization efficiency of server capacity and capacity scaling, lowering content delivery resource costs and enabling CDNs to provide improved QoE performance for end-users. It helps CDNs better respond to flash crowds and denial of service attacks (cf. e.g., Nygren et al., 2010; Maggs & Sitaraman, 2015). When the content is available from multiple locations, single points of failure are eliminated, which improves reliability. The ability to match content requests with the “best” server for each request helps content providers reduce end-to-end delays and ensure a more consistent and higher QoE for their end-users (cf. Dilley et al., 2002; Chiu, Schlinker, Radhakrishnan, Katz-Bassett, & Govindan, 2015). CDN providers make use

[3] QoE is a holistic concept that describes the subjectively perceived quality of a user when consuming content or applications over the Internet.

[4] For example, TCP performance degrades as the distance between communicating hosts increases and in the presence of packet losses along the end-to-end path (cf. Leighton, 2009). Also, TCP's flow-rate fairness principal may be exploited by applications (e.g., using swarming techniques to simultaneously operate multiple TCP connections) resulting in some applications capturing a disproportionate share of available capacity, while leaving other applications starved of capacity (cf. e.g., Briscoe, 2007).

[5] Over time, the Internet suite of protocols have expanded to include better support for content delivery and other types of QoS differentiated packet transport services. For example, multicast enhances content delivery capabilities by allowing a single packet that is to be sent to multiple destinations to be replicated only at branching points, thereby reducing the packet transport resources required (cf. Cisco, 2001). Alternatively, newer protocols like DiffServ and IntServ provide support for better-than-best-effort packet delivery (cf. Cisco, 2005). Although such capabilities already exist in the Internet, they are neither uniformly available nor implemented in a standard way across ISPs which complicates efforts to make use of them.

[6] Cache servers are dedicated servers used to store content for subsequent delivery. In general, one can distinguish between three general types of cache servers involved in CDN service provisioning: front-end, back-end, and origin servers. The "origin server" refers to the server where the original or master copy of the content is stored. A key role for CDNs is to place additional copies of that content on other cache servers elsewhere in the Internet to improve content delivery services. The "front-end" servers are the ones that end-users communicate with; and these may communicate with appropriate "back-end" servers that are located within the CDN network and provide additional CDN functionality as we explain. For further details on CDN servers, see Leighton (2009), Nygren, Sitaraman, and Sun (2010), or Flach et al. (2013).

[7] Determining the “best” server may depend on different metrics in the quality/cost-space in which different delivery features may be emphasized. As will be described in Section 4, the location of the server is important, and "location" may be defined in multiple ways. While in the case of CDNs with a network of dedicated servers, redirection decisions are typically done via the Domain Name System (DNS), distributed naming schemes are used in peer-to-peer CDNs.

[8] Dynamic content, as opposed to static content, needs to be frequently updated, and as a consequence may not be cacheable. For example, a telephone call or data associated with a highly interactive website (e.g., for gaming or providing real-time stock price data) may not be cacheable. Other content like magazine stories, movies, picture files, or software updates are static content that, once stored in a CDN's cache, does not require updating. For further discussion on the challenges of supporting dynamic content, see, for example, Leighton (2009) or Nygren et al. (2010).