Chapter 1
CDN Modeling and Performance

Benjamin Molina
Universitat Politecnica de Valencia, Spain

Carlos E. Palau
Universitat Politecnica de Valencia, Spain

Manuel Esteve
Universitat Politecnica de Valencia, Spain

ABSTRACT

Content Distribution Networks (CDN) appeared a decade ago as a method for reducing latencies, improving performance experienced by Internet users, and limiting the effect of flash-crowds, so as balance load in servers. Content Distribution has evolved in different ways (e.g. cloud computing structures and video streaming distribution infrastructures). The solution proposed in early CDN was the location of several controlled caching servers close to clients, organized and managed by a central control system. Many companies deployed their own CDN infrastructure– and so demonstrating the resulting effectiveness. However, the business model of these networks has evolved from the distribution of static web objects to video streaming. Many aspects of deployment and implementation remain proprietary, evidencing the lack of a general CDN model, although the main design concepts are widely known. In this work, the authors represent the structure of a CDN and the performance of some of its parameters, using queuing theory, simplifying the redirection schema and studying the elements that could determine the improvement in performance. The main contribution of the work is a general expression for a CDN environment and the relationship between different variables like caching hit ratios, network latency, number of surrogates, and server capacity; this proves that the use of CDN outperform the typical client/server architecture.

DOI: 10.4018/978-1-4666-1794-0.ch001

Copyright © 2012, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

INTRODUCTION

Few things compare with the growth of the Internet over recent years. A key challenge for Internet infrastructure has been delivering increasingly complex data of different types and origin to a growing user population. The need to scale led to the development of clusters (Mendonça et al, 2008), global content delivery networks (Verma, 2002) and, more recently, P2P structures (Androutsellis-Theotokis et al, 2004). However, the architecture of these systems differs significantly, and the differences affect their performance, workloads, and the role that caching can play (Gadde et al, 2000;Sariou et al, 2002).

Content Delivery Networks (CDNs) are overlay networks across the wide-area Internet which consist of dedicated collections of servers, called surrogates, distributed strategically throughout the Internet. The main aim of the surrogates is to be close to users and provide them with content in a low-latency mode. The surrogates are normally proxy caches that serve cached content directly with a certain hit ratio; the uncached content is previously obtained (if possible) from the origin server before responding. When a client makes a request for content inside a CDN, it is directed to an optimal surrogate, which serves this content within low response time boundaries – at least compared to contacting the origin site (Cardellini et al, 2003). CDNs such as Akamai (Akamai, 2011) or Limelight Networks (Limelight Networks, 2011) are nowadays used by many websites as they effectively reduce the client-perceived latency and balance load (Johnson et al, 2000). They accomplish this by serving content from a dedicated, distributed infrastructure located around the world and close to clients. The content is replicated either on-demand, when users request it, or replicated beforehand, by pushing the content on the content servers (Dilley et al, 2002; Verma et al, 2002). CDN services can improve client access to specialized content by assisting in four basic areas:

- *Speed*, reducing the response and download times of site objects (e.g. streaming media), by delivering content close to end users.
- *Reliability*, by delivering content from multiple locations; a fault-tolerant network with load balancing mechanisms can be implemented.
- *Scalability*, both in bandwidth, network equipment and personnel.
- *Special events*, by incrementing capacity and peak loads for special situations by distributing content as it is needed (Yoshida, 2008).

CDNs improve performance and availability of web and some media content by pushing the content towards the network edges and providing replication and replica location services. Intelligent replica placement improves response time by serving content from a topological location near the client (in terms of network hops), avoiding the congested backbone networks and network access (Mao et al, 2002). Replica location services direct requests for objects to nearby replicas by means of redirections through DNS, based on extensive measurements and monitoring of network performance (Shaikh et al, 2001). The overall performance of a CDN is largely determined by its ability to direct client requests to the most appropriate server (Johnson et al, 2000; Doyle et al, 2002; Khan et al, 2008). Content providers, such as websites or streaming video sources, contract with commercial CDNs to host and distribute content (Cranor et al, 2001). They are attractive for content providers because in some cases the responsibility is offloaded to the CDN infrastructure. Most CDNs have servers in ISP points of presence, so clients can access topologically nearby clients with very low latencies. They are capable of sustaining large workloads and flash-crowds due to a large number of servers, or few but powerful servers (Dilley et al, 2002). The main features of a CDN are:

26 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/cdn-modeling-performance/66991

Related Content

Big Data Analytics and Internet of Things in Industrial Internet in Former Soviet Union Countries
Vardan Mkrttchian, Leyla Ayvarovna Gamidullaeva, Svetlana Panasenkoand Arman Sargsyan (2019). *Handbook of Research on Big Data and the IoT (pp. 359-378).*
www.irma-international.org/chapter/big-data-analytics-and-internet-of-things-in-industrial-internet-in-former-soviet-union-countries/224279

Advanced Java ME Programming
Wen-Chen Hu (2009). *Internet-Enabled Handheld Devices, Computing, and Programming: Mobile Commerce and Personal Data Applications (pp. 306-332).*
www.irma-international.org/chapter/advanced-java-programming/24708

Intelligent Infrastructure of Route Scheduling for Smart Transportation Systems in Smart Cities
Shiplu Das, Buddhadeb Pradhan, Shivam Sharma, Bishwanath Jana, Gobinda Dasand Prasit Chakraborty (2023). *Handbook of Research on Network-Enabled IoT Applications for Smart City Services (pp. 174-188).*
www.irma-international.org/chapter/intelligent-infrastructure-of-route-scheduling-for-smart-transportation-systems-in-smart-cities/331332

Data-Driven Mall Advertising
Jiaxing Shen, Yi Lauand Jiannong Cao (2019). *Smart Marketing With the Internet of Things (pp. 123-138).*
www.irma-international.org/chapter/data-driven-mall-advertising/208509

Exploring Internet and Politics: E-Mailing Lists as Political Spaces for Social Movements
Andrea Calderaro (2012). *E-Politics and Organizational Implications of the Internet: Power, Influence, and Social Change (pp. 259-276).*
www.irma-international.org/chapter/exploring-internet-politics/65219