By George Pallis and Athena Vakali

Insight and Perspectives for CONTENT DELIVERY NETWORKS

Striking a balance between the costs for Web content providers and the quality of service for Web customers.

More efficient content delivery over the Web has become an important element of improving Web performance. Content Delivery Networks (CDNs) have been proposed to maximize bandwidth, improve accessibility, and maintain correctness through content replication [11]. With CDNs, content is distributed to cache servers located close to users, resulting in fast, reliable applications and Web services for the users.

More specifically, CDNs maintain multiple Points of Presence (PoP) with clusters of (the so-called surrogate) servers that store copies of identical content, such that users' requests are satisfied by the most appropriate site (see the figure here). Typically, a CDN topology involves:

- A set of surrogate servers (distributed around the world) that cache the origin servers' content;
- Routers and network elements that deliver content requests to the optimal location and the optimal surrogate server; and
- An accounting mechanism that provides logs and information to the origin servers.

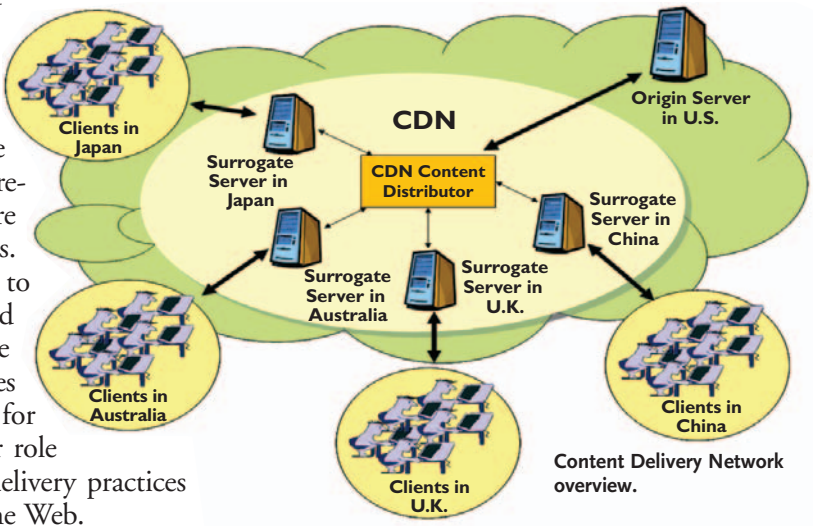
Under a CDN, the client-server communication is replaced by two communication flows: one between the client and the surrogate server, and another between the surrogate server and the origin server. This distinction into two communication flows reduces congestion (particularly over popular servers) and increases content distribution and availability. To maintain (worldwide) distributed copies of identical content, the practice for a CDN is to locate its surrogate servers within strategic data centers (relying on multiple network providers), over a globally distributed infrastructure. In this context, the most indicative advantages from using CDNs are:

- Reducing the customer's need to invest in Web site infrastructure and decreasing the operational costs of managing such infrastructure;
- Bypassing traffic jams on the Web, since data is closer to user and there is no need to traverse all of the congested pipes and peering points;

- Improving content delivery quality, speed, and reliability; and
- Reducing the load on origin servers.

Organizations offering content to a geographically distributed and potentially large audience (such as the Web), are attracted to CDNs and the trend for them is to sign a contract with a CDN provider and offer their site's content over this CDN. CDNs are widely used in the Web community, but a fundamental problem is that the costs involved are quite high. The sidebar "CDNs: Current Status" lists the most popular CDN providers and gives a historical background for the CDN evolution.

Since CDNs are in a rather recent and evolving status, it is important to understand their value and their implications. In [11], we presented a survey of CDN architecture and popular CDN service providers. The purpose of that survey was to understand the CDN framework and its usefulness. Here, we identify the most characteristic current practices and present an evolution pathway for CDNs, in order to understand their role in the recent evolution of content delivery practices over distributed environments and the Web.



Content Delivery Network overview.

CDNs in Practice

Several issues are involved in CDN content delivery since there are different decisions related to where to locate surrogate servers, which content to outsource, and which practice to use for (selected content) outsourcing. It is obvious that each decision for these issues results in different costs and constraints for CDN providers. Critical issues involved in content delivery practices are summarized here.

Surrogate Servers Placement. Choosing the best location for each surrogate server is important for each CDN infrastructure since the location of surrogate servers is related to important issues in the content delivery process. Determining the best network locations for CDN surrogate servers (known as the Web server replica placement problem) is critical for content outsourcing performance and the overall content distribution process. CDN topology is built such that the client-perceived performance is maximized and the infrastructure's cost is minimized. Therefore, effective surrogate server placement may reduce the number of surrogate servers needed and the size of content (replicated on them), in an effort to combine the high quality of services and low CDN prices. In this context, several placement algorithms have been proposed (such as Greedy[1], which incrementally places replicas, Hot Spot [10], which places replicas near the clients generating the greatest load, and Tree-based[2] replicas). These algorithms specify the locations of the surrogate servers in order to achieve improved performance with low infrastructure cost. Earlier experimentation has shown that the greedy placement strategy can yield close to optimal performance [10].

Content Selection. The choice of the content that should be outsourced in order to meet customers' needs is another important issue in the content selection problem. An obvious choice is to outsource the entire set of origin servers' objects to other surrogate servers (the so-called entire replication). The greatest advantage of entire replication is its simplicity, however, such a solution is not feasible or practical because although disk prices are continuously dropping, the sizes of Web objects increase as well (such as audio or video on demand). Moreover, the problem of updating such a huge collection of Web objects is unmanageable. Therefore, the challenge of the content selection problem is to find a sophisticated management strategy for replication of Web content.

A typical practice is to group Web content based on either correlation or access frequency and then replicate objects in units of content clusters. Two types of content clustering have been proposed:

- Users' sessions-based: the content of the Web log

[1]Weisstein, E. Greedy Algorithm. A Wolfram Web Resource; mathworld.wolfram.com/GreedyAlgorithm.html.
[2]Li, B. et al. On the optimal placement of Web proxies in the Internet. In *Proceedings of the 18th IEEE INFOCOM Conference* (New York, Mar. 1999), 1282–1290.

files[3] is exploited in order to group together a set of users' navigation sessions showing similar characteristics. Clustering users' sessions is useful for discovering both groups of users exhibiting similar browsing patterns and groups of pages having related content based on how often URL references occur together across them.

- URL-based: Web content is clustered using the Web site topology (which is considered as a directed graph), where Web pages are vertices and hyperlinks are arcs. The Web pages (URLs) are clustered by eliminating arcs between dissimilar pages. In [1], the authors identify the most popular objects from a Web site, (the so-called hot data), and replicate them in units of clusters where the correlation distance between every pair of URLs is based on a certain correlation metric. Furthermore, several coarse-grain dynamic replication schemes where the Web content is replicated on per-group granularity are proposed in [3]. By using these replication schemes, the performance of the Web services can be significantly improved. The relevant evaluation experiments showed that a clustering-based replication can reduce client download time and server load by 4.6 to 8 times compared to the entire replication since the popularities of Web objects are quite localized. A disadvantage of these schemes is the high complexity of the processes often involved.

Content Outsourcing. Under a CDN infrastructure with a given set of surrogate servers and a chosen content for delivery it is crucial to decide which content outsourcing practice to follow. To date, three distinct content outsourcing practices have appeared.

Cooperative push-based: Content is pushed (proactively) from the origin Web server to CDN surrogate servers. Initially, the content is prefetched (loaded in cache before it is accessed) to the surrogate servers and then, the surrogate servers cooperate in order to reduce the replication and update cost. In this scheme, the CDN maintains a mapping between content and surrogate servers, and each request is directed to the closest surrogate server (that has the requested object), or otherwise, the request is directed to the origin server. Several replication strategies for cooperative push-based schemes over CDNs have been studied in [7], where it is noted that greedy-global heuristic algorithms are the best choice in making the replication decisions between cooperating surrogate servers. This approach has been proposed at a more theoretical

[3]Web log files provide information about activities performed by a user from the moment the user enters a Web site to the moment the same user leaves it.

CDNs: Current Status

Most Popular CDN Providers

Akamai Technologies (www.akamai.com) is the market leader (80% of the overall CDN market) in providing content delivery services. It owns more than 12,000 servers over 1,000 networks in 62 countries.

Mirror Image Internet, Inc. (www.mirror-image.com) supports surrogate servers located in 22 cities around the world (North America, Europe, and Asia), which provide a range of value-added services, from content distribution to media streaming and managed caching.

Inktomi, a Yahoo Company (www.inktomi.com) provides managed services for global load balancing, failover, content delivery, and streaming media using more than 1,000 surrogate servers worldwide.

LimeLight Network (www.limelightnetworks.com) provides a suite of services (including music download and subscription services, video game developers and distributors, movie/video download services, and so forth) and supports surrogate servers located in 72 locations around the world (Asia, the U.S., and Europe).

Historical Background

1998—First CDNs appear. Companies realize they could save money by putting more of their Web sites on a CDN, getting increased reliability and scalability without expensive hardware.
1999—Several companies (such as Akamai and Mirror Image) become the specialists in providing fast and reliable delivery of Web content, earning large profits.
2000—In the U.S. only, CDNs are a huge market generating $905 million with the expectation to reach $12 billion by 2007.
2001—The flash crowd event [2] (numerous users access a Web site simultaneously), such as the one that occurred Sept. 11, 2001 when users flooded popular news sites with requests about the terrorist attacks in the U.S., resulted in serious caching problems since sites had typically become unavailable. Flash events transfer more dollars to CDN sales income, since CDNs provide the desired level of protection to Web sites against them.
2002—Large-scale ISPs (such as AT&T) tend to build their own CDN functionality, providing customized services.
2004—More than 3,000 companies use CDNs, spending more than $20 million monthly (see www.irg-intl.com). Marketing research [1] shows that CDN providers have doubled their revenue derived from their streaming media operations in 2004 compared to 2003. Furthermore, many CDN providers are trying to move Web services (such as Microsoft .NET and Java 2 Platform Enterprise Edition) closer to users.
2005—CDN revenue for both streaming video and Internet radio is estimated to grow at 40%, spending more than $450 million for delivery of news, film, sports, music, and entertainment [1]. C

References

1. *CDN Market Share: A Complete Business Analysis 2004 and 2005.* AccuStream iMedia Research; www.researchandmarkets.com.
2. Jung, Y. et al. Flash crowds and denial of service attacks: Characterization and implications for CDNs and Web sites. In *Proceedings of the 11th International World Wide Web Conference*, (Hawaii, May 2002), 293–304.

level, since it has not yet been adopted by a CDN provider [1, 3].

Uncooperative pull-based: Clients' requests are directed (by using either DNS redirection[4] or URL rewriting[5] mechanisms [3]) to their closest surrogate server. If there is a cache miss and the requested content is not found, the request is directed either to a peering surrogate server of the underlying CDN or to the origin server. More specifically, the surrogate servers, which serve as caches, pull content from the origin server when a cache miss occurs. A problem in this practice is that CDNs do not always choose the optimal server from which to serve the content (as pointed out in [6]). However, many popular CDN providers use uncooperative pulling (such as Akamai and Mirror Image), since the cooperative push-based schemes are still at the experimental stage.

Cooperative pull-based: Client requests are directed through DNS redirection to their closest surrogate server. The key in the cooperative pull-based CDNs (such as Coral[6]) is that the surrogate servers are cooperating with each other in case of cache misses. Specifically, using a distributed index, the surrogate servers find nearby copies of the requested objects and store them in their caches.

CDN Pricing. Commercial-oriented Web sites turn to CDNs to contend with the high traffic problems while providing high data quality and increased security for theirs clients in order to increase their profit and popularity. CDN providers charge their customers—owners of Web sites—according to their traffic (delivered by their surrogate servers to the clients).

There are technical and business challenges in pricing CDN services. The services that are usually delivered by a CDN infrastructure include video on demand, electronic books, and news services. But how should CDN services be priced? Pricing of CDN services is a relatively new and unexplored issue, however, the use of analytical models to address the optimal prices of such services is discussed in [4]. This work concluded the prices of CDNs will decline (and at the same time will accelerate the content delivery process on a Web site) based on the recent trends (such as decreasing bandwidth costs) and their impact on CDN pricing policies. Moreover, from a recent CDN market report,[7] it is evident that CDN prices are quite high (since the average cost per gigabyte of streaming video transferred in 2004 was $1.75, whereas the average price to deliver a gigabyte of Internet radio was $1). The most indicative factors affecting the pricing of CDN services include:

- Bandwidth cost;
- Variation in traffic distribution;
- Size of content replicated over surrogate servers;
- Number of surrogate servers;
- Reliability and stability of the whole system; and
- Security issues of outsourcing content delivery.

According to the marketing practices, cost reduction occurs when an information technology investment enables a firm to produce more (of a given service) with fewer resources. Hence, an obvious solution in order to decrease the CDN pricing services would be to increase the bandwidth, but such a choice involves increasing economic cost. However, the higher bandwidth would temporarily solve the problems since it would only allow users to create more resource-hungry applications, further congesting the network. Therefore, the bandwidth limitation induces high communication and economic costs for CDN clients.

CDNs: How to Proceed

It is interesting to identify a pathway for CDN evolution since CDNs are still evolving and there are certain requirements that should be met. Here, we propose particular techniques toward improving CDN quality of service and performance. The following ideas serve as a guideline for potential practices that could be integrated into the existing CDN framework.

Exploit Caching under CDNs. Content selection and outsourcing are mostly related to the client-perceived services by a CDN. Since caching over the Web has been a more mature practice (than CDNs), it is interesting to understand if (and how) employing particular caching-related processes on a CDN would result in better performance and content accessing. Considering caching under CDNs is a simple idea since surrogate servers are equipped with caches that can and should be exploited. Some ideas have already been highlighted and in an effort to further develop the initial caching on CDNs points, the following issues appear to be critical.

Web Prefetching: A process of deducing client's future requests for Web objects by moving popular requested objects into the cache prior to an explicit request for them. The potential main advantages of adopting prefetching over a CDN infrastructure include preventing bandwidth underutilization and

[4]DNS performs the mapping between a surrogate server's symbolic name and its numerical IP address.
[5]The origin server redirects clients to different surrogate servers by rewriting the dynamically generated pages' URL links.
[6]The Coral Content Distribution Network; www.coralcdn.org/overview.
[7]CDN Market Share: A Complete Business Analysis 2004 and 2005. AccuStream iMedia Research; www.researchandmarkets.com.

reducing a significant part of the latency involved. The practice of prefetching in CDNs has been discussed in [12], where the costs and benefits of prefetching in CDNs are highlighted. These results have shown that CDNs can achieve significant benefits at modest costs by focusing on the most popular long-lived objects. More specifically, the long-term prefetching increases disk space costs but it benefits the CDN infrastructure since it improves the hit rate (a variable that actually reflects the users' satisfaction from the system).

Surrogate Server Cache Segmentation. Each cache of a surrogate server may be partitioned logically in several domains to provide more flexible memory management. This practice will be beneficial for reducing the CDN costs since an "intelligent" cache segmentation (a cache may be partitioned on semantic domains that have a specific meaning) on surrogate servers will increase cache hits and reduce accessing costs. The cache segmentation practice in CDNs may be based on the similar practice in conventional Web information management systems, which has been proven to significantly improve performance on the Web [8]. Furthermore, cache segmentation is quite promising for CDNs, since the cache segments may grow and shrink deliberately (according to request streams) and also at each segment a separate replacement policy may be applied.

Data Mining Approach	CDN Practices Involved	Advantages				
		Save Bandwidth	Reduce Traffic	Reduce Number of Surrogate Servers	Reduce Transfer Rate	Improve Security
Similarity-based Clustering	Content personalization, Prefetching, Cache's segmentation	√	√	√	√	
Link-based Clustering	Content personalization, Prefetching	√	√	√	√	
Model-based Clustering	Content personalization, Prefetching, Cache's segmentation	√	√	√	√	
Bayesian Networks	Content personalization, Prefetching	√	√		√	√

Popular data mining practices and their role in CDNs.

Meet CDN User Preferences. Meeting the user preferences is crucial for CDNs and an initial practice is to consider content personalization: adopt a content management task by which the content is personalized to meet the specific needs of each individual user (or group of users). Such a practice in CDNs may be inspired by the Web personalization system presented in [9], where the user preferences were automatically learned from Web usage data by using data mining techniques.

Some indicative objectives of content personalization over CDNs are highlighted next:

- Deliver the appropriate content to the interested users in a timely, scalable, and cost-effective manner;
- Increase the quality of the published content by ensuring it is accurate, authorized, updated, easily searched and retrieved, as well as personalized according to various users and audiences;
- Manage the content throughout its entire life cycle from creation, acquisition, or migration to publication and retirement; and
- Meet security requirements since introducing content personalization on CDNs will facilitate the security issues raised in [2] such as authentication, signing, encryption, access control, auditing, and resource control for ensuring content security and users' privacy.

Employing Data Mining over CDNs. Data mining techniques seem to offer an effective benefit for CDNs, since CDNs manage large collections of data over highly distributed infrastructures. In this context, data mining practices have been related to CDNs in [1], and these practices could provide effective ways of dealing with the difficulties (such as traffic, billing) of large-scale data management involved over a CDN. Therefore, CDN developers and customers may exploit data mining solutions in order to improve CDN pricing, topology, and content outsourcing.

In response to the question "Why use data mining over CDNs?" the following replies often reoccur:

To detect relevant objects: So that push-based CDN schemes (or prefetching) are facilitated. The relevant objects could be identified by employing well-known clustering techniques that are mostly similarity-based (use distance metrics such as Euclidean, cosine) [5];

To identify a CDN topology: Since link-based clustering techniques [5] might be used by considering the Web graph properties so the location of surrogate servers might be identified by Web graph clustering;

To determine clusters of pages: To address the content selection problem by selecting clusters of content for content outsourcing. Various mining techniques such as model-based clustering (use probability distributions for each cluster of pages) may be used to facilitate content outsourcing;

To define clusters of users: In order to facilitate con-

tent personalization by using existing practices (as the ones based on belief functions, Bayesian networks, or Markov models) for classifying users over clusters.

The table here highlights some particular data mining practices and the issues involved from the CDN side in an effort to understand the importance and the challenge in adopting such practices under a CDN framework.

Conclusion

CDNs are still in an early stage of development and their future evolution remains an open issue. It is essential to understand the existing practices involved in a CDN framework in order to propose or predict the evolutionary steps. The challenge is to provide a delicate balance between costs and customers satisfaction. In this framework, caching-related practices, content personalization processes, and data mining techniques seem to offer an effective roadmap for the further evolution of CDNs. C

Meeting the user preferences is crucial for CDNs and an initial practice is to consider content personalization: adopt a content management task by which the content is personalized to meet the specific needs of each individual user (or group of users).

References

1. Chen, Y. et al. Efficient and adaptive Web replication using content clustering. *IEEE Journal on Selected Areas in Communications 21*, 6 (Aug. 2003), 979–994.
2. Fink, J. et al. Putting personalization into practice. *Commun. ACM 45*, 5 (May 2002), 41–42.
3. Fujita, N. et al. Coarse-grain replica management strategies for dynamic replication of Web contents. *Computer Networks 45*, (2004), 19–34.
4. Hosanagar, K. et al. Optimal pricing of content delivery network services. In *Proceedings of the 37th International Conference on System Sciences* (Big Island, Hawaii, Jan. 2004).
5. Jain, A. et al. Data clustering: A review. *ACM Computing Surveys 31*, 3 (Sept. 1999), 264–323.
6. Johnson, K.L. et al. The measured performance of content distribution networks. *Computer Communications 24*, 2 (Feb. 2001), 202–206.
7. Kangasharju, J. et al. Object replication strategies in content distribution networks. *Computer Communications 25*, 4 (Mar. 2002), 367–383.
8. Katsaros, D. and Manolopoulos, Y. Caching in Web memory hierarchies. In *Proceedings of the 19th ACM Symposium on Applied Computing*, (Nicosia, Cyprus, Mar. 2004), 1109–1113.
9. Mobasher, B. et al. Automatic personalization based on Web usage mining. *Commun. ACM 43*, 8 (Aug. 2000), 142–151.
10. Qiu, L. et al. On the placement of Web server replicas. In *Proceedings of the 20th IEEE INFOCOM Conference* (Anchorage, Alaska, Apr. 2001), 1587–1596.
11. Vakali, A. and Pallis, G. Content delivery networks: Status and trends. *IEEE Internet Computing 7*, 6 (Nov./Dec. 2003), 68–74.
12. Venkataramani, A. et al. The potential costs and benefits of long term prefetching for content distribution. *Computer Communications 25*, 4 (Mar. 2002), 367–375.

George Pallis (gpallis@ccf.auth.gr) is a Ph.D. candidate in the Department of Informatics at Aristotle University of Thessaloniki in Greece.

Athena Vakali (avakali@csd.auth.gr) is an assistant professor in the Department of Informatics at Aristotle University of Thessaloniki in Greece.