

The Future of Global Load Balancing: Customization through Virtualization

Arijit Ghosh

February 2010



Introduction

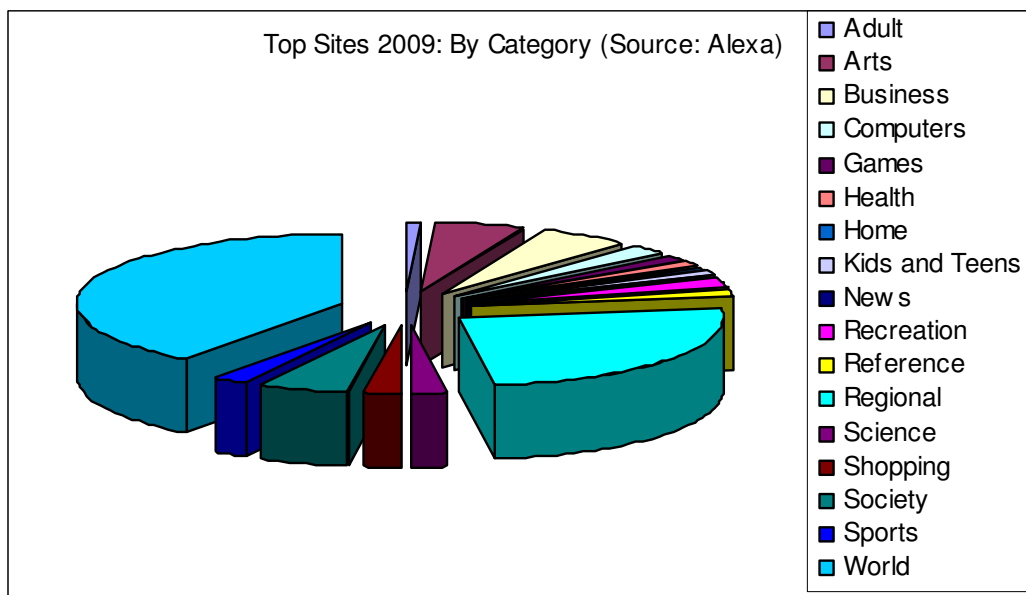
Towards the end of the last millennium, the Internet exploded into a platform to deliver different kinds of content, including images, audio and video to end users spread out across the world. Traditional content delivery networks (CDNs) were created around the same time to address the three main issues of content providers:

- How can I deliver a great end-user experience?
- How can I deliver to a global audience 24x7?
- How can I scale to serve millions of users?

However, in the past few years, the variety of both content and content providers has increased dramatically. The Internet is now used to deliver a wide range of content for a variety of applications - from software-as-a-service to social networking sites. Content providers range from global Fortune 500 companies to teenagers producing user-generated content in their back yard. Because there is such a wide variety of diverse content and applications, the need for customization and personalization for content delivery is significant. Content providers have different, often orthogonal, objectives: some business, some social and even some political. As such, providers are now asking the question, "How can the CDN be customized to better serve the content provider's unique needs?" This paper discusses how content delivery methods can be tailored to meet specific objectives without compromising the general requirements for performance, scalability and reliability.

The Content Explosion

A decade ago, when CDN services were first introduced by a small handful of companies, the reach of the Internet as well as the content on the Internet was relatively limited. According to Internet World Stats (<http://www.internetworldstats.com/stats.htm>), about 361 million people across the world had access to the Internet in 2000. At the time, this represented 25% of the world population. In the past decade, Internet penetration has grown by a staggering 380% with some parts of the world like Africa and the Middle East registering quadruple digit growth. Within the same time frame, the type of content offered has diversified dramatically. A decade ago, the predominant type of content was limited mostly to news and commercial products. Now, there are many more diverse categories of content served over the Internet, as illustrated in the chart below.



Challenges of Content Explosion

The massive explosion of content presented numerous challenges that CDNs had to address:

- How can the user-perceived response time be reduced?
 - Online games, for example, require latencies of less than 20 ms. This problem is compounded by the fact that users can be in different networks across the globe, each with different latency characteristics.
- How can the system scale to thousands and even millions of customers? What happens when all of the users try to access content at the same time, creating what is known as a “flash crowd” effect? What if it is not possible to predict when this will happen?
- How can content availability be guaranteed?
 - The Internet was designed for best-effort delivery with no guarantees. But for today’s content, such as Internet TV, best effort is simply not good enough.
- How can a CDN be customized to meet the unique needs of different content providers?
 - Content providers often have very unique needs which are much different from other providers. For example, a content provider might want to offer a “premium” end-user experience to a certain geographical location, and a “standard” end-user experience to the rest of the world

Over the past years, traditional CDNs have done an excellent job of addressing the first three issues described above. The approach is to replicate content across thousands of servers located in hundreds of datacenters around the world, and transparently redirect user requests to the best available server. This redirection is done by the Global Load Balancer (GLB). The GLB runs an optimization algorithm that maximizes an objective function under the constraints imposed by the state of the servers and the network. Server state is determined by CPU load, memory availability, disk utilization, and system resource availability. Network state is determined by available bandwidth, latency between users and servers, and rate of packet loss. The optimization objective of the GLB depends on the application and/or the content. Common objectives include reduction in response time, increased scalability, improved availability, increased resource utilization, and maximized throughput.

Providing customization as a central feature of the GLB, and hence by extension the CDN, has so far been largely ignored. This is primarily for two reasons:

1. It is a very difficult problem to solve. Building a GLB that can optimize multiple, often conflicting, objectives under a set of constraints is mathematically impossible to solve.
2. The need for customization has only recently become very important, as the spectrum of content types continues to grow rapidly.

The Need for Customization

As previously indicated, different content providers may have different socio-economic and political objectives in addition to their basic requirements for performance, scalability and reliability. This depends on the nature of the content they are providing, as well as the profile of their end-users. Some owners are very sensitive to maintaining social integrity, and may not want their content to reside on servers with morally questionable content. For example - The goal of the children's website *MrsP.com* is to give kids the wonderful experience of having a trusted, skilled storyteller read them classics of literature. The audience being children, it is very important that the integrity of the content is maintained and that by mistake, they are not redirected to an adult content or gambling site.

Content providers that offer premium, fee-based services are concerned with offering the highest possible end-user experience to keep subscribers from churning. However, many other content providers are simply interested in delivering content at the lowest possible cost. Many other content providers fall in between these two extremes.

Increasingly, content owners are using the Internet to get initial feedback on their pilot content before producing additional content. During that period, they might want to restrict access to end-users within a specific demographics.; while performance and quality experience is important in this example, so is the ability to limit content access to that specific target demographic.

Blocking content access for security reasons is also common practice. During the 2004 Presidential elections, former President Bush's website (www.georgebush.com) blocked viewers from outside the United States to minimize the chances of a network attack. In situations like these, the ability of the CDN to provide geo-blocking capabilities is as important as performance.

A content provider's regional focus is also important. CDN service providers charge content owners based on server and/or bandwidth usage. Imagine the website of a business located in Boston that only serves the people living in that metropolitan area. To maximize the end-user experience to their target market and reduce delivery costs, they might want their content to be delivered from servers located only in this region. Delivering a high-quality experience to other parts of the country or the world would unnecessarily increase delivery costs, as they do not provide services outside of their region.

The Solution: Virtualization

In order to satisfy the need for customization, a global load balancer, along with virtualization technologies can be deployed to optimize multiple, often conflicting, objectives under constraints imposed by the servers and networks. It is impossible for a single GLB solution to solve this problem. The optimal way to address this problem is to build a customizable GLB that can be configured for a unique set of features as mandated by the content provider.

If the GLB is considered as a computer resource and if virtualization is offered on top of it, it is possible to create multiple "virtual instances" of the GLB, each with its unique set of features. A GLB virtualization technology will work by restricting the basic GLB functionality, like load balancing, to the actual GLB software. The virtualization layer will be responsible for providing the more esoteric features, such as geographical restrictions and server grouping (based on cost metrics, for example).

A virtualized GLB offers many advantages:

- Enables the content providers to create a customized load balancer which serves their unique needs. Giving the control to providers for dictating the behavior of GLBs and by extension, the CDN, will lead to increased customer satisfaction and hence, increased revenue.
- Enables instantiation of virtual GLBs to serve the customer's needs. This dramatically reduces the investment of the CDN providers by eliminating the need to build multiple GLBs.
- Provides extensibility to the GLB. It is virtually impossible for anyone to predict how the GLB might be used in the future. While designing an all-inclusive feature suite is unrealistic, designing for extensibility is certainly possible. In the future, with the evolution of the user market, the objectives of the content owners might change too. The load balancer must be designed in such a way that it is easy to add newer objectives.

Thus, virtualization is a relatively inexpensive way to incorporate customization as a central feature of a CDN.

Conclusion

The Internet as a content delivery platform has evolved considerably over the past several years. Content from a wide variety of providers with disparate needs and objectives places unique demands on the CDN. Traditional CDNs are unable to account for the level of personalization and customization required by the content providers of today. Building a CDN that simultaneously serves the conflicting needs of all providers is of course impossible. However, an alternative approach based on virtualizing the global load balancer allows customization at unprecedented levels. With the flattening of the Internet economy, it is imperative that businesses provide their customers with a better end-user experience and more control.

About CDNetworks

CDNetworks is a top tier, full-service, global content delivery network (CDN), providing technology and services that enable fast, efficient, reliable delivery of content anywhere in the world. CDNetworks offers a comprehensive suite of services and solutions for caching, live/on-demand video streaming, large file downloads, application acceleration, and whole site acceleration. Some of the world's top companies in media, entertainment, technology, retail, and online gaming rely on CDNetworks to maximize their user experience while minimizing delivery costs. With more than 80 points of presence across 5 continents, CDNetworks provides true global coverage and unparalleled network performance backed by 24/7/365 service and support. Founded in 2000, CDNetworks has offices in the US, Korea, China, Europe, and Japan.

CDNetworks Global Offices



Korea

Handong Bldg., 828-7 Yeoksam-dong, Gangnam-gu, Seoul
Tel. +82-2-3441-0400



USA

130 Rio Robles, San Jose, CA 95134
37 W. 20th St. Suite 1105 NY, NY 10011
Tel. +1-408-228-3700



Japan

MY ARK Nihonbashi Bldg., 3F, 10-16 Tomizawa-cho,
Nihonbashi, Chuoh-ku, Tokyo
Tel. +81-3-6667-6655



China

A-1502, Keijidalou, 900 Yi shan Road, Shanghai
Tel. +86-21-5423-4802



EMEA

8, rue de l'Isly, 75008 Paris, France
Tel. +33-0-1-75-43-81-92

www.cdnetworks.com