*XXVIII Simpozijum o novim tehnologijama u poštanskom i telekomunikacionom saobraćaju – **PosTel 2010**, Beograd, 14. i 15. decembar 2010.*

CONTENT DELIVERY NETWORKS AND THE FUTURE INTERNET

Zoran Miličević[1], Zoran Bojković[2]
[1]Telecommunication and IT Department Genaralstaff of the SAF in Belgrade
[2]University of Belgade

Abstract: *The future Internet architecture should preserve the features of best effort while offering differential quality of service (QoS). In order to provide additional functionality and flexibility, an efficient solution is to build the content delivery networks (CDNs) with the specific advantage of optimizing the Internet. After showing in brief the state of the art in the current area, the evolution of content delivery technologies, together with content delivery network functions is described. The final part of the paper deals with potential solutions to the problems in IP networks.*

Key words: *content delivery technology, IP network, future Internet, content distribution.*

1. Introduction

Content delivery network (CND) is a comprehensive, end-to end solution for optimizing global networks for Web content delivery. Users requesting information from a Web site may well have those requests served from a location closer to them than the originals server on which it is generated. By serving content from points a lot closer to the user, a CDN reduces the likelihood of hot spots by dispersing the different points of convergence and by distributing the workload among multiple servers. Delivering content from the edge of the network instead of the original server has the added benefit of additional reliability. The probability of the lost packets is decreased and the performance of streaming audio and video is improved [1]. CDNs deploy servers in multiple geographically diverse locations in order to redirect users requests to the nearest available servers. End users observe higher QoS, while content providers offer more reliable and larger volumes of the service. At the same time, Internet service providers (ISPs) can also benefit from deploying CDN servers in their networks as the total amount of the traffic transmitted in the backbone is reduced [2].

From the perspective of the user and that of users ISP, the question often arises is whether or not the user will be able to gain the benefit of the fast access to an edge servers if the only CDN provider in the region is affiliated with a competing ISP. These questions raise issues that cannot be addressed entirely by technology. Issues can be addressed only through a business arrangement among the different ISPs, CDN provides, as well as content owners. These questions are being addressed through the information of Content Alliance. In August 2000 CiscoSystems announced the formation for the Content Alliance to speed the adoption of compatible CDN technology formed to help develop standards and protocols to advance content networking. The Alliance generates proposals for standards and depends on traditional standard bodies such as the Internet Engineering Task Force (IETF) to gain broad industry acceptance. The initial focus of the group was a term that describes the process that enables the CDNs of multiple independent service providers to work in cooperation. In addition, the Alliance is also focused on defining specification to address issues of authorization of the use of content among networks and sharing of logging or billing information for charge settlement. Content peering creates the ability to deliver the benefits of content delivery networks to global user base regardless of where the server is hosted and by whom.

Content peering requires the CDNs to share information in three areas: Content distribution, Content request – routing and accounting. By content distribution we mean the process of moving files to the remote delivery devices. Content request – routing is the process whereby a viewer's page request is redirected to the appropriate delivery devices. Finally, accounting represents the process for collecting usage and billing data.

The rest of this paper is organized as follows. At first, the evaluation of content delivery technology is presented. Secondly, we briefly describe technologies with sufficient coverage of content delivery networks, together with the corresponding network functions. Potential solutions to problems in IP network conclude the paper.

2. Evolution of the content delivery technology

As the Web sites grew and more functions were added, the problems of content management and the addition of dynamic features became more and more challenging. Application servers emerged and were used for content management as shown in Figure 1.

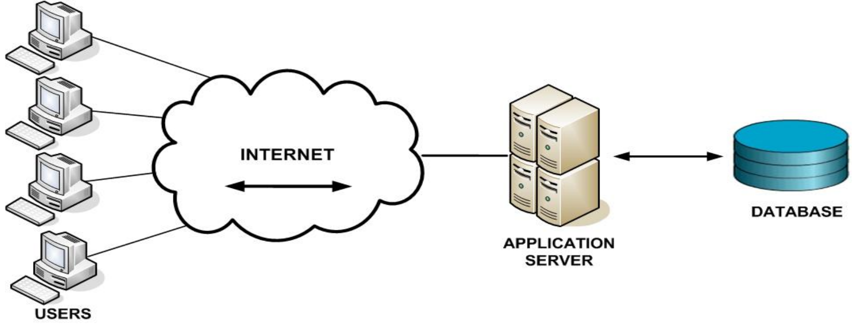

Figure1. *Application server position when used for content management.*

These servers were placed between the user and back – end business system application database and modern legacy servers. In that way, the back – end systems could continue to operate in the way they were designed, and continue to address the functions for which they were intended. The application server was used as a translator. On one side, it understood the Web – based structure of the user requests, and on the other hand the native structure of the serving database or server. In recent years, this intermediate layer, occupied by the application servers, has grown in function. This layer, the components of which are often referred to as middleware – has been significant development in an attempt to minimize the complexity of client programs and improve performance. Functions of security were also added to this layer to ensure the security of date and user traffic.

The growth of e – business and other e – enabled services has placed demands on content hosting and delivery that have resulted in a more complex infrastructure, built to deliver personalized and dynamic content. Within a content hosting data center, it is now common to find components that include routers, switches, firewalls, reverse proxy caches, devices that do load balancing, Web servers, application servers, database service and storage applications. Each component may be duplicated to ensure redundancy, provide adequate capacity and improve reliability and availability. An example for the architecture of a Web hosting service is shown in Figure 2.

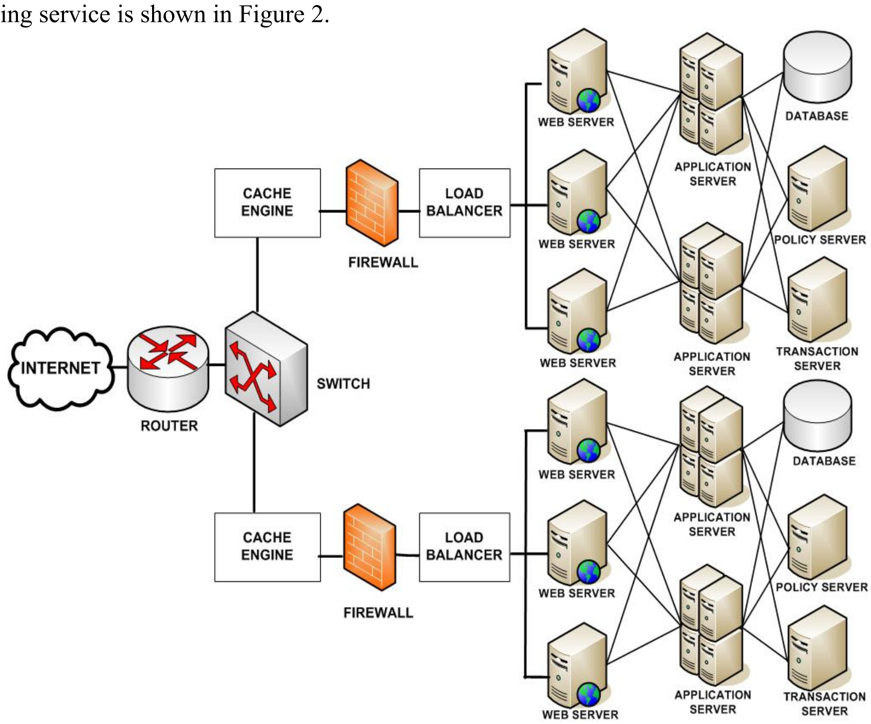

Figure2. *Architecture of a Web hosting service.*

Requests coming in from the Internet are first directed to the policy server, which validates the request against set of rules and authorized user lists. Valid requests are then sent to the load balancer for distribution to a Web server. These Web servers interact with an application server, which in turns interacts with the appropriate content source: a database server, a transaction server or some other server type.

When viewed as a whole, this architecture is just an extension of the client/server approach, where the server function is being delivered through the combination of many components acting as a single resource to the client. One can easily see that it is still a centralized approach to the delivery of content, and a centralized approach is subject to many weaknesses. For example a centralized approach is not scalable and does not address latency introduced with the network. Furthermore, a centralized approach limits global reach. It means that users in other parts of the world may have adequate infrastructure for accessing local content, but the links back to the country of the hosting service may be limited.

3. Technologies with sufficient coverage of the content delivery networks

The technologies can be classified into infrastructure and service as it is shown in Figure 3.

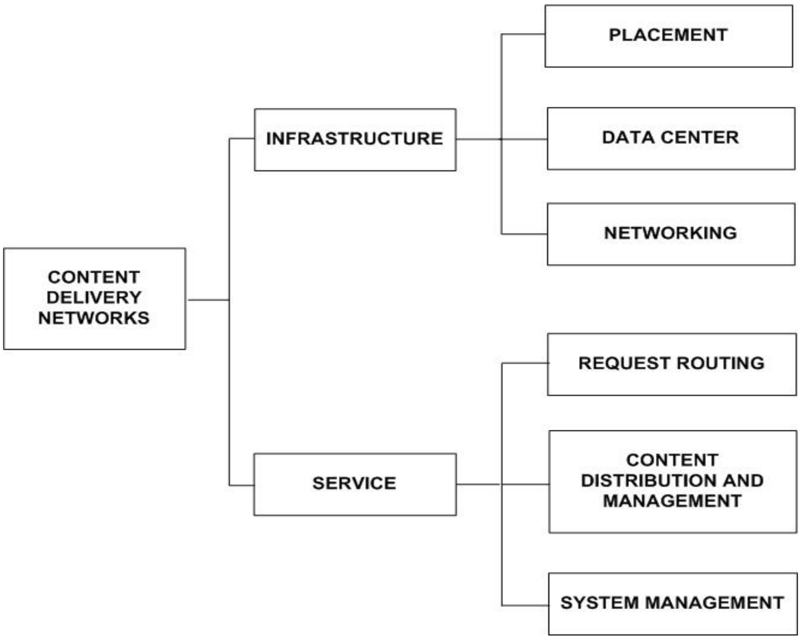

Figure3. *Technologies with coverage of content delivery networks.*

The infrastructure part involves three separate but relevant technologies like placement, data center and networking. They work together in order to find out a good trade – off between QoS and cost. The questions often arise are: how many data centers should the system have, where should the servers be allocated, how many servers should be deployed and organized in a given data center and finally how should the relationship between the data centers and Internet Service Providers (ISPs) be established?

The service part is composed of three main technologies: request routing, content distribution and management and system management.

The problem of how many data centers are enough has made it even more important to decide whether we should adopt highly distributed content delivery networks or big data center CDNs. The general view is that the larger the number of data centers, the better the user experience, but the higher the cost. A recent study indicated that the number of nodes can be reduced without noticeable degradation of the system performance [3]. It should be mentioned the Leighton has noted that highly distributed CDNs can also reduce cost by deploying servers in same regions that can host their servers for free [4]. In a data center, it is important to achieve efficient task migration among the servers in a data center structure. Several related technologies including virtualization are involved in order to reduce the configuration and accelerate deployment speed.

Request routing is responsible for routing client requests to an appropriate server for delivery of contents. The request routing system uses a set of metrics such as network proximity, client perceived latency, distance, and replica server load in an attempt to direct users to the most suitable server to best serve the request. The common request routing mechanism used in most commercial CDNs is Domain Name System (DNS) – based request routing. Its advantage is in maintaining lower overhead in the original server than Uniform Resource Location (URL) rewriting and having higher security as it conceals the source to end users. Because of this highly distributed design the task of network maintenance and management becomes very challenging. Sophisticated algorithms are required to shuffle data among the servers across the public Internet. Moreover, edge side includes technology and has been adopted to achieve dynamic web page acceleration where the dynamic web page is separated into two parts: static and dynamic. The static part is cased in an edge server, while the CDN edge server retrieves the dynamic part from the original server through an optimal path in order to reduce total response time. A critical challenge is how to combine or find other mechanisms to achieve efficient request routing. Thus, CDN request routing needs an integrated solution to accommodate heterogeneous systems. Moreover, several network applications (for example, online gaming) are latency-sensitive, and need more efficient and accurate routing strategies [5].

Content distribution and management includes content outsourcing, content delivery and content management technologies. The latter is largely dependent on the techniques for cache organization (caching techniques, cache maintenance and cache update). The recent challenge of content distribution in CDNs focuses on the acceleration of dynamic contents and computing applications, while have been considered uncacheable. The general solution is to optimize the path between the source and edge servers. One feasible way is to allocate them in the same data center. This approach cannot offer desirable performance for end users worldwide. Thus, the significant challenge is to find potential solutions to cache or partly cache dynamic contents or applications so that edge servers can be widely distributed to enhance the user experience.

System management is a significant challenge for the service provider to manage a large-scale distributed system. The traditional mechanisms typically focus on the management of servers. However, CDN servers are deployed in multiple Internet Service Providers (ISPs). So, it is essential to manage multiple resources including servers and networks. Efficient systems management can reduce investment in CDNs and becomes a significant challenge. An efficient tool was introduced in [6] to dynamically redirect client requests and achieve a route injection mechanism to efficiently change the content distribution path, thus reducing total traffic cost. System management includes operational support systems (OSSs) and business support systems (BSSs). An OSS mostly deals with supporting process such as maintaining inventory, providing services, configuring components, and managing faults. A BSS typically deals with customers, supporting processes such as taking orders, processing bills, and collecting payments.

4. Content delivery network functions

In order to present content delivery network functions, we use CDN distributed model shown in Figure 4.

This model divides the functions into three tiers: content generation tier, integration tier and an assembly and delivery tier.

The primary function of the content generation tier is the generation of the requested information from the legacy systems and transactions servers. Within this tier, a policy – based server may also be used to administer rules and security policies for users and user requests. The policy server works in conjunction with the firewall to administer the security polices. The application server understands the different data structures of the different servers that are used to serve content. The user request received by an application server is translated into a set of requests that must be sent to the appropriate content server, in a structure that these servers understand. The response must then be converted from the native structure of the serving platform to one that can be processed by the Web – site functions. Earlier versions of CDNs were capable of handling only static pages. The current versions of CDNs are now capable of serving personalized and dynamic pages through the use of a specification called

Edge Side Included (ESI). This is an open language for creating a uniform programming model that facilitates interoperation of ESI – compliant systems from different vendors.

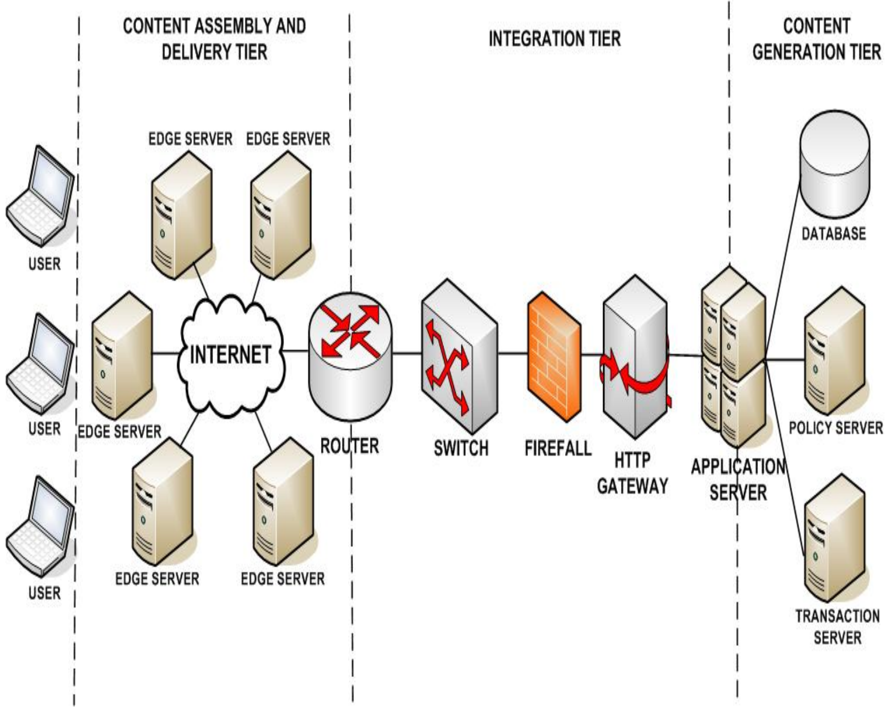

Figure4. *Content delivery network distributed model.*

ESI enabled a Web site developer to break down Web pages into fragments of different cacheable profiles. Maintaining these fragments as separate elements on the content delivery network enables dynamically generated content to be cached, then reassembled and delivered from the edge to the network. The freshness of the information is managed through a process of sending invalidation messages from the application server to the edge servers, informing them overwrite out-dated objects residing on them. In this way, changing content can be controlled much as it was when only pages were served. In a CDN, the initial request is always sent to the edge server, which checks its internal cache to see if this contains the page as a result of a previous request. What is cached, how long it is stored in cache, and other general rules for caching content are defined by the content provider via a metadata configuration file.

Integration Tier lies between the components that provide content generation and those that do page assembly and delivery. The two main components of this tier are the Hyper Text Transport Protocol (HTTP) gateway and firewall. The HTTP gateway is a fancier version of the Web server of the centralized model. The gateway's basic functions are to forward requests from the edge servers – for fragments and fields – to the applications and then send back to the results.

Content assembly and delivery tier is responsible for moving the content closer to the user, resulting in much improved user experience. This tier forms a distributed and fully managed edge network. The bigger network, the better the user community is served. The edge network may be further divided into a logical hierarchy consisting of edge servers and regional edge servers forming a core. The edge servers generate and serve content. If they do not have a particular file in cache, they may request copies from the core servers, which in turn may request the information from other regional servers. Using this method, the traffic to the original site is further reduced and issues of flash crowds become less a factor.

5. Potential solutions to problems in IP networks

We will start by analyzing the existing problems in the current IP networks. Then, we will discuss the potential solutions.

The design principles of the current Internet can be characterized by layering, packet switching, a network of collaborating networks, and finally intelligent end systems as well as the end – to – end argument [7]. The critical issue facing the Internet is to support differentiated QoS for heterogeneous applications in a flexible and inexpensive way. However, the current IP networks were originally designed for providing the best effort applications not covered by QoS guarantees and other control mechanisms. However, the emerging applications like online gaming; social networking sites (SNSs), videoconferencing, live streaming and video sharing require varying amounts of reliability, functionality, speed, efficiency, cost effectiveness and scalability. Therefore, it is necessary to bridge the gap between the emerging heterogeneous applications with various demands of service capacity and IP networks.

Despite its simplicity and scalability, the IP network lacks QoS differentiation, traffic control and management mechanisms. This lack brings significant challenges to Internet service. The question often arises in recent years is how to improve the current Internet architecture to accommodate the emerging applications in academic and industrial communities.

Today, there are two alternative design principles to develop the future Internet: incremental design and clean – state design. In incremental design, the system is moved from one state to another with incremental patches. Network researchers have focused on solutions that incrementally improve the Internet with the implicit assumption that radical new solutions are not needed or have not chance of being deployed. On the other hand, to care the future needs, the Internet has to be extended. Newely, it has to be redesigned for present requirements, ensuring at the same time enough flexibility to adequately incorporate the future requirements. This is clean-state design. Due to the industrial and economic reasons, it is impossible to enable such a number of more than 13000 competing Internet service providers to abandon the current Internet infrastructure and establish a new one. Thus, the incremental design principle will still inhabit the mainstream in coming years.

The OSI stack of the Internet can be characterized in two logical layers: infrastructure layer and service layer. Infrastructure layer traditionally processes data at layer 1 through 3, centered on the routing, forwarding, and switching of frames and packets. On the other hand, service layer includes layer 4 through 7 and deals with the routing and forwarding of requests and responses for content.

To preserve stateless best effort IP networks requires putting more state – relative functions in the upper (service layer) to solve the flexile QoS control problem over the global Internet. The way the IP layer provides only simple and general service can reduce cost, facilitate network upgrades, and enable new applications to be added without the need for changes to the existing network.

An efficient way is to be build a virtual network on top of a generic IP transport layer in order to add on additional functionality, security, and flexibility to IP networks. Such overlay networks provide flexibility, traffic control and resource management with the advantages of easy extension across a heterogeneous network platform significant changes to the underlying technology.

6. Conclusion

Content Delivery Networks (CDNs) are highly scalable and provide improved performance and reliability that directly benefit the user. The distributed model of their architecture protects against issues of the flash crowds and network hot spots. On the other side, the future Internet architecture should preserve the features of the best effort while offering differentiated QoS and various amounts of accessibility, reliability, flexibility and security. The CDN technique was developed to offer the service of the large – scale content delivery based on IP network and improve system performance in order to satisfy the QoS requirements of heterogeneous Internet applications. The development over the past decade shows that CDNs can efficiently satisfy the demands of emerging applications by adopting innovative architecture and technologies. The formation of the Content Alliance has helped for a cooperative model between independent CDN operators, which has made the content owners live a lot simpler. Through a single contract with a single CDN provider, the content owners gain the benefits of the reach and coverage provided by their CDN provider and the reach and coverage of any other CDN with which there is an agreement. The emerging challenges are server placement and organization, content distribution, request routing and system management.

References

[1] K.R.Rao, Z.S.Bojković, D.A.Milovanović, "*Wireless Multimedia Communications: Convergence, DSP, QoS and Security*", CRC Press, Boca Raton, FL, USA, 2009.

[2] M.S.Blumenthal, D.D. Clark,"Rethinking for the Design of the Internet: Te End-to-End Arguments vs. Brave New World", *ACM Trans. Internet Tech*., vol. 1, no. 1, , pp. 70-109, Avg. 2001.

[3] K.R.Rao, Z.S.Bojkovic, D.A.Milovanovic,"*Introduction to Multimedia Communications: Applications, Middleware, Networking*", Wiley, New Jersey, USA, 2006.

[4] T. Leighton, "Improving Performance on the Internet", *Commun. ACM*, vol. 52, no. 2, pp. 44-51, Feb. 2009.

[5] S. Agarwal, J.R.Lorch, "Matchmaking for Online Games and Other Latency – Sensitive P2P Systems", *Proc.ACMSIGCOMM*, pp.315-326, Avg. 2009.

[6] Z. Zhang et al.,"Optimizing Cost and Performance in Online Service Provider Networks", *Proc. NSDI*, p.15, April 2010.

[7] A.Feldmann,"Internet Clear – State Design: What and Why?", ACMSIG-*COMM Comp. Commun. Rev*., vol. 37, no. 3, pp. 59-64, July 2007.

Sadržaj: *Da bi se obezbedila dodatna funkcionalnost i fleksibilnost za budući Internet, kao jedno od efikasnih rešenja predlaže se uvođenje mreže za isporuku multimedijalnog sadržaja. Na taj način postiže se optimizacija Interneta u pogledu uvođenja različitih klasa kvaliteta servisa. Posle kratkog prikaza rezultata istraživanja u ovoj oblasti, dat je opis mreže za isporuku multimedijalnog sadržaja kao i način njenog funkcionisanja. Drugi deo rada bavi se mogućnostima potencijalnih rešenja vezanih za probleme tehnologije isporuke multimedijalnog sadržaja u budućim IP mrežama.*

Ključne reči: *mreže za isporuku multimedijalnog sadržaja, IP mreža, budući Internet, kvalitet servisa.*

MREŽE ZA ISPORUKU MULTIMEDIJALNOG SADRŽAJA I BUDUĆI INTERNET

Zoran Miličević, Zoran Bojković