Hybrid Content Networking Architecture

Muhammad Rehan Sami
Research Assistant
Computer Engineering Department
King Fahd University of Petroleum & Minerals
KFUPM # 354, Dhahran, 31261, Saudi Arabia.
Email: sami@ccse.kfupm.edu.sa

Abstract

Content delivery networks (CDNs) can be defined in many ways. Basically CDNs are intelligent networks that understand the type of content request, where to find it, and how to deliver it in an efficient way. Now the network to be intelligent enough to do all this requires specialized techniques both at the application level as well as the network level. Caching of requested objects enhances the performance of content delivery by reducing the traffic to and from the origin server. This paper presents the importance of web caching in content networking and emphasizes on the fact that web caching techniques are significant for deploying high performance content delivery networks. The author proposes a hybrid content networking architecture that involves both the overlay approach (using web caching) as well as the network approach (using layer 5 switching/routing). In order to reduce the cost of deploying such architecture especially at the network level, the idea of using software routing/switching is proposed in this paper.

Keywords: Content Networking, Web caching, Content Delivery Networks

1. Introduction

With the rapid growth of the Internet and the increase in demand for content rich applications, it is becoming more and more necessary to provide the users with efficient and speedy network services. To do so, there are many approaches. One approach requires less change in the existing Internet architecture and the deployment of which is less complicated as it uses the current infrastructure as its base. These are the Content Delivery Networks (CDNs) that deploy techniques to speed up content delivery over the Internet.

The main focus of Content Delivery networks is to accelerate the delivery of normal web content and conserve Internet bandwidth. It is a technique deployed to push content from the origin server to geographically dispersed locations, usually in proximity with the clients. This requires not only clever and efficient placement of content on the network but also intelligent

routing of packets to assure that the request made by a client is routed to the closest node on the network that contains the requested content.
Content Delivery networks are optimized to deliver specific content, such as static Web pages, transaction based Web sites, streaming media, or even real-time video or audio. Its purpose is to quickly give users the most current content in a highly available fashion. Among the different approaches of deploying CDNs, the overlay approach introduces application-specific servers or caches at various points in the network to handle the distribution of specific content types [1]. In this regard, web caching is an important technique to improve the performance and scalability of the Web by increasing document availability and enabling download sharing. This paper conducts a survey regarding web caching, its different techniques and its importance when it comes to deliver content rich data to today's Internet user.
The remainder of the paper is organized as follows. The next section gives an overview of CDNs and their classification. Section 3 introduces the concept of web caching and describes different techniques of caching web objects. Section 4 describes the role of web caching in content networks and proposes key content networking models that can be implemented using web caching techniques. Finally section 5 concludes the paper.

2. Overview of Content Delivery Networks

There are many ways to define content delivery networks. In easy words, a CDN pushes the web content to locations near to the users. This in turn minimizes the hop count and avoids possible bottlenecks.

There are two general approaches to content delivery networks [1]. These are:

1. The overlay approach
2. The network approach

The first approach refers to the placement of servers and web caches within the network to serve client requests. These servers communicate with the clients and also with each other in order to serve client requests. On the other hand, the network approach involves intelligent network devices (switches & routers) that are not only capable of forwarding packets with respect to the IP address but are also decisive enough in locating the appropriate server for the particular request [2].

2.1 Classification of CDNs

Content delivery networks can be classified on the basis of their attributes in two dimensions [3]:

1. Content aggregation
2. Content placement

The above two classes can be further divided into two parts each. The content aggregation class is elaborated in terms of semantic and syntactic aggregation whereas the placement

class is classified into the content-sensitive and content-oblivious placements. Figure 1 shows this classification.

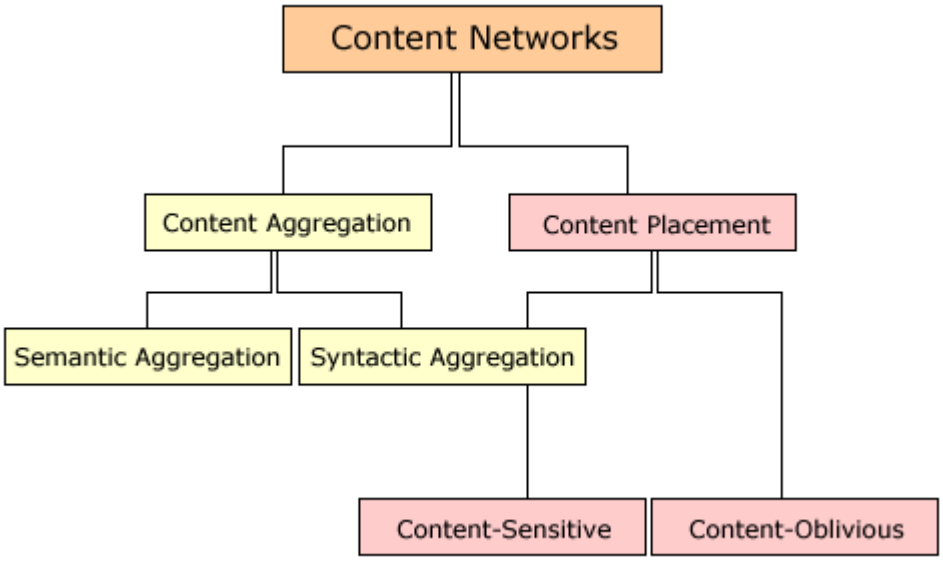

Figure 1: Taxonomy of Content Networks.

2.1.1 Content Aggregation

Content aggregation is defined as a process of assigning individual content to content-groups.

2.1.1.1 Semantic Aggregation

Semantic aggregation involves mapping of individual content with respect to some meaning full value. Then all the contents with the same external mapping value are grouped together in one content group. For example, an aggregation that uses taxonomy of vehicles, involves in the first step mapping of contents related to trucks and jeeps to the value “trucks” and “jeeps” respectively. In the next step, as both trucks and jeeps are four-wheelers, they are grouped together in the four-wheeler aggregate.

2.1.1.2 Syntactic Aggregation

In this type of aggregation, the contents are mapped to a value that does not have any external meaning. For example, a possible syntactic identifier of any content can be its file name. Any mathematical calculation such as a hash of the file name in bits could be used to map the content but it will not have any meaning nor it would share and common feature with other contents in the group.

2.1.2 Content Placement

Placement of contents or a group of content in a network can either be dependent on the content itself or may have nothing to do with the nature of the content.

2.1.2.1 Content Sensitive Placement

In this type of placement, the location of the content in the network is a function of the content. Content routing is very efficient in such a case because this placement method yields a content hierarchy. For example, sports contents can be placed on a sport subnet, football contents on the subnet of sports and finally English League contents on the football subnet. The resulting tree would make routing easy because the route to any given content from the root is fixed [3].

2.1.2.2 Content Oblivious Placement

For content oblivious placement, content groups can be placed any where in the network regardless of the nature of the content. This requires more learning by the network to locate the requested content. This can be done by a centralized server that maintains the locations of the content groups. All requests must first go to the server and then redirected accordingly (Napster).

3. Web Caching

Web caching is a popular technique to improve the performance of the World-Wide Web. This is done by increasing document availability and allowing download sharing. Web caching requires specialized servers called the web proxy servers that have application level software and perform the following basic tasks:

- Accept HTTP request from clients.
- Grabs the requested object from the origin server.
- Caches the request in local memory/cache.
- Send back the requested object to the client.

This reduces the traffic flow to and from the origin server and speeds up the request latency. Figure 2 shows the nature of traffic flow if there are no proxy servers in the network.

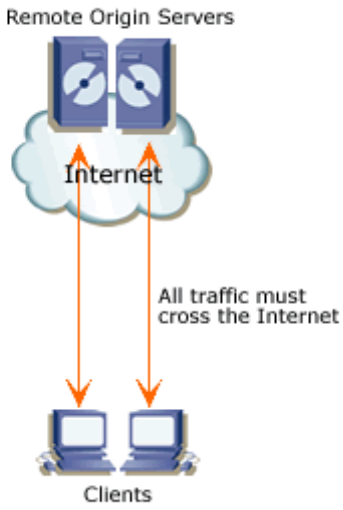

Figure 2: Without Caching.

It can be seen from the figure above that the traffic has to cross the internet at every request made by the clients to the remote origin servers. This noticeably increases the traffic on the links and causes delays in the servers' response time. On the other hand, by introducing proxy servers, all the traffic that has to cross the Internet in order to reach the content provider's origin server can be redirected to local caches from where the request can be served in lesser time. Figure 3 shows how this is done. No doubt there will be some traffic between the origin servers and the cache servers in order to get the cached object for the first time.

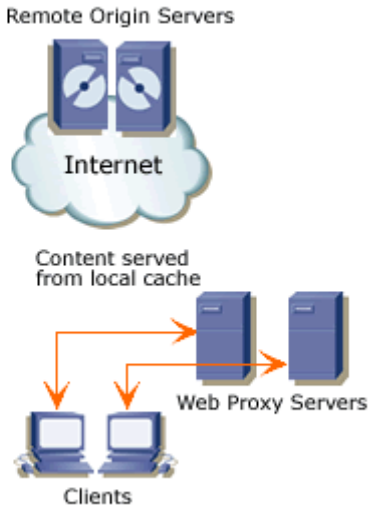

Figure 3: With Caching.

Top

3.1 Web Caching with respect to classification of CDNs

A network with Web caching proxy servers is an example of the Syntactic Content-Oblivious Network. In this case the content is the URL and the network nodes are the proxy servers. Because no content aggregation is performed, it is a syntactic content network. Also any content could be placed on any proxy server; hence the placement of contents is also oblivious in nature.

3.2 Local vs. Distributed/Cooperative Web Caching

What was described in the previous section refers to local proxy web caching [4]. There is another approach to web caching known as the distributed or cooperative web caching. This mechanism employs the sharing of documents between caches. This in turn improves the performance of the system [5].

Grouping multiple web cache servers into a cluster of caches that behaves as one single large cache is of great advantage. This not only increases the capacity of the servers linearly but also reduces the number of cache miss [6]. If the requested object is not available on the local server, the request is forwarded to a cluster of cooperative caches. If any one of these servers has the copy of the object, the request results in a cache hit.

3.3 Web Caching Techniques

There are four Web caching techniques, namely Internet Cache Protocol (ICP), Cache Digest, Layer 5 switch and Load Balancing-Layer 5. This section provides an overview of these techniques in order to understand them more.

3.3.1 Internet Cache Protocol (ICP)

The Internet Cache Protocol or simply the ICP is the most popular caching protocol that allows communication among cooperative proxy Web caches [7]. ICP is an application layer protocol running on top of UDP. In ICP, the client first sends a request to its default configured proxy cache server. If this server does not have the requested object, it broadcasts an ICP query message to all other cooperating cache servers. If any one of these cooperating servers can serve the request, it sends back an ICP hit message. The configured cache server then sends an HTTP request of the object to the responding cooperating server. On receiving the object, the configured server stores a copy in its own cache and sends the object to the client. If none cooperative server responds within the time out period, the configured server fetches the object from the origin server. Squid proxy server is an example of a Web cache server using the ICP protocol [8] [9].

3.3.2 Cache Digest

Digests are compressed contents. The cache digest protocol allows cooperating cache servers to exchange their content digests [10]. The idea of compressed contents comes from the use of a directory-based approach. In the directory-based approach, cache servers make information about their cache content available to peers in order to avoid the query/response delay. But if the servers go for an uncompressed directory, this will consume a lot of bandwidth. Compressed representation of the cached content is therefore used. Cache Digest

eliminates the need for per-client ``Do you have it?" queries and network delays associated with them. The client response time improves. For example, Cache A that has a digest of cache B knows what documents are likely to be (or not to be) in cache B. Consequently, cache A can decide whether to fetch an object from cache B without contacting cache B first.

Cache Digest uses HTTP to transfer the directory information.

3.3.3 Layer 5 Web Caching

This type of web caching makes use of the network devices to redirect HTTP traffic to cache servers. It introduces the concept of “transparent” caching as the web browsers are unaware of the proxy server that is serving the request [11].

This can be done by a layer 4 switch but using a layer 4 device poses some drawbacks. A layer 4 switching device is unaware of the application layer protocols such as HTTP, FTP etc. It only takes into consideration the layer 4 ports and forwards packets based on the port numbers. This drawback can be removed by employing a much intelligent device. Layer 5 switches, like their L4 counterparts, provide fast switching. L5 switches make routing decisions based on the information in the application layer protocol header [12], for example the URL requested and not just only the layer 4 port numbers. Using L5 switches that are separate from the caches themselves allow the client load to be dynamically spread over multiple caches, scaling process power which, in turn, can reduce response time.

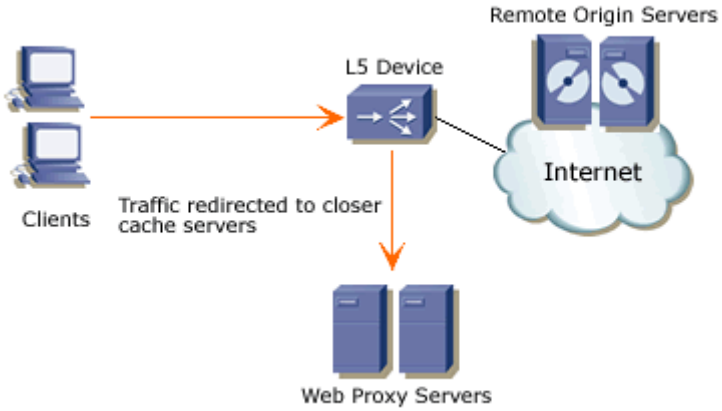

Figure 4: L5 Caching.

3.3.4 Load Balancing-Layer 5 Web Caching

The Load balancing L5 caching scheme in addition to the transparency of simple L5 scheme, introduces one further step to enhance performance [13]. LB-L5 directs client requests to the server with the intention of balancing the server’s workload. It also reduces response time by redirecting the request to a server that is closest to the client. It uses the link delay information to do so.

Top

4. Web Caching and the CDNs

Content Delivery Networks may also be defined as dedicated networks of servers located strategically throughout the Internet for distributing web content to end-users. Typically, CDNs use caching technology to store frequently used content closer to users at the so-called "edge" of the Internet.

The different web caching techniques discussed in the previous section can all be applied in a network separately or together. As described in section 2, the two approaches to CDNs namely the overlay approach and the network approach both incorporate the above mentioned caching techniques. For example, in overlay networks, proxy servers can be placed intelligently [14] [15] to serve the clients and these servers can communicate with one another and with the clients, using these techniques. On the other hand, the L5 caching technique is a direct application of the network approach in which the switch or the router is smart enough to redirect traffic to the appropriate server.

4.1 Hybrid Content Networking Model

Using different caching techniques, a hybrid content networking model can be viewed as one in which a L5 device is at the edge of the network close to the clients. This L5 device redirects client requests to cooperative web cache servers that communicate with each other through some cache protocol like ICP or cache digest.
Figure 5 illustrates the hybrid model. One drawback of such an approach is that it poses a single point of failure. If the L5 device is not functional, the clients can not be served any request as their request will never reach the cache servers. Moreover as the number of clients increase, the L5 switch is subject to more and more load. This can increase both bandwidth consumption and also the response time.

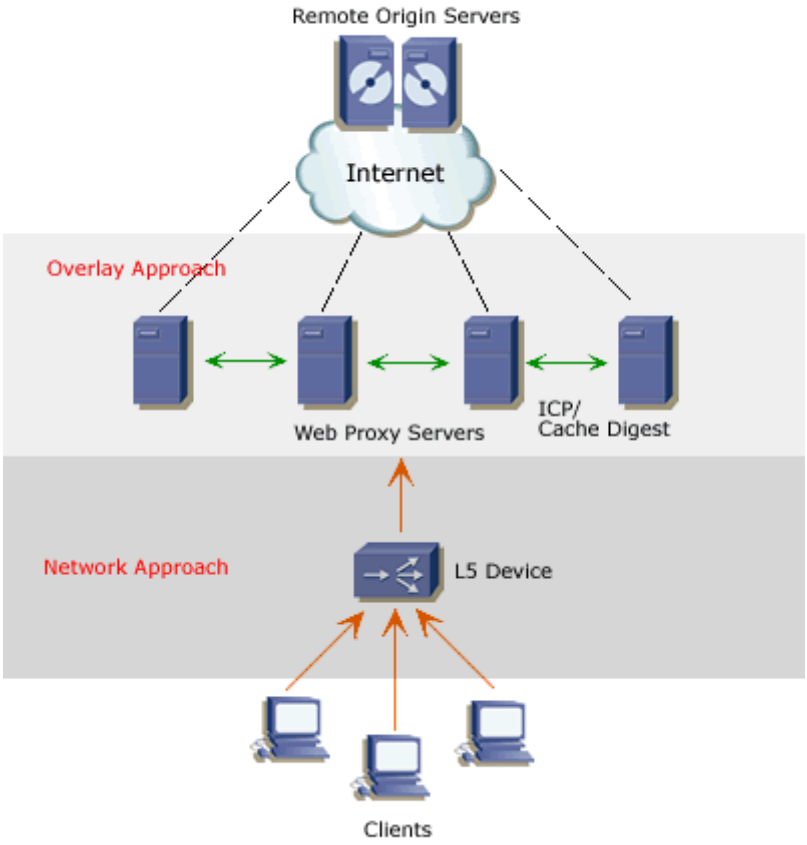

Figure 5: Hybrid Content Networking Model.

4.2 Hybrid Content Networking Model with Load Balancing

To remove the above mentioned drawbacks, a load balancing scheme could be introduced. The idea behind this is to add more L5 switches at the edge of the network so that the client requests can have multiple paths (redundancy) to avoid single point of failure. Also, load can be reduced on the L5 switch as other switches can operate in cooperation with each other. This requires some communication protocol between the switches that allows them to talk to each other and exchange information such as link utilization, server proximity etc. figure 6 illustrates this idea.

Top

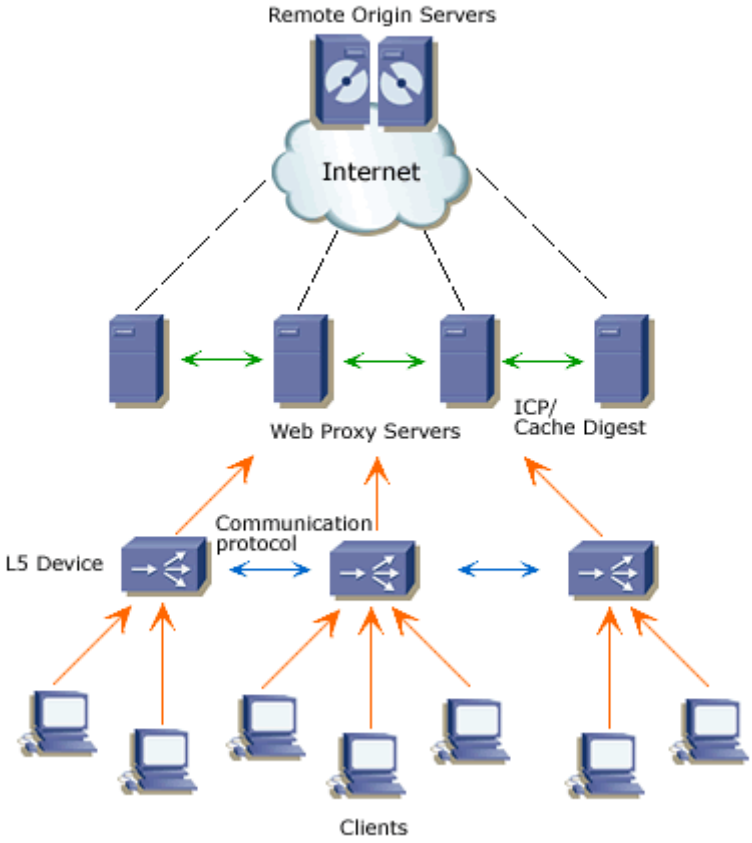

Figure 6: Hybrid Content Networking Model with Load Balancing.

This approach is no doubt costly as it requires more resources in terms of layer 5 switches and these devices are costly.

4.3 Modified Content Networking Model

A more cost effective version of this idea can be implemented using software switches/routers. A software router can be defined as a general-purpose off-the-shelf computer that executes a computer program capable of forwarding IP datagrams among network interface cards (NIC) attached to its I/O bus. But software routers have a disadvantage that they use a single CPU and a single shared bus to process all packets [16]. This can pose performance degradation. However due to the ease with which they can be programmed for supporting new functionality, they can be implemented at the edge of the network to provide layer 5 switching/routing to redirect client requests to the appropriate cache servers.

To increase the performance of the hybrid CDN model with software routers at the edge is an open area of research. The main issues that come into mind are:

Top

- Performance of the software router.
- Communication among software routers for load balancing.
- Implementation of routing protocols.
- Redirection of client request to the appropriate cache server.

It is to address whether a general purpose machine is capable of performing all the above mentioned tasks with minimum performance degradation when compared to high performance layer 5 hardware switches/routers.

Another advantage of using PC-based switches is that along with layer 5 switches, these machines can also perform caching at the first level i.e. these switches may have their own caches to copy most frequently requested objects. In this way these general purpose computers can be used as a L5 device as well as a proxy server all at the same time. This seems to be similar to the existing NAT/proxy configuration that most enterprises and campus network administrators use to connect their networks to the Internet. The only difference being the L5 switching that not only differentiates traffic from IP address and port number but also considers the application layer protocol.

5. Conclusion

This paper gives an introduction to the basic concepts of content networking and classifies them according to content grouping and placement. It gives a detailed overview of web caching, its importance and types and its role in delivering content rich data to the users. It has been seen from available literature and current research that web caching can play a vital role in improving the performance of the Internet and speeding up data delivery.

All the popular web caching techniques can be used in deploying cost effective content delivery networking solutions. This paper proposes a hybrid content networking architecture that involves both the overlay approach (using web caching) as well as the network approach (using layer 5 switching/routing). In order to reduce the cost of deploying such architecture especially at the network level, the idea of using software routing/switching is proposed in this paper.

References

[1] Irwin Lazar & William Terrill. “Exploring Content Delivery Networking”. *IT Professional*, Volume: 3, Issue: 4, Page(s): 47 –49, Jul/Aug 2001.

[2] Barani Subbiah & Zartash Uzmi. “Content Aware Networking in the Internet: Issues and Challenges”. IEEE International Conference on Communications, Volume: 4, Page(s): 1310 –1315, 2001.

[3] H.T.Kung & C.H.Wu. “Content Networks: Taxonomy and New Approaches”. To appear in The Internet as a Large-Scale Complex System. Oxford University Press, 2002.

Top

[4] Hassanein et. al. “Performance Comparison of Alternative Web Caching Techniques”. Proceedings of the 7th IEEE International Symposium on Computers & Communications, 2002.

[5] Wolman et. al. “On the scale and performance of cooperative web proxy caching”. Proceedings of the 17th Symposium on Operating Systems Principles, December 1999.

[6] James Feenan, Patrick Fry & Ming Lei. “Clustering Web Accelerators”. Proceedings of the 4th IEEE International Workshop on Advanced Issues of E-Commerce and Web-Based Information Systems, 2002.

[7] D. Wessels & K.Claffy. “RFC 2186: Internet Cache Protocol (ICP), Version 2”. National Laboratory for Applied Network Research/UCSD, September 1997.

[8] D. Wessels & K.Claffy. “ICP and the Squid Web Cache”. IEEE Journal on Selected Areas in Communication, Volume: 16, Issue: 3, Page(s): 345-357, April 1998.

[9] “Squid Web Proxy Cache”. www.squid-cache.org

[10] A. Ruosskov & D. Wessels. “Cache digests”. Proceedings of the 3rd International WWW Caching Workshop, June 1998.

[11] B. Williams. “Transparent Web Caching Solutions”. Proceedings of the 3rd International WWW Caching Workshop, June 1998.

[12] Apostolopoulos et. al. “Design, implementation and performance of a content-based switch”. Proceedings of the 19th Annual Joint Conference of the IEEE Computer and Communications Societies, Volume: 3, Page(s): 1117 -1126, 26-30 Mar 2000.

[13] Z. Liang, H. Hassanein & Patrick Martin. “Transparent Distributed Web Caching”. Proceedings of the IEEE Conference on Local Computer Networks, Page(s): 225 -233, Nov 2001.

[14] S. Ratnasamy et. al. “Topologically-Aware Overlay Construction and Server Selection”. Proceedings of the 21st Annual Joint Conference of the IEEE Computer and Communications Societies. Volume: 3, Page(s): 1190 -1199, 2002.

[15] Sherlia Shi & Jonathan Turner. “Placing Servers in Overlay Networks”. Technical report, Department of Computer Science, Washington University in St. Louis, 2002.

Top

[16] Oscar-Iván Lepe-Aldama & Jorge García-Vidal. “A Performance Model Of A PC Based IP Software Router”. IEEE International Conference on Communications, Volume: 2, Page(s): 1230 -1235, 2002.

Top