ISBN 978-981-11-7861-0
Proceedings of 2018 the 8th International Workshop on Computer Science and Engineering (WCSE 2018)
Bangkok, 28-30 June, 2018, pp. 131-135
doi: 10.18178/wcse.2018.06.023

Service Load Management Mechanism for CDN

Xinhua E [1+], Binjie Zhu [2]

[1] Faculty of Information Technology, Beijing University of Technology, Beijing, 100124, China
[2] China Mobile Group Beijing Company Limited, Beijing, 100007, China

Abstract. CDN is a kind of overlay network that was used to the content acceleration on Internet. How to ensure service quality is an important issue in CDN. Differentiated service is a way to ensure the QoS in other basic network. In this paper, the multi-queue management and scheduling mechanism was introduced to the service load management in CDN, so it can supports differentiated service for the content service resources. Each content object has a different quality mark. The quality mark of the request content object was identified for classification. Different types have different sub-queues. Queue scheduling module is responsible for scheduling service resources. The objects with high rank mark get more service resources, so get better quality. Finally, the mathematical model of this method was established, and the experimental program was set up for the analysis. The experimental results show that the method supports differentiated services for service resources to ensure the quality, and controls the congestion effectively.

Keywords: load management, management and scheduling, P2P-CDN, differentiated services.

1. Introduction

Content distribution network is a network acceleration technology [1]. A content distribution network model was design in our previous work, shown as Figure 1. It including three layers: CDN management layer, global routing layer, the storage network layer. Management is responsible for the entire CDN overlay network management. CDN management layer includes the source server, redirection system, distribution system, and service management system. Global routing overlay network layer is responsible for the overall route lookup. Storage network layer is responsible for the content of storage and local routing. Every storage server is a peer node in the storage network, in other words, the storage network is a P2P network. P2P network is an effectively way to organize the servers [2] .Storage server is a content service server at the same time. In this paper, we will focus on how to use the service resource in the storage network layer.

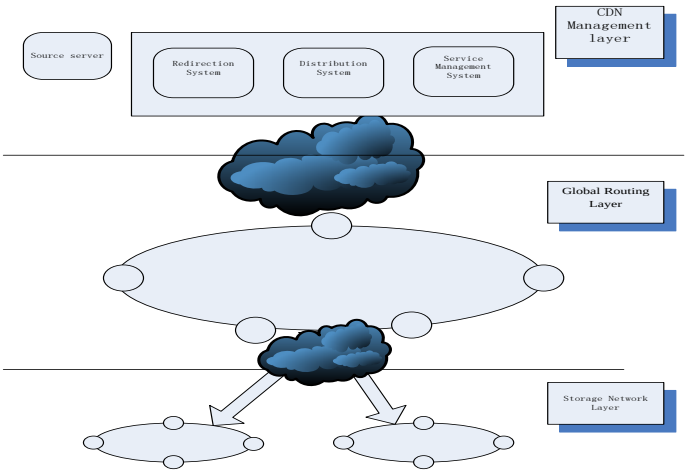

Fig. 1: Architecture of P2P-CDN.

+ Corresponding author. Tel.: +86 13811324402; fax: +86 .01088334372
E-mail address: elson1900@163.com

2. Related Works

For Basic network the definition of QoS is service quality, including the transmission bandwidth, transmission delay, data packet loss rate. In the network with QoS can ensure the transmission bandwidth and reduce transmission latency, lower data rate and delay jitter, packet loss and other measures to improve service quality. Network resources are always limited, as long as there to snatch the network resources, there will be quality of service requirements. For example, a fixed total bandwidth in the network, if certain types of businesses more bandwidth, so other businesses can use less bandwidth and may affect other business use [3].

Differentiated service is a way to ensure the QoS in other basic network. Queue management and scheduling mechanism is a method to supports differentiated service. It has applications to network routers and communications networks [4, 5]. Here we introduce such a mechanism to service load management in CDN overlay network.

3. Multi- Queue Based Differentiated Services in CDN

The contents were distributed to the cache servers. The user requests are directed to the nearest server. If the server has not the request content, find another server. The figure 2 is the CDN request routing process of the user. Routing modules includes the user, redirection system, and content routing system. User sends a request for content. Redirect system sends the request to the nearest server resources. If the server has not the requested resources, find the right resources through the content routing system. In our design, content routing system uses a P2P way.

In the traditional method, when the server receives the request, first determine whether there are content for local request. If it has the request content, then it responding to request. There are some disadvantages of this method. First, the user's request does not limit the number. Many users share services, resources. When too many users requests, it will result in service degradation. Second, there is no distinction between content services. This does not guarantee the content of the high level of better quality of service. Third, when many users requests in the same time, it will cause congestion.

To avoid the above shortcomings, multi-queue management and scheduling mechanism was introduced to the CDN, shown as figure 2. Through the queue management and scheduling mechanism to achieve differentiated services, to avoid congestion. In the redirection system to guide the user's request for server resources, the first request queue management system by processing the request. Queue management system to decide whether the response request. If it do not accept the request, then the request were directed to the routing system for re-routing. If it accepts the request, the request was scheduling through the queue scheduling mechanism. Content for different levels of resources allocated for different services. It aims to ensure quality of service.

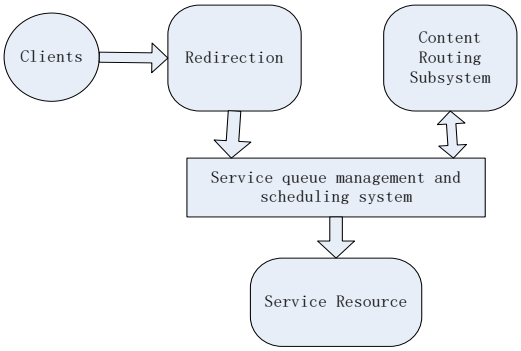

Fig. 2: Multi- queue based differentiated services in CDN.

4. Queue Management and Scheduling Mechanism

As the figure 3 shown, multi- queue based differentiated services includes the following modules: the request queue, classification, queue management, sub-queue, scheduling module, service resource pool. When a user's request comes from the redirection system or content routing subsystem, the request arranged in order of request queue. Classifier identifies the level of content request. According to the mark the request

was redirected to the sub-queue. Sub-queue management system to manage the queue pairs, including the captain and other indicators to decide whether to forward, or inserted into the sub-queue tail. Scheduling module is for the service resource scheduling in accordance with the scheduling algorithm.

Each sub-queue has a queue management system. Queue management module forwards the request after detected an exception arrival. This will ensure that access is guaranteed in the event of unexpected performance. The exception label of each queue to determine forwards is different. High level queue can have a greater arrival rate. By detecting the queue length to determine whether there is an exception.

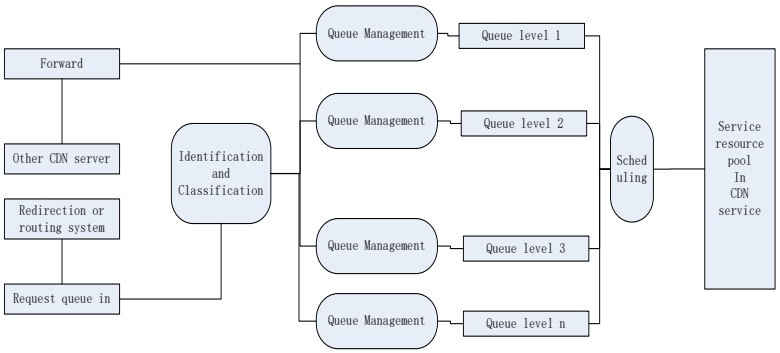

Fig. 3: Queue management and scheduling mechanism in CDN.

Queue scheduling module, according a scheduling method to use the service resources. For example, weighted round robin algorithm can be used as scheduling algorithm. The scheduling principle is also a high level of service queue has more resources.

Service queue management and scheduling process is as figure 4:

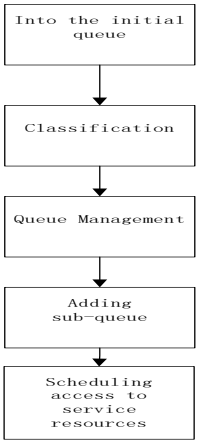

Fig. 4: Queue management and scheduling process.

5. Experiments

5.1 Experimental Setup

The queue is modeled as a multi-service windows mixed M / M / n / m [6], shown as figure 5. In the traditional case, assuming arrival rate A. There are three levels, each level of the request in equal proportions. Then the three sub-queue arrival rate A / 3.

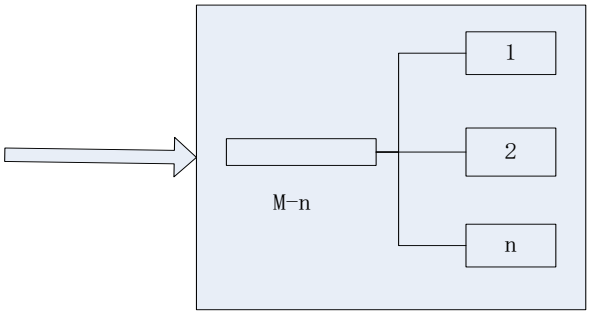

Fig. 5: The model of queue.

The multi-service windows mixed model is calculated as follows:

$$p_0 = \begin{cases} \left[\sum_{k=0}^{n-1} \frac{(n\rho)^k}{k!} + \frac{(n\rho)^n}{n!} \frac{1-\rho^{m-n+1}}{1-\rho}\right]^{-1} & \rho \neq 1 \\ \left[\sum_{k=0}^{n-1} \frac{n^k}{k!} + \frac{n^n}{n!}(m-n+1)\right]^{-1} & \rho = 1 \end{cases} \quad (1)$$

$$P_s = p_m = \frac{n^n \rho^m}{n!} p_0 \quad (2)$$

$$\lambda_1 = \lambda P_s = \lambda p_m = \frac{\lambda n^n \rho^m}{n!} p_0 \quad (3)$$

$$L_f = \frac{\lambda_e}{\mu} \quad (4)$$

$$L_q = \frac{n^n \rho^{n+1} p_0}{n!(1-\rho)^2}[1-(m-n+1)\rho^{m-n}+(m-n+1)p_0] \quad (5)$$

$$L_s = L_q + L_f = L_q + \frac{\lambda_e}{\mu} \quad (6)$$

$$W_s = \frac{L_S}{\lambda_e} = W_q + \frac{1}{\mu} \quad (7)$$

In order to calculate conveniently, we do not consider the case of forward. Therefore, the mechanism of the sub-queue modeled into M / M / s. Assumed arrival rate 0.1, service time is 20s. The total service capacity of s = 15. There are three levels: level 3, level 2, and level 1. Resource scheduling allocation ratio of scene 1 is 5:5:5. Resource scheduling allocation ratio of scene 2 is 3:6:6. Resource scheduling allocation ratio of scene 3 is 3:5:7. Resource scheduling allocation ratio of scene 4 is 3:4:8. Code was shown as figure 6.The parameters set as Table 1.

```
Model:
S=6;R=0.10;T=20;load=R*T;
Pwait=@peb(load,S);
W_q=Pwait*T/(S-load);
L_q=R*W_q;
W_s=W_q+T;L_s=W_s*R;
END
```

Fig. 6: The main code.

Table 1: Parameters set

Scene	Level 1	Level 2	Level 3
1	S=5	S=5	S=5
2	S=3	S=6	S=6
3	S=3	S=5	S=7
4	S=3	S=4	S=8

5.2 Experimental Results

After the establishment of the four scenes, each running program can get an experimental result. The waiting time w_q in the four scenarios were shown as the figure 7. The horizontal axis is the scenario. The vertical axis is the waiting time. Results can be seen from the figure. When the three levels of resources are equal, their waiting times are equal, shown as scene 1. The waiting time of level 2 increased with the resources reduced. The waiting time of the level 1 decreased with the resources increasing.

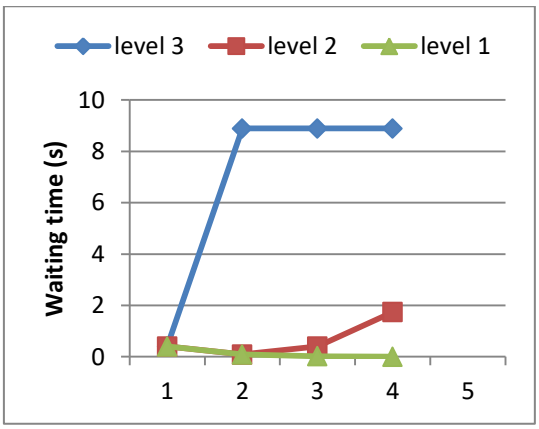

Fig. 7: Comparative analysis of waiting time.

The probability of customer not needs to wait the shown in figure 8. The probability of no waiting is also called pass rate. The pass rate of level 1 increased with the number of S. The pass rates of level 2 decrease as the number of S decreased.

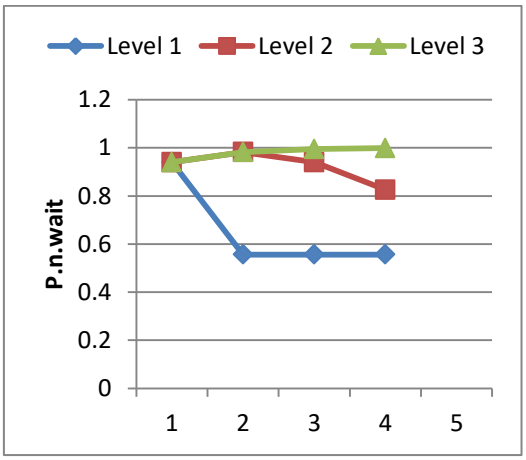

Fig. 8: Comparative analysis of waiting time.

6. Conclusion

Differentiated services model of the CDN services load was researched in this paper. The multi-queue management was introduced to the CDN service load management. Multi-queue management and scheduling was used to achieve the different levels of quality service. The experimental results shows that it could supported differentiated services for service resources to ensure the quality, and controls the congestion effectively through controls the parameter S. In the future, our research will focus on the optimization of queue management and scheduling algorithms for CDN.

7. Reference

[1] SanaaSharafeddine, KarimJahed, Omar Farhat, et al. Failure recovery in wireless content distribution networks with device-to-device cooperation[J]. Computer Networks, 2017, 128(9): 108-122.

[2] Rami Halloush, Hang Liu, Lijun Dong, et al. Hop-by-hop Content Distribution with Network Coding in Multihop Wireless Networks[J]. Digital Communications and Networks, 2017, 3(1): 47-54.

[3] XiaoyingZheng, Ye Xia. Optimizing network objectives in collaborative content distribution [J]. Computer Networks, 2015, 91(14): 244-261.

[4] UttamMandal, PulakChowdhury, et al. Energy-efficient networking for content distribution over telecom network infrastructure[J]. Optical Switching and Networking, 2013, 10(4): 393-405.

[5] LazarosGkatzikis, VasilisSourlas, Carlo Fischione, et al. Low complexity content replication through clustering in Content-Delivery Networks[J]. Computer Networks, 2017,121(5): 137-151.

[6] Bernd Klasen. Efficient content distribution in social-aware hybrid networks [J]. Journal of Computational Science, Pages 209-218, 2013, 4(4): 209-218.