LOAD BALANCING SCHEMES FOR MULTIMEDIA E-SERVICE

Chia-Sheng Tsai and Sh-Bin Cheng

Dept. of Computer Science and Engineering, Tatung University, Taiwan

Abstract

As communication networking becomes vital to people's digital life. Also, various communication protocols are evolved in e-Service for e-Business models. The main push factor behind this network evolution is the real time transfer of multimedia information, such as, audio, image and video. Several protocols, such as RTP (Real Time Protocol) and RTCP (Real Time Control Protocol) , are developed to facilitate the deployment of multimedia services in the Internet. However, as for the intention of deliver high quality of real time multimedia information, data transmission bandwidth becomes a major issue. Insufficient network bandwidth would also result in poor service quality. To overcome this obstacle, this thesis proposes an adaptive load balancing system that targets for intensifying overall network usage in a multi path network environment.

The proposed system uses a modified RTP header field to send packets through multiple paths in an RTP session to improve e-service quality. On the sender side, a suggested load balancing algorithm is used to perform RTP packet distribution. Meanwhile, the receiver side would utilize RTCP to notify senders on packet arrival intervals. Via the communication between sender and receiver, overall network bandwidth fluctuation is thus determined. By knowing the actual available network bandwidth, sender is then able to dynamically adjust packet distribution. Therefore, a load balancing system is achieved. This accomplishment would result in an increase to overall network throughput, thus, end users are able to enjoy stable multimedia services.

1. INTRODUCTION

Today, the efficiency of network technology and related software and hardware has increased greatly that popularizes network multimedia, furthermore the service quality of still images, music, and voice cast can be managed well, but the video service is yet to be improved. For example, distance learning, Video on Demand, and Video Call need to transfer a large volume of data, so how to use network bandwidth efficiently is a big challenge in multimedia transfer.

The communication network traffic is increasing steadily because of the increasing amount of multimedia data, so routers with bandwidth sharing function are available on the market, and are becoming faster in processing speed, having better service quality, and easier management. But the rise of this kind of product is mainly for improving the whole bandwidth in response to the sharp increase of network traffic. However, a perfect network service shall not only provide faster network speed, but also stable and reliable transmission service. Therefore, the self-adaptation load-balancing system proposed in this paper not only improves the network bandwidth, but also ensures the multimedia service quality as an important technology.

With the popularization and increasing prevalence of multimedia applications on Internet, the network application of real-time transmission has become a focus point, and has been widely used by users. At present, the application of real-time transmission on Internet still have problems of insufficient network bandwidth and change of transmission bandwidth, especially on Internet or wireless networks. The bottleneck bandwidth sometimes disables the multimedia data to be transferred smoothly. Therefore, this research proposes a combined multi-transmission path to increase users' available bandwidth. The proposed algorithm can distribute the bandwidth properly, and utilize the mechanism of dynamic detecting network bandwidth to self-adaptation the algorithms, in order to reduce the impact resulted from the variation of Internet. Finally it secures the transmission quality of multimedia data in network.

At present, the load-balancing mechanism is discussed most frequently, it can be approximately divided into two main types: server load balancing, and network load balancing. The main difference between these two types is that the structure of the former is like an exchanger, for example the Scalable Web Server Clustering Technologies [1] can integrate the flow and capacity of different servers effectively, which not only improves the whole efficiency of server, but also has the ability of network preparation. As for the latter one, its structure is like a router, and is suitable for multi-circuit load balancing in enterprise LAN and WAN.

Recently, most of common network load balancing products are using Round-Robin or Weight Round-Robin or Fewest connections with limits collocating TCP[2] or UDP [3] to reach load-balancing mechanism. However, these Quantum fixed load-balancing mechanisms cannot be dynamic with the alteration of network environment to alter the distribution of network flow dynamically. In order to solve the impact resulted from the variation of network bandwidth, it is proposed to use the communication protocol controlled by network congestion for detecting the network's available bandwidth in real-time, and altering the transmission flow of multimedia data dynamically. However, reducing the network transmission flow would decline the multimedia service quality. Here, this research is to try to find out the way to detect the changes of Internet bandwidth under the Internet environment and to touch off the load balancing algorithm automatically.

2. System Model

Because the sending end uses RTP packet as the key instrument for bandwidth detection, in order to avoid instant congestion in network which would result in inaccurate bandwidth measurement. Thus, the measurement of available bandwidth in this paper is detecting 50 RTP packets every time as shown in Fig 1, and then seeks average value. The detecting cycle length can be adjusted according to required transmission medium status, as for multimedia transmission, it is better to know the transport delay status in a very short period of time, so the cycle length shall be reduced, and the value can be obtained quickly by decreasing the times of detection at every turn to judge whether there is any packet loss based on Sequence number of consecutive RTP list head. The available bandwidth of network is measured periodically, and multi-packet can be used to measure the bandwidth in average for reducing the impact of sudden congestion in network and improve the accuracy of bandwidth measurement. In case a network disconnection is detected, set the bandwidth of the interface at 0 which means this network interface cannot provide transmission service until a network connection is recovered. And the receiving end will carry out bandwidth measurement periodically, figure out the available bandwidth of the path, and collocates the foresaid dynamic load balancing algorithm, so that the quantum of each transmission interface can be determined, eventually complete the operation of whole system.

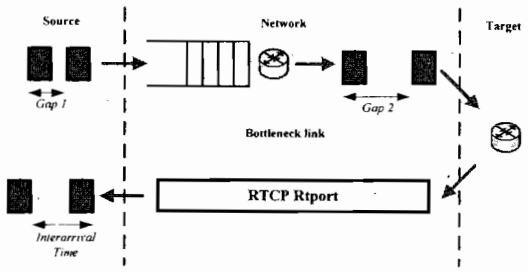

Fig. 1 Inter-arrival time was probed by RTCP packets

We can use the method of obtaining minimal value under limiting conditions to get the optimal solution of packet distribution quantum. The coefficient β can be obtained below:

$$\beta(\lambda,\mu_i)=\frac{\frac{1}{\lambda}\left(\sum_{i=1}^{N}\sqrt{\mu_i}\right)^2}{\frac{1}{\lambda^2}\left(\sum_{i=1}^{N}\mu_i\right)^2-\frac{2}{\lambda}\left(\sum_{i=1}^{N}\mu_i\right)+1}$$

The acquired coefficient β is put into following equation to obtain the quantum of each transmission interface. Its functional equation is as follows.

$$\phi_i(\lambda,\mu_i)=\frac{\mu_i}{\lambda}\left(1-\sqrt{\frac{\lambda}{\mu_i\beta}}\right)$$

ϕ_i : Transmission interface i distributing quantum of load balancing.

μ_i: Transmission interface i service rate.

λ: Source packet arrival rate.

β: coefficient of Lagrange Multiplier.

3. Simulation and Results

Due to the network is changeful, for example, the changes of ambient data stream, bottleneck bandwidth and each connection data volume of the same connection would directly influence the usage of other simultaneous connections. Especially changes occurred in multimedia network environment would influence the experience of customers of both ends on usage. Therefore, this research conducted experiment to dynamically calculate the available network bandwidth of system based on the algorithm described in Section 2 and the system would also update and re-allocate the network flow quantum of load balancing immediately in accordance with practical network bottleneck bandwidth to ensure the multimedia service quality to be free from the obstruction of network environment.

3.1 Experimental Model

The experimental environment of this study can be divided into two parts: application and development of analytical software part, and operating system hardware structure part.

A. Developing analytical software:

A. System development software: Borland C++ 6.0 [6]
B. Network development suite: Indy.Sockets 9.0 [7]
C. Network simulation software: The National Institute of Standards and Technology, NIST Net 2.0.7 (Network Emulator)[8]

B. Operating system hardware structure:
A. Traffic Generator: Windows XP
B. Gateway-1 ~ 3: Linux Red Hat 7.0
C. Packet Scheduler: Windows XP
D. Performance Analyzer: Windows XP

C. Scenario

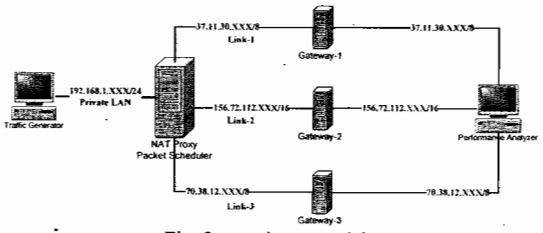

Fig. 2 experiment model

First, Traffic Generator generated RTP multimedia data stream, and distributed as Traffic Model based on indices of Poisson probability. Furthermore, the self-adaptation load balancing algorithm mentioned in last chapter was embedded into Packet Scheduler, and then all RTP packets generated would pass through Packet Scheduler for arrangement and distribution of packet flow to realize the load balancing of network packet accordingly. Next, NIST Network Emulator at Gateway-1~3 was installed to simulate packets passing through Internet. The key was to emulate the bottleneck bandwidth of bottleneck connection by Gateway. Finally, Performance Analyzers was used to monitor and analyzes the arrival state of RTP packets at the receiving end. The configuration of experimental environment is as Fig. 2.

3.2 Experimental methods

This system is designed aiming at transferring multimedia, so it combines UDP communication protocol of network layer with RTP communication protocol.

The multimedia data transmission in implementation adopts UDP communication protocol, mainly because of the large data volume of multimedia transmission and real-time of multimedia transmission data. Therefore, in order to enable an amount of multimedia data to be sent to the receiving end timely for service quality, UDP communication protocol is adopted, while inevitable packet loss was neglected. As for TCP communication protocol, though it can ensure the data packet not to be lost, it is the fatal shortcoming of real-time data transmission because it would retransfer when any packet loss occurred resulted from network congestion. This action must keep posterior packets waiting until previous packets are surely disposed, so the real-time multimedia play quality will be severely influenced.

Additionally, in order to detect the available bandwidth of network, the system added RTP communication protocol to UDP communication protocol. This study amended the fields in RTP list head. Thus, the system can deduce the required message of network bandwidth through the message in the fields, and the supplementary system uses dynamic bandwidth detection to ultimately achieve optimized load balancing self- adaptation.

The round-robin (RR) algorithm is often used as a simple-yet-effective method of distributing requests to a single-point-of-entry to multiple servers. It's used by DNS servers, peer-to-peer networks, and many other multiple-node clusters/networks. In a nutshell, round-robin algorithms pair an incoming request to a specific machine by cycling through a list of servers capable of handling the request. It's a common solution to many network load balancing needs,

Weighted round-robin (WRR) is one way addressing these shortcomings. In particular, it provides a clean and effective way of solving the first-half of the problem by focusing on fairly distributing the load amongst available resources, verses attempting to equally distribute the requests. In the server environment, a difference in server capacities and processing capabilities can result in a more-marked difference in the performance of the different servers when compared to the effect of a difference in the requests themselves. In a weighted round-robin algorithm, each destination is assigned a value that signifies, relative to the other servers in the list, how that server performs. This "weight" determines how many more or fewer requests are sent that server's way; compared to the other servers on the list.

There are two emulations and comparisons in this chapter. Emulation I evaluates the packet delay variation of three systems including RR, WRR and Self-adaptation load-balancing under a stable network environment, it is used to judge which load balancing algorithm is most suitable for providing multimedia transmission. Emulation II evaluates and compares bandwidth measurement and the load-balancing system improved by this research under the bandwidth congested network environment, to find out which one is suitable for the variable environment of Internet bandwidth. It is mainly used for observing and comparing packet delay and variation.

3.3 Considerations

The experiment mainly analyzed Jitter variation degree of packets in RR, WRR and Self-adaptation

load-balancing system, under the condition of using three path bandwidths at the same time, analyzes and compare the efficiency of three load-balancing systems, it mainly demonstrates which algorithm is suitable for multimedia transmission.

Jitter variation ratio will change due to different transmission interfaces have different number of queue, high quantum set not always reduce packet delay variance ratio, sometimes may receive packets continuously and then another packet after an interval of a long time. As a result, the degree of variation would increase. Finally, Jitter variation degree is analyzed through Root mean square error (RMSE).

The RMSE is a frequently-used measure of the differences between values predicted by a model or an estimator and the values actually observed from the thing being modeled or estimated.

$$RMSE = \frac{\sqrt{\sum_{i=0}^{N}\left(((Recvtime(j)\text{-}Sendtime(j))\text{-}(Recvtime(i)\text{-}Sendtime(i)))\right)^2}}{N}$$

When transferring multi path data, Jitter variation of packets becomes more severe, and it relatively influences the play quality of multimedia data. Therefore, the experiment simulated when service rate of transmission path is respectively fixed at 20, 30 and 50 packets per second, and simulate using RR, WRR, Self-adaptation load-balancing and the algorithm proposed herein, observe the degree of variation of RTP packet delay. Table 1 shows the experimental parameters.

Table 1 System experimental parameter table

Parameter of Experiment I	
Link-1 RTP packet service rate	20
Link-2 RTP packet service rate	30
Link-3 RTP packet service rate	50
Gateway-1 bottleneck service rate	1000
Gateway-2 bottleneck service rate	1000
Gateway-3 bottleneck service rate	1000
RTP packet arrival rate	70
Number of test RTP packet	300

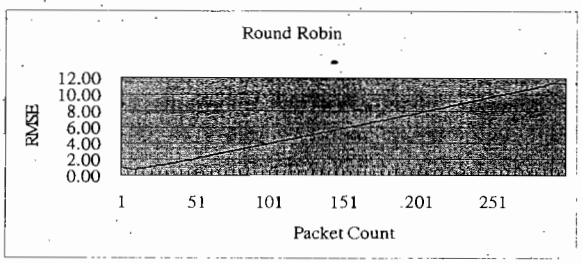

Fig 3 RR Jitter variation in stable networking bandwidth

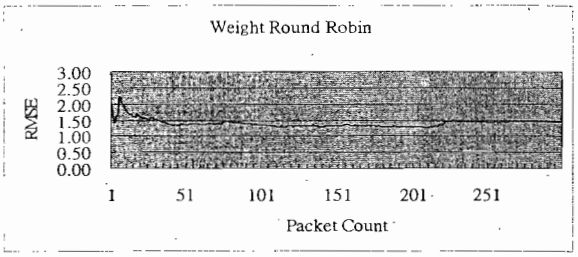

Fig. 4 WRR Jitter variation in stable networking bandwidth

The experiment result showed that Self-adaptation load-balancing algorithm as shown in Fig.4 which can allow the load-balancing system to minimize the amount of variation of packet delay, and this characteristic can be used for multimedia transmission, so that users can use multimedia data smoothly, and the multimedia quality is improved.

4. CONCLUSIONS

In this paper, a suite of theoretical technology is brought forward using RTP/RTCP to measure and feedback the network bandwidth value, and combining the theoretical technology of load-balancing system to achieve self-adaptation of load-balancing mechanism. Through this mechanism, the multimedia transmission mechanism can combine multiple network path bandwidth under inadequate single network bandwidth to increase user's available bandwidth, and guarantee the service quality of multimedia data at the same time. It can also avoid termination of service resulted from over congestion or interruption on single network. Therefore, this mechanism is conducive to multimedia transmission mechanism application, especially for those users who have inadequate single path bandwidth. Moreover, the bandwidth detection mechanism raised herein still has some points shall be improved. Due to the detection principle of instant bandwidth is estimating present available bandwidth through messages returned from RTCP, the system return interval is also in relation to the sensitivity and accuracy of bandwidth decision mechanism. The shorter the RTCP return time is, the more accurate the estimated bandwidth will be, but relatively more bandwidth will be used, so there is a correlation between the sensitivity and occupied bandwidth. Both characteristics should be comprised at the same time, and the design of bandwidth detection should be improved in the future.

5. REFERENCES

[1] Schroeder, T.; Goddard, S.; Ramamurthy, B.; "Scalable Web Server Clustering Technologies" Network, IEEE Volume 14, Issue 3, May-June 2000 Page(s):38 - 45.

[2] Jon Postel, "Transmission Control Protocol", IETF RFC 793, September 1981.

[3] Jon Postel, "User Datagram Protocol", IETF RFC 768, August 1980.

[4] Chang, L.H.; Tai, C.F.; Wang, D.J.; Lai, K.C.; "Dynamic load balancing for wired and wireless Internet access", Circuits and Systems, 2004. Proceedings. The 2004 IEEE Asia-Pacific Conference on Volume 2, 6-9 Dec. 2004 Page(s):889 - 892.

[5] Robert L. Carter; Mark E. Crovella, "Measuring Bottleneck Link Speed in Packet-Switched Networks", BU-CS-96-006, Boston University, March 16, 1996.

[6] Borland C++ 6.0, http://www.borland.com/

[7] Indy Sockets library, http://www.indyproject.org/index.en.aspx.

[8] The National Institute of Standards and Technology, NIST Net 2.0.12, http://snad.ncsl.nist.gov/nistnet/