中華民國八十六年全國計算機會議

Demand-Based Mirror System

需求驅動自動轉存系統

林家邦　蔡奕楷　林宏達　曾黎明
Chia-Ban Lin　Yi-Kai Tsai　Hong-Dar Lin　Li-Ming Tesng

國立中央大學資訊工程研究所

{lucas,cuma,honda}@dslab.csie.ncu.edu.tw

摘要

FTP 物件的 mirror 傳統上是由管理者來做. 由管理者 mirror 的檔案可能不會被使用者用到因為管理者並不知道使用者的需求. 然而, WWW 快取代理者會根據使用者需求來保存檔案. 為了解決 FTP mirroring 的缺點, 我們使用有使用者需求特性的 WWW 快取代理者來建造 FTP 伺服器. 該伺服器會 mirror 常用的檔案並且可以減少 FTP 的交通量.

ABSTRACT

The mirroring of FTP object is traditionally made by administrator. The files mirrored by administrators might not even used by users because administrators do not know the user demand. On the other hand, WWW caching proxy keeps files according to user demand. To solve the drawbacks of FTP mirroring, we use the user-demand characteristics of WWW caching proxy to build an FTP server that mirrors FTP objects without the need of an administrator. The mirror on demand FTP server can mirror popular files and reduce FTP traffic outside local network.

Keywords: Internet; World Wide Web; Proxy; FTP; Cache.

1 Introduction

The FTP service is one of the most commonly used services across the Internet. Its main function is to let users download packages, documents, and files in which users are interested across the Internet. However, with the population growth of the Internet society, the traffic growth due to the growing population of the Internet has become an important issue. This problem is especially important in areas where the communication link from local network to the Internet is limited, such as TANET in the Taiwan area. Current techniques to solve such traffic problem includes proxy-caching[6],[7],[8] and mirroring[9].

Caching technique is often used to reduce WWW traffic on the communication link and shorten response time in the WWW requests. This method requires users to set their browsers right so that the browsers can redirect user's requests to the caching proxy server. The caching proxy server, on demand of user's requests, relay requests to WWW servers and caches and forward the results to users if the requested objects are not in the caching proxy server, or response the requested objects to users without contacting the WWW servers in the remote side if the requested objects are already in the caching proxy server. When the redirected user requests can be satisfied by the caching proxy server, the user gets shorter response time and the traffic load is reduced on the communication link. When the caching proxy server's capacity is full, the caching proxy server will purge out

some objects from the proxy.

A mirror site stores files that will likely to be used by users across the Internet. If the files that users want are located at remote site, it will spend much time and cause certain amount of traffic load to download files. If the traffic capacity of the communication link to the Internet is limited, such as that of TANET, it will become an important issue to reduce the traffic load on the communication link. Mirroring are often used to solve this problem. FTP server administrators use a MIRROR[4] program to copy files at remote FTP servers to the local FTP servers (which is called mirror site) so that users can get the wanted files at local site, thus reduce the traffic load on the limited communication link.

The problems of the mirror system are:(1) it requires a administrator to decide which files to copy and which files not to copy, (2) the adminitrator might copy files that are not popular any more, thus wasting communication resource on copying the not needed files, (3) the users may not even know the existence of such mirror site, thus mirroring is in vain and (4)searching the needed files among files on the mirror site is time-consuming. The main purpose of this paper is to solve the problems of such mirror system. We combine the strength of caching and mirroring and build a system that mirror files on user's demand without the need of administrators. The roadmap of this paper is organized as follows. Section 2 will discuss the strength and weakness of caching and mirroring. We propose an improved architechture of mirror system that in section 3. Section 4 reports the current usage statistics of our system. Finally in section 5, we concludes our work and discuss future work.

2 Mirroring and Caching

In this section, we examined the difference between mirroring and caching techniques. As stated previously, mirroring and caching are two commonly used techniques to alleviate the traffic load on the communication link to the internet. The objects on the caching proxy server are stored according to user demand. Each WWW/FTP object on the proxy server are at least accessed once by users. Traditional caching proxy server will not stored objects that are not demanded by users. On the other hand, objects on the mirror site might not be accessed by users, because the administrators do not know what user's requests are now. Therefore, caching has on-demand characteristics while mirroring do not.

On the other hand, objects on the mirror site has longer TTL (time-to-live) value. Because FTP objects are not likely to change very often, the mirrored copies of FTP objects can stayed on the mirror site longer without worrying their validity (or freshness). The objects on the caching proxy server are often expired or purged out quickly from the proxy server because of the cache replacement. Thus the mirrored copies of FTP objects are utilized for longer time than the WWW cached objects. It is important to decide which FTP objects to replicate on the local disk because that if we make the wrong decision, the mirrored

Table 1 Difference between mirroring and caching

	Mirroring	Caching
On-Demand characteristics	×	✓
File copy decision by Administrator	✓	×
Long TTL on local storage disk	✓	×
Visible by users	✓	×

objects wastes disk space and will consumes disk space for much longer time that of WWW objects.

Another difference between FTP mirroring and WWW caching is that mirrored files can be searched and are visible to users while cached objects on proxy server are transparent to WWW users. The visibility of mirrored objects can be searched via search engine such as Archie[1]. Table 1 list the difference between mirroring and caching.

3 System architecture of Mirror system

In order to improve the tradition mirror system, we combine the strength of WWW caching and FTP mirroring to let the system mirror FTP objects according to user demand without the help of administrators. Our system utilization the user-demand characteristics of WWW caching proxy server to help replicating FTP objects. Our system contains the following components: (1).WWW caching proxy server (or cache server), (2).Local FTP server, and (3). proxy-ftp client. Figure 1 shows the architecture of our demand-based mirroring system that combining the WWW caching proxy server and FTP mirror server.

3.1 WWW caching proxy server

WWW caching proxy server can be used to cache FTP objects. The user redirect FTP retrieval requests to WWW cache server, thus cache server can therefore cache the FTP objects. Because of demand-based characteristics of user FTP retrieval requests, then we can use the user requests information to copy FTP objects to local FTP server. Currently we used Squid Internet Object Cache[10] as our cache server and we modify it so that it can replicate FTP objects to local FTP server.

3.2 Local FTP Server

The local FTP server has large disk space capable of storing FTP objects across the internet. The FTP objects are mirrored to this FTP server via our proxy cache server. This made the invisible FTP objects on the WWW cache server visible on the local FTP server. Search engines like Archie[1] can gather information on this local FTP server and make human search FTP objects more easily. This reduce the traffic to fetching FTP objects across the communication link of which bandwidth is limited.

3.3 Popularity of FTP objects

Although the storage space for FTP objects mirrored by cache server is large, it

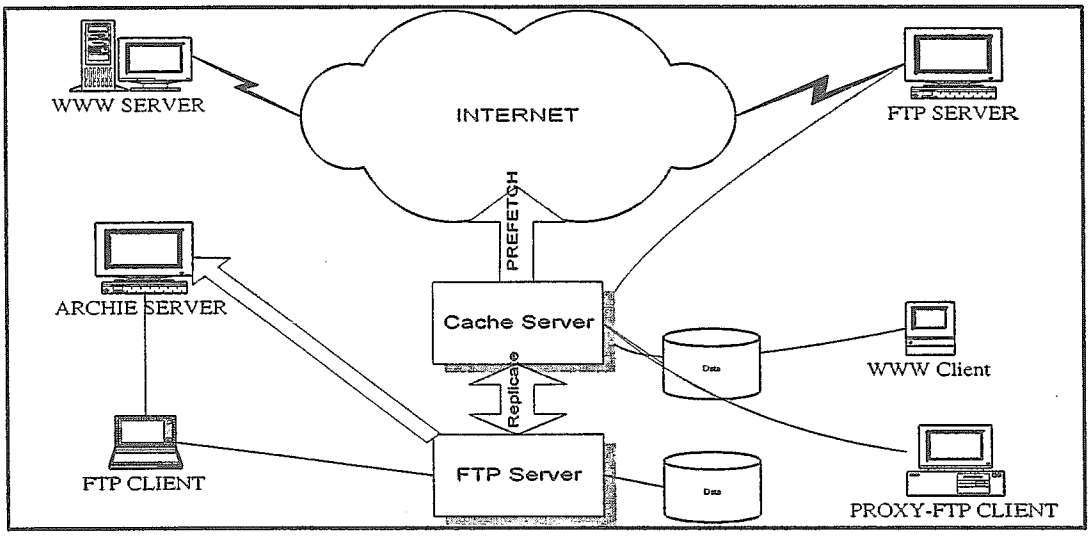

Figure 1. Architecture of Demand based mirroring system.

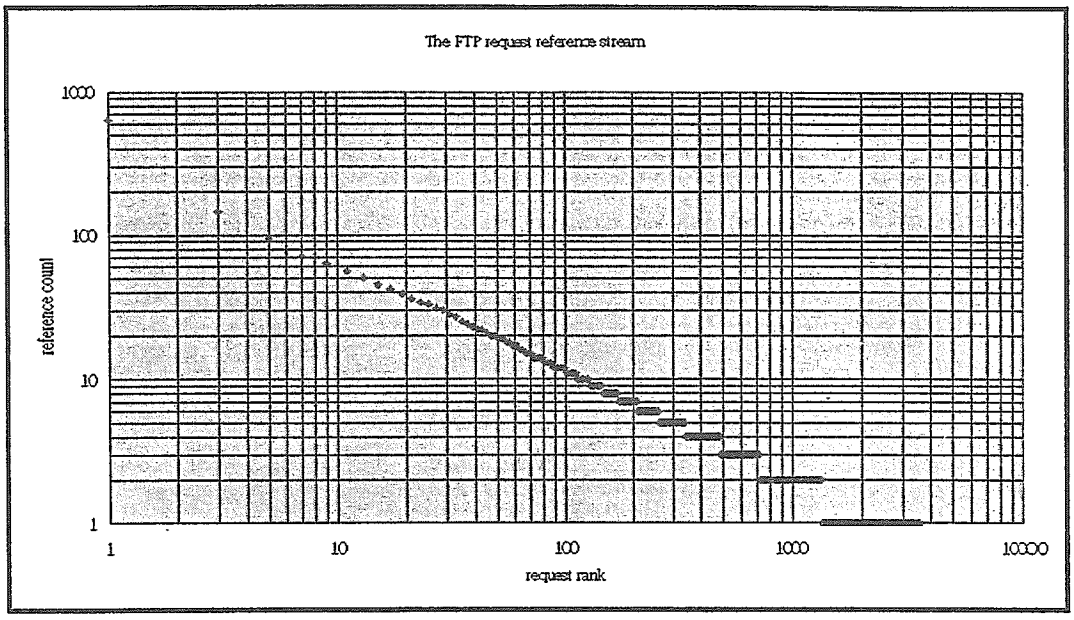

Figure 2. The Popularity of FTP objects memory area.

is possible that local FTP server run out of disk space in the long run of mirroring FTP objects via cache server. When the disk space are full, the system must purge some FTP objects out of the disk space. There are several algorithms for choosing objects out of the disk space. Traditionally in virtual memory system [2,3], LRU (Least Recently Used) algorithm is generally used for choosing a page to be purged out of the physical memory area.

The reason for using LRU page replacement algorithm is that program behavior of accessing memory exhibits strong temporal locality characteristics. However, in the FTP retrieval behavior, we don't know if the temporal locality characteristics still exists in the FTP objects retrieval behavior. Instead of using LRU algorithm, we investigate another approach: frequency based algorithm. The frequency-based algorithm choose the objects/pages which are least frequently used. We investigate the FTP reference stream to see if the frequency based algorithm fits in FTP objects accessing behavior. We use the proxy server's log to get the FTP objects reference stream, and then we count how many times each FTP objects are accessed. The reference stream are logged from 04:30 December 10 1996 to 04:30 December 27 1996. The number of times of FTP objects retrieval is 12035 and the number of FTP objects accessed is 3896. Figure 2 shows the sorted popularity each FTP objects. We found that the popularity of FTP objects confirms to the Zipf's law[5]. The Zipf's law can be showed by the following equation:

$$F(X) \propto \frac{1}{X}$$

X represents the rank of objects, and F(X) represents the frequency of the Xth object being accessed. FTP objects access behavior shows Zipf's law. From Fig 2, we know that out of the 3896 files being accessed, more than 2000 files are only accessed once while the top 100 files are accessed at least 10 times. Thus the popularity of FTP objects can be used as an index for choosing objects to be purged. If we accumulate access times from the hottest FTP objects to the least popular objects, we can know how many FTP objects will account for how many FTP requests. Figure 3 shows that top 12.5% files account 60% of the requests and top 37.5% files can account for 80% of the FTP requests.

3.4 Proxy-FTP client

Current common FTP client software make connection directly to FTP

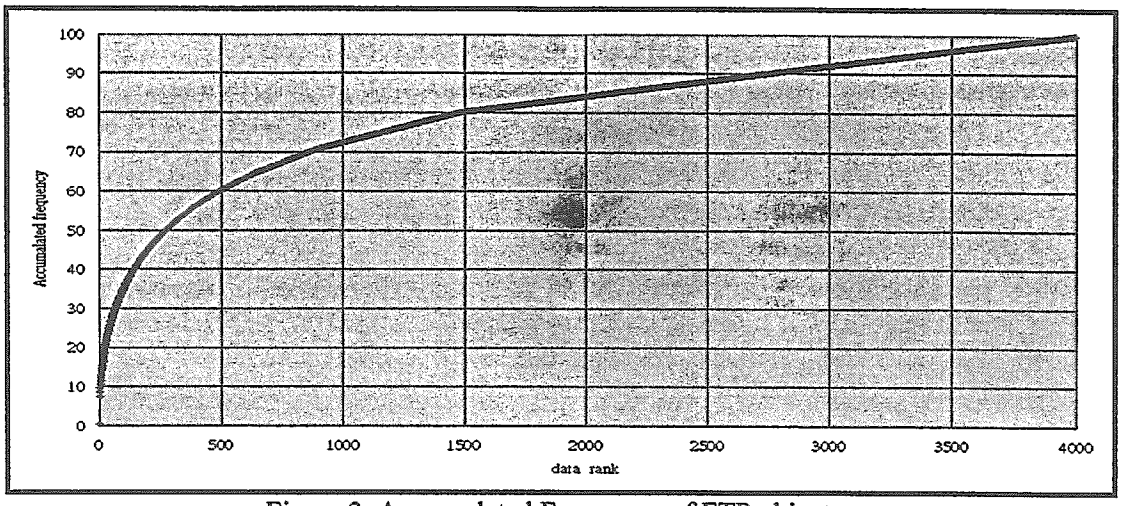

Figure 3. Accumulated Frequency of FTP objects

servers without proxying. Therefore, we design a FTP Client program that redirects requests to WWW proxy server, which will hopefully reduce FTP traffic. The Proxy-FTP client talks to WWW proxy server with HTTP protocol, and can be used to download public files (files on anonymous FTP servers).

4 System usage report

We set up a FTP server that mirrors FTP objects automatically with the help of WWW proxy server. The mirrored copies on the local FTP server have more user-demand characteristics than those of mirrored by administrators. The mirrored files on FTP server are highly utilized. We examined the number of times being accessed of each FTP object on the FTP server. If a object being accessed is repeated accessed again and again, we know that this object is very popular, in such case the mirroring of such a file can reduce much bandwidth because the local FTP server satisfy the demand of requesting such a file. We examine each day's FTP traffic serviced by local mirror server and the redundant traffic. Redundant traffic is caused by accessing files that are already accessed. Figure 4 shows the FTP traffic serviced by mirror server on FTP objects outside from TANET and the redundant traffic. Lighter color bars represent FTP traffic serviced and darker color bars mean redundant traffic. For 11 days of FTP objects retrieval, 1070 MB of traffic is serviced and 905 MB of traffic is redundant.

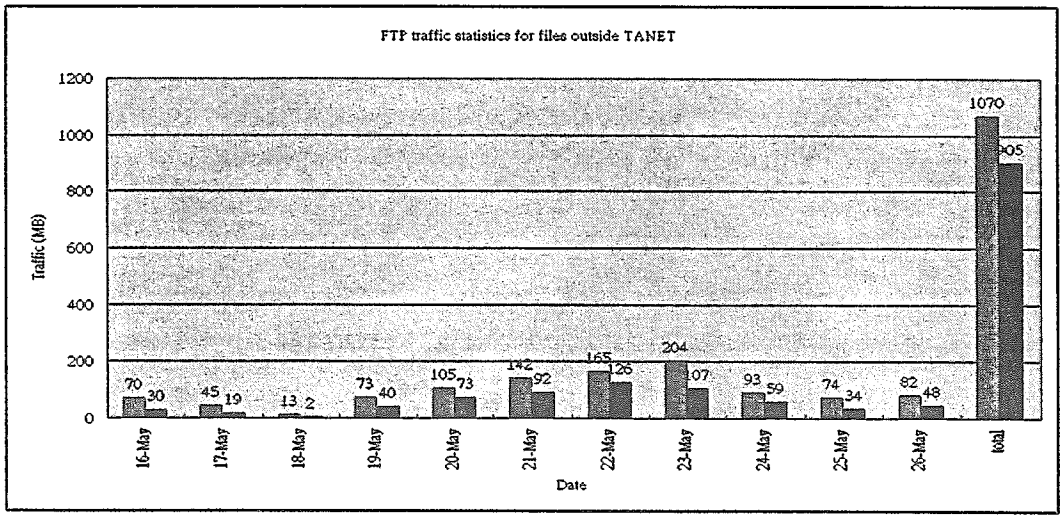

Figure 4.

5 Conclusion and Future work

In this paper, we discuss the improved mirror system that is demand based. The mirror on demand FTP server mirror files with the help of WWW cache server. A Proxy-FTP client is designed to utilize proxy server for FTP objects retrieval. We find that mirrored copies on the system are popular and thus help alleviate traffic demand on communication link.

In this paper, we also discuss the frequency based algorithm for purging FTP objects out of our FTP mirror server.

6 Reference

[1] Alan Emtage and Peter Deutsch. "Archie - an electronic directory service for the internet" In proceeding of the *USENIX Winter conference*, January 1992.

[2] Jeffery Spirn. "Distance string models for program behavior" IEEE computer, November 1976.

[3] R. Mattson, J. Gecsei, D. Slutz, and I. Traiger. "Evaluation techniques and storage hierarchies", IBM systems Jurnal, 9:78-117,1970

[4] Lee Mcloughlin lmjm@doc.ic.ac.uk "mirror - mirror pacakge on remote sites"

[5] G.K. Zipf, "Human Behavior and the Principle of Least-Effort", Addison-Wesley, Cambridge, MA, 1949.

[6] Mosaic-x@ncsa.uiuc.edu "Using proxy gateways. World Wide Web" available from http://www.ncsa.uiuc.edu/SDG/Software/Mosaic/DOCS/ proxy-gateways.html

[7] Mathew A. Blaze. "Caching in large-scale distributed file systems", Technical Report 39792, Princeton University, 1993

[8] Peter Danzig, Richard Hall, and Michael Schwartz. "A case for caching file objects inside internetworks", Technical Report CU-CS-642-93, University of Colorado, Boulder, 1993

[9] James Gwertzman, "Autonomous Replication in Wide-Area Internetworks" Technical Report TR-17-95, Harvard University, Apr 1995.

[10] SQUID Internet Object Cache http://squid.nlanr.net/Squid/