An Investigation of File Access Patterns in Public FTP Archives*

DRAFT

Silvano Maffeis
maffeis@ifi.unizh.ch

Institut für Informatik der Universität Zürich (IFI)
Winterthurerstr. 190, CH–8057 Zürich, Switzerland

July 21, 1992

Abstract

New file transfer protocols and global filesystems are a great need and challenge in the presence of drastically growing networks. As a first step towards designing and realizing such new protocols and filesystems, we present results obtained from an investigation of access to public files which took place over three months. This work visualizes first results on the popularity of public `ftp` files, on common operations (deletions, updates and insertions) to public files and on encountered file-sizes. The results show that most file transfers relate to only a small part of the files in an archive and that a considerable part of the operations to public files are updates of a file. Further results are presented and interpreted in the paper. An index for measuring locality of reference to a resource is also proposed.

Keywords: file transfer, popularity of files, locality of reference, file-sizes, replication

*This work was supported by *Siemens AG, ZFE*, Germany and *Schweizer Bundesamt für Konjunkturfragen*, Grant No. 2255.1

1 Introduction

The number of hosts in the `Internet` increases drastically from year to year [6], accompanied by an increase of file transfer traffic. Today there are thousands of anonymous FTP sites with millions of files provided on them. Under these conditions, the traditional FTP utility [9] shows severe inefficiencies like duplicate file-transmissions [4], long access times and inadequate file locating mechanisms. Therefore, several improvements and redesigns of FTP have been proposed, for example [10, 2, 5] and [7].

In this paper we analyze FTP data-traffic on a *per-file basis*. The purpose of the analysis is to gather characteristic information on file-accesses in public FTP archives, which is useful for designing new file transfer protocols and global file systems or for simulating data access and transfer in large networks.

The rest of the paper is structured as follows: In section 2 we define which specific analyses were performed and how the data was collected. Section 3 presents results concerning the popularity of anonymous FTP files and proposes an index for characterizing locality of reference. Section 4 presents our results relating to the most important operations on file-archives, namely creation of new files, deletion of files and updating of files. File-sizes are examined in section 5. Finally, section 6 summarizes the main findings and concludes the paper.

2 Conducted Analyses

The intent of our work was to find answers to the following questions:

1. What can be said about the popularity-distribution of public files? Here we aimed at deriving statements of the form "X% of the filetransfers from a certain archive relate to Y% of its files".
2. How are the frequencies of operations like creation of new files, deletion of files and updates of files interrelated? Is the common assumption that *updates* to files happen very infrequently valid?
3. How are file-sizes distributed over public FTP archives and how does the average filelength evolve over time?
4. What can be said about the sizes of the most frequently retrieved files?

To address these questions two distinct classes of measurements were carried out:

2.1 Analyzing `FTP-log` files

The FTP server process `ftpd` logs FTP commands issued by anonymous users to a specific logfile. With the help of some `perl` programs we wrote for this purpose, statements concerning the access-rates to files could be derived. This class of measurements was complicated by the fact that many different formats of FTP-logfiles exist.

Logfiles were located by having the `archie` [3] service search for appropriate filenames in the `Internet` network. Also, some operators of well-known archivesites participated directly by granting us the access to their FTP-logfiles. We tried to find a few characteristic sites showing a minimum of 2000 file transfers per month and which were located at different geographical places in the `Internet`. The following archive sites' log-files were analyzed:

ftp.nisc.sri.com
iam.unibe.ch
nic.stolaf.edu
nic.switch.ch
phloem.uoregon.edu
rusinfo.rus.uni-stuttgart.de
vax.ph-cip.uni-koeln.de

Information relative to accesses to singular files is usually contained in a FTP-logfile. Therefore question number one and four of section 2 could be addressed by evaluating several such logfiles.

2.2 Analyzing ls-lR.Z files

Most of the large and frequently accessed public archives maintain what is called an ls-lR.Z index file. At fixed time intervals (mostly every two weeks) the whole FTP–archive is browsed by locally issuing the UNIX ls -lR command. The output of this procedure is written to disk, compressed and stored as ls-lR.Z file, usually near the root-directory of the archive. Users interested in the contents of a certain archive may retrieve and inspect this file.

The information of interest we obtained evaluating ls-lR.Z files concerns the directory-structure of an archive, files stored in the archive at the time the index was created, file-sizes and modification times of files. This part of the research pointed at questions number two and three of section 2.

Internet sites appropriate for participating to this class of measurements were selected by having the archie service search for hosts containing ls-lR.Z files. We obtained more than 200 domainnames from which we initially selected 25. The discriminating criteria for the selection were

- size of the ls-lR.Z file: we required a minimum size of 40 kilobytes
- update-frequency: the ls-lR.Z should be updated at least every two weeks
- we also tried to select participating sites residing at various geographical locations and serving different purposes: academic, commercial and nameserving (NIS-hosts).

We applied these criteria in a rather intuitive manner. Table 1 contains the domain-names of the Internet-sites used in this analysis and the size of the related ls-lR.Z files.

3 Popularity of Public Files

In this section we examine the popularity of files in public FTP–archives. By "popularity of a file" we mean the relative amount of times the file has been retrieved from the surveyed archive during a certain period of time, compared to the other files in the archive. In this part of the study we assume that the archives consist only of files which got retrieved at least once during the surveyed time-period. For a surveyed archive $\mathcal{A}$ consisting of N files f_i, with $i = [1..N]$, we assign each f_i its access-probability p_i, which denotes the likelihood for file f_i being accessed by a filetransfer from archive $\mathcal{A}$ during the surveyed timeperiod τ.

3.1 Popularity-graphs and LORI

We characterize the popularity of Internet files by drawing what we call a *popularity-graph (PG)* (figure 1). The y–axis of a PG denotes the percent-ratio of filetransfers from an archive during

domain-name	*ls-lR.Z (kb)*	*num files*
akiu.gw.tohoku.ac.jp	50	2197
archive.cis.ohio-state.edu	142	8713
arthur.cs.purdue.edu	601	37943
colonsay.dcs.ed.ac.uk	180	11798
csc2.anu.edu.au	59	3473
decuac.dec.com	93	6520
etlport.etl.go.jp	107	5993
gatekeeper.3com.com	52	3727
gdr.bath.ac.uk	120	7975
iros1.iro.umontreal.ca	183	14523
labrea.stanford.edu	45	3266
liasun3.epfl.ch	224	15329
metro.ucc.su.oz.au	82	5100
nic.funet.fi	1685	112052
nic.switch.ch	85	4704
nuri.inria.fr	63	3975
pprg.eece.unm.edu	69	4246
procyon.cis.ksu.edu	82	4134
relay.cs.toronto.edu	89	5889
reseq.regent.e-technik.tu-muenchen.de	302	15600
rs3.hrz.th-darmstadt.de	224	14368
shiva.com	92	5790
unixd1.cis.pitt.edu	53	3355
wuarchive.wustl.edu	1592	122407
zaphod.ncsa.uiuc.edu	60	3529

Table 1: Internet sites which participated to the `ls-lR.Z` evaluation