File Access Patterns in Public FTP Archives and an Index for Locality of Reference*

Silvano Maffeis
maffeis@ifi.unizh.ch

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

IFI TR 92.13

August 1992

Abstract

Global filesystems and new file transfer protocols are a great need and challenge in the presence of drastically growing networks. In this paper 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 file-archives and on encountered filesizes. An index for measuring locality of reference to a resource is also proposed. The results show that most file transfers relate to only a small fraction of the files in an archive and that a considerable part of the operations to public files are updates of files. Further results are presented and interpreted in the paper.

Keywords: file transfer, popularity of files, locality of reference, filesizes, 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 [7, 5], accompanied by an increase of file transfer traffic. Today there are thousands of anonymous FTP sites with millions of files stored on them. Under these conditions, the traditional FTP utility [10] 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 [11, 2, 6] and [8]. 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 global filesystems or new file transfer protocols and for simulating data access 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.

The tables and figures referenced in the text are included at the end of 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 assumption that updates to files happen very infrequently valid?
3. How are filesizes distributed in public FTP archives?
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: we analyzed FTP-logfiles and FTP-indexfiles.

2.1 Analyzing FTP-logfiles

Many FTP server processes (`ftpd`) log FTP commands issued by anonymous users to a specific logfile. With the help of some `perl` [12] 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 archive sites participated directly by granting us 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:

domain-name	*location*	*purpose*
ftp.nisc.sri.com	USA California	commercial
iam.unibe.ch	Switzerland, SWITCH	scientific
nic.stolaf.edu	Northfield MN, MRNet	educational, mailing lists archive
nic.switch.ch	Switzerland, SWITCH	NIS, mirrors etc.
phloem.uoregon.edu	Oregon, USA	network services archive
rusinfo.rus.uni-stuttgart.de	Germany, BELWUE	different info-services
vax.ph-cip.uni-koeln.de	Germany, WIN	educational

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

2.2 Analyzing FTP-indexfiles

Most of the large and frequently accessed public archives maintain an index of all files in the archive. 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 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, filesizes 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 in 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

- the size of the FTP-indexfile: we required a minimum size of 40 kilobytes
- the update-frequency: the indexfile 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).

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 the "popularity of a file" we mean the number 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 respect only 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)* (see figure 1). The y–axis of a PG denotes the percentage of filetransfers from an archive during the surveyed timeperiod τ (usually one month), and corresponds to the *ascendingly* ordered and then cumulated p_i values. The x–axis expresses the percentage of files in the archive and thus is equivalent to i/N. Statements of the form "Y% of the filetransfers (value on the y–axis) from archive $\mathcal{A}$ relate to X% of the files (value on the x–axis) in archive $\mathcal{A}$" can be expressed using a PG.

A PG is an illustrative tool for graphically demonstrating the locality of access (i.e. the popularity) for files in a certain archive, but it is difficult to compare many archives by just relying on their PGs. We therefore propose a numerical index we found useful for measuring and comparing the locality of reference in file-archives. In virtually all archives, some files are more frequently accessed than others, leading to different p_i values. According to information theory, if file f_i is accessed with probability p_i, then the *entropy* of the accesses to the files f_i in archive $\mathcal{A}$ during period τ is

$$H_{\mathcal{A},\tau} = -\sum_{i=1}^{N} p_i \log_2 p_i$$

Thus the average information we obtain from a single access to $\mathcal{A}$ is $H_{\mathcal{A},\tau}$. The *locality of reference index* (LORI for short) we propose is the *redundancy* of the accesses to the files f_i in archive $\mathcal{A}$ during period τ, thus

$$LORI_{\mathcal{A},\tau} = R_{\mathcal{A},\tau} = \frac{\log_2 N - H_{\mathcal{A},\tau}}{\log_2 N}$$

The LORI is computed by normalizing the difference between $H_{\mathcal{A},\tau}$ and the highest possible entropy, which is $\log_2 N$. This way a LORI expresses the popularity distribution of the files in a surveyed archive. A LORI can also be used to characterize locality of reference to other resources, such as network-nodes in a distributed system or mainmemory-pages. In an extreme case a LORI may be equal to 1.0 meaning that 100% of the accesses related to just *one* single file in the archive. The locality of reference is thus maximal. A LORI of 0.0 denotes no locality of reference, meaning that every file shows the same popularity. In this extreme case, the associated PG lies on the 45 degrees line of figure 1.

To give a notion of which LORIs relate to low or high locality of reference, some sample PGs are drawn in figure 2. We propose following interpretation for the LORIs gathered in this study:

$0 < \text{LORI} \leq 0.05$	low locality of reference
$0.05 < \text{LORI} \leq 0.10$	medium locality of reference
$0.10 < \text{LORI} \leq 0.20$	high locality of reference
$0.20 < \text{LORI}$	very high locality of reference

3.2 PG and LORI: the results

Table 2 presents the LORIs of all surveyed sites. Table 3 answers the question of how many files were accessed in 50% respectively 80% of the file-transfers. On an unweighted average of table 3, we see that 50% of all filetransfers related to 12% of the most popular files and 80% of the transfers to 42% of the most popular files.

Figure 3 exemplifies PGs computed for three hosts found in table 2. The pictured data was collected during one month (April 1992). The associated LORIs can be found in table 2. All PGs of the figure show what we defined a *high locality of reference*. It is interesting to note the knee in the region of 50% accessed files. We found a similar phenomenon in PGs relating to other hosts which also offer NIS services to the net. The knee in the graph is due to the fact that many files like mapping tables etc. are fetched regularly, but only very few times per month from a NIS host. Mirroring [9] can also induce a kinked PG. Future file transfer protocols and filesystems, for example ones involving file-replication [8] will have to consider such and other described effects.

4 Operations on File-Archives

This section presents results which relate to question number two posed in section 2. Table 4 overviews the relative amount (percentage) of deletions (*del*), updates (*upd*) and insertions (*ins*) to the surveyed filearchives for a timeperiod of two weeks. The two Δ–values denote the net amount of files resp. directories added to the archive during the time-period. *num files* stands for the number of files in the archive at the beginning of the time-period.

It is interesting to remark that real updates, meaning that a file is altered but keeps its original name and directory-location, happen relatively frequently to public files. On an average, 12.6% of the surveyed operations were pure updates of files. We found that many of the updated files related to archive index-files (`ls-lR.Z`, `INDEX`, `Index`, ...), different `README`–files, but also documentation files and technical reports. On an average, the number of files in an archive seems to increase by about 3% per month.

Future file transfer protocols and file-systems will have to take into account the considerable amount of file-updates and the fast rate at which public file archives tend to grow.

5 Filesizes

Measurements in the SPRITE distributed file system [1] reported about the advent of multimegabyte files in todays filesystems. Similarly, we were interested on the distribution of filesizes in public FTP-archives.

Table 5 contains the average filesize (*avg*) and the standard-deviation (*stddev*) of the encountered file-lengths. The 97.5% quantile column in table 5 demonstrates what the least size of the 2.5% largest files was, the 99.5% quantile relates to the least size of the 0.5% largest files. (*max*) stands for the largest encountered file and is thus equivalent to the 100% quantile. The table was computed using the first series of `ls-lR.Z` files we collected for our study.

Figure 4 tracts a histogram of files with length $\leq$ 600kb for the archive `nic.switch.ch`. We found that most of the surveyed sites had very similarly distributed filelengths, with more than 50% of their files being smaller than 100 kb. The figure depicts some kind of exponential distribution of filelength-occurrences, with files becoming disproportionately rare as the filesize increases.

The peak at value 524288 bytes (512 kb) is caused by the fact that 512 kb is a popular chunk-size for splitting large files. For example, the huge `X-Windows` distribution is split into many chunks of 512 kb each to ease its retrieval and mirroring.

We further found that 50% of the file-data on host `nic.switch.ch` was scattered over 95% of the files, and that consequently the other 50% of the file-data was contained in only 5% of the largest files (figure 5). This means that the largest part of the archived bytes is made up by very few, very large files ($\geq$ 500 kb per file). The rest of the data is scattered over a huge amount of smaller ($<$ 500 kb per file) files. The advent of multi-magabyte files could be confirmed by our investigation, but as may be seen by the 97.5% and 99.5% quantile only about 1% of the files today spawn multiple megabytes of data. Other sampled archives showed a very similiar allotment of their data.

5.1 Filesizes of the most popular files

The former part of this section deduced statements concerning sizes of public files by analyzing `ls-lR.Z` index files. Interesting results can also be obtained by evaluating filesizes reported by the FTP-logfiles described in section 2.1. Here we can address question number four posed in section 2 by measuring what the sizes of the most often retrieved files are.

Figure 6 shows which percentage of filetransfers (value on the y-axis) involves what maximal filelength (value on the x-axis). We see that 80% of the file transfers relate to files smaller than 130 kb. Thus, the transfer of large files is very rare and the absolutely largest fraction of transferred files is smaller than 100 kb. The kinks at file size 512 kb is caused by the fact that 512 kb is a popular size for splitting large files. This effect also appeared in figure 4.

For part of the surveyed hosts, table 6 reflects the maximal encountered filesize in 50% resp. 80% of the filetransfers of April 1992. Again we can note that most of the transferred files are small in size and that the transfer of very long files happens very infrequently.

6 Summary and Conclusions

Our measurements of FTP traffic on a per-file basis addressed four basic 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 assumption that updates to files happen very infrequently valid?
3. How are filesizes distributed in public FTP archives?
4. What can be said about the sizes of the most frequently retrieved files?

Analyzing FTP-indexfiles and FTP-logfiles we found the following answers:

1. Over all surveyed FTP–archives, 50% of the filetransfers related to only 12% of all retrieved files, and 80% of the filetransfers related to only 42% of all retrieved files. The LORI values computed for the filetransfers happening during a timeperiod of one month showed what we interpreted as a *high locality of reference* for most of the surveyed hosts.
2. On an average over a sample period of two weeks, we found that 32.5% of the operations to public files were deletions of files, 54.9% insertions of new files into an archive and 12.6% of the operations related to updates of files. The number of files in an average archive seems to increase by about 3% per month.
3. The average size of a file stored on the surveyed archives amounted to 60 kb and showed a standard deviation of 344 kb. We found that only about 1% of the files had a size larger than 1 mb, but 50% of the data was spent over only 5% of the largest files.
4. We found that about 50% of the file retrievals related to files smaller than 25 kb, and 80% of the retrievals to files smaller than 100 kb for the surveyed hosts. Thus, the transfer of multimegabyte files happened very infrequently for all the surveyed hosts.

Observing these results, we feel that future file transfer protocols and large scale filesystems could effectively make use of *object-replication* for substantially increasing availability of and decreasing access times to stored data. The fact that most of the filetransfers relate to relatively very small files and that file-updates happen considerably often asks for adapted caching mechanisms and replication protocols. Since filearchives have to store a wide variety of filesizes, adequate storage-management mechanisms will also be needed in future large scale filesystems and transfer protocols.

For this study the data related to file-transfers has been collected from more than 25 `Internet` archives over three months.

Acknowledgements

We would like to thank Apostolos Lytras (University of Zurich) for helping us collecting the data to this study and for reviewing the first drafts of the paper. Clemens Cap (University of Zurich) also reviewed the first drafts and also provided useful suggestions concerning the LORI index. Our gratitude also goes to the administrators of the surveyed `Internet` sites, for maintaining the index- and logfiles we evaluated in this study. Kurt Jaeger (University of Stuttgart, Germany) and Thomas Lenggenhager (SWITCH Switzerland) also participated in this study by granting us access to FTP-logfiles and other interesting data.

Tables and Figures

domain-name	*indexfile (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 considered by the `ls-lR.Z` evaluation.

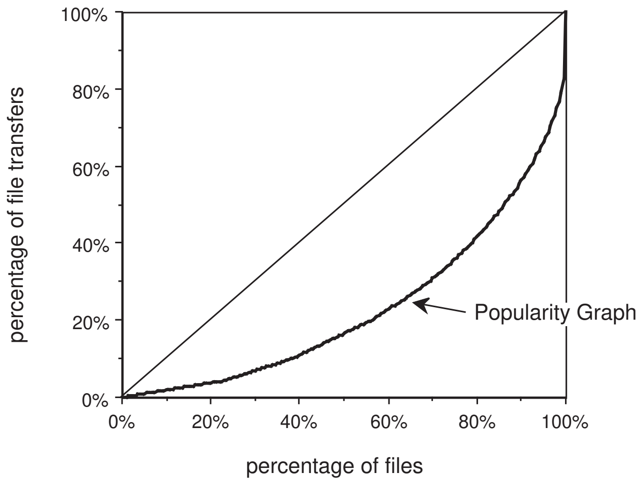

Figure 1: A sample popularity graph.

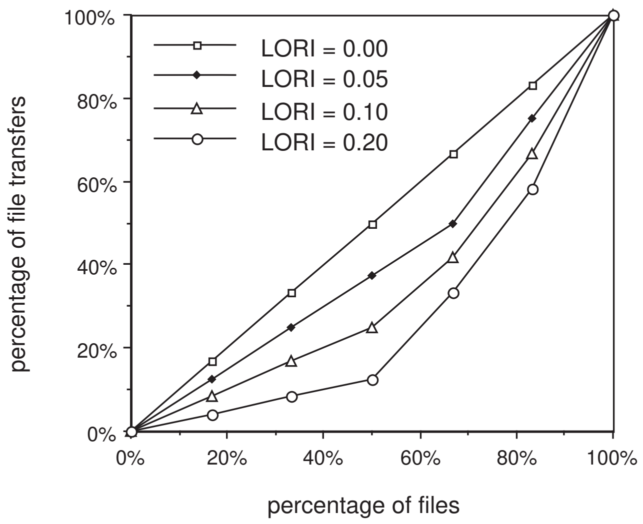

Figure 2: Sample popularity graphs and LORIs.

domain-name	*surveyed period*	*LORI*
ftp.nisc.sri.com	Jan 92 – May 92	0.13, 0.12, 0.09, 0.13, 0.11
iam.unibe.ch	Apr 92, May 92	0.10, 0.11
nic.stolaf.edu	Mar 92 – Jun 92	0.10, 0.11, 0.13, 0.13
nic.switch.ch	Jan 92 – May 92	0.14, 0.13, 0.17, 0.16, 0.19
rusinfo.rus.uni-stuttgart.de	Apr 92 – Jun 92	0.10, 0.09, 0.10
vax.ph-cip.uni-koeln.de	Jan 92 – Apr 92	0.06, 0.06, 0.06, 0.07

Table 2: LORIs computed for the surveyed hosts on a monthly basis.

domain-name	*surveyed period*	*50% of file transfers* / *80% of file transfers*
ftp.nisc.sri.com	Jan 92 – May 92	9%, 11%, 14%, 8%, 11%
		35%, 39%, 47%, 32%, 38%
iam.unibe.ch	Apr 92, May 92	13%, 10%
		42%, 35%
nic.stolaf.edu	Mar 92 – Jun 92	15%, 13%, 8%, 10%
		47%, 40%, 38%, 35%
nic.switch.ch	Jan 92 – May 92	18%, 18%, 13%, 14%, 11%
		58%, 57%, 50%, 50%, 41%
rusinfo.rus.uni-stuttgart.de	Apr 92 – Jun 92	8%, 9%, 8%
		29%, 33%, 30%
vax.ph-cip.uni-koeln.de	Jan 92 – Apr 92	18%, 18%, 20%, 18%
		48%, 49%, 49%, 48%

Table 3: Portion of files accessed in 50% resp. 80% of the file transfers.

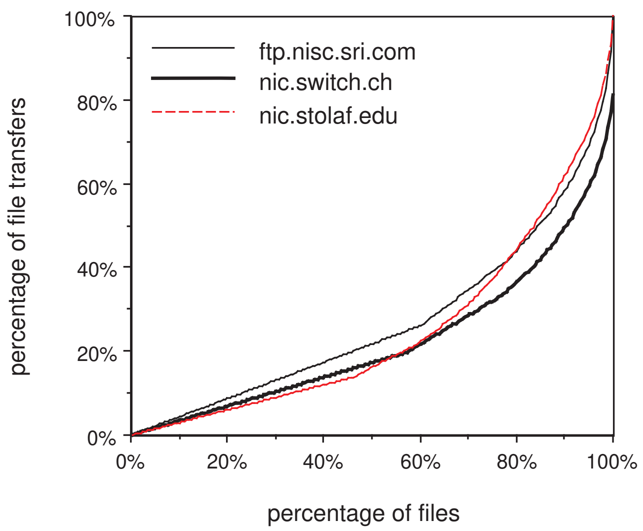

Figure 3: A selection of real popularity graphs and LORIs.

domain-name	*del*	*upd*	*ins*	Δ *files*	Δ *dir*	*num files*
akiu.gw.tohoku.ac.jp	18.40%	4.60%	77.00%	191	9	2197
archive.cis.ohio-state.edu	33.81%	21.44%	44.74%	53	0	8713
arthur.cs.purdue.edu	50.00%	0%	50.00%	0	0	37943
colonsay.dcs.ed.ac.uk	13.75%	6.47%	79.78%	245	4	11798
decuac.dec.com	0.26%	9.06%	90.68%	689	18	6520
etlport.etl.go.jp	55.09%	31.56%	13.34%	-291	-31	5993
gatekeeper.3com.com	33.23%	4.09%	62.69%	274	6	3727
gdr.bath.ac.uk	17.92%	30.19%	51.89%	36	0	7957
iros1.iro.umontreal.ca	0.18%	24.95%	74.86%	808	19	14523
labrea.stanford.edu	7.90%	63.16%	28.95%	8	1	3266
liasun3.epfl.ch	72.02%	2.27%	25.71%	-3210	-126	15329
metro.ucc.su.oz.au	0.87%	4.07%	95.06%	324	20	5100
nic.funet.fi	25.00%	19.07%	55.94%	4678	174	107374
nuri.inria.fr	35.00%	8.46%	56.54%	168	10	3975
pprg.eece.unm.edu	9.19%	32.17%	58.64%	226	8	4246
procyon.cis.ksu.edu	2.24%	84.33%	13.43%	15	0	4134
relay.cs.toronto.edu	38.56%	16.45%	44.99%	34	0	5889
rs3.hrz.th-darmstadt.de	28.11%	18.65%	53.24%	353	36	14386
shiva.com	30.75%	1.59%	67.66%	718	72	5790
unixd1.cis.pitt.edu	1.98%	67.18%	30.84%	131	3	3286
wuarchive.wustl.edu	41.43%	2.86%	55.71%	4645	120	122407
zaphod.ncsa.uiuc.edu	14.47%	31.83%	53.70%	122	5	3529
average (weighted)	32.5%	12.6%	54.9%			

Table 4: Operations on public file archives. Surveyed timeperiod: June 9 – June 30, 1992.

domain-name	*avg (kb)*	*stddev (kb)*	97.5%	99.5%	*max (kb)*
akiu.gw.tohoku.ac.jp	140.5	442.7	823.2	2634.3	8575.8
archive.cis.ohio-state.edu	71.7	180.6	524.3	813.4	7287.6
arthur.cs.purdue.edu	37.9	149.0	274.3	777.6	6575.6
colonsay.dcs.ed.ac.uk	21.1	135.4	78.3	910.9	7715.5
csc2.anu.edu.au	72.7	272.6	799.7	1808.4	6715.6
decuac.dec.com	44.7	221.3	369.1	1419.8	6061.3
etlport.etl.go.jp	81.0	513.3	614.7	3184.1	16929.1
gatekeeper.3com.com	87.6	307.6	524.3	1341.2	8786.3
gdr.bath.ac.uk	50.8	152.2	251.7	708.5	6252.1
iros1.iro.umontreal.ca	28.0	157.5	512.0	524.3	8585.2
labrea.stanford.edu	83.2	319.1	524.2	1519.9	7721.7
liasun3.epfl.ch	33.2	261.4	195.3	911.2	13221.0
metro.ucc.su.oz.au	97.9	473.6	524.3	1938.8	16695.2
nic.funet.fi	66.9	301.8	471.2	1380.0	66749.4
nic.switch.ch	100.6	415.5	620.1	1639.5	20408.0
nuri.inria.fr	244.3	726.3	1835.5	4404.9	19671.0
pprg.eece.unm.edu	188.7	1165.0	736.8	3946.2	67156.6
procyon.cis.ksu.edu	60.6	177.4	464.0	1456.8	4160.0
relay.cs.toronto.edu	90.2	469.5	639.6	2075.1	19700.0
reseq.regent.e-technik.tu-muenchen.de	42.0	233.7	203.7	1228.8	11890.0
rs3.hrz.th-darmstadt.de	61.2	383.8	524.3	1474.6	24526.4
shiva.com	18.6	97.4	112.1	377.6	43295.3
unixd1.cis.pitt.edu	58.6	318.0	315.8	1082.1	14123.0
wuarchive.wustl.edu	43.6	275.9	371.2	659.3	30121.2
zaphod.ncsa.uiuc.edu	73.4	274.4	803.4	1795.8	67156.6
overall values (weighted)	60.2	344.8	432.7	1564.1	20496.0

Table 5: Average filesize, standard-deviation and maximum filesize as measured on June 9, 1992.

domain-name	50%	80%
iam.unibe.ch	61.2 kb	160.2 kb
nic.stolaf.edu	6.1 kb	109.4 kb
nic.switch.ch	11.6 kb	102.2 kb
rusinfo.rus.uni-stuttgart.de	3.2 kb	28.8 kb
vax.ph-cip.uni-koeln.de	38.7 kb	97.3 kb

Table 6: Maximal filelength for 50% resp. 80% of the filetransfers in April 92.

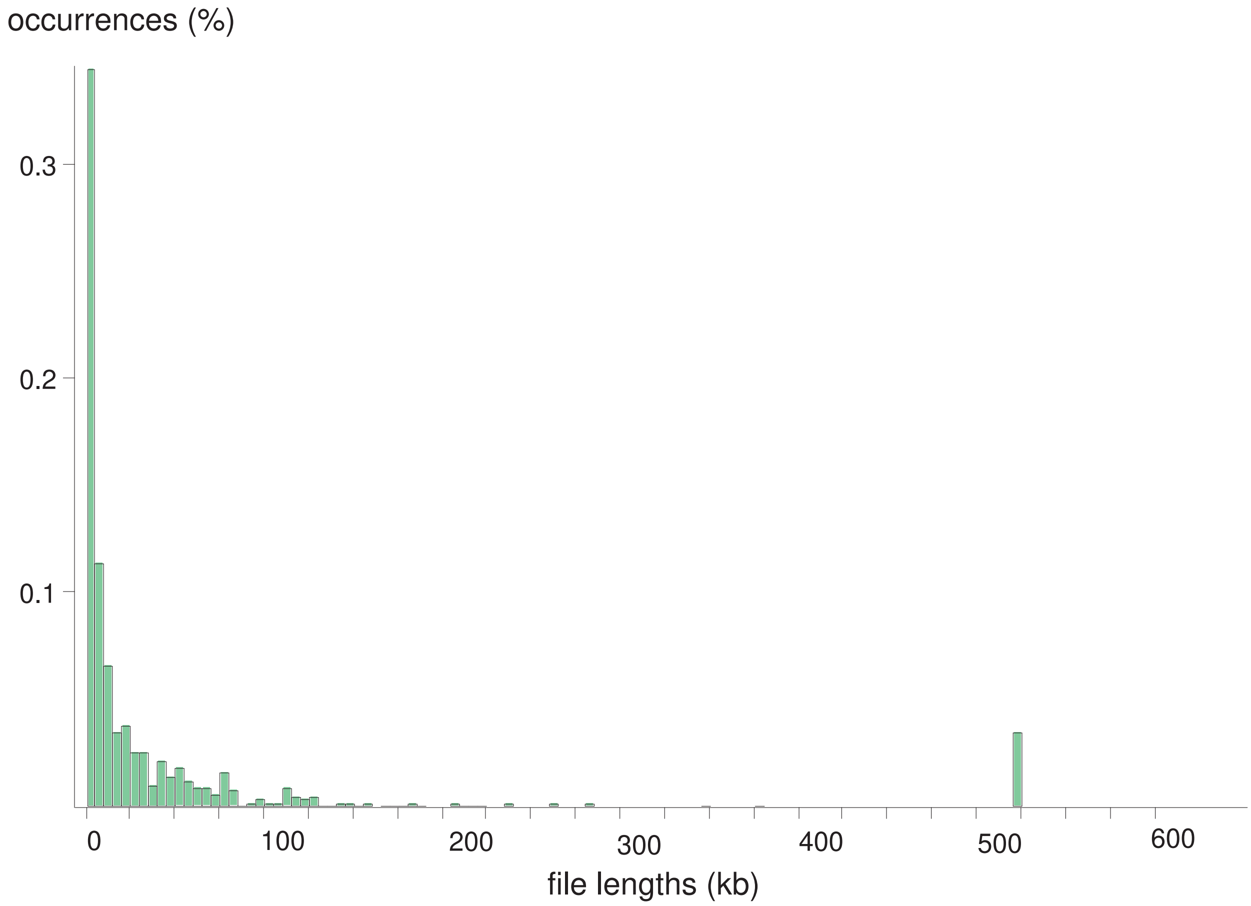

Figure 4: Histogram for files with length ≤ 600kb for archive nic.switch.ch. As seen on June 9, 1992.

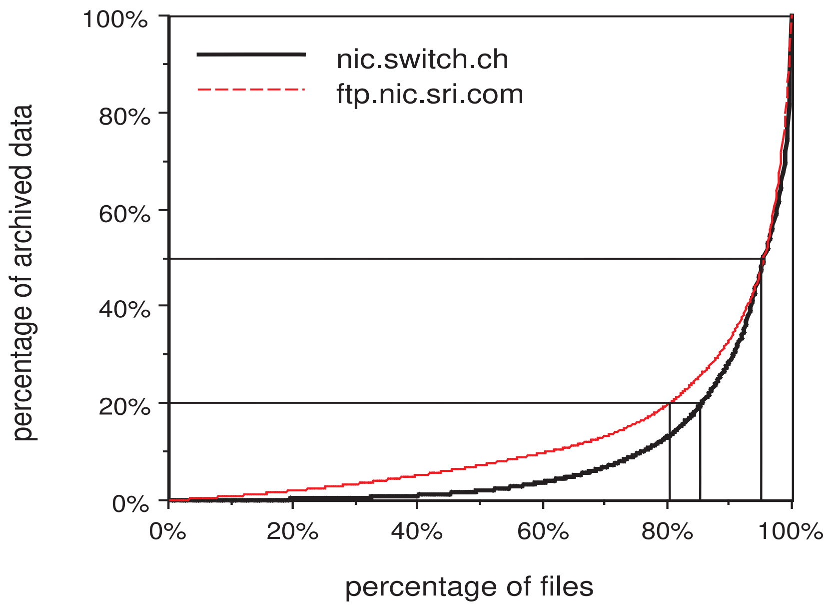

Figure 5: Allotment of file-data. Cumulated presentation for archive nic.switch.ch as seen on June 9, 1992.

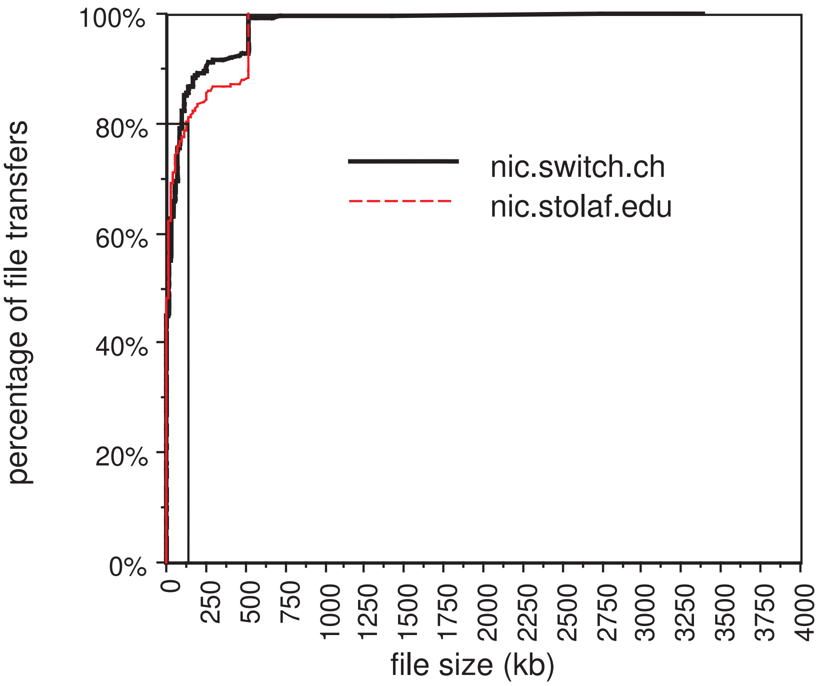

Figure 6: Histogram for files with length $\leq$ 600kb for archive nic.switch.ch. As seen on June 9, 1992.

References

[1] Baker, M. G., et al. Measurements of a Distributed File System. In *Proceedings of the 13th ACM Symposium on Operating System Principles* (Pacific Grove, CA, Oct. 1991).

[2] Cate, V. Alex – A Global File System. In *Proceedings of the USENIX File System Workshop* (Ann Arbor, Michigan, May 1992), pp. 1–11.

[3] Emtage, A., and Deutsch, P. archie – An Electronic Directory Service for the Internet. In *Conference Proceedings Usenix* (San Francisco, CA, Jan. 1992).

[4] Ewing, D. J., Hall, R. S., and Schwartz, M. F. A Measurement Study of Internet File Transfer Traffic. Tech. Rep. CU-CS-571-92, University of Colorado, Jan. 1992.

[5] Ganatra, N. K. Census: Collecting Host Information on a Wide Area Network. Master's thesis, University of California, Santa Cruz, June 1992.

[6] Lindgreen, M. R. EUnet Distributed Archive: Specification 1992 (Draft), 1992.

[7] Lottor, M. K. Internet Growth (1981–1991). RFC 1296, SRI International, Jan. 1992.

[8] Maffeis, S., and Cap, C. H. Replication Heuristics and Polling Algorithms for Object Replication and a Replicating File Transfer Protocol. Tech. Rep. IFI-92.06, CS Dept. University of Zurich, June 1992.

[9] McLoughlin, L., et al. *Mirror – Mirror Packages on Remote Sites.* (UNIX man–pages).

[10] Postel, J., and Reynolds, J. File Transfer Protocol (FTP). RFC 959, USC Information Science Institute, Oct. 1985.

[11] Schwartz, M. F., et al. Supporting Resource Discovery Among Public Internet Archives Using a Spectrum of Information Quality. Tech. Rep. CU-CS-487-90, University of Colorado, Sept. 1990.

[12] Wall, L., and Schwartz, R. L. *Programming perl.* A nutshell handbook. O'Reilly Sebastopol, 1991.