Long-term File Activity Patterns in a Unix Workstation Environment

Tim Gibson

University of Maryland - Baltimore County

Baltimore, MD

26 March 1998

Overview

- Methodology

- Tools

- Results
  - Daily
  - Long-term

- Conclusions

Methodology

- Traces at UMBC Computer Science Department
- Trace length, 287 days
- Four file systems
  - 7.5 GB
  - 230,000 files
  - ~500 users
- Traces collected for tertiary storage migration simulator.

Long-Term File Reference Patterns - Gibson & Miller

Tools - Overview

- Traces collected daily at 1:00 AM
- Traces collected with a modified "find" program
- Traces are based on physical inode number
- Full filenames and pathnames are retained
  - path and file name components can be optionally scrambled with MD-5 for privacy.
- Daily traces are "differenced"
- The new "difference" file only has entries for those files which were altered.

Tools - Advantages and Disadvantages

- Advantages
  - Relatively fast (traces 300 files per second, "differences" 15,000 files per second)
  - Low long-term data storage requirements
  - No database to update during trace collection or "difference generation"
  - Easily portable, all programs are standard C++
- Disadvantages
  - misses short-term activity
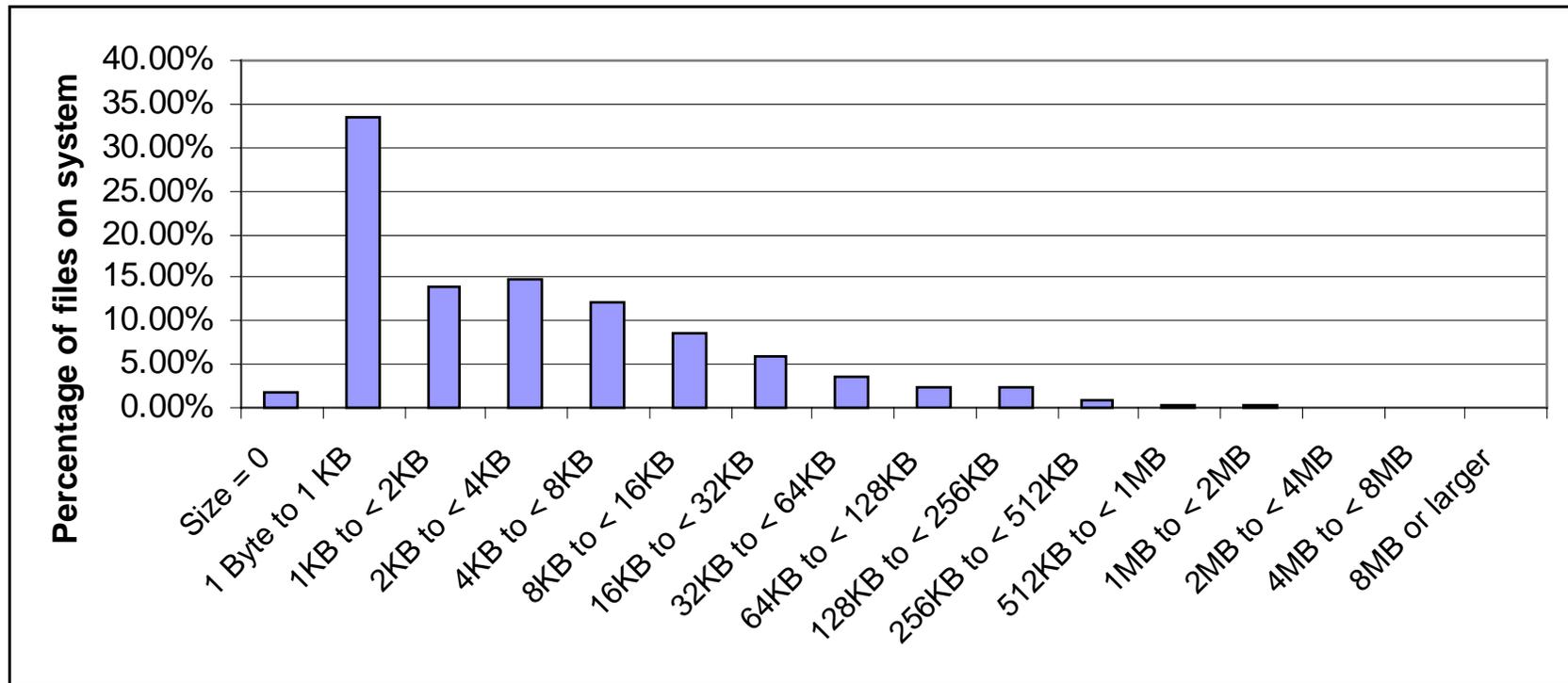 - misses number of times a file in accessed or modified, only notices that is was accessed or modified

Overall file system activity
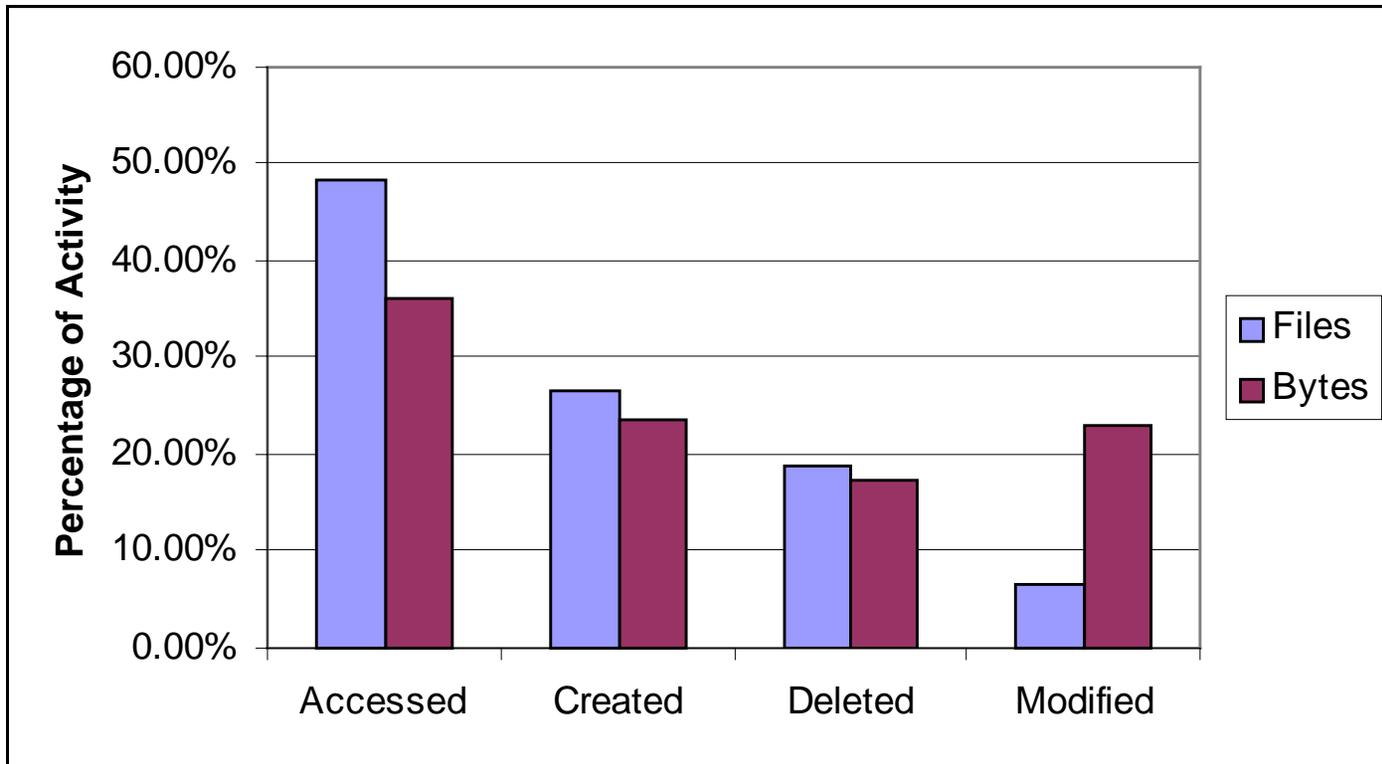
Average daily system activity
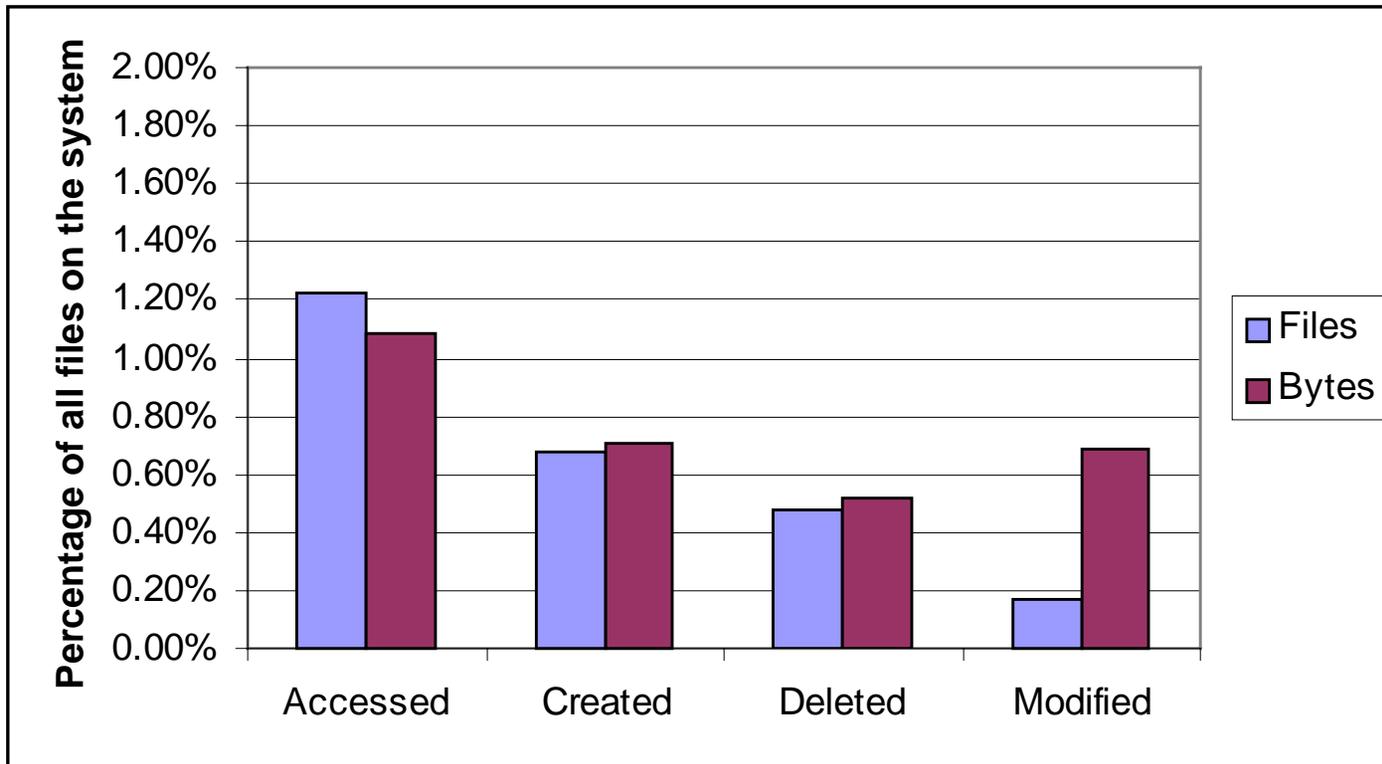
Size distribution of files on system

Long-Term File Reference Patterns - Gibson & Miller

Summary of modification observations

- Few files are modified
- Modifications equally distributed between
  - file remaining same size
  - file growing
  - file shrinking
- When growing, files increase by less than 32 KB 93% of the time
- When shrinking, 65% of all reductions are 512 bytes of less.

Files "used"
daily versus long-term

Long-Term File Reference Patterns - Gibson & Miller
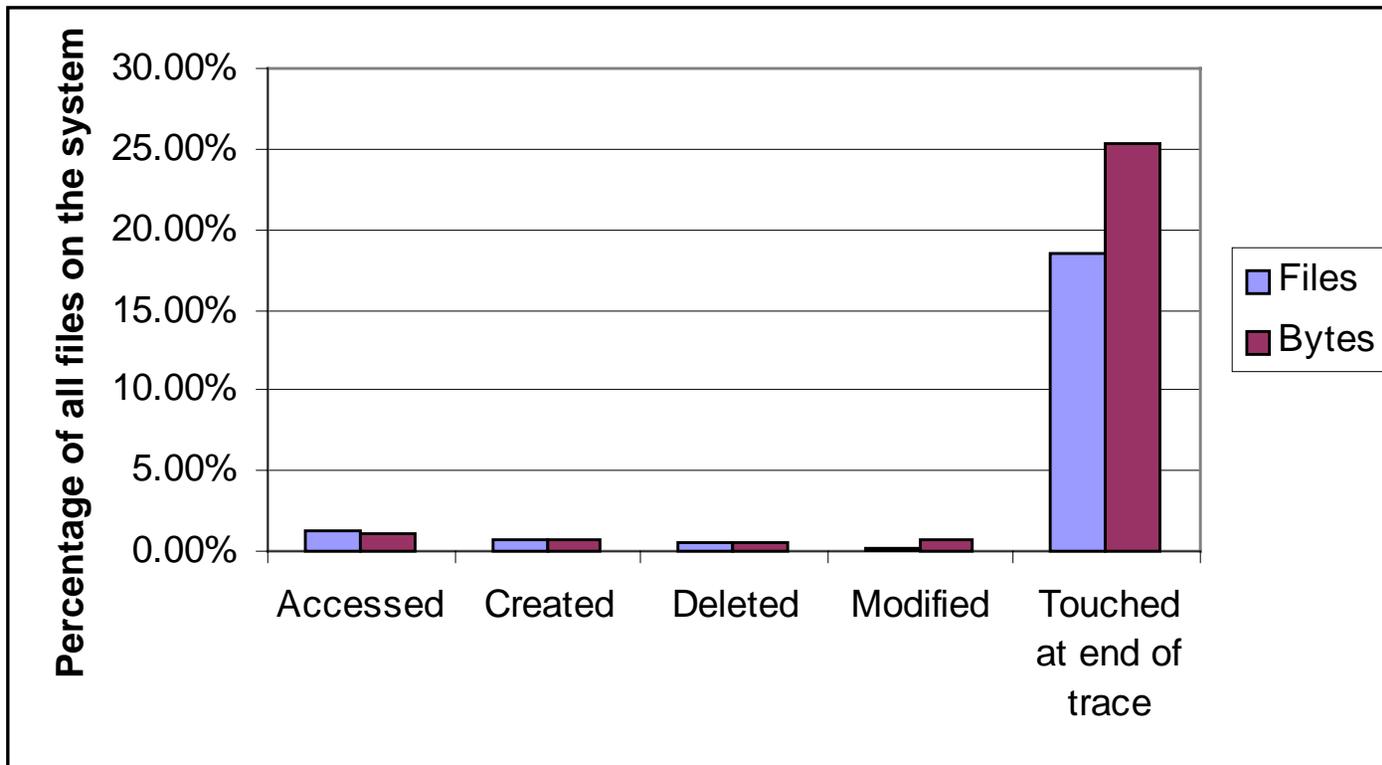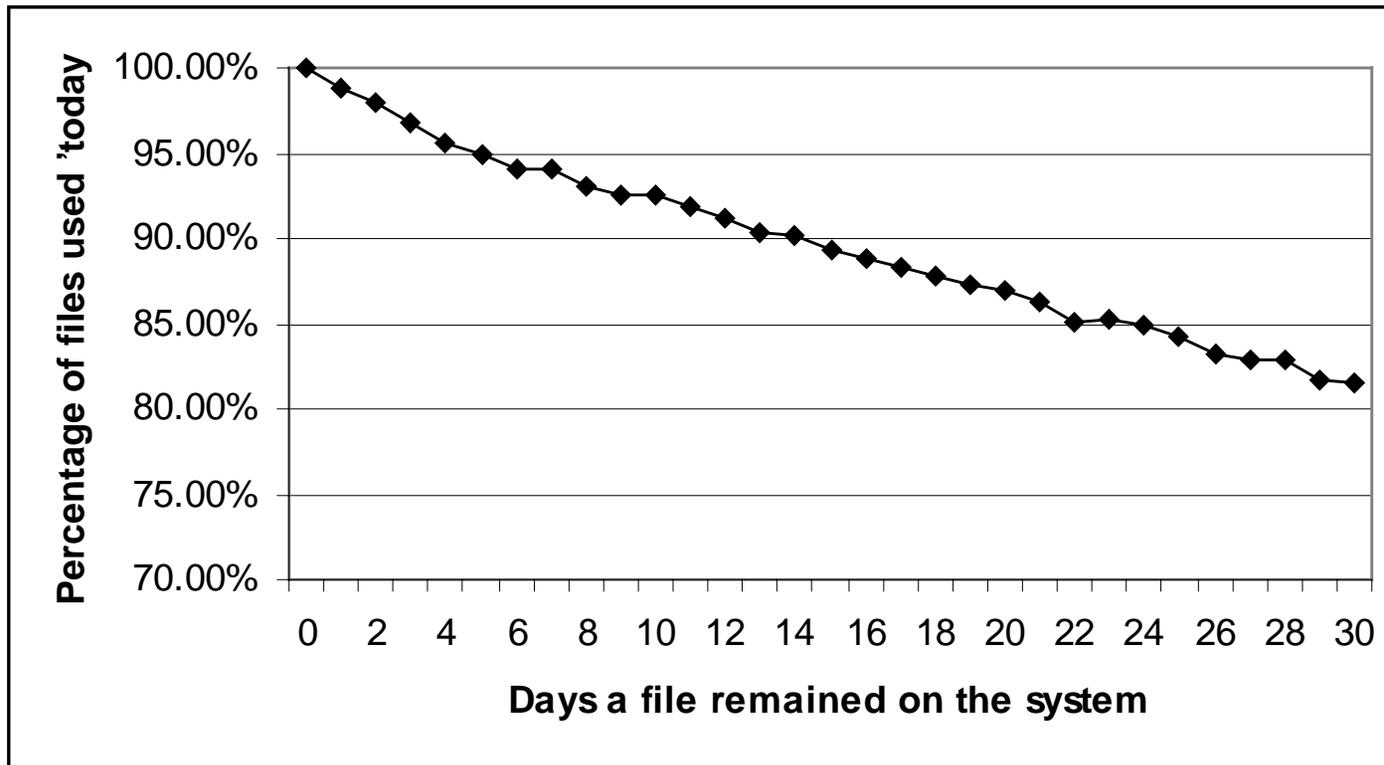
File reference locality (1)
files accessed or modified and remaining on system for 30 days (average)

File reference locality (2)

*deletion rate for "used" files as a percentage
of all files with the same lifetime until deletion*
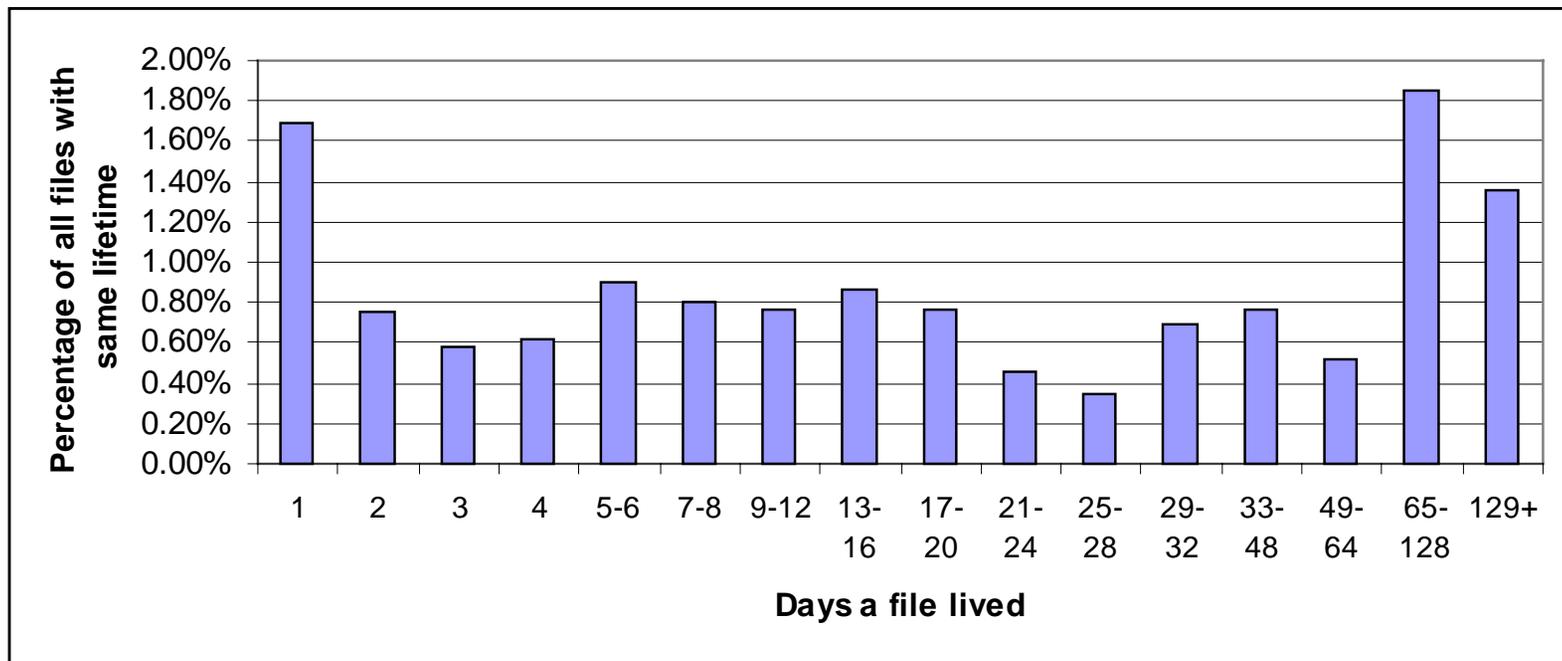
File reference locality (3)
Inter-reference periods
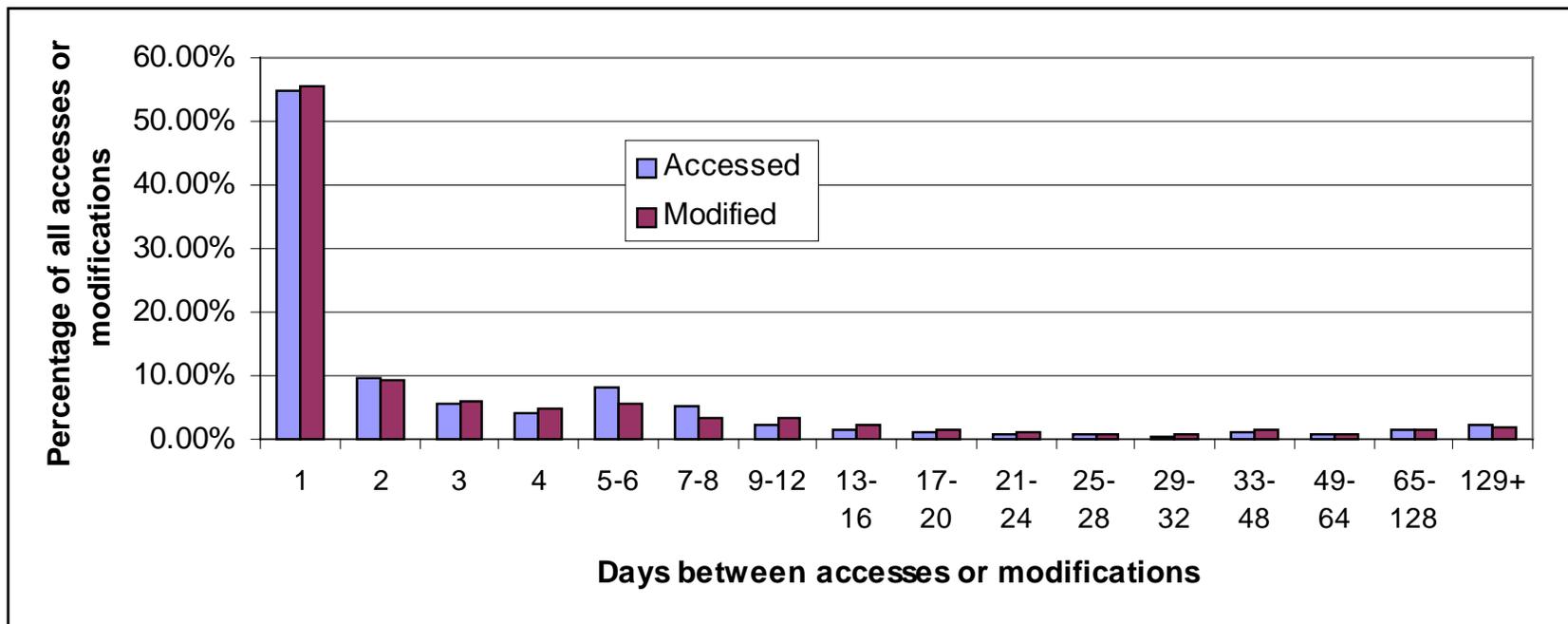
File reference locality (4)
times files were accessed or modified as a percentage of all files created

Long-Term File Reference Patterns - Gibson & Miller
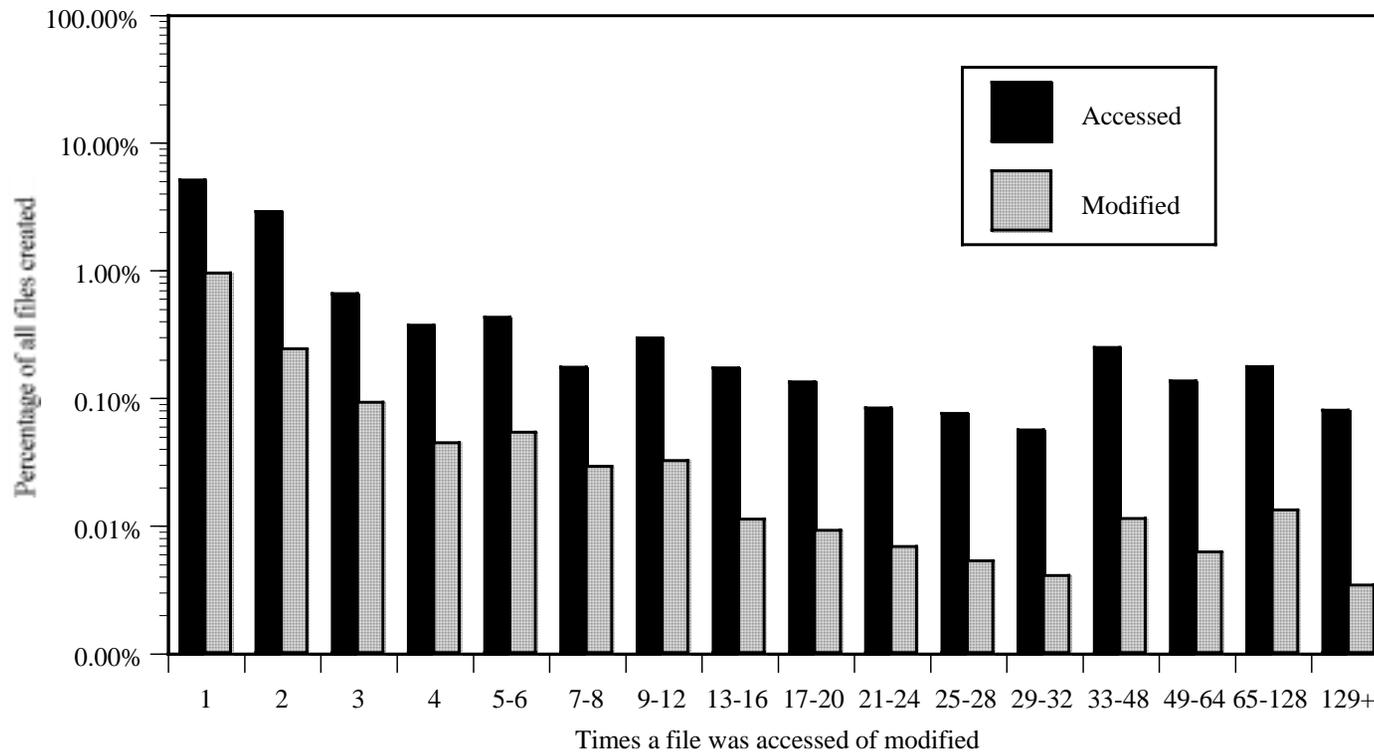
File reference locality (5)
repeated file accesses as a percentage of files accessed "today"
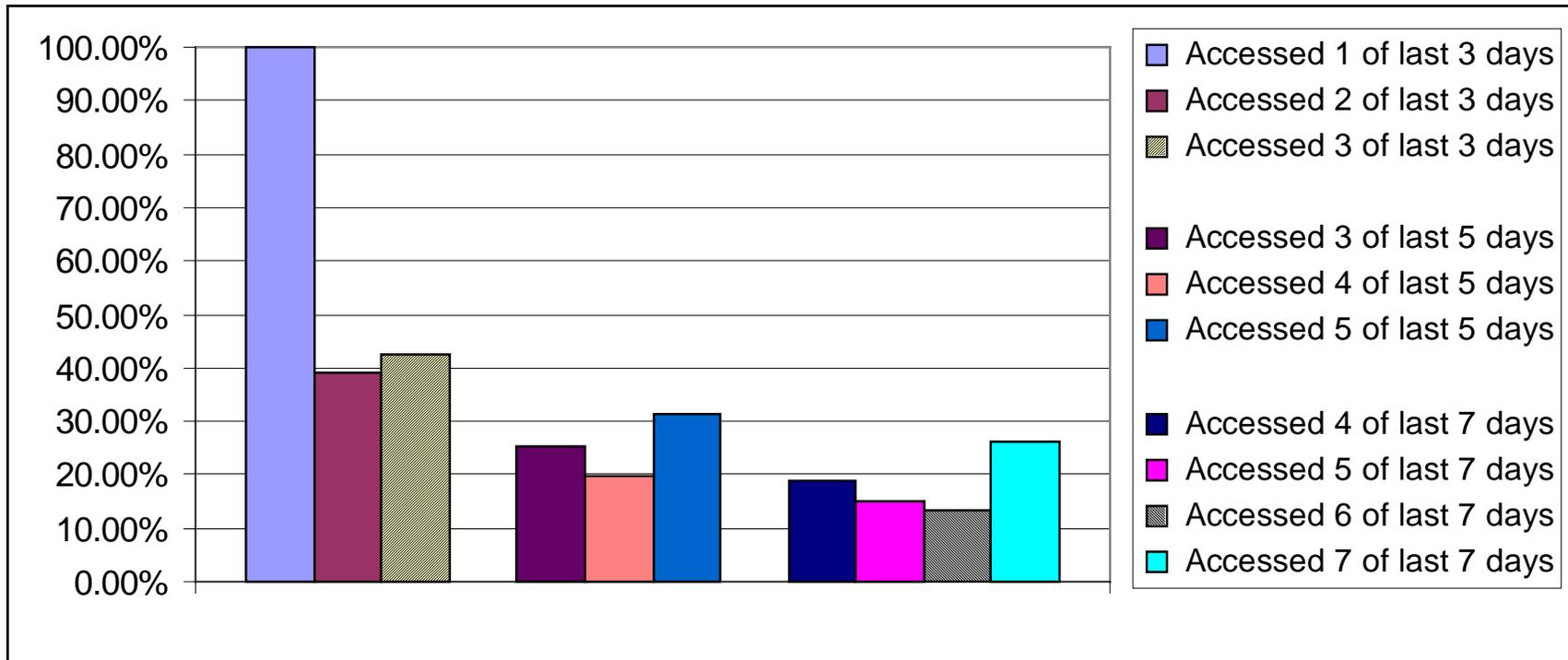
Conclusions

- Most Unix workstation files are small, less than 8 KB

- Accesses are most common transaction

- Most files are never used (*i.e.*, accessed or modified)

- Files that are used exhibit reference locality because:

  - they are used repeatedly until they are either deleted or lapse out of use

  - are deleted at a significantly lower rate than files which are not used

Future Research and Preliminary Results

- Finish tertiary storage simulator to evaluate different hardware combinations
- Collect data from other operating systems

- Completed simulator for different migration algorithms
- Developed an easily implementable migration algorithm based on file activity that provides an order of magnitude improvement in the file miss rate
- Examined the relationship between file system activity at the inode versus filename level
- Proved self-similarity in short-term file system activity

Long-Term File Reference Patterns - Gibson & Miller

Contacting Us

Tim Gibson

Lieutenant-Colonel, U.S. Army

tgibso2@cs.umbc.edu

URL=http://www.cs.umbc.edu/~tgibso2

Dr. Ethan Miller

Assistant Professor, Department of CS & EE

University of Maryland - Baltimore County

elm@cs.umbc.edu

URL=http://www.cs.umbc.edu/~elm

Backup slides follow

The Process

```
┌─────────────────────┐                                    ┌──────────────────────────┐
│  Raw File System    │                                    │        Rebuild           │
└─────────────────────┘                                    └──────────────────────────┘
     │            │
     │  Ftrace    │
     │  Sort      │
     ▼            ▼
┌─────────────────────┐     ┌──────────────────────────┐
│ Sorted Trace File   │ ──▶ │ Two Sorted Trace Files   │
└─────────────────────┘     └──────────────────────────┘
                                 │           │
                                 │  Fdiff    │
                                 │  Sort     │
                                 ▼           ▼
                            ┌──────────────────────────┐
                            │    Sorted Fdiff File     │
                            └──────────────────────────┘
                                        │
                                        ▼
                            ┌──────────────────────────┐
                            │        Simulator         │
                            └──────────────────────────┘
```

Long-Term File Reference Patterns - Gibson & Miller
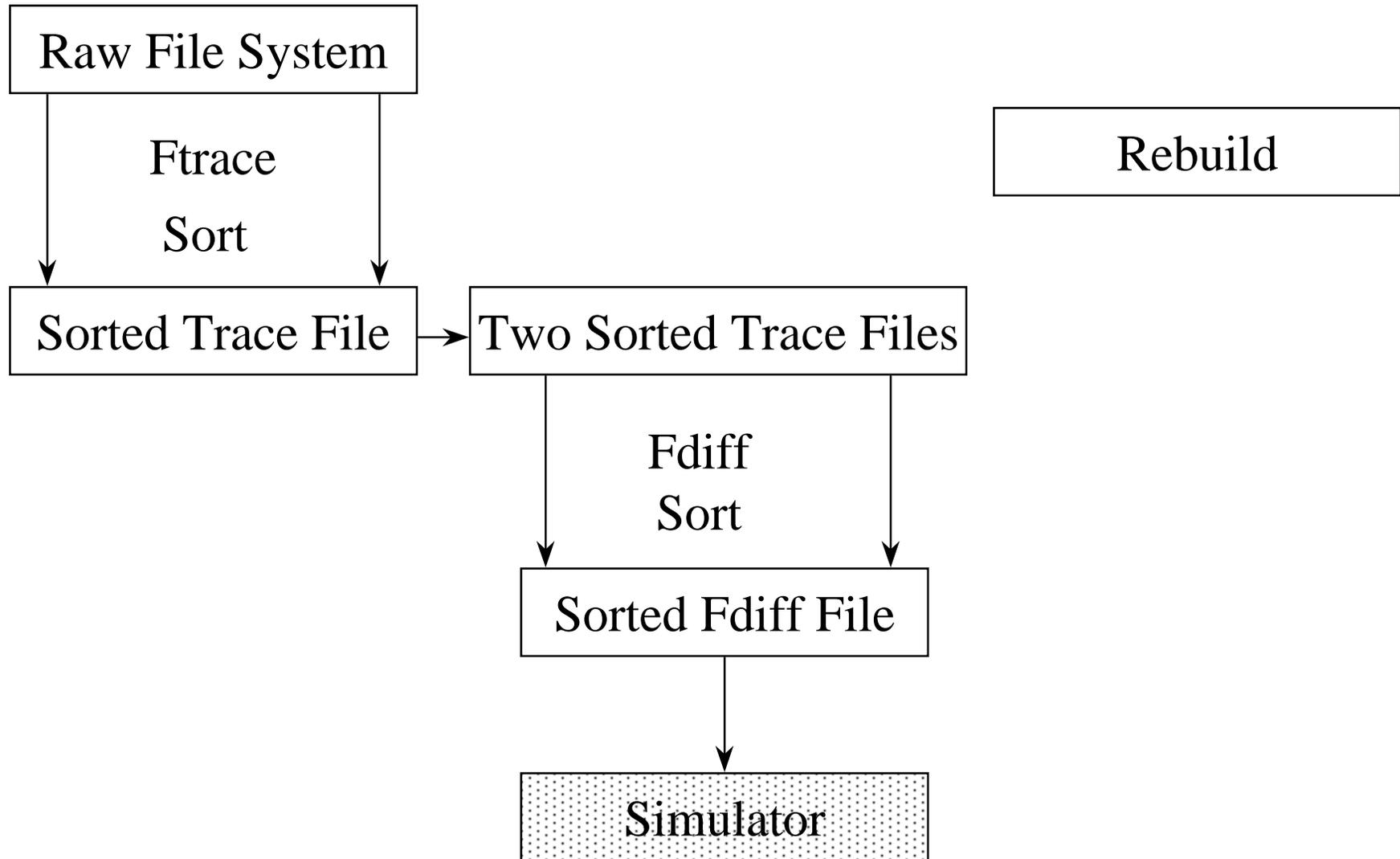
File System Changes Tracked by Fdiff

- Access *
- Creation *
- Deletion with Inode Reuse *
- Deletion w/o Inode Reuse *
- Modified & Increased *
- Modified & Equal Size *
- Modified & Decreased *

- Ctime Only *
- Change of owner
- Change of Group
- Name Add *
- Name Delete *

*Tracks count, sum (bytes), average size, max size, min size, stdev

Long-Term File Reference Patterns - Gibson & Miller

Composition of a **ftrace** Record

69948 43c 2 2 32a48fe8 3284d07f 3284d07f /usr/home/tgibson/ftrace/lib/Makefile

Index Node (I-node)	69948
Size	43c
User Number	2
Group Number	2
Access Time	32a48fe8
Ctime	3284d07f
Modify Time	3284d07f
Name	/usr/home/tgibson/ftrace/lib/Makefile

| /usr	/home	/tgibson
/n2pFhKkzgNY	/C+370VWI1EA	/8vzDuNIoWFI/

fdiff Format

N 0 0 845917257 845918025 /usr/home/tgibson/ftrace/junk/test1.srt /usr/home/tgibson/ftrace/junk/test2.srt
A 70014 849646632 3284d4e9 10ab
A 70015 849646632 3284d4ea 102
A 70016 849646632 3284d4eb 54d
A 70017 849646632 3284d4ed 755
A 70018 849646632 3284d4ee a1
A 70019 849646632 3284d4ee a1
E 70034 849646632 32a49429 32a49428 e818
A 77984 849646622 324023ea 4893
A 78001 849646633 32a49303 202
E 78002 849646631 32a49429 32a49427 466c
A 78004 849646622 3284d54a 2f6
S 78041 849646638 32a49360 32a4942e 2000
L 78061 849646576 32a493af 32a493f0 3c58
D 78062 845917971 bc53 /n2pFhKkzgNY/C+370VWI1EA/8vzDuNIoWFI/JNsBhKWReJ4/SRXVxNgREKU
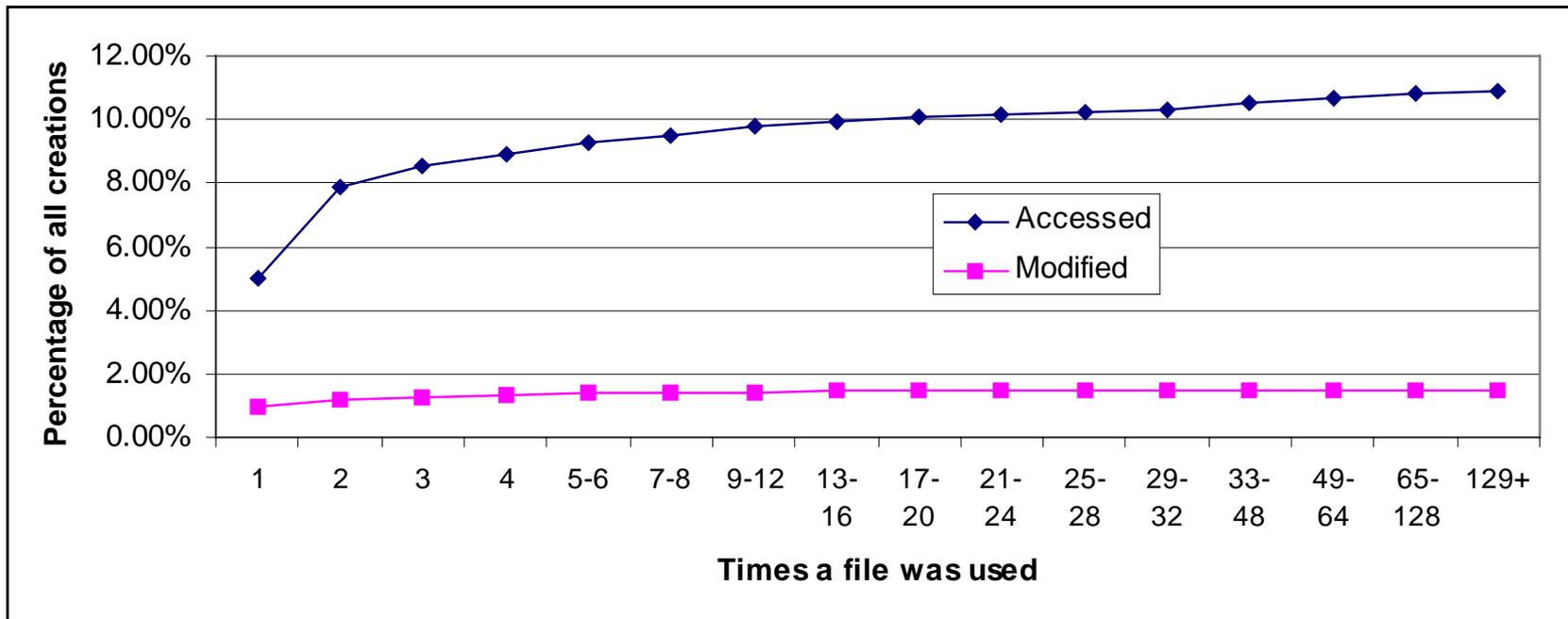D 78065 845917658 3acb /n2pFhKkzgNY/C+370VWI1EA/8vzDuNIoWFI/JNsBhKWReJ4/7JlkA7FnuEo

File lifetimes
"used" versus unused cumulatively scaled
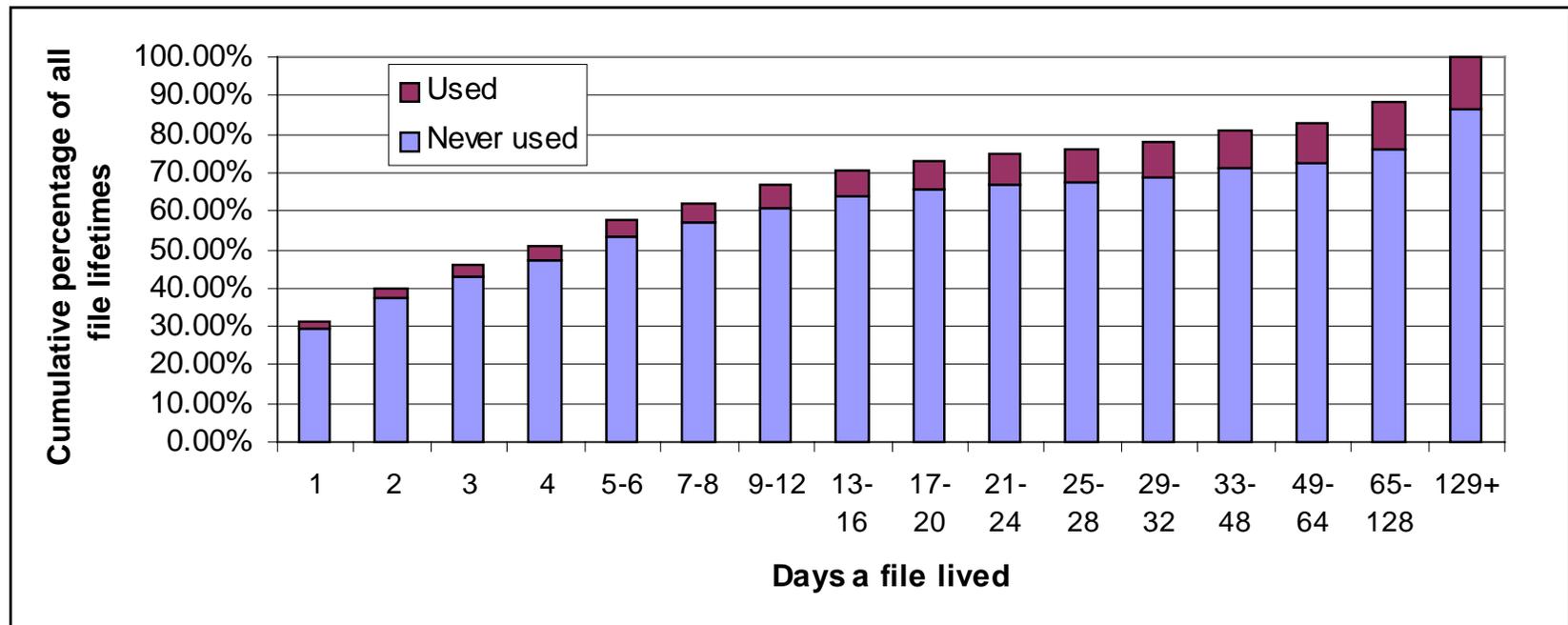
File reference locality (6)

*times **all** files were accessed or modified as a cumulative percentage of all files created*

Systems Studied

	Aberdeen Proving Ground	University Computing Services	Computer Science Dept. (Long Period)	Computer Science Dept. (Short Period)	Univ. of CA
Files on System (average)	72,000	1,320,000	230,000	690,000	2,300,000
Capacity (avg. percent full)	3.6 GB (80%)	35 GB (70%)	11 GB (70%)	28 GB (65%)	84 GB
FS type	User and System	User only	User only	User and System	User and System
FS traced (#)	7 user, 1 system	9 user	4 user	6 user, 2 system	49 mixed
System Type	Admin.	Mixed	Development	Development	Development
Number of Users	300 full-time	2,000 email	~15,000	~500	~500
Trace Length	210 days	239 days	287 days	157 days	207 days

Long-Term File Reference Patterns - Gibson & Miller