

A Pointer-Based File Management System to Reduce Redundancy and Storage Overhead

Braden D. Licastro

Allegheny College, USA

First Annual Computer Science 580 Research Symposium

Tuesday, April 23, 2013



ALLEGHENY COLLEGE

What is a file system?

File System (N.)

A type of data store which can be used to store, retrieve and update a set of files.

Name ▲	Size	Type:	File Folder
99998.txt	1 KB	Location:	C:\
99999.txt	1 KB	Size:	488 KB (500,059 bytes)
100000.txt	1 KB	Size on disk:	390 MB (409,608,192 bytes)
mkfile.bat	1 KB	Contains:	100,002 Files, 0 Folders
source.txt	1 KB		

Common file systems

Common file systems:

- ntfs
- exFAT
- fat
- fat32
- ext3

Common file systems

Common file systems:

- ntfs
- exFAT
- fat
- fat32
- ext3

Why so many? There is no "perfect" file system. Each system has features tailored to a specific need.

Common file systems

Common file systems:

- ntfs
- exFAT
- fat
- fat32
- ext3

Why so many? There is no "perfect" file system. Each system has features tailored to a specific need.

Lamborghini file system impractical

Venti File System



Plan 9 Operating System

Venti File System

Plan 9 Operating System

- Uses Venti file system
- Venti monitors block level storage
- Checks parts of files for similarities

Venti File System

```

/% exit

-----

Preparing menu...
The following tasks are done:
  configfs - choose the type of file system to install
  partdisk - edit partition tables (e.g., to create a plan 9 partition)
  prepdisk - subdivide plan 9 disk partition
  fmtfossil - initialize disks for a fossil server
  mountfs - choose the type of file system to install
  configdist - configure the distribution
  mountdist - locate and mount the distribution

The following unfinished tasks are ready to be done:
  copydist - copy the distribution into the file system

Task to do [copydist]:

-----

I

```

Installing file system 3%

gnot	prompt: fsys main create /active/dist/replica sys sys d775
l 3031	prompt: fsys main create /active/dist/replica/client sys sys d775
m	prompt: fsys main create /active/dist/replica/client/plan9.db
i	sys sys 664
s 1375	prompt: fsys main create /active/dist/replica/client/plan9.log
c	sys sys a664
e	prompt: % 9660srv
	% mount /srv/9660 /n/distmedia /dev/sdD0/data

Published Papers

A fast filtering scheme for large database cleansing.

- Database focused algorithms
- Data redundancy reduction
- Matches data, not files
- Completely removes duplicates

Table 1: Four records in the same window

Record	Name	Gender	Dept.
<i>A</i>	li zhao	M	CS
<i>B</i>	li zhai	M	CS
<i>C</i>	li zhao	M	CS
<i>D</i>	sun peng	M	CS

Published Papers

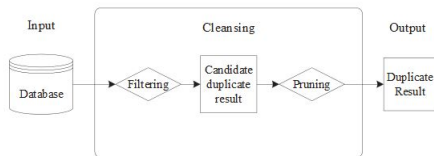
A data de-duplication access framework for solid state drives.

- Tuned for solid state drives
- Strives to reduce file storage overhead
- Minimal read and write operations needed
- Tested on large computer cluster
- Finds candidate duplicates from calculated score

Published Papers

A data de-duplication access framework for solid state drives.

- Tuned for solid state drives
- Strives to reduce file storage overhead
- Minimal read and write operations needed
- Tested on large computer cluster
- Finds candidate duplicates from calculated score



Implementation

Pointer based file system implementation

- Create or modify existing system
- Hash entire file to show matches
- Reduce duplicates to one file, replace removed with pointer
- Automatically overwrite pointers if file is modified
- Track all non-system files in simple database

Implementation

Pointer based file system implementation

- Create or modify existing system
- Hash entire file to show matches
- Reduce duplicates to one file, replace removed with pointer
- Automatically overwrite pointers if file is modified
- Track all non-system files in simple database

Files	
PK	Field
PK	id
	FileHash
	FilePath

Design Challenges

Design Challenges

- Address storage costs over benefit
- Minimal performance impact
- Initial setup feasibility
- Implementing new system vs building on existing

Technical Challenges

Technical Challenges

- Testing results
- Bench-marking performance
- Comparison with other popular file systems
- Determining optimal database for this application

End