538W: Data at Scale

An I/O Patterns and Measurement Study for User Applications

Nodir Kodirov   Omar AlOmeir

Dec. 11, 2013

Introduction

- Inspired by A File is not a File paper

- Important to file system designers and application developers

- Conducted experiments that are focused on IO patterns

- The scripts from A File is not a File are not compatible, old, and produce too much data

- Developed our own tools

Scope

- Replicate results from A File is not a File.
  - DTrace and AppleScript

- Go further and look at browsers (based on popular class request).

Start - Simple C

```c
#include <stdio.h>
#include <stdlib.h>

int main(int argc, char *argv[])
{
    FILE *file_to_open = fopen("test.txt", "ab+");

    if(file_to_open != NULL)
        printf("Successfully opened file.\n");
    else
        printf("Could not open a file.\n");

    fprintf(file_to_open, "text we want to write");
    fclose(file_to_open);

    return 0;
}
```

Text Editor Workloads

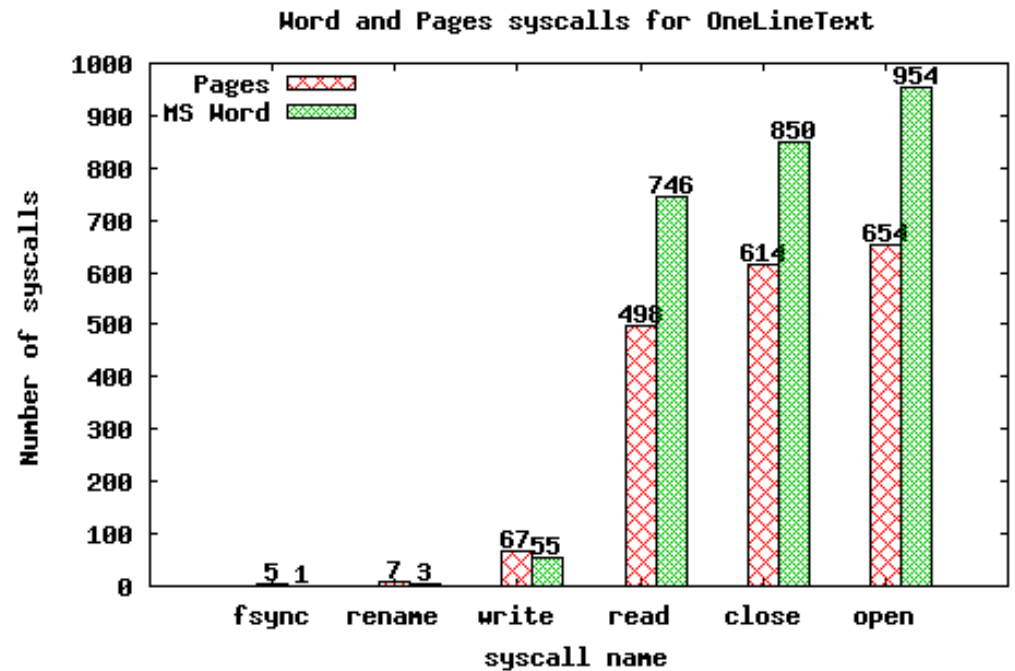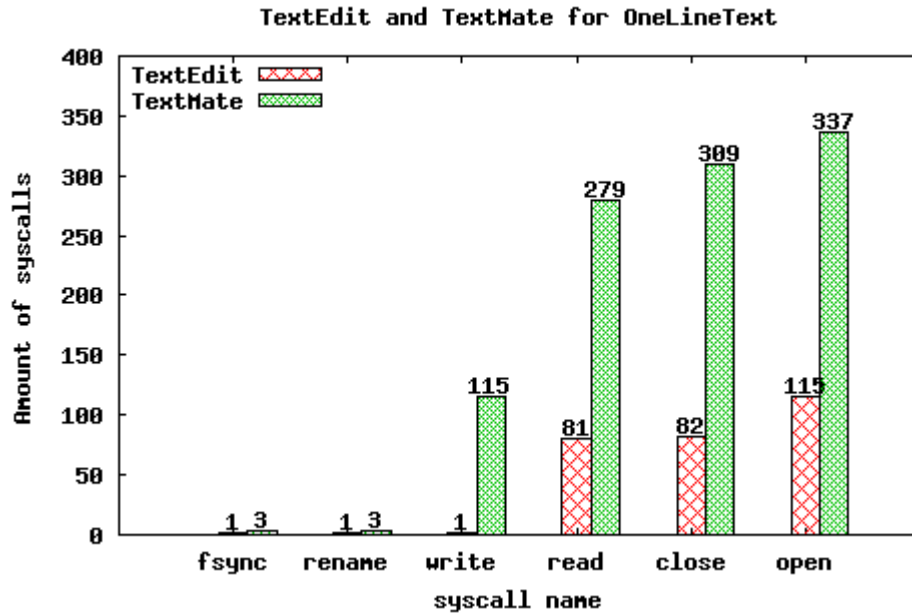| Application	One Line of Text	500 Lines of Text	5 JPGs (~2 MB)
TextEdit/ TextMate	Create a file Write 1 line of text Save as .txt Quit	Create a file Write 500 line of text save as .txt quit	———
Pages	Create a file Write 1 line of text Save as .txt Quit	Create a file Write 500 line of text save as .txt quit	Create a file Insert 5 JPG images save as .docx quit
Microsoft Word	Create a file Write 1 line of text Save as .txt Quit	Create a file Write 500 line of text save as .txt quit	Create a file Insert 5 JPG images save as .docx quit

Sample text: The quick brown fox jumps over the lazy dog.

Browser Workloads

| Application	Hello World Page	Hello with JPG Page (21 KB)
Safari	Launch Open a "Hello World" HTML page Quit	Launch Open an HTML page with a JPG image Quit
Google Chrome	Launch Open a "Hello World" HTML page Quit	Launch Open an HTML page with a JPG image Quit

Demo

OneLineText

TextEdit and TextMate for OneLineText

Amount of syscalls

syscall name	TextEdit	TextMate
fsync	1	3
rename	1	3
write	1	115
read	81	279
close	82	309
open	115	337

Word and Pages syscalls for OneLineText

Number of syscalls

syscall name	Pages	MS Word
fsync	5	1
rename	7	3
write	67	55
read	498	746
close	614	850
open	654	954

4KB threshold

			< 4KB	> 4 KB
500Line	Word	Read	92	654
		Write	0	56
	Pages	Read	184	316
		Write	1	86
5JPGs	Word	Read	127	657
		Write	0	58
	Pages	Read	223	399
		Write	57	177

Browsers
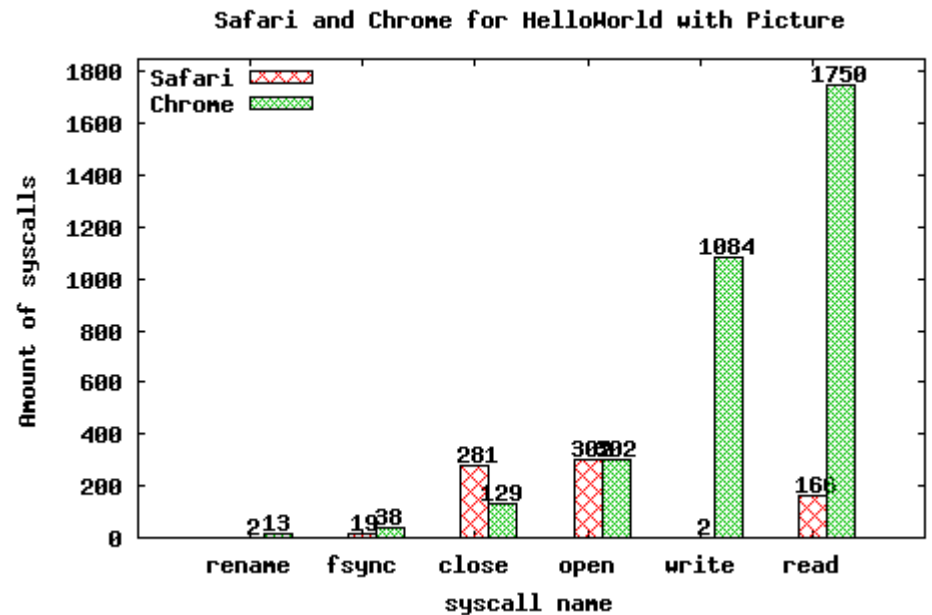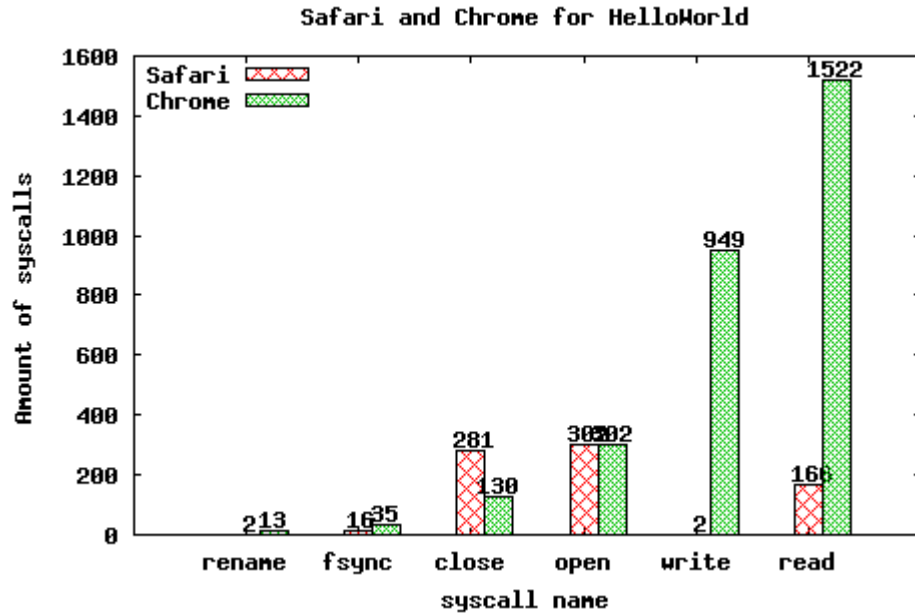
Future work

- Virtual Machine environments
  - how many of those syscalls actually hit the disk
- Investigate browsers
  - Checked random Wiki page, a lot of syscalls, not much meaningful
  - impact of JavaScript to execution, more workloads;
- Variance mining
  - study app logged behavior with inferred models
  - got a grasp, but need to run more experiments

Conclusions

- Replicated the results - not quite same

- File systems expose new interfaces

  - applications express their exact needs to FS

- We learned

  - DTrace, AppleScript, gnuplot, Synoptic

- We had fun :-)