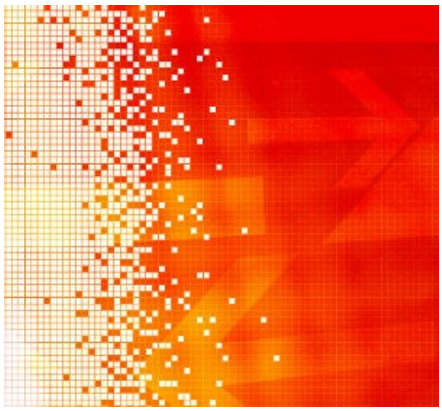
POSIX IO Must Die!

POSIX IO is becoming a serious impediment to IO performance and scaling. POSIX is one of the standards that enabled portable programs and POSIX IO is the portion of the standard surrounding IO. But as the world of storage evolves with greatly increasing capacities and greatly increasing performance, it is time for POSIX IO to evolve or die.

By Jeffrey B. Layton
Tuesday, March 2nd, 2010



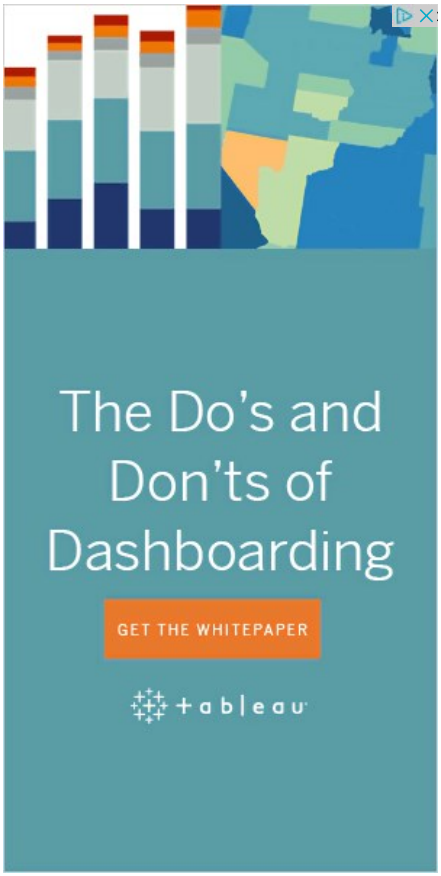
POSIX (Portable Operating System Interface for Unix) is a set of standards that define the application programming interface (API) as well as some shell and utility interfaces. It was developed primarily for *nix operating systems but actually any operating system can utilize the standards. POSIX IO (not an official name) is the portion of the standard that defines the IO interface for POSIX compliant applications. Function such as read(), write(), open(), close(), lseek(), fwrite(), fread(), and so on, are defined including their errors. But these definitions were first codified in 1988 – 22 years ago!

During this time storage has changed dramatically. We now have thousands of systems with performance in the TeraFLOPS range including some people who have clusters of this size in their homes (nudge, nudge, wink, wink) and there are PetaFLOPS systems at strategic sites today. These systems can have hundreds of thousands of processors with a large percentage perhaps performing IO. That means there is the potential for a great deal of IO happening at the same time, usually to a single shared file system, including the possibility of a large number of nodes all writing to the same file.

Sitting in the middle of this is POSIX with an interface that has not appreciably changed in 22 years! For several years people have been asking for changes or relaxation of POSIX standards to improve IO performance. The reasons for these requests are fairly simple – to improve the IO performance of applications. Concurrently a change or relaxation in POSIX IO could the development of new storage mechanisms to improve not only application performance but management, reliability, portability, and scalability.

POSIX

POSIX is one of the big reasons that the world of *nix allows you to take a program from one operating system to another operating system that are both POSIX compliant. You are not limited to *nix operating systems since the POSIX standard is open to anyone.



Software

Stick a Fork in Flock: Why it Failed
CentOS 5.6 Finally Arrives: Is It Suitable for Business Use?
Rooting a Nook Color: Is it Worth It?

System Administration

Scripting, Part Two: Looping for Fun and Profit
Command Line Magic: Scripting, Part One
Making the Evolutionary Leap from Meerkat to Narwhal

Storage

Extended File Attributes Rock!
Checksumming Files to Find Bit-Rot
What's an inode?

The original POSIX standard was developed by IEEE and was labeled as “IEEE Std 1003.1-1988.” Prior to 1997 it had several sections including:

- POSIX.1: Core Services (This includes IO port interface and control)
- POSIX.1b: Real-time extensions (IEEE Std 1003.1b-1993)
- POSIX.1c: Threads extensions (IEEE Std 1003.1c-1995)
- POSIX.2: Shell and Utilities (IEEE Std 1003.2-1992)

After 1997 the Austin Group which is a combination of the Open Group and the international ISO group has been responsible for the reorganization of the POSIX standard as well as revisions to the standard. Consequently, the POSIX standard is an international standard.

While the title of the article says that “POSIX IO Must Die” POSIX is a very important standard. It defines much of the general behavior we have come to know in our Linux systems. It also allows us to take programs written for Linux and run them on AIX, HP-UX, BSD, and even Mac OS X (this assumes that all dependent libraries are available on these systems but that’s outside the scope of POSIX). It allows the world to write standard libraries that use POSIX interfaces and make them available for applications. Without POSIX writing applications would be much more difficult.

For those reading this article that are a bit younger probably don’t remember the days when there was really no standard and writing programs for different operating systems was a very difficult process. As a famous person once described it, “… cats and dogs living together! Mass hysteria!…” Taking a program written on a VAX with VMS and then running it on a Unix system (a sane decision if you ask me) was problematic because of the lack of common interface standards. I remember writing an application on a VAX system in graduate school and then running it on a larger *nix based system because it was faster. I spent a great deal of time bugging the system support staff about porting simple routines because of the lack of POSIX compatibility between the two operating systems.

At the same time, the POSIX standard, while evolving, is 22 years old! (I love the rule of three). POSIX has become this extremely large cruise ship that people love to travel upon because the food and the entertainment are always consistent and well defined. However if the food or the entertainment on the cruise ship aren’t to your liking or are preventing you from really having a good time, then it can seem quite limiting. This is exactly the case with POSIX IO for applications and organizations wanting high performance IO.

Changes for Better Performance

As systems started to scale to large numbers of processors and larger problems were tackled, it was soon realized that storage systems were becoming bottlenecks. However, the problem didn’t necessarily lie with the file system but with the standards for interfacing the applications with the storage. This was particularly noticed for applications where there were many “writers” to a common shared file system.

A few years ago, a sub-group of the Open Group was created called the High End Computing Extensions Working Group (HECEWG). The goal of this group was to create a set of extensions or relaxations to POSIX that allowed applications to basically have better IO performance including better scaling. The business case for this is presented in, “A Business Case for Extensions to the POSIX I/O API for High End, Clustered, and Highly Concurrent Computing”. The group came up with a few proposals for changes for the Open group that can be summarized as:

- Allowing changes to the *stat()* function to dramatically improve performance when discovering information about the files in a file system
- Opening a large number of files using a shared file system
- Opening a single file from a large number of nodes on a shared file system
- Creating a list of IO functions that you can send to the file system for fulfillment (reduces the number of individual IO operations)

Another document co-authored by several members of the HECEWG, gives a longer list of changes and efforts surrounding the need for improved IO performance. From the document, **Relaxation of POSIX Semantics for parallelism**

- “Scalable metadata operations in a single directory”
- “NFSv4 security and the pNFS effort to allow NFS to get more native file system performance through separation of data and control which enables parallelism”
- “I/O Middleware enhancements to enable dealing with small, overlapped, and unaligned I/O”
- “Tighter integration between high level I/O libraries and I/O middleware”

These proposed changes are very important even if you are not an HPC user. Let’s look at the first item to explain why.

Metadata Operations

The first proposed change, scalable metadata operations in a single directory, affects a very large number of people, not just HPC. As an experiment, run the following command on your system in the root directory (“/”).

```
% time find . -type f | wc -l
```

This command will count all the files from the current directory (“.”) on down the tree. If you do this from the root you will get a count of all the files on your system. For my system, the result was the following:

```
606914
```


Mobile

Putting Text to Speech to Work

Look Who's Talking: Android Edition

Upgrading Android: A Guided Tour

HPC

A Little (q)bit of Quantum Computing

Emailing HPC

Chasing The Number

```
real    1m43.424s
user    0m0.796s
sys     0m3.024s
```

So it took almost 1.5 minutes to count all the files and there were 606,914 files on my home system. Just a few years ago this would have been perhaps 100,000 or so. Now imagine a single file system having to keep track of about a half a million files without making a mistake or having any corrupt data. This is just for a desktop.

In the HPC world there are applications that can produce millions of files in a single directory per node. Moreover there are file systems with well over 1 PB (Petabyte) of data and hundreds of applications running at the same time all producing data to a single shared file system. In the middle of this ballet a user runs the command, “ls -lsa” to see if the file that his application is writing to is changing size. For this command, the file system has to walk the entire directory tree. Then it has to read the metadata associated with the appropriate files of which several applications may be reading or writing at a particular time. Then the results are formatted and presented to the user. It can take a great deal of time to perform all of these operations using lots of CPU time and putting the file system under a great deal of stress. While these metadata operations are happening the storage system has to perform producing high levels of throughput and IOPS.

Comments on "POSIX IO Must Die!"

← Older Comments


What’s up, its nice article on the topic of media print, wwe all be familiar with media is a impressive source of facts.

Stop by my web blog; cheap car insurance in ga

March 23rd, 2016 at 11:55 pm


Hi there, I found your blog by the use of Google whilst searching for a comparable subject, your web site got here up,
it seems great. I’ve bookmarked it in my google bookmarks.

Hello there, just was alert to your blog via Google, and found that it’s really informative. I am gonna wath out for brussels. I will appreciate when you continue this
in future. A lot of folks will probably bee benefited from your
writing. Cheers!

Also visit my homespage – cheap car insurance texas

March 24th, 2016 at 8:18 am


Stunning quest there. What occurred after?

Good luck!

Here is myy website: cheap car insurance michigan

March 24th, 2016 at 6:06 pm


I have to thank you for the efforts you hafe
put in penning this blog. I really hope to check out the same high-grade content by you later
on as well. In fact, your creative writing abilities has inspired me tto get
my own, personal website now ;)

Here is my homepage – cheap car insurance florida

March 25th, 2016 at 7:58 am


I every time emailed this website post page to all my associates, as if like to read it then my friends will too.

My page … cheapest car insurance company

March 25th, 2016 at 12:11 pm

[redacted]

I have read sso many content regarding the blogger lovers however this paragraph is truly a good paragraph, keep it up.

Feel free to suf to my page cheap car insurance in michigan

March 25th, 2016 at 2:04 pm

[redacted]

Superb website youu hve here but I was curious about if you knew of any user discussion forums thnat cober the same topis talked about in this article?

I'd really loive to be a part of community where I ccan gget comments from other experienced people that share the same interest.
If you have any recommendations, please llet me know.
Manyy thanks!

Visit my web blog: Cheap Car Insurance

March 26th, 2016 at 11:14 am

[redacted]

Thank you for some other informative web site. The place else could I am getting that type of information written in suich a perfect way?
I have a project that I am simply now running on,and I've been at the look out for such info.

Here is my blog post; very cheap car insurance

March 26th, 2016 at 9:39 pm

[redacted]

Stop by my blog post … Jeremiah

March 28th, 2016 at 7:01 pm

[redacted]

Feel free to surf to my weblog … facebook.sercanbozkurt.com

March 28th, 2016 at 7:52 pm

[redacted]

Also visit my web page: hop over to here to pre-book erotic massage

March 29th, 2016 at 3:34 am

[redacted]

My webpage – raovat.Chopho.Vn

March 29th, 2016 at 5:09 am

[redacted]

This paragraph is in fact a pleasant one it assists new the web people, who are wishing in favor of blogging.

Also visit my page Cheap Car Insurance

March 29th, 2016 at 8:52 am

[redacted]

It's really very complicated in this busy life to listen news on TV, so I only use web for that reason, and obtain the most up-to-date news.

Also visit my homepage – cheap car insurance

April 1st, 2016 at 12:51 pm

[redacted]

Problems with blood clotting and regularly consume garlic is not for those who, even if only as garlic tablets, nausea, vomiting, use the
recommended blood-thinning medication. You will continue to experience all the eulogized features
of Branded Tablet PCs but without the exorbitant prices.
This tablet does not work well with nitrates or alpha-blockers, so avoid
Caverta if you are taking aforementioned medicines.

April 5th, 2016 at 3:24 am

[redacted]

[url=http://viagraonlineapotheke.com/cialis_generika.html]was ist cialis generika[/url]

In unserer Apotheke kann man Potenzmittel rezeptfrei kaufen und anonym bleiben. Wir stellen Ihnen ein reiches Sortiment an Präparaten und Arzneimitteln, die dabei helfen, den Männern die vergangene Festigkeit und Stärke wiederzubekommen. Mit unseren Präparaten werden Sie sich an die „alten Zeiten“ erinnern und alle Komplexe loswerden. Wir bieten Ihnen Viagra, Levitra, Cialis, Tadalafil, Sildenafil, Vardenafil, Dapoxetin an. Wir sind 24 Stunden am Tag erreichbar und Sie können eine Bestellung jederzeit aufnehmen. Sie können die bestellten Arzneimittel sowie bargeldlos, als auch mit Bargeld bezahlen.

[url=http://viagraonlineapotheke.com/super_kamagra.html]super kamagra[/url]

In unserer Internet-Apotheke kann man Präparate für Potenzsteigerung ohne Rezept kaufen, und zwar dabei anonym bleiben. Wir stellen Ihnen ein reiches Sortiment an Präparaten und Arzneimitteln, die dabei helfen, den Männern die vergangene Festigkeit und Stärke wiederzubekommen. Unsere Mittel helfen Ihnen dabei, sich an „die alten Zeiten“ zu erinnern und sich von den Komplexen zu befreien. Wir bieten Ihnen Viagra, Levitra, Cialis, Tadalafil, Sildenafil, Vardenafil, Dapoxetin an. Wir arbeiten rund um die Uhr und Sie können zu jeder Uhrzeit eine Bestellung aufnehmen. Die Zahlung von den Bestellungen ist in der Regel bargeldlos, aber Sie können auch mit Bargeld bezahlen.

April 5th, 2016 at 8:47 pm

[redacted]

For boosting your chances of gaining you should choose this solution to make sure that you can be in win-win situation. how to get usa targeted facebook likes

April 6th, 2016 at 11:48 pm

[redacted]

Great, thanks for sharing this post.Really thank you! Much obliged.

April 8th, 2016 at 3:35 am

[redacted]

[url=http://canadianpharcharmyonlineusa.com/]canadian pharcharmy online[/url]
canadian pharcharmy online
http://canadianpharcharmyonlineusa.com/ canadian pharcharmy online

April 17th, 2016 at 10:32 am

[redacted]

Very quickly this site will be famous among all blogging and site-building visitors, due to it’s pleasant posts

May 8th, 2016 at 3:22 pm

[redacted]

Have you been inclined to get popularity shortly? Online community site Twitter can be the proper way for you. After that, you require to get cheap twitter followers to become recognized to this world. buy followers on twitter

May 12th, 2016 at 8:09 am

[redacted]

Please visit the sites we stick to, which includes this one, because it represents our picks in the web.

May 15th, 2016 at 12:53 am

[redacted]

We prefer to honor a lot of other internet internet sites around the web, even if they aren?t linked to us, by linking to them. Beneath are some webpages worth checking out.

May 29th, 2016 at 8:29 am

← Older Comments

Leave a Reply

You must be logged in to post a comment.