State of the Union

When you don't need Union Mounts

Jan Blunck

Novell
`jblunck@suse.de`

30. October 2009

522
Mount Union
TO
16

What Union?

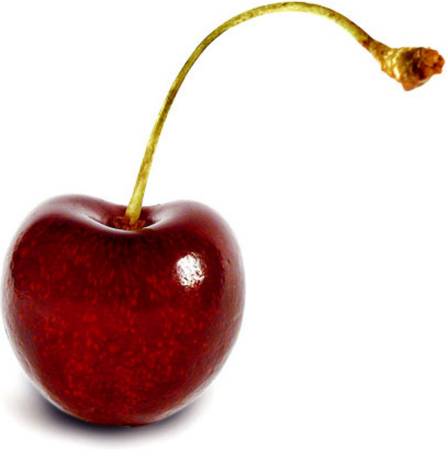
- Not the European Union … this is Dresden not Brussels
- This is about Filesystems
- In particular about Filesystem Namespace Unification

What Union?

- Not the European Union … this is Dresden not Brussels
- This is about Filesystems
- In particular about Filesystem Namespace Unification

What Union?

- Not the European Union ... this is Dresden not Brussels
- This is about Filesystems
- In particular about Filesystem Namespace Unification



Outline

- Introduction
- Where is the Problem?
- Unioning Filesystems
  - UnionFS
  - Another UnionFS
  - UnionFS-FUSE
  - mini_fo
- Union Mount
- You probably don't need Union Mounts
  - Device-Mapper Snapshot
  - Delta Filesystem
  - CLIC Filesystem
  - SquashFS Fake Write Support
  - Shared root filessytem - NFS Root
  - Shared root filessytem - XIP
- Whats left to do than?
- Thanks

Disclaimer

I'm the author of the VFS based Union Mount patches.
That somehow makes me biased.
I'll try my very best though ...

Where is the Problem?

POSIX Requirements

- seek to `cookie`

POSIX is missing

- whiteout filetype `DT_WHITEOUT`
- topology of mount tree
- open (directories) by inode number

Where is the Problem?

POSIX Requirements

- seek to `cookie`

POSIX is missing

- whiteout filetype `DT_WHITEOUT`
- topology of mount tree
- open (directories) by inode number

Where is the Problem?

- NFS Sucks

ftp://ftp.lst.de/pub/people/okir/papers/2006-OLS/nfs-sucks-slides.pdf

Where is the Problem?

- NFS Sucks

`ftp://ftp.lst.de/pub/people/okir/papers/2006-OLS/nfs-sucks-slides.pdf`

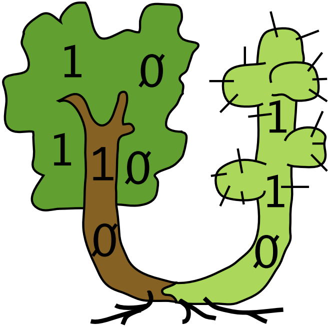
In the Linux kernel there are basically two layers that implement filesystem features:

- in the Virtual Filesystem (VFS)
- in a low-level Filesystem

Both layers come with their own responsibilities!

Union FS

- UnionFS is the best-known and longest-living implementation so far
- It has its origin in the FiST stackable filesystem project at SUNY Stony Brook
- The project is led by Erez Zadok, professor at Stony Brook



Union FS

Major Features

- Allows merging of up to 128 read-only or read-writable branches
- Allows multiple writable branches
- Supports copy-up to higher-priority branches
- Remove-all unlink() semantics plus whiteout
- Append and prepend semantics

Current Status

- Latest version is Unionfs 2.5.3, released in September 2009
- Support for Linux 2.6.9, and 2.6.18 to 2.6.32
- Although it was in -mm, it was NAKed by VFS maintainers

Union FS

Future Directions

- unionfs-lite
  - Supports only two branches (one read-only, one read-writable)
  - Uses native low-level filesystem whiteout support
- Native low-level filesystem whiteout
  - Native additional “filetype” `DT_WHT`
  - Support for tmpfs, ext2/3/4 ...
  - Add ioctl() interface to have userspace control over whiteouts
- Keep up-to-date with latest Linux kernel releases
- Upstream ?

Another UnionFS

- Started as a UnionFS fork; rewritten from scratch in 2006
- Lead developer is Junjiro Okajima

Another UnionFS

Major Features

- Supports thousands of branches
- NFS exportable through external inode number table
- Pseudo Link
- Direct branch access
- Different policies for creat and copy-up

Another UnionFS

Current Status

```
Date: Fri, 10 Apr 2009 13:41:55 -0400
From: Christoph Hellwig <>
Subject: Re: [RFC Aufs2 #5 28/29] export lookup functions

On Sat, Apr 11, 2009 at 02:26:33AM +0900, hooanon05@yahoo.co.jp wrote:
> I have been asked to include aufs into mainline from several people
> several times. As long as you have strong NACK for aufs and reject all
> union-type filesystems, I have to give up unwillingly and will answer
> them "Aufs was rejected. Let's give it up."

Yes, that's the case.
```

UnionFS-FUSE

- Developed by Radek Podgorny and Bernd Schubert
- FUSE based approach

Major Features

- feature complete
- Live CD
- USB media
- Copy On Write

Current Status

- Last release March 2009

mini_fo - Mini Fan-Out Overlay Filesystem

- Developed by Markus Klotzbuecher

Major Features

- Only two branches
- Optimized for embedded usage

Current Status

- Used by OpenWRT
- Last release in October 2005

Union Mount

- Started in 2004 by Jan Blunck
- Help from Bharata B Rao, Miklos Szeredi, David Woodhouse, Valerie Aurora (formerly Henson)

Major Features

- VFS based
- Limited feature set

`http://valerieaurora.org/union/`

Union Mount

Current Status

- Userspace `readdir()` support failed
- Focus on upstream acceptance
  - Get directory reading right
  - Play well with existing VFS namespace concepts
  - Document how the locking works
- Whiteout Support (used by UnionFS, too)
- Writable Overlays
  - Only two branches
  - Whiteout/Fallthrough support for EXT2, JFFS2 and tmpfs

You probably don't need Union Mounts

- What is available for LiveCDs?
- ... USB media?
- ... shared root filesystem?

You probably don’t need Union Mounts

- What is available for LiveCDs?
- ... USB media?
- ... shared root filesystem?

Device-Mapper Snapshot

- Block based
- Multiple Layers/Snapshots
- Efficient
- Upstream

Delta Filesystem

- FUSE based
- Block based/File based
- Uses two directories

`http://lwn.net/Articles/321391/`

CLIC Filesystem (former DoenerFS)

- FUSE based
- Similar to Delta Filesystem
- Compression
- Boottime optimized

http://lizards.opensuse.org/2009/04/28/whats-behind-lzma-compressed-livecds/

SquashFS Fake Write Support

- Make SquashFS write to tmpfs
- Not faster that CLIC Filesystem
- NAK from Phillip, because of VFS union mount support

`http://lizards.opensuse.org/2009/05/15/livecd-performance-clicfs-vs-squashfs/`

Shared root filessytem - NFS Root

- Works

Shared root filessytem - XIP

- Works as well
- Uses bind mounts
- Problem: You need a zSeries Mainframe running z/VM

`http://linuxvm.org/presentations/`

Shared root filessytem - XIP

- Works as well
- Uses bind mounts
- Problem: You need a zSeries Mainframe running z/VM

`http://linuxvm.org/presentations/`

What would happen if we would have Union Mounts?

- Every guest can modify everything ...
- How do you merge back changes?
- You will need common ancestor ...

Thanks