ZACK'S KERNEL NEWS

BtrFS as a Linux "Next Generation" Filesystem

Chris Mason has migrated BtrFS development onto its own git repository and is starting to lean toward submitting the code into the main kernel tree. Several technical problems, not least of which is that the on-disk format hasn't stabilized yet, are standing in the way of this.

Without a stable format, users will have to be informed clearly that the earlier formats will not be supported after a certain time. And there will always be those users who didn't get the memo and are therefore out of luck when they need to access their data later. Chris's plan is to get as close as possible to a final disk format and to include backward compatibility on any changes made after that. If successful, the effect on users would be much less severe.

Andrew Morton has come out strongly in favor of a quick merge. The filesystem itself includes loud warnings at run time, and Andrew feels that BtrFS has a bright future and should be put in a prominent position as soon as possible so as to attract a body of contributors.

Adrian Bunk is skeptical of that theory, citing ext4 as an example of a filesystem that was accepted early into the kernel and did not experience the accelerated development for which Andrew was hoping. Serge E. Hallyn objected and said that because BtrFS is really cool, having it available in mainline would have a special excitement factor that ext4 lacked. But Adrian pointed out that anyone could install BtrFS without it being in the mainline tree.

The discussion went back and forth, and at some point Theodore (Ted) T'so piped up with his take on the politics. Essentially, he said, Adrian was opposed to early merging, whereas Andrew was in favor of it. And because Linus Torvalds had already expressed his new preference to merge drivers earlier than in the past, it seemed to Ted that BtrFS would indeed be merged as Andrew suggested, and that Adrian's objections were swimming too much against the tide to achieve their goal. But he did at least partially acknowledge Adrian's point, that ext4 development went slower than expected after it had been merged.

Ted also mentioned a little backstory about BtrFS, saying, "… about a year ago (on November 12-13, 2007), a small group of key filesystem developers, which included engineers employed by HP, Oracle, IBM, Intel, HP, and Red Hat, and whose experience included working with a large number of filesystems – ext2, ext3, ext4, OCFS2, Lustre, BtrFS, AdvFS, ReiserFS, and XFS – came together for a two-day 'next generation filesystem' (NGFS) workshop.

"At the end of the workshop, there was unaminous agreement (including from yours truly) that (a) Linux needed a next-generation filesystem to be competitive, (b) Chris Mason's BtrFS (with some changes/enhancements discussed during the workshop) was the best long-term solution for NGFS, and (c) because creating a new enterprise filesystem always takes longer than people expect, and even then, it takes a while for enterprise users to trust a new filesystem for their most critical data, ext4 in the next generation of filesystems was needed as the bridge to the NGFS."

He added, "It is fair to say that BtrFS isn't just a private project of a single Linux kernel developer, but rather the design has been discussed and reviewed by a large number of experienced filesystem architects."

It does seem clear that unless something weird happens, BtrFS will be in a kernel near you quite soon.

Linux Foundation TAB Election Results

Jonathan Corbet announced the results of the recent election of members of the Linux Foundation Technical Advisory Board (TAB).

James Bottomley, Kristen Carlson Accardi, Chris Mason, Dave Jones, and Chris Wright were all elected to two-year positions on the Board. A single one-year position had opened up when Olaf Kirch resigned, and the vote to replace him resulted in a tie between Theodore T'so and Christoph Hellwig. In proper democratic fashion, they broke the tie with a single coin toss, which went to Christopher.

Status of UWB, WUSB, and WLP Subsystems

David Vrabel, maintainer of the Ultra-Wideband (UWB) radio, Certified Wireless USB (WUSB), and WiMedia LLC Protocol (WLP) subsystems, felt that the code was ready to be included in the mainline tree, and he created a git repository from which folks can pull. Without much discussion, some folks were a little unclear as to whether David was submitting his code for inclusion or just for review. Barring any big technical issue, it does seems as though the code will be accepted soon.

The Linux kernel mailing list comprises the core of Linux development activities. Traffic volumes are immense, often reaching ten thousand messages in a given week, and keeping up to date with the entire scope of development is a virtually impossible task for one person. One of the few brave souls to take on this task is Zack Brown.
Our regular monthly column keeps you abreast of the latest discussions and decisions, selected and summarized by Zack. Zack has been publishing a weekly online digest, the Kernel Traffic newsletter for over five years now. Even reading Kernel Traffic alone can be a time consuming task.
Linux Magazine now provides you with the quintessence of Linux Kernel activities, straight from the horse's mouth.

Character Devices in User Space

Tejun Heo has been working on CUSE (character devices in user space), which is similar to FUSE and is based on the same code, with the main differences occurring at initialization. This effort spawned a number of requests for a corresponding BUSE, block devices in user space, but Tejun said applying this technique to block devices wouldn't really produce much of an improvement over loopback mounting and would require much more than the slight modification of FUSE necessary for CUSE. He did admit that loopback over FUSE was problematic and that probably anything would be better than that.

Junjiro R. Okajima said that his own ULOOP driver, which essentially implemented this very thing, already existed. Mike Hommey also suggested DUSE, device mapper in user space, although he followed his own post with a link to the DmUserspace project that did this already. Until now, it seems that CUSE has really been the odd one out.

Tejun announced an OSS (open sound system) proxy that uses CUSE as its back end and is intended as a replacement for the old and largely removed OSS drivers that are currently emulated in ALSA. Tejun explains that the problem with OSS under ALSA is that if the sound card doesn't support multiple audio streams, users must choose whether to use ALSA or OSS at any given moment. With his OSS proxy, which is really an emulation tool, the sound card can support ALSA and OSS simultaneously.

Adrian Bunk pointed out that after six years of effort at replacing all OSS drivers with ALSA, ALSA now supports nearly all applications. He suggested that if Tejun knew of an application that didn't work under it, he should fix the ALSA support rather than writing an OSS emulator. Greg Kroah-Hartman said that even if the OSS implementation was redundant, the underlying CUSE project was still useful for a number of other projects. And Tejun defended his OSS project, saying that at the very least, old binaries were lying around, as well as old code bases that wouldn't work with ALSA and wouldn't be updated. An OSS emulator would let people use those old games and tools.

SR-IOV Support

Yu Zhao coded up support for single-root I/O virtualization (SR-IOV), which allows multiple concurrently running operating systems on a given piece of hardware to share the same PCI device without exploding. SR-IOV is a nice piece of a puzzle that will one day let users do a lot of cool stuff.