Contribution ID: **102** Type: **not specified**

Atomic file writes: Who really wants this?

Tuesday, 21 September 2021 08:30 (30 minutes)

I would like to chair a discussion at LPC to discuss atomic file writes for userspace applications. Do we want to expose such a capability to programs, and if so, how?

I propose filesystem implementations provide a general-purpose interface in software. As proposed, the FIEXCHANGE_RANGE system call requires the ability to exchange the contents of two files, with a promise that once we commit to the exchange, it must either succeed completely.

Atomic file writes can be performed by creating a temporary file, cloning the contents, making arbitrary updates to the temporary file, and calling FIEXCHANGE_RANGE to commit the changes. There are no restrictions on length, number of updates, etc.

The ability to exchange the contents of files atomically is a requirement for online repair of XFS metadata; upon finishing the functionality I realized that we could expose it to userspace to provide atomic file updates.

NOTE: This is a separate topic from enabling userspace to access hardware atomic writes. That is a simple matter of making the advertised device capabilities (and alignment/size restrictions) discoverable and adding a flag to io_uring/pwritev2 for directio writes.

I agree to abide by the anti-harassment policy

I agree

Primary author: WONG, Darrick (Oracle)

Presenter: WONG, Darrick (Oracle)

Session Classification: File Systems MC

Track Classification: File Systems MC