“Bad” Storage vs File Systems

Problem Statement

- Many file systems and application authors primarily test and develop using traditional storage devices
- When writing, HDD's / SSD's remap bad media to spare/replacement sectors
  - Writes rarely fail unless the device fails catastrophically (head crash, firmware show stops)
- Cloud storage / software defined storage / iSCSI devices fail differently
  - A single write can fail, while other writes immediately before and after succeed
  - Perhaps writes to one set of stripes will fail, while writes to other stripes succeed
- Ext4 is getting patches to make it more robust to these sorts of failures
  - But what about data blocks? Metadata blocks are < 1% of all the blocks of storage
  - Ready to bet userspace application writers are correctly handling these failures?

Possible solutions for discussion

- If we can’t trust applications to handle these failures correctly maybe allow sysadmins to request that data write failures should be treated like fs errors?
  - Force the file system to be shut down or remounted read-only
  - Force a reboot / panic (to allow failover to backup servers)
- How to specify what should happen?
  - Mount options to specify the policy?
  - eBPF programs so the policy can be more granular (based on uid/gid of writer, based on the file)?
  - Using an fsnotify scheme to send notification to userspace?
- Should we establish a standard interface across multiple file systems?