
RAI 1

Dirty deals:
the story of a data corruption bug in OpenZFS

Rob Norris, Klara Inc.

Hello!

#robnfacts

- He/him
- Australian
- 🇦🇺🏳️‍🌈🏴
- One wife, two cats, three kids
- 1989 - 1999: Kid messing with computers
- 1999 - 2023: Linux sysadmin
- 2023 - ....: OpenZFS developer
- 2023 - ....: one FreeBSD server
- Hundreds of side-projects and dumb experiments

November 2023

📅 November 2023

- 🦃 Thanksgiving long weekend (USA & Canada)
- Home alone (sort of)
- OpenZFS 2.2.0
- FreeBSD 14

🪲 Bug report

#15526 some copied files are corrupted (chunks replaced by zeros)
https://github.com/openzfs/zfs/issues/15526

- Compiling a Go program
- Reading and writing the same files in parallel
- Reading all-zeroes instead real data

🪲 Bug report

#15526 some copied files are corrupted (chunks replaced by zeros)
https://github.com/openzfs/zfs/issues/15526

- OpenZFS 2.2.0
- Coreutils 9.2

No one ever got fired for blaming block cloning

tonyhutter commented on Nov 23, 2023 Contributor •••

I'm able to reproduce in Fedora 37 with 6.5.11 kernel, coreutils 9.1 and zfs-2.1.13. So doesn't look like this is 2.2.x only.

☺ 😕 5 👀 10

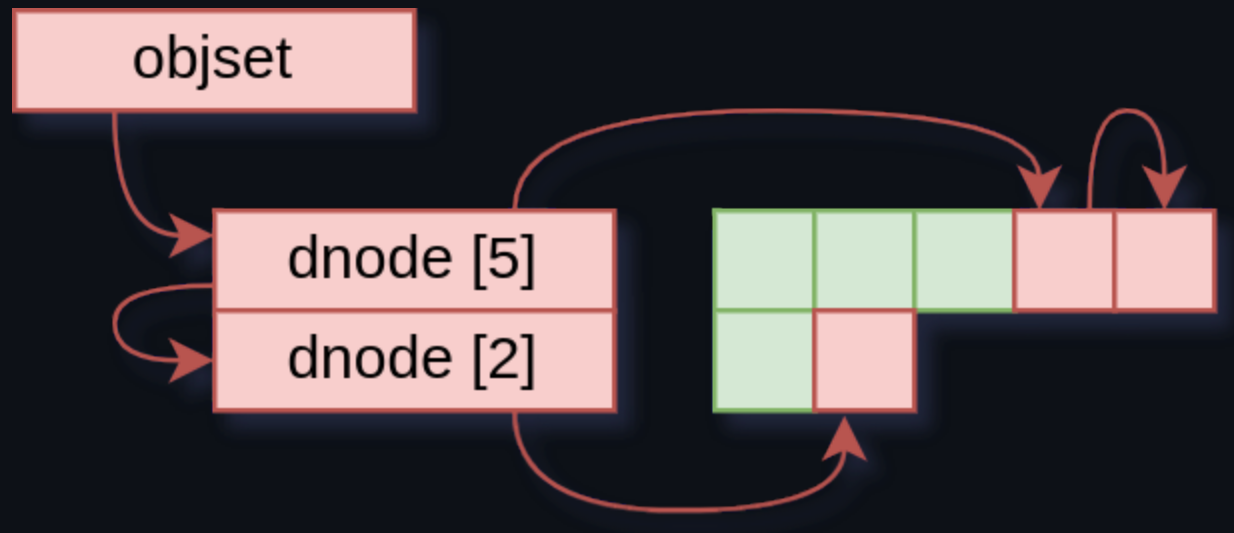
DATA CORRUPTION

Filesystem lesson

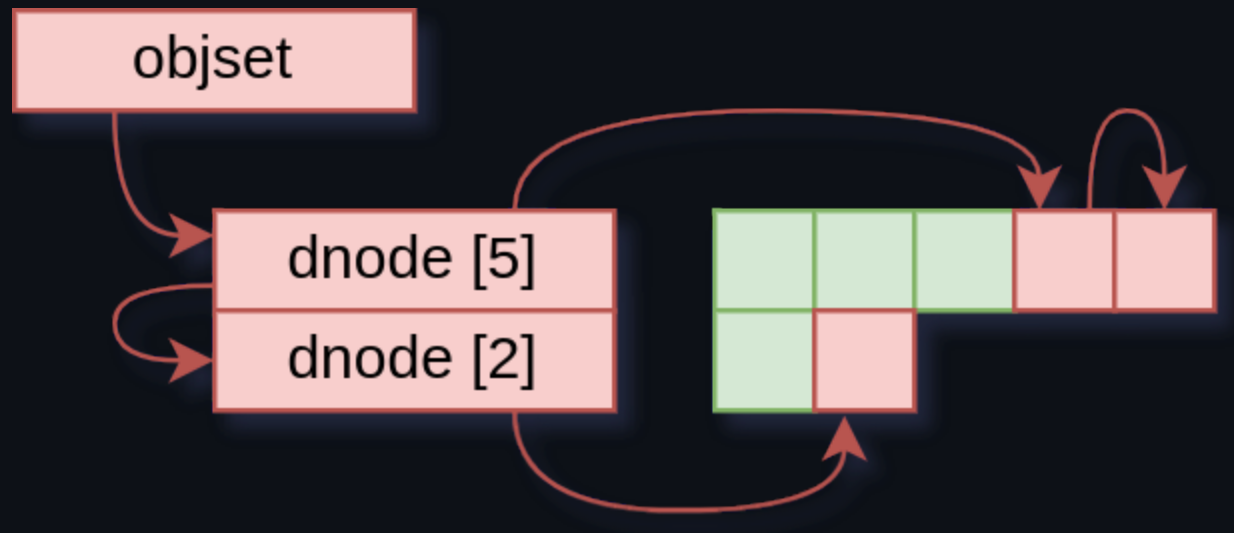
What is a file, really?

Not all data is data

💾 The fastest disks are memory

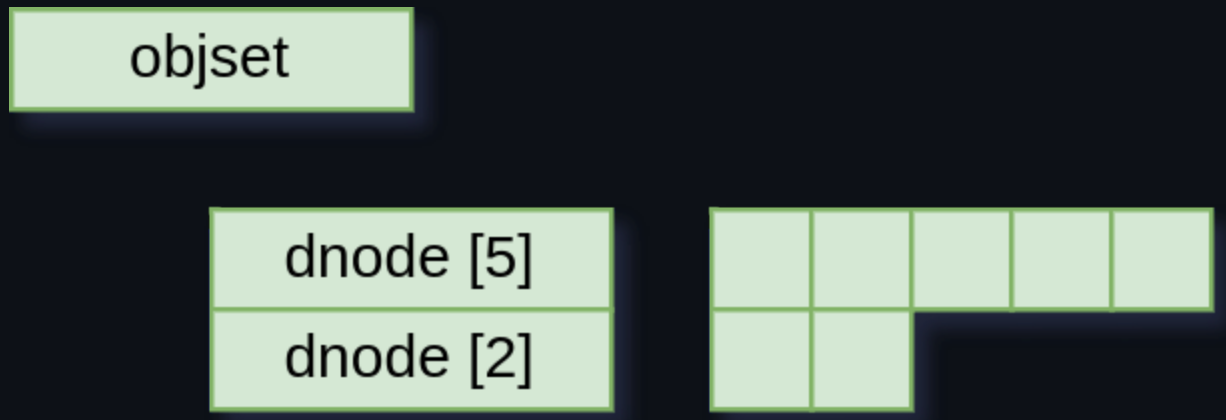


Data first



Data first

Data first



Beware of holes

Hole detection: `lseek()`

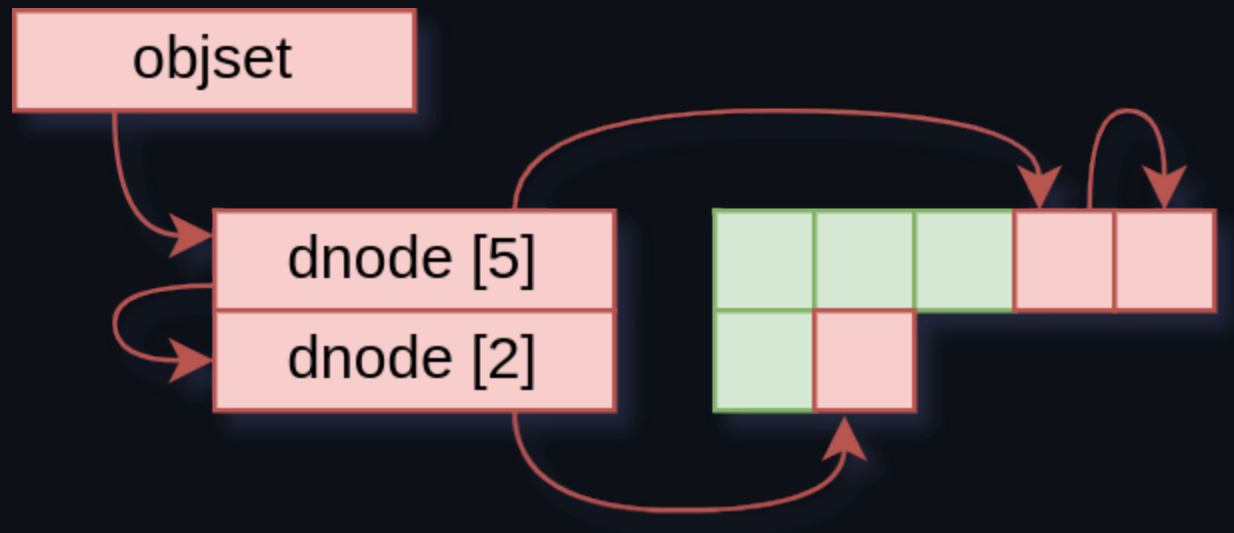
```c
#include <unistd.h>

off_t lseek(int fildes, off_t offset, int whence);
```

- `SEEK_HOLE`: move to start of next hole
- `SEEK_DATA`: move to start of next real non-hole data

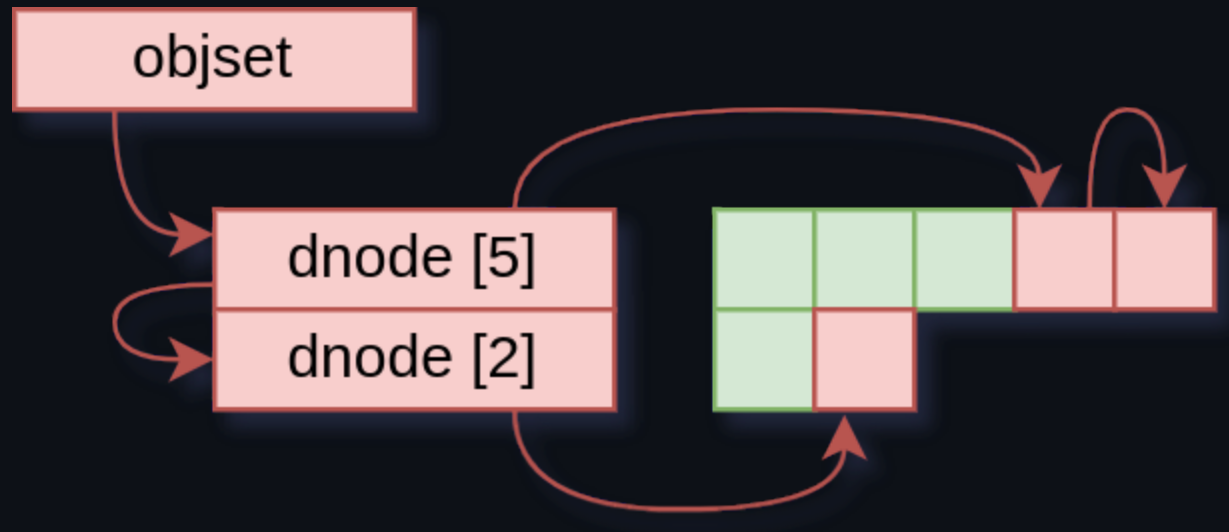
🔍 Hidden dirt

- Holes only exist once stored
- Can't tell if a dirty block will be stored as a hole

🔍 Finding dirt

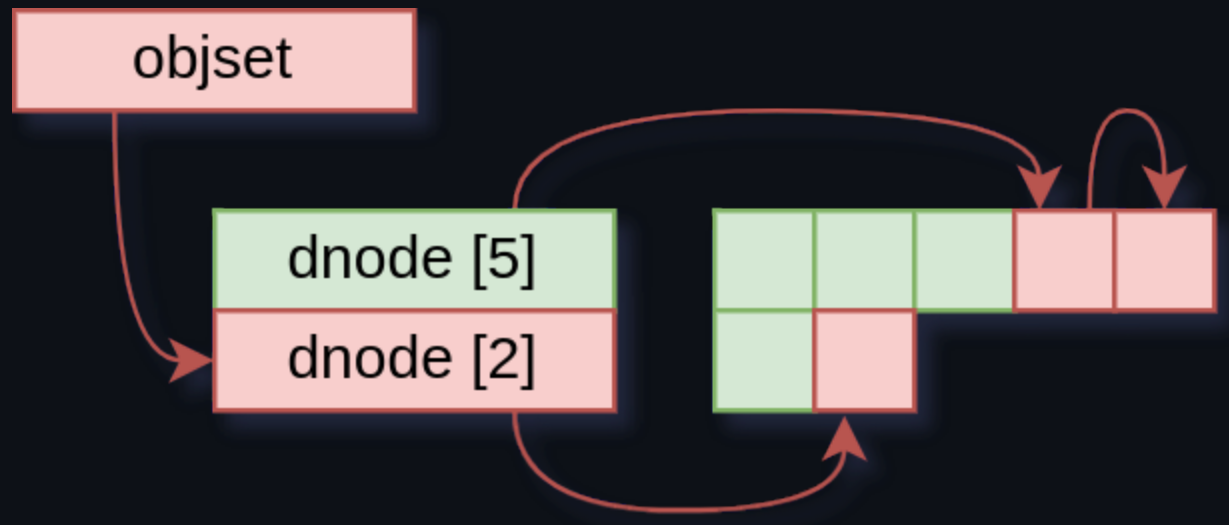


🔍 Finding dirt



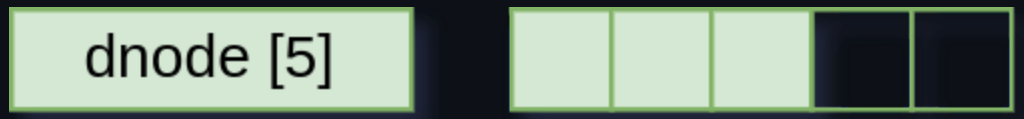

```
multilist_link_active(&dn->dn_dirty_link)
```

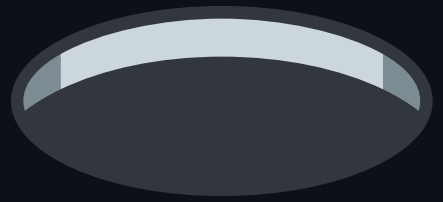
🔍 Finding dirt



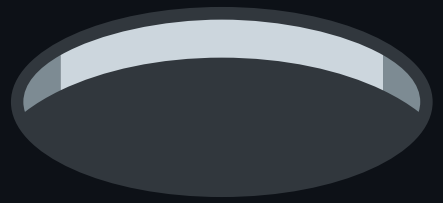

```
multilist_link_active(&dn->dn_dirty_link)
```

🔍 Finding dirt





Hole



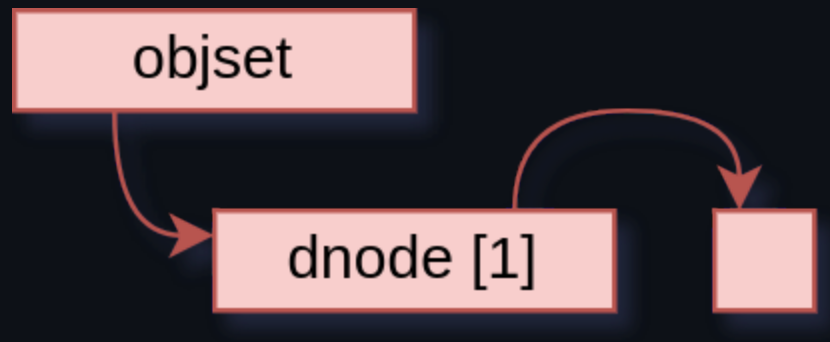
copying

Hole copying

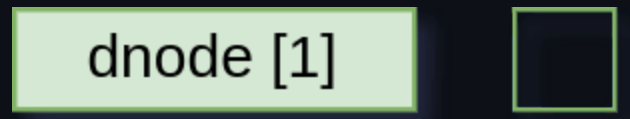
objset

dnode [0]

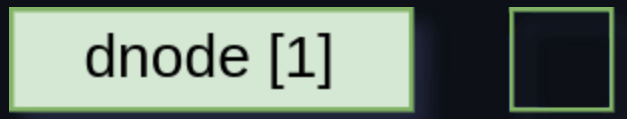
Hole copying



Hole copying



Hole copying




```
off_t offset = lseek(srcfd, 0, SEEK_DATA);
if (offset < 0 && errno == ENXIO) {
        ftruncate(dstfd, srclen);
}
```

Plugging the hole

⛳ Plugging the hole

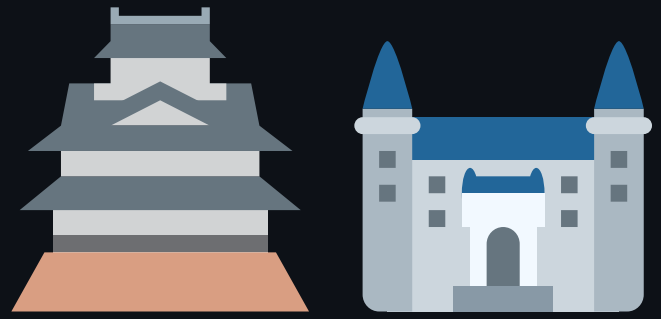
```
diff --git module/zfs/dnode.c module/zfs/dnode.c
index 029d9df8a..7ae74ad13 100644
--- module/zfs/dnode.c
+++ module/zfs/dnode.c
@@ -1786,7 +1793,8 @@ dnode_is_dirty(dnode_t *dn)
        mutex_enter(&dn->dn_mtx);

        for (int i = 0; i < TXG_SIZE; i++) {
-               if (multilist_link_active(&dn->dn_dirty_link[i])) {
+               if (multilist_link_active(&dn->dn_dirty_link[i]) ||
+                   !list_is_empty(&dn->dn_dirty_records[i])) {
                        mutex_exit(&dn->dn_mtx);
                        return (B_TRUE);
                }
```

⛳ Plugging the hole

- 2023-11-28: Patch posted and accepted (GH#15571)
  - Backport for 2.2.x (GH#15579)
  - Backport for 2.1.x (GH#15578)
  - Backport for 2.0, 0.8, 0.7, 0.6 available
- 2023-11-28: OpenZFS 2.2.2 released
- 2023-12-01: FreeBSD errata issued (FreeBSD-EN-23:16.openzfs)
  - 14.0-RELEASE-p1
  - 13.2-RELEASE-p6
  - 12.4-RELEASE-p8

⛳ Plugging the hole

- 2023-12-03: Illumos patched (#16087)
- 2024-02-12: Ubuntu 23.10 updated (OpenZFS 2.2.0)
- 2024-03-12: Ubuntu 22.04 updated (OpenZFS 2.1.5)
- 2024-03-13: Ubuntu 20.04 updated (OpenZFS 0.8.3)



History lesson

🏯 History lesson

- Reproduced back to ZFS-on-Linux 0.6.5
- Reproduced in FreeBSD 12 & 13
- Reproduced in Illumos (current)
- Attempts to confirm in old Sun ZFS difficult

History lesson

2006: proto-bug introduced in Sun ZFS

```
        for (i = 0; i < TXG_SIZE; i++) {
-               if (dn->dn_dirtyblksz[i])
+               if (list_link_active(&dn->dn_dirty_link[i]))
                        break;
        }
```

History lesson

Mar 2017: `66aca24 SEEK_HOLE should not block on txg_wait_synced()`

```
-       for (i = 0; i < TXG_SIZE; i++) {
-               if (list_link_active(&dn->dn_dirty_link[i]))
-                       break;
+       if (dn->dn_dirtyctx != DN_UNDIRTIED) {
+               for (i = 0; i < TXG_SIZE; i++) {
+                       if (!list_is_empty(&dn->dn_dirty_records[i])) {
+                               clean = B_FALSE;
+                               break;
+                       }
+               }
```

History lesson

Nov 2017: `454365b Fix dirty check in dmu_offset_next()`

```
-       if (dn->dn_dirtyctx != DN_UNDIRTIED) {
-               for (i = 0; i < TXG_SIZE; i++) {
-                       if (!list_is_empty(&dn->dn_dirty_records[i])) {
-                               clean = B_FALSE;
-                               break;
-                       }
+       for (i = 0; i < TXG_SIZE; i++) {
+               if (list_link_active(&dn->dn_dirty_link[i])) {
+                       clean = B_FALSE;
+                       break;
```

History lesson

Mar 2019: `ec4f9b8 Report holes when there are only metadata changes`

```
        for (i = 0; i < TXG_SIZE; i++) {
                if (multilist_link_active(&dn->dn_dirty_link[i])) {
+
+                       list_t *list = &dn->dn_dirty_records[i];
+                       [checks against dn_dirty_records]
```

History lesson

May 2019: `2531ce3 Revert "Report holes when there are only metadata changes"`

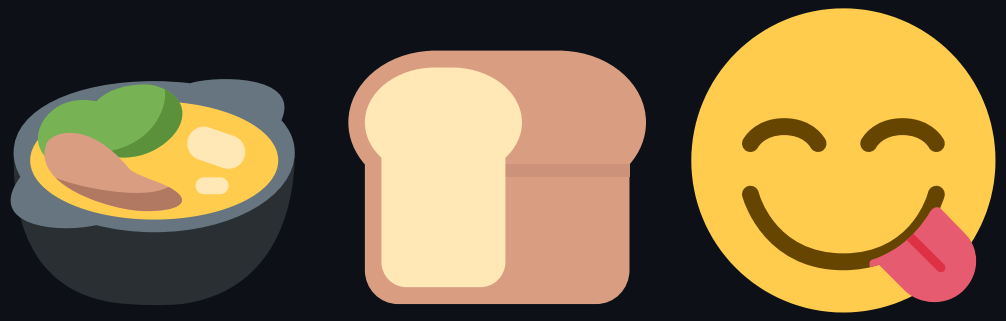
```
        for (i = 0; i < TXG_SIZE; i++) {
                if (multilist_link_active(&dn->dn_dirty_link[i])) {
-
-                       list_t *list = &dn->dn_dirty_records[i];
-                       [checks against dn_dirty_records]
```

History lesson

Nov 2021: `de198f2 Fix lseek(SEEK_DATA/SEEK_HOLE) mmap consistency`

```
-       for (i = 0; i < TXG_SIZE; i++) {
-               if (list_link_active(&dn->dn_dirty_link[i]))
-                       break;
+       if (dn->dn_dirtyctx != DN_UNDIRTIED) {
+               for (i = 0; i < TXG_SIZE; i++) {
+                       if (!list_is_empty(&dn->dn_dirty_records[i])) {
+                               clean = B_FALSE;
+                               break;
+                       }
+               }
```

Instant community



People need to eat

Conclusion