Eric Schrock ZFS Hot Spares

June 06, 2006 03:48 PM UTC

It's been a long time since the last time I wrote a blog entry. I've been working heads-down on a new project and haven't had the time to keep up my regular blogging. Hopefully I'll be able to keep something going from now on.

Last week the ZFS team put the following back to ON:

```
PSARC 2006/223 ZFS Hot Spares
PSARC 2006/303 ZFS Clone Promotion
6276916 support for "clone swap"
6288488 du reports misleading size on RAID-Z
6393490 libzfs should be a real library
6397148 fbufs debug code should be removed from buf_hash_insert()
6405966 Hot Spare support in ZFS
6409302 passing a non-root vdev via zpool_create() panics system
6415739 assertion failed: !(zio->io_flags & 0x00040)
6416759 ::dbufs does not find bonus buffers anymore
6417978 double parity RAID-Z a.k.a. RAID6
6424554 full block re-writes need not read data in
6425111 detaching an offline device can result in import confusion
```

There are a couple of cool features mixed in here. Most importantly, hot spares, clone swap, and double-parity RAID-Z. I'll focus this entry on hot spares, since I wrote the code for that feature. If you want to see the original ARC case and some of the discussion behind the feature, you should check out the original zfs-discuss thread.

The following features make up hot spare support:
Associating hot spares with pools

Hot spares can be specified when creating a pool or adding devices by using the spare vdev type. For example, you could create a mirrored pool with a single hot spare by doing:

```
# zpool create test mirror c0t0d0 c0t1d0 spare c0t2d0
# zpool status test
  pool: test
 state: ONLINE
 scrub: none requested
config:

        NAME        STATE     READ WRITE CKSUM
        test        ONLINE       0     0     0
          mirror    ONLINE       0     0     0
            c0t0d0  ONLINE       0     0     0
            c0t1d0  ONLINE       0     0     0
        spares
          c0t2d0    AVAIL

errors: No known data errors
```

Notice that there is one spare, and it currently available for use. Spares can be shared between multiple pools, allowing for a single set of global spares on systems with multiple spares.
Replacing a device with a hot spare

There is now an FMA agent, zfs-retire, which subscribes to vdev failure faults and automatically initiates replacements if there are any hot spares available. But if you want to play around with this yourself (without forcibly faulting drives), you can just use 'zpool replace'. For example:

```
# zpool offline test c0t0d0
Bringing device c0t0d0 offline
# zpool replace test c0t0d0 c0t2d0
# zpool status test
  pool: test
 state: DEGRADED
status: One or more devices has been taken offline by the adminstrator.
        Sufficient replicas exist for the pool to continue functioning in a
        degraded state.
action: Online the device using 'zpool online' or replace the device with
        'zpool replace'.
 scrub: resilver completed with 0 errors on Tue Jun  6 08:48:41 2006
config:

        NAME           STATE     READ WRITE CKSUM
        test           DEGRADED     0     0     0
          mirror       DEGRADED     0     0     0
            spare      DEGRADED     0     0     0
              c0t0d0   OFFLINE      0     0     0
              c0t2d0   ONLINE       0     0     0
            c0t1d0     ONLINE       0     0     0
        spares
          c0t2d0       INUSE     currently in use

errors: No known data errors
```

Note that the offline is optional, but it helps visualize what the pool would look like should and actual device fail. Note that even though the resilver is completed, the 'spare' vdev stays in-place (unlike a 'replacing' vdev). This is because the replacement is only temporary. Once the original device is replaced, then the spare will be returned to the pool.
Relieving a hot spare

A hot spare can be returned to its previous state by replacing the original faulted drive. For example:

```
# zpool replace test c0t0d0 c0t3d0
# zpool status test
  pool: test
 state: DEGRADED
 scrub: resilver completed with 0 errors on Tue Jun  6 08:51:49 2006
config:

        NAME              STATE     READ WRITE CKSUM
        test              DEGRADED     0     0     0
          mirror          DEGRADED     0     0     0
            spare         DEGRADED     0     0     0
              replacing   DEGRADED     0     0     0
                c0t0d0    OFFLINE      0     0     0
                c0t3d0    ONLINE       0     0     0
              c0t2d0      ONLINE       0     0     0
            c0t1d0        ONLINE       0     0     0
        spares
          c0t2d0          INUSE     currently in use

errors: No known data errors
# zpool status test
  pool: test
 state: ONLINE
 scrub: resilver completed with 0 errors on Tue Jun  6 08:51:49 2006
```

```
config:

        NAME        STATE     READ WRITE CKSUM
        test        ONLINE       0     0     0
          mirror    ONLINE       0     0     0
            c0t3d0  ONLINE       0     0     0
            c0t1d0  ONLINE       0     0     0
        spares
          c0t2d0    AVAIL

errors: No known data errors
```

The drive is actively being replaced for a short period of time. Once the replacement is completed, the old device is removed, and the hot spare is returned to the list of available spares. If you want a hot spare replacement to become permanent, you can zpool detach the original device, at which point the spare will be removed from the hot spare list of any active pools. You can also zpool detach the spare itself to cancel the hot spare operation.
Removing a spare from a pool

To remove a hot spare from a pool, simply use the zpool remove command. For example:

```
# zpool remove test c0t2d0
# zpool status
  pool: test
 state: ONLINE
 scrub: resilver completed with 0 errors on Tue Jun  6 08:51:49 2006
config:

        NAME        STATE     READ WRITE CKSUM
        test        ONLINE       0     0     0
          mirror    ONLINE       0     0     0
            c0t3d0  ONLINE       0     0     0
            c0t1d0  ONLINE       0     0     0

errors: No known data errors
```

Unfortunately, we don't yet support removing anything other than hot spares (it's on our list, we swear). But you can see how hot spares naturally fit into the existing ZFS scheme. Keep in mind that to use hot spares, you will need to upgrade your pools (via '`zpool upgrade`') to version 3 or later.

Next Steps

Despite the obvious usefulness of this feature, there is one more step that needs to be done for it to be truly useful. This involves phase two of the ZFS/FMA integration. Currently, a drive is only considered faulted if it 'goes away' completely (i.e. `ldi_open()` fails). This covers only subset of known drive failure modes. It's possible for a drive to continually return errors, and yet be openable. The next phase of ZFS and FMA will introduce a more intelligent diagnosis engine to watch I/O and checksum errors as well as the SMART predictive failure bit in order to proactively offline devices when they are experiencing an abnormal amount of errors, or appear like they are going to fail. With this functionality, ZFS will be able to better respond to failing drives, thereby making hot spare replacement much more valuable.