intrbiz

CEPH

A Frustrating Journey In Shared Storage

Wolves LUG - 2020

Chris Ellis - @intrbiz

What Is Ceph

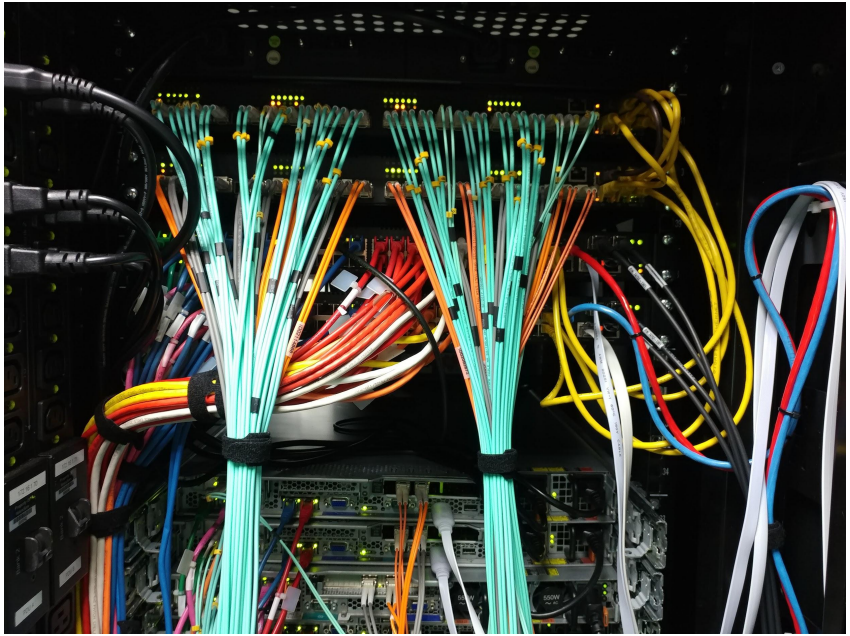
- Ceph is a software defined storage solution
  - AKA: an open source SAN
- It turns a bunch of HDDs or SSDs into:
  - Block storage
  - Shared file system
  - Object storage (S3)
- No fancy hardware needed, just:
  - Linux
  - Ceph
  - HDDs and SSDs
  - Fast network

Why Ceph

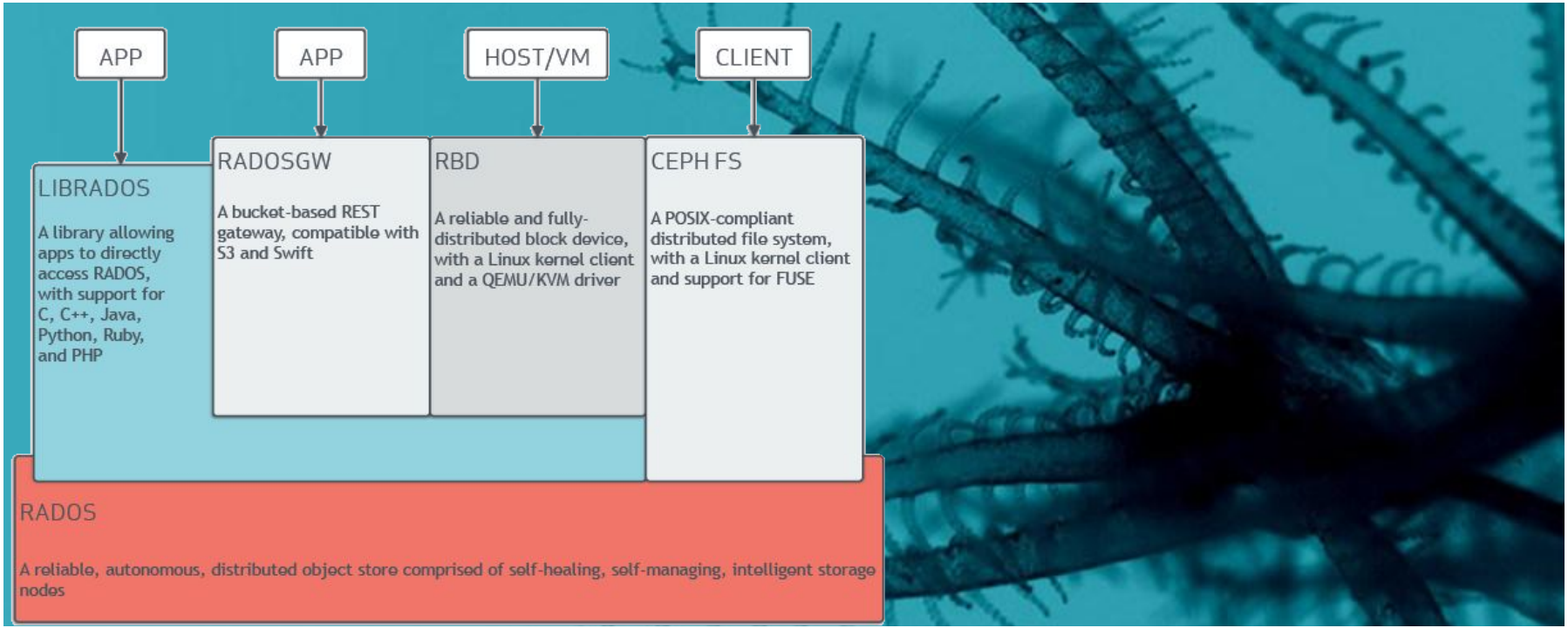
- Live Migration
  - Main driver for me was to be able to live migrate a VM from host to host
    - Live migration requires shared storage between the VM hosts
  - Snapshots and Clones
    - Can create a VM image in seconds, since they COW from a backing snapshot
- Performance
  - The VMs disk image is spanned across many disks and servers
- Availability and Redundancy
  - A VMs disk is spread across many disks and servers
- Commodity
  - Turn a bunch of disks into a fully features SAN

What We Need


System x3650 M4
System x3650 M3
IBM
System x3650 M3
IBM
System x3650 M3
IBM



Awake The Kraken...



Cool, It Works…?

- All excited that I'd got it all working
  - VMs running on top
- Let's test disk IO
  - Got a whole 8MB/s
  - @ 40% IO Wait
- Oh :(
- But why

Devil In Detail

- Tried many things:
  - Rebuild networking several times
  - Added SSD cache tier
  - Flipped between different storage engines
- None made a big difference
- Key issue is how Ceph writes to disks
  - Writes need to be atomic, so they go direct to the disk
    - Bypass kernel page cache and disk cache
  - Ceph uses a journal for this, in my initial setup this was on each disk
  - Disk didn't have a battery backed cache
  - IE: on spinning rust VERY VERY slow

Devil In Detail

- Solution to this is SSDs as a journal for multiple disks
- No two SSDs are equal under this write situation
- A SSD that is fast under normal write situations can be dog slow
  - Crucial MX500, normal IO 500MB/s, direct: 2MB/s
  - Even NVMe doesn't fix this
  - Not as simple as consumer vs enterprise
- It's all down to firmware and show the SSD manages MLC vs SLC
- After lots of research and testing some SSDs I got
  - I found some that are fast under the right: Samsung SM863
    - Even better I found someone selling a bunch of them cheap on ebay
- Rebuilt the cluster with 2 SSD journals per chassis (7 HDDs to 1 SSD)

Cool, It Works… Yup!

- Can now hit 800MB/s from a VM
- Don't really get more than 1-5% IO wait