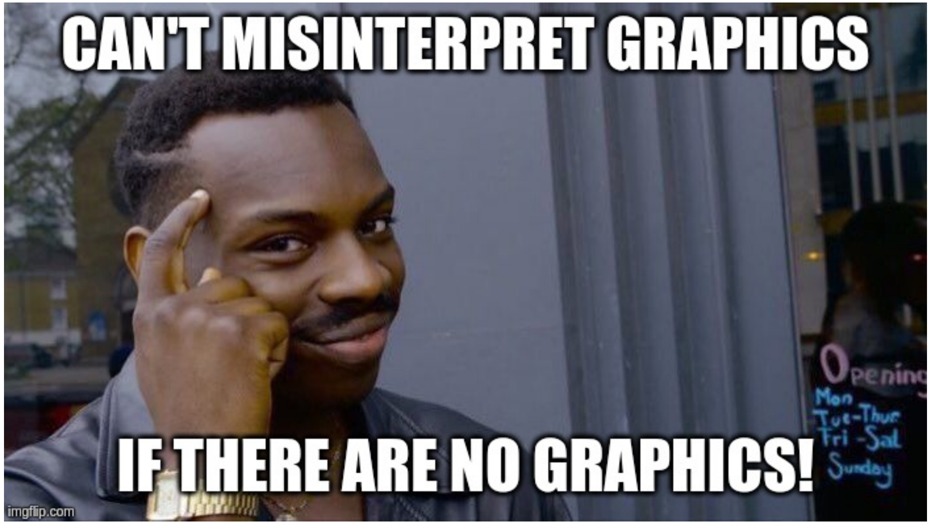
A Tale of Two Clusters

CernVM-FS and CephFS in Context

Jesse Williamson, CernVM Workshop 2019

A Tale of Two Clusters



CernVM-FS

CernVM-FS

- Developed at CERN primarily for the task of disseminating HEP software

- Emphasis on many readers of the *same data*

- Environments where HTTP is the most suitable transport

- > 1 billion files in management, millions in a single directory
    - many directories
    - distributed across ~100,000 clients

CernVM-FS

- CVMFS is very good at handling *wide-area* replication of data
    - under the hood, Merkel trees to provide *stratums*

- Many caching opportunities and optimizations for read-only data that changes comparatively infrequently
    - with CVMFS, you don’t *write*, you *publish* (via cvmfs-gateway)

- (traditionally) single-publisher model with FS *view*

Ceph

Ceph

- developed at University of Santa Cruz
  - Weil, Brandt, Miller, Long, Maltzahn. "Ceph: a scalable, high-performance distributed file system". OSDI '06 2006.

- ~98K commits, 750 contributors (GitHub, 06/2019)

- part of OpenStack

Ceph

- Ceph provides a *foundation* for distributed storage, beginning with a virtual storage abstraction of physical storage topology called RADOS, mapping storage into "pools" using CRUSH to distribute data by computing where to store Ceph objects across the virtual projection.

RADOS: Reliable Autonomic Distributed Object Store
CRUSH: Controlled Replication Under Scalable Hashing

Ceph

- Server damons themselves implement RADOS (OSDs and MONs), providing scaling, metadata management, and allows *different views* of a Ceph cluster to be provided.

    - librados: C++ and C, bindings for Python, etc.
    - RBD: cluster as block storage
    - RGW: S3, Swift
    - CephFS: POSIX filesystem (FUSE and kernel module)

CephFS

CephFS

- FUSE and kernel module

- POSIX fileystem view, atomic operations

- MDS servers: metadata caching and synchronization

- clients "open files" via the MDS
    - read and write operations scale linearly with # of OSDs

- transactional writes: ACID, isolated by OSDs

ACID: Atomicity, Consistency, Isolation, Durability
MDS: MetaData Server
OSD: Object Storage Daemon

CephFS

- Multiple MDS servers allow scaling read/write operations:

    - directory fragmentation allows splitting (partitioning) of a directory's metadata across multiple servers

    - subtree pinning: "pins" a subtree to a particular MDS server rank

CephFS

- FileStore:
    - legacy back-end, built over a regular filesystem (e.g. xfs)

- BlueStore:
    - current back end (since Luminous, August 2017)
    - motivation:
        - no way to provide transactional rollback via POSIX;
        - POSIX readdir() is not ordered, building dir trees expensive;
        - double-journalling caused negative effects on throughput;

Weil, Sage. "BlueStore: A New, Faster Storage Backend for Ceph". Vault (conference) 2016.
https://twitter.com/liewegas/status/725429304117497856

CephFS / BlueStore

- directly manages storage device (partitions or block devices, no FS)

- specifically designed for OSDs: ~2x write performance improvement
    - “https://ceph.com/community/new-luminous-bluestore/”

- efficient copy-on-write: cloning for snapshots and erasure-coded pools

- checksumming, compression, etc.

- tools: ceph-bluestore-tool, etc.

CephFS / BlueStore

- embedded RocksDB kv store for to handle metadata (e.g. mapping object names to block locations)

- includes a RocksDB environment (BlueRocksEnv) over a
"mini filesystem" used internally, BlueFS

- BlueStore can (as of Nautilis) show fine-grained disk usage information

Ceph @CERN

- (at least) 8 production clusters, ~17PB
  - using RBD (Ceph block storage)

- Collet, van der Ster, Cameselle, Lamanna. CERN IT-ST, 2019.
  "https://per3s.sciencesconf.org/data/pages/2019_per3s_jcollet.pdf"
    - interesting performance analysis
    - colors! diagrams! ;-)

- Dan van der Ster and Teo Mouratidis. Cephalocon 2019:
  "Ceph Operations at CERN: Where Do We Go From Here?”

Ontological Interlude

Ontological Interlude

- to understand ontology, you must first understand ontology
    - multiple papers spend *pages* on this

Ontological Interlude

- to understand ontology, you must first understand ontology
    - multiple papers spend *pages* on this

- the field that answers questions of being or existence

- the ontology of a given *domain* describes the constituents of that reality in a systematic way

- natural *kinds* like crows and cows are distinct

- artifactual *kinds* like cups and char*-s are distinct

Ontological Interlude

- See your nearest philosopher!

...who may be closer than you think, but first…

Do philosophers exist..?

Ontological Interlude

Do mountains exist?

Ontological Interlude

Do mountains exist?

...well, *of course*, that’s silly!

Ontological Interlude

Do mountains exist?

...well, *of course*, that’s silly!

...if you don’t think mountains exist, **try ignoring one**!

Ontological Interlude

What about the “foot” of a mountain?

Ontological Interlude

What about the “foot” of a mountain?

Of course it exists-- everyone knows that!

Ontological Interlude

What about the “foot” of a mountain?

Of course it exists-- everyone knows that!

...but *where is it*..?

Smith and Mark, “Do Mountains Exist?”. https://doi.org/10.1068/b12821

Ontological Interlude

- a mountain is a kind of *locality*: it reflects human perception

- an etymologist studying ladybugs needs “mountains”

- the ladybugs being studied do not

- So, ontology involves not only what you *call* something, but what you *mean* by something.

Smith and Mark, “Do Mountains Exist?”. https://doi.org/10.1068/b12821

Distributed Filesystems

...do exist-- but who knows what they *look like*?

- strongly and weakly-consistent I/O?

- how is the underlying storage treated?
    - is it persistent? (e.g. memcached vs. leveldb vs. ...)

- what view of the storage is offered?

- how is data transported?

Distributed Filesystems

- Many considerations influence the total design of storage systems, its environment(s), its hardware, its software.

- Omnipresent pathological situations-- scaling directories with many files, active updates, use as home directories, etc.

- CERN and other demanding environments will always require special considerations.

- One size will never fit all: always understand how a given storage solution sees the world, and **consider it in context**.

CernVM-FS might be right when...

- You want a central authority to publish data

- You have many files to distribute identical copies of

- HTTP is the preferred transport across your enviornment

- You want to distribute software/data globally

- Comparative operational simplicity is important

CephFS might be right when...

- Your needs model “a filesystem”

- Control of topology is important

- You need strong consistency, and plan on frequent writes or updates from multiple clients

- You want flexible ways of handling physical storage

- You want commercial support

Some Resources: CVMFS

https://cernvm.cern.ch/portal/filesystem

https://github.com/cvmfs

Blomer, Buncic, and Fuhrmann. 2011. "CernVM-FS: Delivering Scientific Software to Globally Distributed Computing Resources". DOI: https://doi.org/10.1145/2110217.2110225

Some Further Resources: Ceph

https://github.com/ceph
https://ceph.com/resources/

Weil, Leung, Brandt, and Maltzahn. 2007. “RADOS: a scalable, reliable storage service for petabyte-scale storage clusters.” DOI: https://doi.org/10.1145/1374596.1374606

Weil, Brandt, Miller, and Maltzahn. 2006. “CRUSH: Controlled, Scalable, Decentralized Placement of Replicated Data.” DOI: 10.1109/SC.2006.19

Weil, Sage. 2017. “Bluestore: A New Storage Backend for Ceph – one year in”. https://events.static.linuxfound.org/sites/events/files/slides/20170323%20bluestore.pdf

Fisk, Nick. "Mastering Ceph", Packt Publishing 2019. ISBN-10: 1789610702

CVMFS and CephFS at CERN

https://cern.service-now.com/service-portal/service-element.do?name=Ceph-Service

https://ceph.com/community/new-luminous-scalability/

“Characterization of OSD performance in a Ceph cluster”
Collet, van der Ster, Cameselle, Lamanna. CERN IT-ST, Per3S 2019.
https://per3s.sciencesconf.org/data/pages/2019_per3s_jcollet.pdf

"Ceph Operations at CERN: Where Do We Go From Here?"
Dan van der Ster and Teo Mouratidis. Cephalocon 2019.

“Evolution of a CernVM-FS Infrastructure at CERN”
Enrico Bocchi, CernVM Workshop 2019.
https://indico.cern.ch/event/757415/contributions/3421573/

And, of course: https://cernvm.cern.ch/

A Tale of Two Clusters

THANK YOU!

Questions?

Jesse Williamson / nerd dot cpp at gmail dot com