Ceph

scalable, unified storage
files, blocks & objects

Tommi Virtanen / @tv / DreamHost

OpenStack Conference 2011-10-07

Storage system

Open Source

LPGL2
no copyright assignment

Incubated by DreamHost

started by Sage Weil at UC Santa Cruz,
research group partially funded by tri-labs

50+ contributors

around the world

Commodity hardware

No SPoF

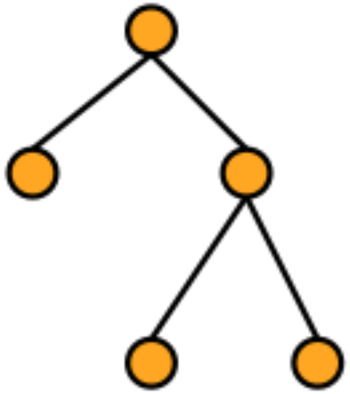
No bottlenecks

Smart storage

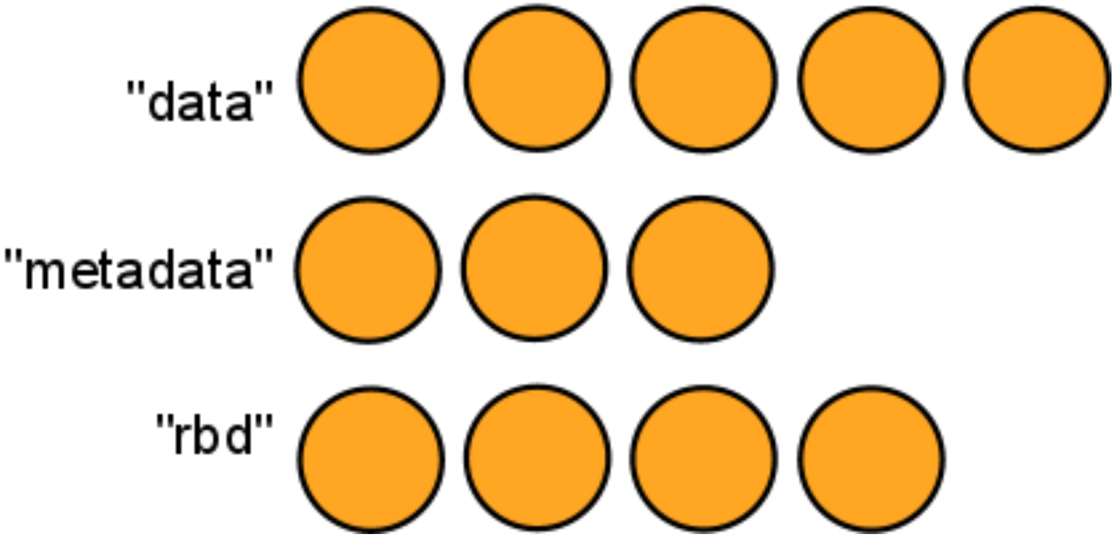
peers detect, gossip, heal

Block devices

Distributed file system



Objects

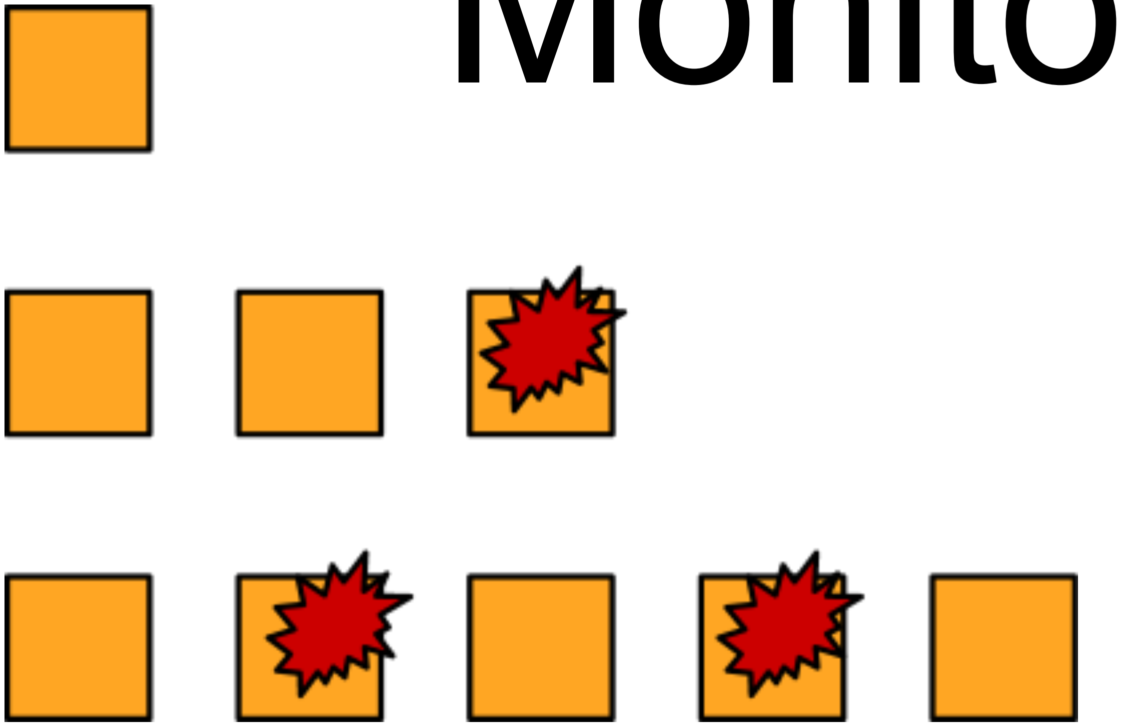


"data"

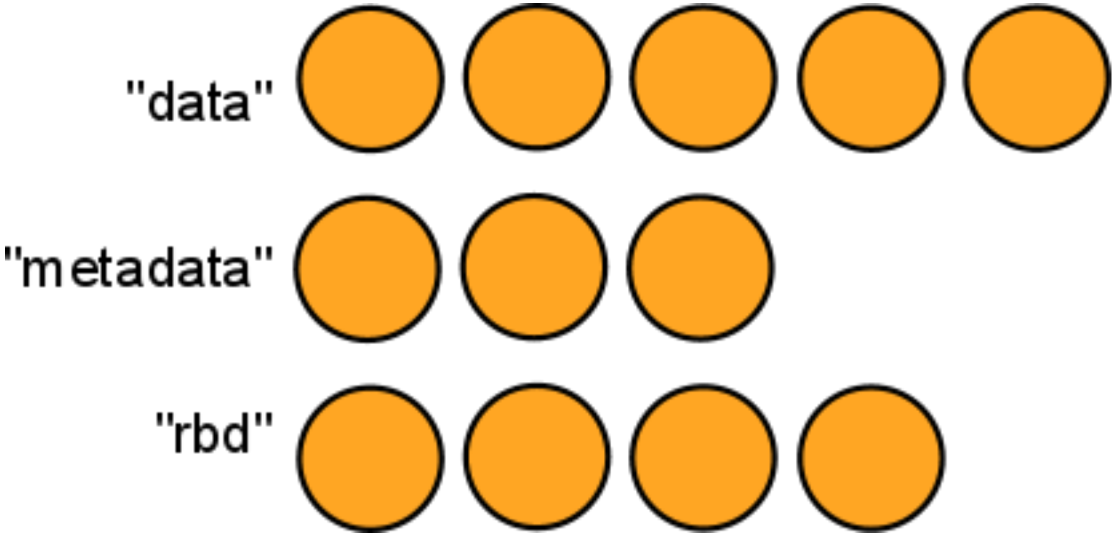
"metadata"

"rbd"

Monitors



Object storage



pool, name

→ data (bytes),

metadata: key=value, k2=v2, ...

librados (C)

libradospp (C++)

Python

PHP

your favorite language here

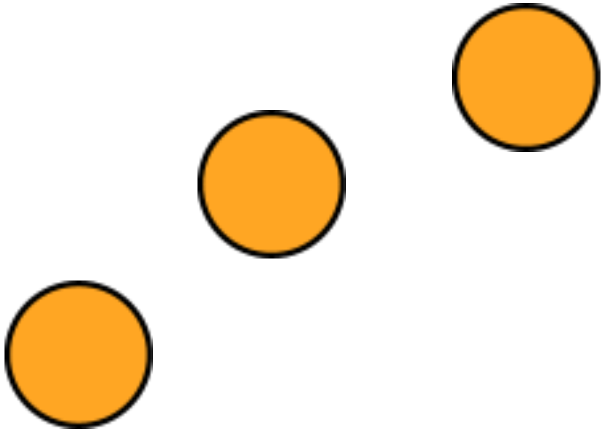
Smart client

talk to the cluster, not to a gateway
compound operations
choose your consistency (ack/commit)

Pools

replica count,
access control,
placement rules,
...

zone
row
rack
host
disk



CRUSH

deterministic placement algorithm
no lookup tables for placement
DC topology and health as input
balances at scale

Autonomous

others say: expect failure
we say: expect balancing
failure, expansion, replica count, ...

btrfs / ext4 / xfs / *

really, anything with xattrs
btrfs is an optimization
can migrate one disk at a time

process per X

X = disk, RAID set, directory

tradeoff: RAM & CPU vs fault isolation

RADOS gateway

adds users, per-object access control
HTTP, REST, looks like S3 and Swift

i <3 boto

use any s3 client
just a different hostname
we'll publish patches & guides

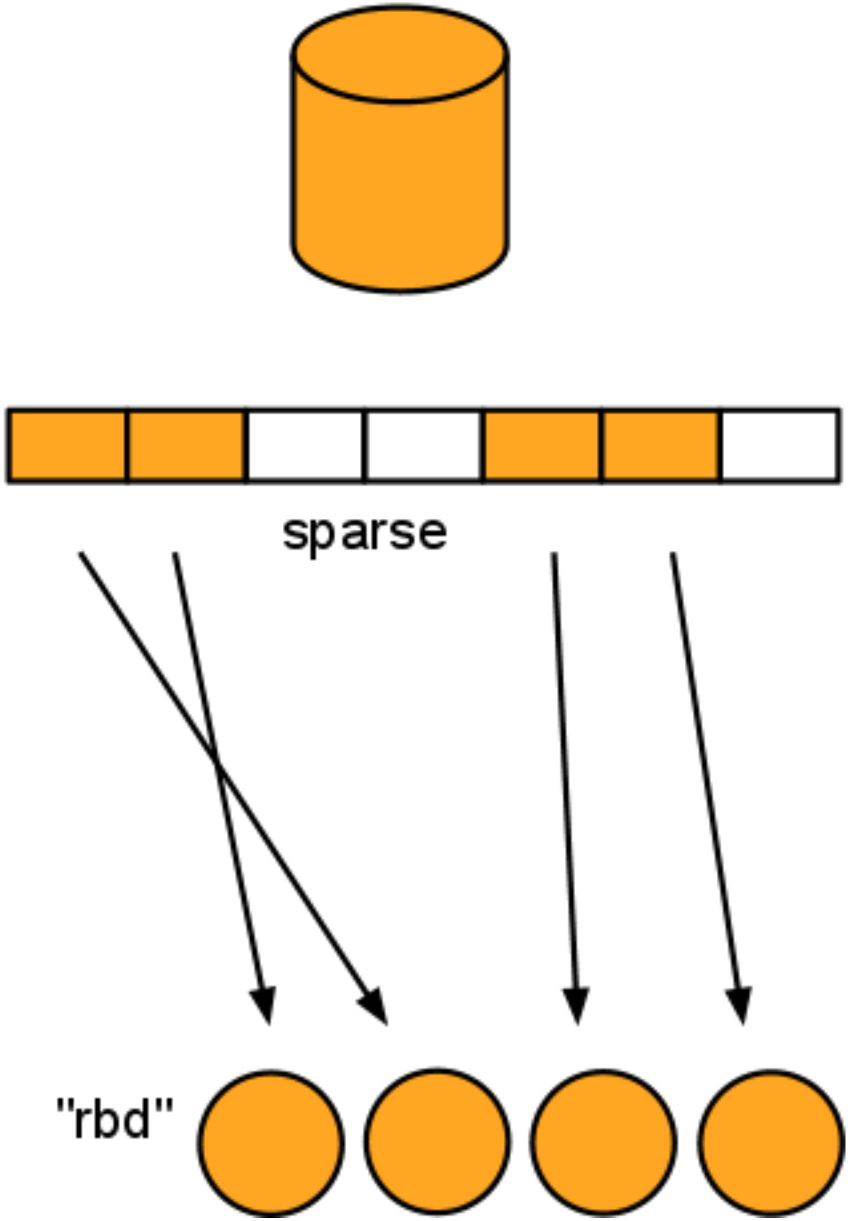
RBD

RADOS Block Device

Block device

Chunks
(4MB, configurable)

Objects



Live migration

one-line patch to libvirt
don't assume everything is a filename

Snapshots

cheap, fast

rbd create *mypool/myimage@mysnap*

Copy on Write

layering aka base image
soon

/dev/rbd0
/dev/rbd/*

rbd map *imagename*

QEmu/KVM driver

no root needed
shorter codepath

Ceph Distributed Filesystem

mount -t ceph

or FUSE

High Performance Computing

libcephfs

no need to mount, no FUSE
no root access needed
also from Java etc
Samba, NFS etc gateways

Hadoop shim

replaces HDFS,
avoids NameNode and DataNode

devops
devops
devops

Chef cookbooks

Open Source
on Github
soon

Barclamp

Open Source
on Github
soon

devving to help ops

new store node
hard drive replacement
docs, polish, QA

Questions?

ceph.newdream.net

github.com/NewDreamNetwork

tommi.virtanen@dreamhost.com

P.S. we're hiring!

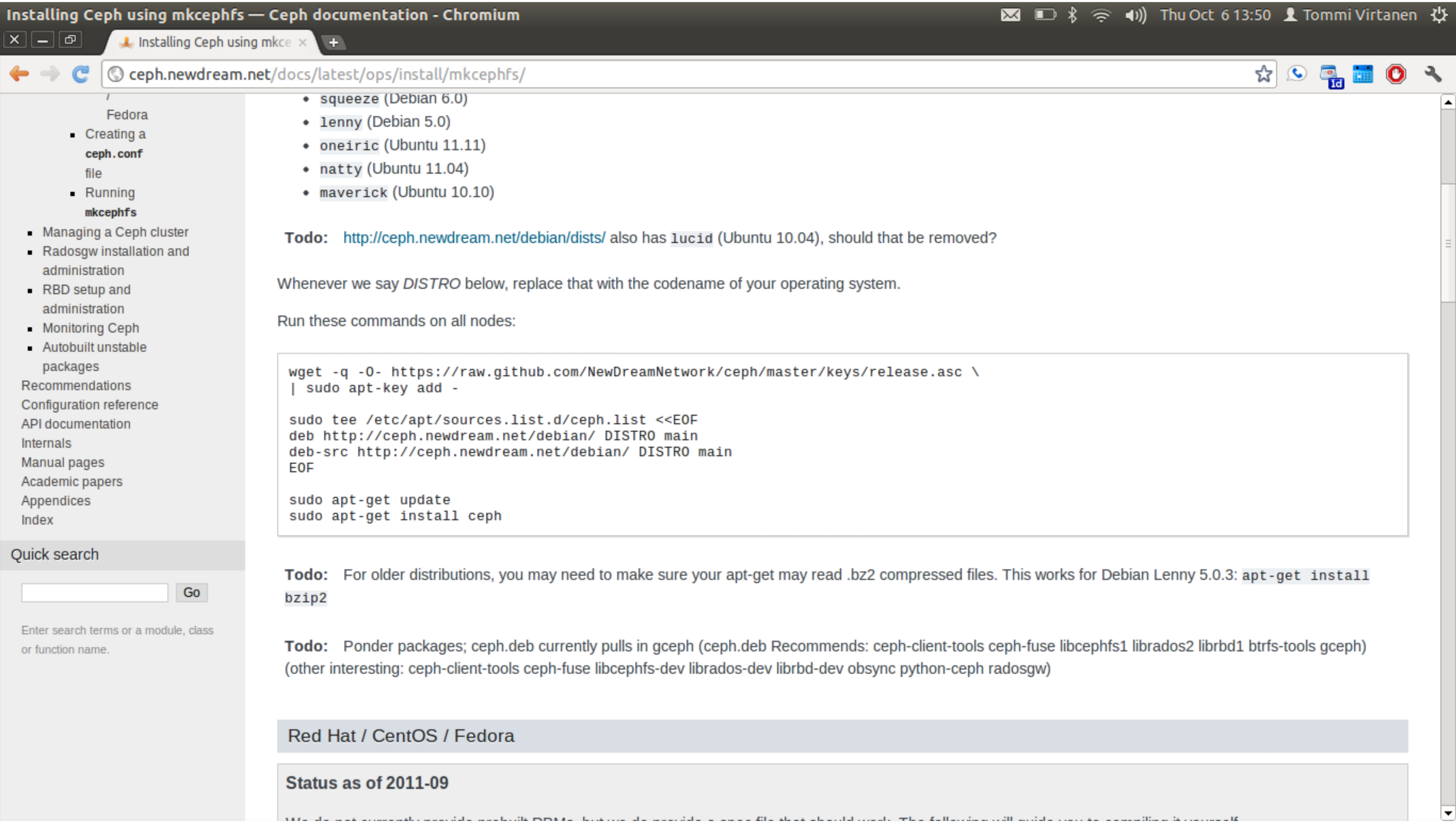
Bonus round

Want iSCSI?

export an RBD
potential SPoF & bottleneck
not a good match for core Ceph
your product here

s3-tests

unofficial S3 compliance test suite
run against AWS, codify responses


Installing Ceph using mkcephfs — Ceph documentation - Chromium
Thu Oct 6 13:50
Tommi Virtanen
Installing Ceph using mkce
ceph.newdream.net/docs/latest/ops/install/mkcephfs/
Fedora
Creating a ceph.conf file
Running mkcephfs
Managing a Ceph cluster
Radosgw installation and administration
RBD setup and administration
Monitoring Ceph
Autobuilt unstable packages
Recommendations
Configuration reference
API documentation
Internals
Manual pages
Academic papers
Appendices
Index
Quick search
Go
Enter search terms or a module, class or function name.

- `squeeze` (Debian 6.0)
- `lenny` (Debian 5.0)
- `oneiric` (Ubuntu 11.11)
- `natty` (Ubuntu 11.04)
- `maverick` (Ubuntu 10.10)

Todo: http://ceph.newdream.net/debian/dists/ also has `lucid` (Ubuntu 10.04), should that be removed?

Whenever we say *DISTRO* below, replace that with the codename of your operating system.

Run these commands on all nodes:

```
wget -q -O- https://raw.github.com/NewDreamNetwork/ceph/master/keys/release.asc \
| sudo apt-key add -

sudo tee /etc/apt/sources.list.d/ceph.list <<EOF
deb http://ceph.newdream.net/debian/ DISTRO main
deb-src http://ceph.newdream.net/debian/ DISTRO main
EOF

sudo apt-get update
sudo apt-get install ceph
```

Todo: For older distributions, you may need to make sure your apt-get may read .bz2 compressed files. This works for Debian Lenny 5.0.3: `apt-get install bzip2`

Todo: Ponder packages; ceph.deb currently pulls in gceph (ceph.deb Recommends: ceph-client-tools ceph-fuse libcephfs1 librados2 librbd1 btrfs-tools gceph) (other interesting: ceph-client-tools ceph-fuse libcephfs-dev librados-dev librbd-dev obsync python-ceph radosgw)

Red Hat / CentOS / Fedora

Status as of 2011-09

We do not currently provide prebuilt RPMs, but we do provide a spec file that should work. The following will guide you to compiling it yourself.

Teuthology

study of cephalopods
multi-machine dynamic tests
Python, gevent, Paramiko

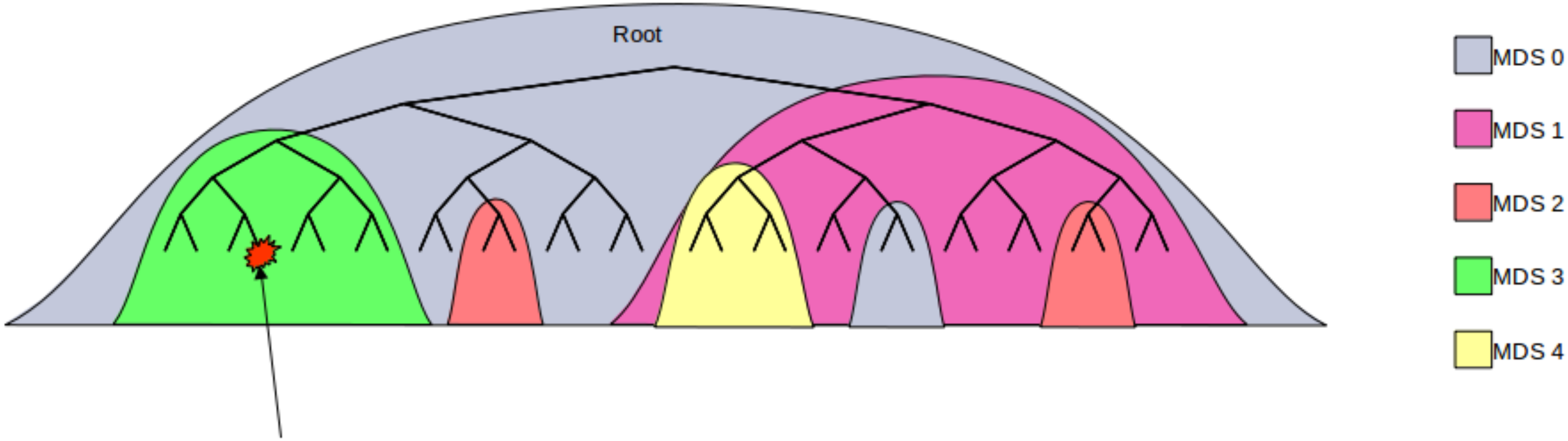
```
cluster.only('osd').run(args=['uptime'])
```

```
roles:
- [mon.0, mds.0, osd.0]
- [mon.1, osd.1]
- [mon.2, osd.2]
- [client.0]
```

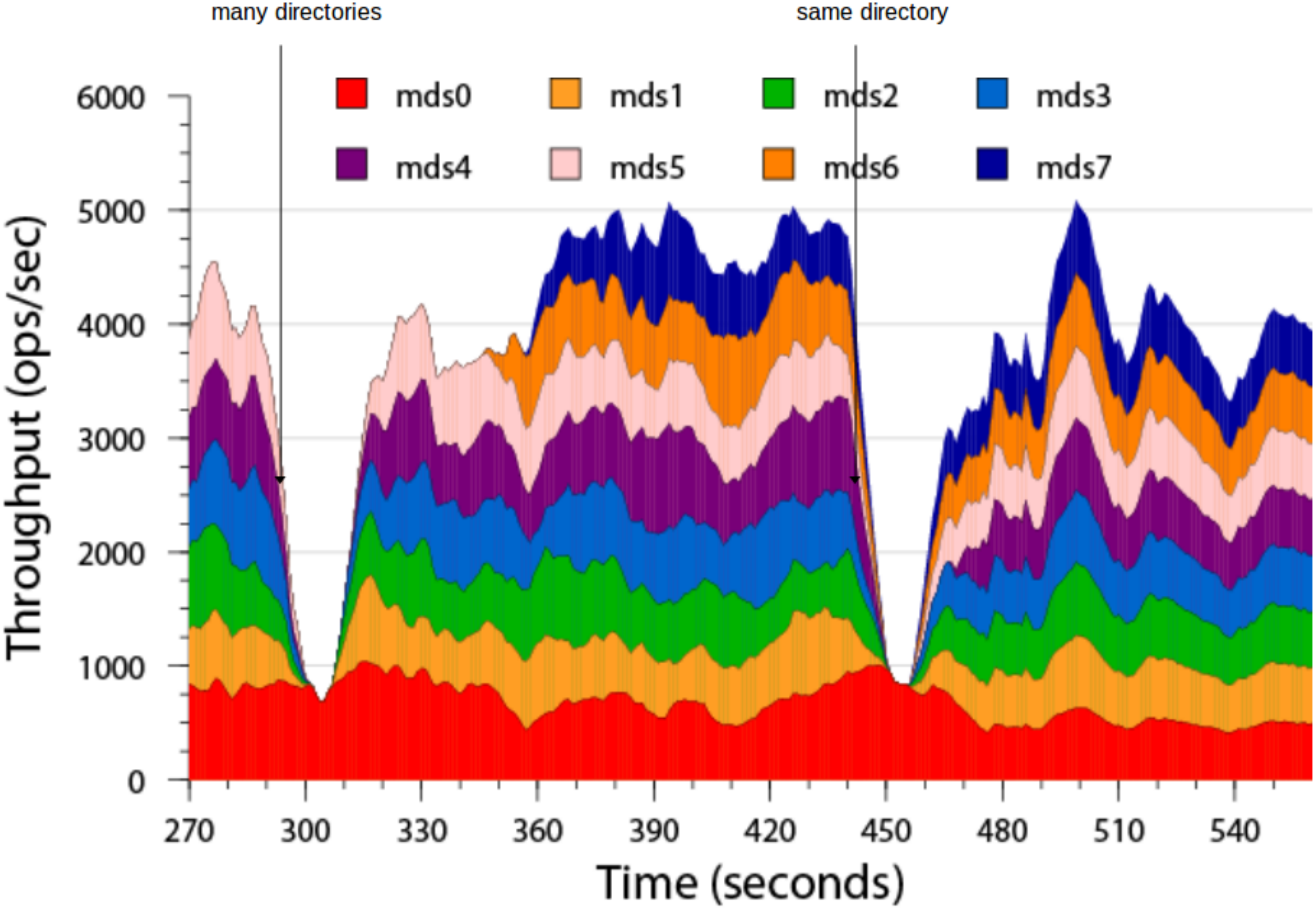
```
tasks:
- ceph:
- trashosds:
    op_delay: 1
    chance_down: 10
- kclient:
- workunit:
    all:
      - suites/bonnie.sh
```

ceph-osd plugins

SHA-1 without going over the network
update JSON object contents



Busy directory fragmented across many MDS’s


many directories
same directory
mds0
mds1
mds2
mds3
mds4
mds5
mds6
mds7
Throughput (ops/sec)
6000
5000
4000
3000
2000
1000
0
270
300
330
360
390
420
450
480
510
540
Time (seconds)

Questions?

ceph.newdream.net

github.com/NewDreamNetwork

tommi.virtanen@dreamhost.com

P.S. we're hiring!