

System Administration in a Linux Web Hosting Environment

Terri Haber

DreamHost History

- Founded in 1997 by Dallas Kashuba, Josh Jones, Michael Rodriguez and Sage Weil
- One server hidden under a friend's desk
- Owners had no money!

DreamHost Today

- Over 2,000 machines in 3 datacenters
 - 1,247 shared web service 1U and 2U servers
 - 162 shared mail machines
 - 334 shared mysql
 - Over 4,000 VPS guests on 223 hosts
 - 146 Netapps
 - 34 Sunfire X4500s
- Over 70 employees in 2 locations
- Over 780,000 domains

Why Linux?

- No Linux, no DreamHost!
- We like to do things our way!
- Easier, inexpensive to add features!
- More profit to share!
- More fun!

Software Tools

- PowerDNS
- Netsaint
- Subversion/Trac
- Servicectl
- Usagewatch
- Webpanel

PowerDNS

- BIND is great until 235,000 domains
- Uses MySQL rather than flat files to manage entries
- We use six PDNS servers (3 for backup) and 4 separate MySQL machines for DNS

Netsaint

- Makes 24/7 monitoring easier
- Sends emails to a mailing list
- Alerts active watcher to problems
- Sorts multiple action items by severity

Subversion/Trac

- Recently converted from CVS
- Wrapper script takes comments from developer, sends changes and comments to an internal mailing list.
- Moving from using in-house bug tracker to Trac

Servicectl

- In-house libraries built with OO Perl
- Script matches database entries with libraries
- Simplifies apache/user configs
- Automation

Usagewatch

- Servicectl library
- Keeps track of user resource usage and progress
- Sends warnings, balances users on machines, moves problem users to virtual private servers

WebPanel

- Probably the #1 reason so many customers are so happy with us
- Power over their own accounts
 - Add users, domains
 - Edit DNS
 - Pay the bill
 - Add/edit cron jobs
 - One-click installs
 - WebFTP
 - And much more!

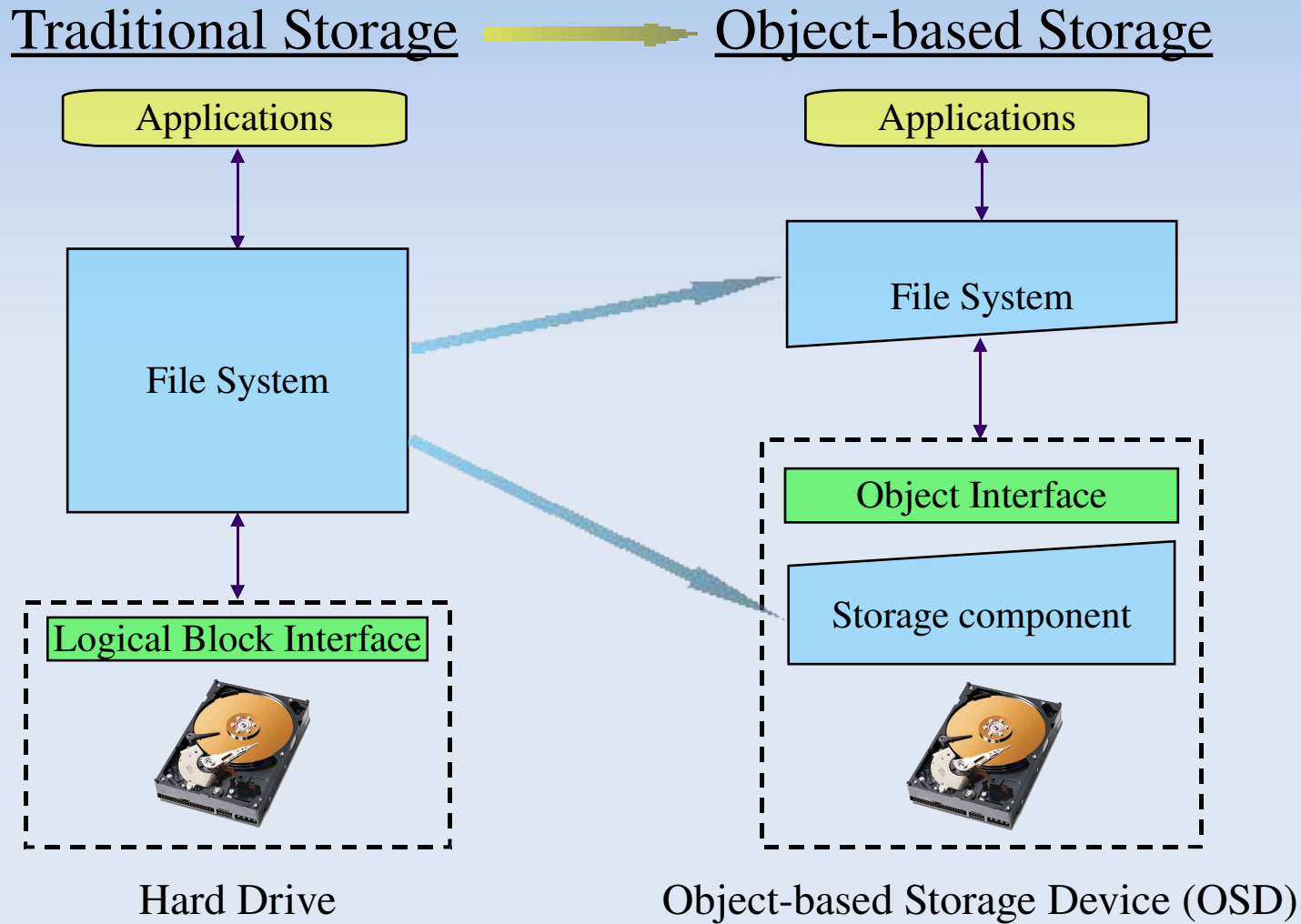
Ceph – Petascale Storage

- Distributed network file system designed to provide excellent performance, reliability, and scalability.
- Developed by Sage Weil, originally partially funded by a grant from the Department of Energy.

Advantages of Ceph

- Seamless scaling
- Strong reliability and fast recovery
- Adaptive Metadata Server (MDS)
- Intelligent disks (OSDs)

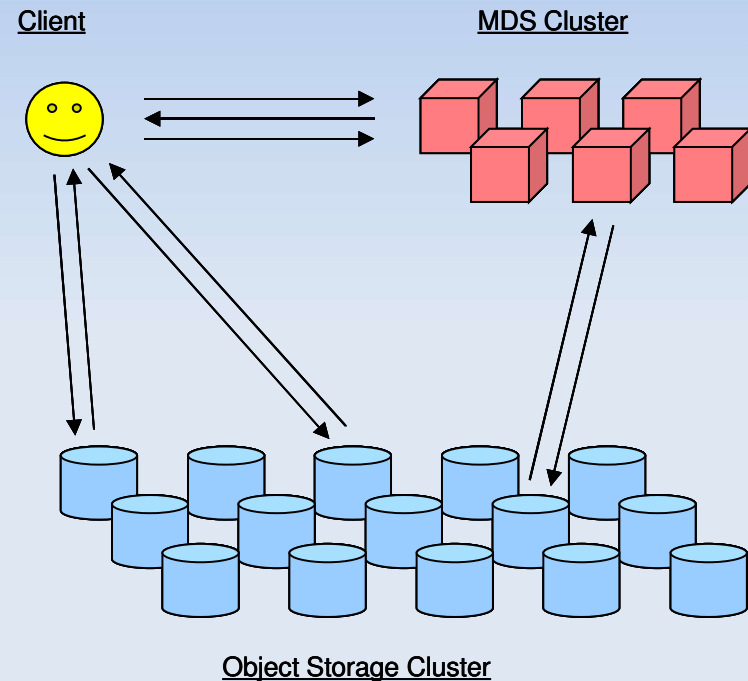
Ceph: Object-Based Storage



Ceph: A simple example

- `fd=open("/foo/bar",O_RDONLY);`
 1. Client: requests open from MDS
 2. MDS: reads directory "/foo" from OSDs
 3. MDS: issues "capability" for "/foo/bar"
- `read(fd,buf,10000);`
 1. Client: calculates name(s) and location(s) of data object(s)
 2. Client: reads data from OSDs
- `close(fd);`
 1. Client: relinquishes capability to MDS

- MDS stays out of I/O path
- Client calculates data location instead of looking it up

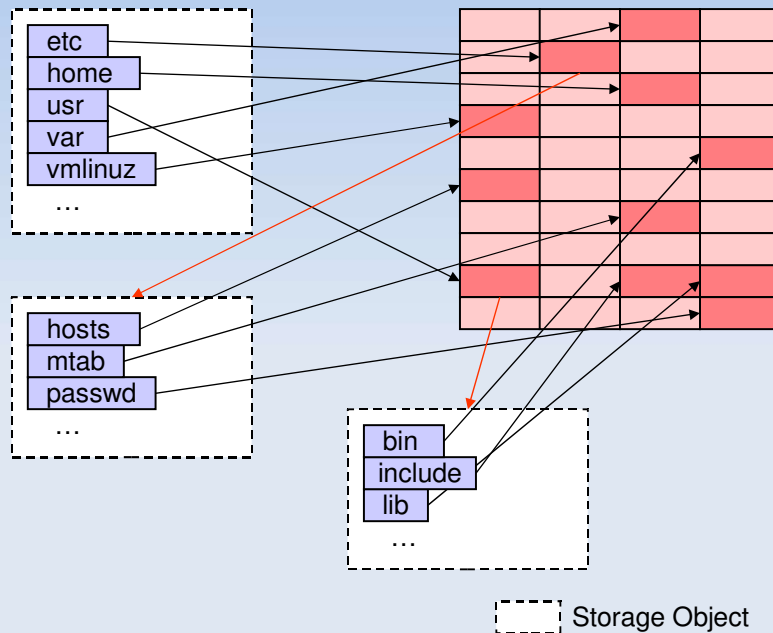


Ceph Metadata

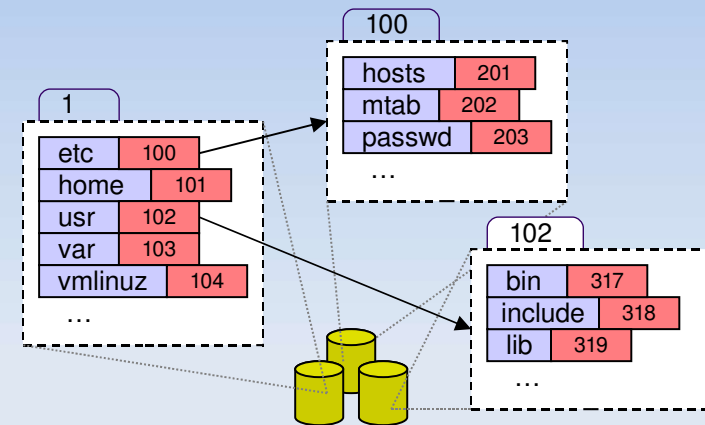
- Conventionally
 - Directory contents (filenames)
 - Per-file metadata (inodes)
 - Ownership, permissions
 - File size
 - Block list—where the data is stored on disk
- In Ceph
 - We can eliminate block lists
 - Objects
 - Robust data distribution function
 - Inodes are small and simple

Ceph Metadata Storage

Conventional Approach



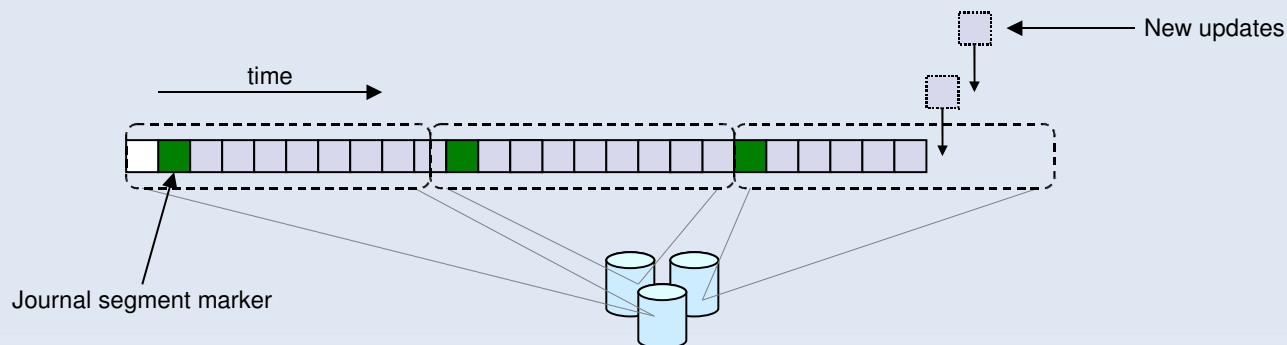
Embedded Inodes



- Embed inodes inside directories
 - Store with the directory entry (filename)
 - Avoid random access to inode tables, which degrades performance
 - Good prefetching when locality is present in workloads
- Each directory stored as an object

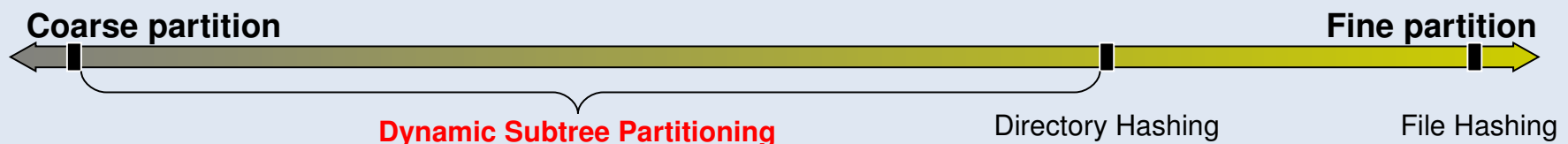
Ceph Journaling

- Metadata updates streamed to a journal
 - Striped over large objects
 - Efficient, sequential writes
 - Journal grows very large (hundreds of megabytes)
 - Many updates combined into small number of directory updates
 - Per-segment dirty list, making trimming very efficient
- Enable failure recovery
- Efficient I/O to object storage subsystem

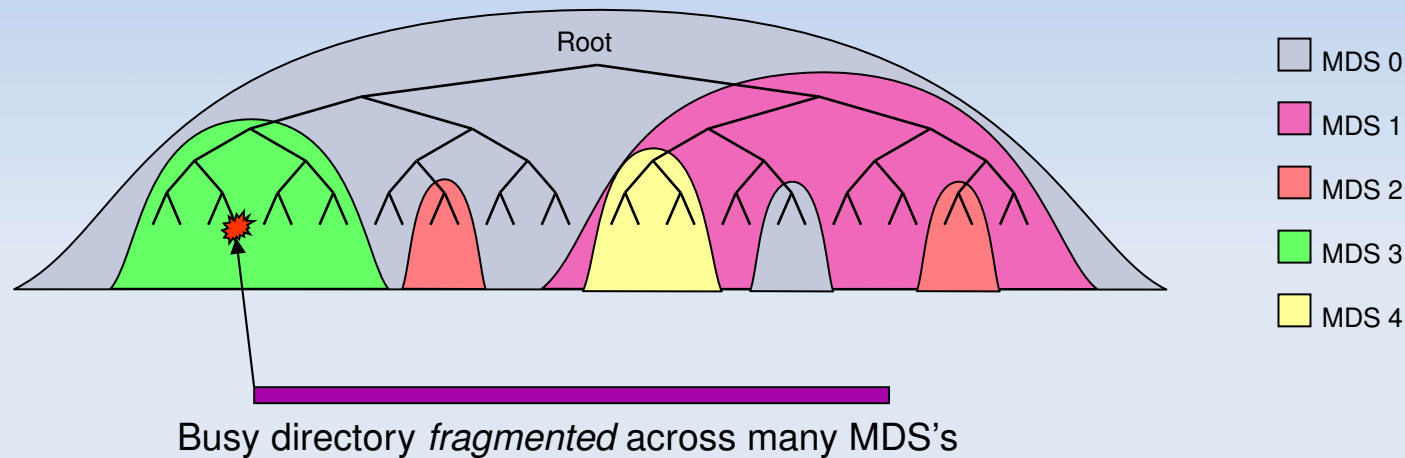


Ceph Partitioning

- Adaptive Partitioning
- Dynamically distribute arbitrary subtrees of the hierarchy
 - Coarse partition preserves locality
 - Adapt distribution to keep workload balanced
 - Migrate subtrees between MDSs as workload changes
- Adapt distribution to cope with hot spots
 - Heavily read directories replicated on multiple MDSs
 - Large or heavily written directories are *fragmented* for load distribution and storage



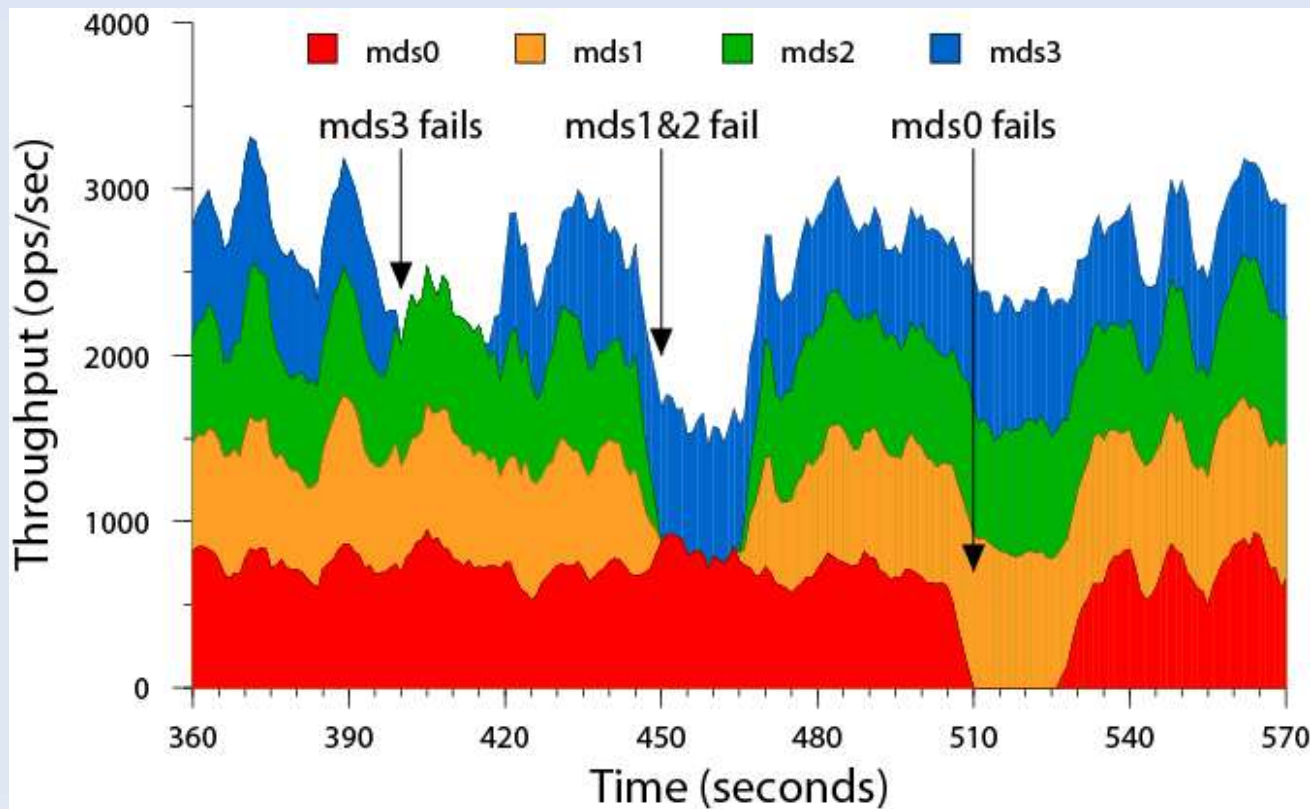
Ceph Metadata Partition



- Scalability
 - Arbitrarily partitioned metadata
- Adaptability
 - Cope with workload changes over time, and hot spots

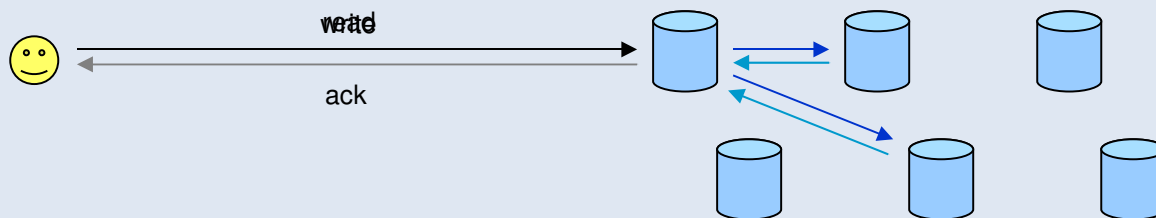
Ceph Failure Recovery

- Nodes quickly recover
 - 15 seconds—unresponsive node declared dead
 - 5 seconds—recovery
- Subtree partitioning limits effect of individual failures on rest of cluster



RADOS – Data Replication

- Each object belongs to a PG
- Each PG maps to a list of OSDs
- Clients interact with the first OSD (“primary”)
 - Reads are satisfied by the primary
 - Writes are forwarded by the primary to all replicas
 - Leverage local OSD interconnect bandwidth
 - Simplifies client protocol, replica consistency
 - Low incremental cost for replication levels > 2



Ceph Today

- Working snapshots
- Asynchronous metadata operations
- Improved threading
- Online scrubbing

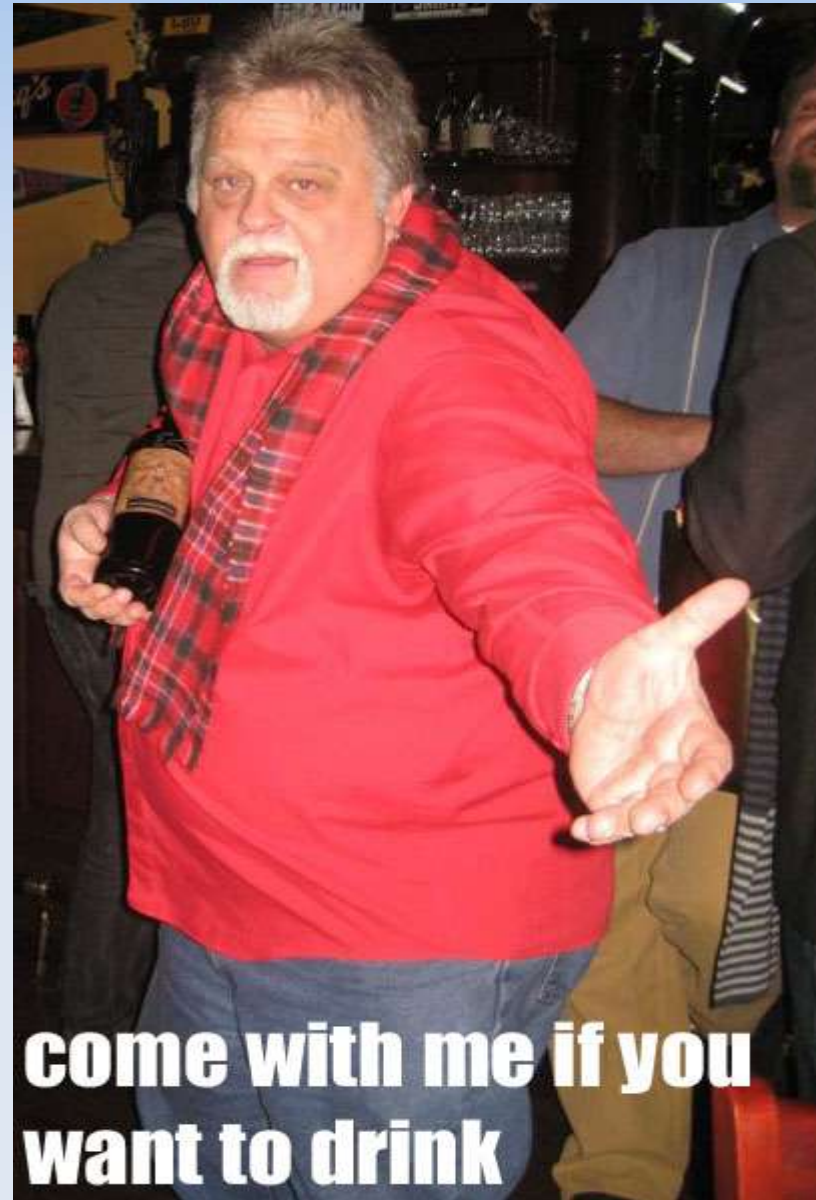
The Future of Ceph

- apt-get install ceph
- Strong, scalable security
- Stabilization

More information

- dreamhost.com
- blog.dreamhost.com
- [@dreamhost](https://twitter.com/dreamhost) on Twitter
- ceph.newdream.net
- [#ceph](https://irc.oftc.net) on irc.oftc.net

Thank you!



**come with me if you
want to drink**