LCI workshop
on high performance clustered computing

Linux Clusters Institute: OpenStack Underlying Technologies and Theory

Yale, August 13th – 17th 2018

John Michael Lowe | Senior Cloud Engineer
Indiana University
jomlowe@iu.edu

Jetstream

Underlying Technologies and Theory

- OpenStack Foundation
- Theory
  - Pets and Cattle
  - RPC vs RESTful
  - CAP Theorem
  - Message Busses
  - JSON
- Technologies
  - Ceph
  - MySQL/Galera
  - HAProxy
  - Keepalived VRRP
  - VLAN and VXLAN
  - Qemu/KVM

OpenStack Foundation

- Started as merging of Rackspace and NASA projects in 2010
- Modeled on Apache Foundation
- Open, transparent, democratic
- Scientific Computing SIG

Theory

Pets and Cattle

Cows, not pets: pets take great amount of care, feeding, and you name them; cows you intend to have high turnover and you give them numbers.

Image source unknown

RPC and RESTful interfaces

- Remote Procedure Calls
  - distributed computing using function call semantics from procedural languages
  - Origins go back to 1970's ARPANET
  - Difficult to scale, nonstandard wire protocol
- REpresentational State Transfer
  - Co-designed with HTTP 1.1
  - Simple operations that match HTTP methods, errors meaningful
  - Off the shelf load balancing and proxying allow scaling
  - Stateless protocol
  - Platform and language independent

CAP Theorem

- Consistent, Available, Partition Tolerant (Distributed)
- Pick any two!
- These properties govern the design and performance of databases and storage systems

Message Bus

- **RabbitMQ**, ZeroMQ, AMQP, STOMP, etc.
- Messages are passed one to one or one to many containing data structures to issue commands or report information
- RabbitMQ supports multi-master clustering
- Persistent message support requires good small file performance and relatively high IOPS
- In the worst case, an Etch-A-Sketch event is survivable because all the queues will be recreated

JSON

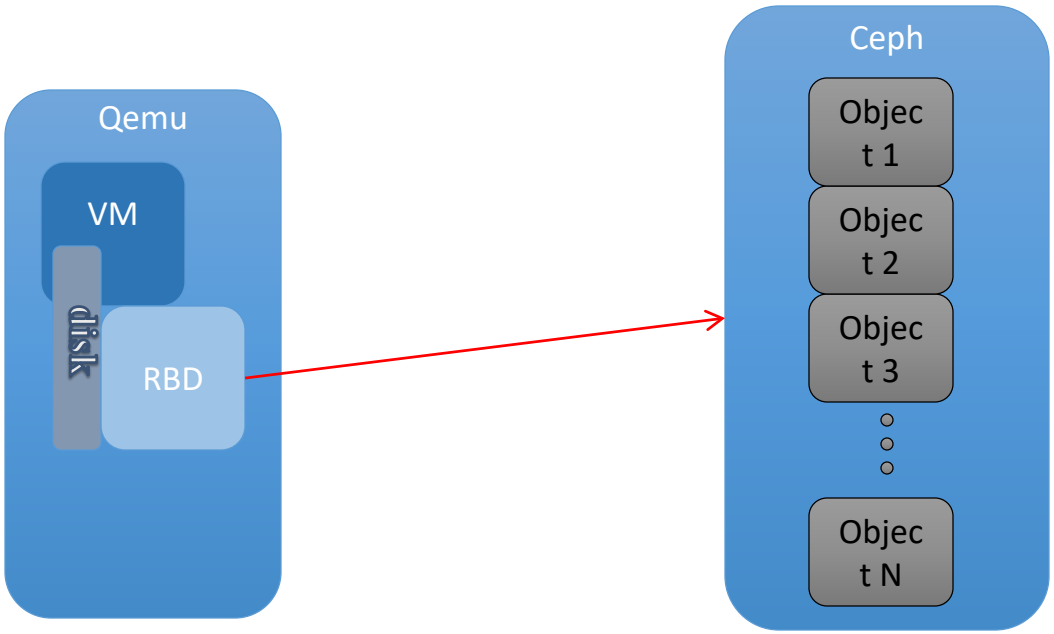
- Java Script Object Notation (no Java Script required)
- Remarkably similar to Python
- All data passed via REST or message bus is encoded this way

Technologies

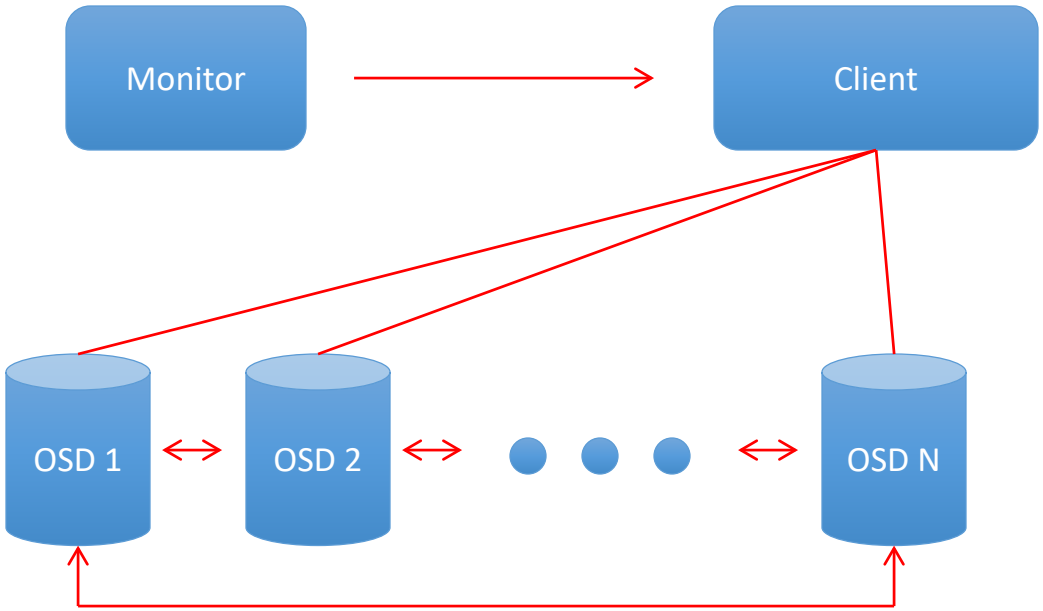
Ceph

- Object storage system with replicas or erasure coding
- Object location is calculated by client using deterministic algorithm based on current cluster version
- Block devices emulated by reading/writing bock sized amounts from objects with offsets
- Can emulate posix semantics by using objects as as inodes and striping files across objects, Lustre like
- Object semantics make Copy On Write easy, low cost instant cloning

Ceph



Ceph



MySQL Galera

- Galera is a plugin for MySQL that allows synchronous multi-master replication
- Must have quorum
- Every OpenStack service has at least one database

HAProxy

- Listens on a port and proxies traffic to some number of backends
- Can do simple TCP all the way up to sophisticated HTTP header operations
- Monitors health and connection counts to backends and front end listeners
- Can terminate SSL/TLS

KeepAliveD

- Uses a back channel to send heartbeats
- When heartbeats are missed it will take over an IP address
- Much simpler and safer than Corosync/Pacemaker
- Used internally by OpenStack network services

VLAN and VXLAN

- VLAN
  - uses 12 bits in a standard ethernet frame
  - up to 4096 virtual layer 2 networks
  - Widely supported
- VXLAN
  - wraps Layer2 in a UDP packet using 24 bits
  - 16M network IDs
  - reduces wrapped network MTU by 50 bytes
  - Emulates broadcast with multicast

VLAN and VXLAN

- VLAN
  - simple and supported by nearly every network
  - requires interaction with switching gear
  - Relatively limited number of networks can be created
- VXLAN
  - much more flexible
  - more networks
  - can be routed on Layer 3
  - no interaction with switching gear
  - needs working multicast

Qemu/KVM

- Qemu emulator, KVM kernel module
- Requires no guest modification
- Very fast when used with paravirtualization, KVM, and VT-x
- Live migration, suspend to disk, native Ceph client supported

Resources

- OpenStack https://www.openstack.org/software/
- Ceph http://docs.ceph.com/docs/master/
- RabbitMQ https://www.rabbitmq.com/
- MySQL/Galera https://www.percona.com/software/mysql-database/percona-xtradb-cluster
- VXLAN https://tools.ietf.org/html/rfc7348

Questions