Architecting Ceph Solutions

TUT1234

David Byte, Sr. Technology Strategist
Agenda

Discuss SUSE goals, process, and artifacts

Discuss some key considerations in designing a solution

Rules of thumb
SUSE Goals

General:
Enable enterprise customers to effectively leverage open source technologies in ways that benefit their business.

Storage Specific:
Provide step-by-step guides to facilitate implementation of Ceph technologies in enterprise environments.
SUSE Solution Designs

The goal for storage designs is to create the building blocks

Hardware implementation guide

+ 

SUSE Product(s)

+ 

Application integration guides
Hardware Guides for SUSE Products

Implementation Guides

- **Hardware settings**
  - Including screenshots
  - Storage controller settings
  - Firmware

- **Documented process**
  - Network design
  - OS install
  - SUSE Enterprise Storage install

- **Performance baseline where applicable**
Document beginning state including pre-reqs e.g. Gateways required

Work done to understand the I/O patterns and recommend proper configuration

Discussion of storage options and how they affect software implementation

Screenshots and step-by-step implementation process
The Process

- Work with partner to define solution design
- Get access to hardware
- Work through hardware config and software config taking screen shots & copious notes
- Write the doc
- Walk through through the document making corrections
- Publish
- Periodically, review and update
So… Where are the guides?
Architecting Clusters
Understanding Storage Needs

- Capacity
- Performance Requirements
- I/O Patterns
- Data Protection
- Client Access Methods
Understand Ceph Architecture

Network
- RBD (iSCSI)
- Block Devices

Applications
- S3 SWIFT
- Object Storage

File Share
- CephFS
- File Interface

RADOS
(Common Object Store)

Cluster
OSD
Storage Server
OSD
Storage Server
OSD
Storage Server
OSD
Storage Server
OSD
Storage Server
OSD
Storage Server
OSD
Storage Server
OSD
Storage Server
OSD
Storage Server
OSD
Storage Server
OSD
Storage Server
OSD
Storage Server
OSD
Storage Server
OSD
Storage Server
OSD
Storage Server

Monitors
HCM
Server

HCM
Server

HCM
Server
Confused About Media Types
What Should I Choose?

Ceph clusters can use all varieties of storage

Spinning Rust or SSD?

SATA, SAS, NVMe, etc?

Ceph is designed for aggregate throughput, not low latency IOPS

   At least for today….

And for goodness sakes, don’t use consumer devices
Evils of Expanders

Bus expanders are common in more storage dense chassis

The more dense the chassis, the greater the chance that the channels get bottlenecked

Mixing device speeds on an expander is a bad idea
Net What???
Speed Matters

10Gb of front-end throughput means you need to plan 30+Gb for the cluster network.

Latency IS the enemy

- LAN
- MAN
- WAN
Differences

What is the difference between 5 10Gb connections in a bond and a single 50Gb?

Load balancing - most load balancing algorithms still limit a single stream to a single link of b/w

Cabling Complexity - fewer cables is generally better

Signaling Rate -
- 10 & 40 share the same signaling rate
- 25, 50, 100 also share the same rate, 2.5 times faster than 10

$/Port
- \((\text{Switch} + \text{NIC cost} \times \text{Switch port count} + \text{Cable})/(\text{NIC port} + \text{Switch port} \times \text{Port count})\)
Topology Choices

Hub-Spoke

Ring

Mesh

Leaf & Spine

Green Spine Switch Management for Datacenter Networks - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/The-Spine-Leaf-Topology-5_fig13_305175609 [accessed 22 Mar, 2019]
Switching Mistakes

Blocking
Not enough uplink
Are they members?
Jumbo is good, usually....
Protocol Gateways

Basically undoing Ceph’s aggregated advantage
Yes, Yes, YES!

- SUSE YES certified hardware provides the best support experience

- Two Levels
  - 1: SLES YES Certification - Base level
  - 2: SES YES Certification – The cluster has been tested as a whole with some fault injection
Processors

We support 64-bit ARM & X86_64
  ● AMD, Ampere, Huawei, Intel, Marvell, etc

For spinning storage
  ● 1x 2GHz thread per device

For SSD
  ● 1-2x 2GHz threads per device

NVME
  ● 2-4x 2GHz threads per device
RAM

Default Values
- spinning=1GB cache per dev
- SSD=3GB cache per dev

Unofficial RAM sizing
- # of OSDs * (2+cache) + 16 rounded up to next logical multiple

Logical multiple = ram channels per socket * number of occupied sockets * ram chip size
Network

Always be redundant (switches and NIC ports)

Cluster network traffic = 3x public

To figure out your size, take the amount of max public throughput per node and multiply by 4.

Example:
- Expected 10Gb public traffic x 4 = 40Gb of required network
- 2x 40Gb connections (failover bond), 2x 25Gb connections (LACP), 4x 10Gb connections (LACP)
Performance of Storage Devices

7.2k SATA < 7.2k SAS < 10k SAS < SATA SSD < SAS SSD < NVME

Delivered perf in a 3x replica Luminous cluster:
7.2k SATA ~ 30MB/s per device
10k SAS ~ 45MB/s per device
SATA SSD ~120 MB/s per device

** YMMV, No Guarantees or Warranty implied or otherwise
Pulling it Together

Lots to consider when architecting a solution

Think about tomorrow as well as today

Make sure you take the workload requirements into account

If using SSDs, look at the service life and create a maintenance program

Engage with SUSE & Partner SE/SA(s)

Make sure Ceph is the right tool for the job
Q & A