Running a Small OpenStack Cluster with a Full NVMe Ceph Cluster

Kevin Honka | @piratehonk@norden.social

Ceph Days Berlin | 2025-11-12

Who am I

- Kevin Honka
- CTO @ AD IT Systems GmbH
- Doing IT related jobs for almost 20 years
- Mastodon: @piratehonk@norden.social

Who is AD IT Systems

- Small Hoster out of Nuremberg, Germany
- Specialized in shop, health and telco applications
- less than 10 employees

Questions

- Who here runs Openstack?

- Who runs Openstack with only Ceph for storage?

The beginning | 2022

- 3 Fulltime Employees
- Proxmox Cluster with Ceph Backend
  - Everything on SATA SSDs
  - 10 old Servers; some older than 10 years
  - everything on the same servers
- One person running puppet

Our Goals

- Reduce manual configuration
- Increase reliability
- make our lives easier

Comments

- Running an OpenStack Cluster with 3 People?! You need atleast 10! - Someone at Red Hat
- Running MySQL on Ceph RBD does not work, the commit latency is way too bad, and too jittery[1][2][3] - Kris Köhntopp

[1] https://blog.koehntopp.info/2022/11/07/bandwidth-iops-and-latency.html
[2] https://blog.koehntopp.info/2022/09/27/mysql-local-and-distributed-storage.html
[3] https://blog.koehntopp.info/2021/02/25/mysql-from-below.html

The new Setup

Openstack

- 3 Controllers
  - 64 Cores
  - 128 GB RAM
- 5 Hypervisors
  - 128 Cores
  - 1TB RAM
  - no local storage

Ceph

- 4 Nodes
  - 32 Cores
  - 128 GB RAM
  - 5 Intel NVMe disks at 3 TB per

Network

- 2x Juniper EX4650-48Y-AFI
- Fiber only, no Copper where possible
- 25 Gbps everywhere
- LACP for every node

Everything is easy in the beginning

- 2 Months to setup Ceph and OpenStack for “pre production” phase
  - cephadm makes Ceph easy
  - kolla-ansible makes OpenStack easyish
- Performance tests are good
- Tests with customer

until it isn't

- Performance Issues
  - CPU Load is ok
  - CPU Wait time is egregiously high
  - IOPS are great
  - IO Latency is bad
- Customer wasn't happy
- We were not happy

Diving deep / CPU Wait time

- thanks to Kris Köhntopp for the pointers
- NUMA issues with VMs

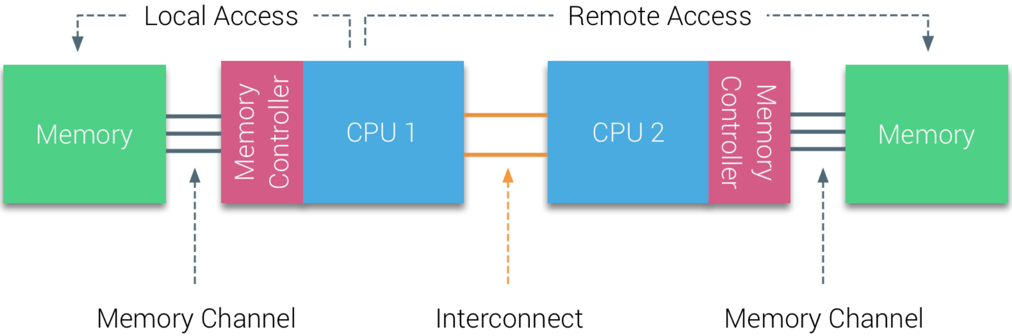

Figure 1: Copyright Frank Denneman

- libvirt has an option for NUMA awareness, it is not set by default!
- newer nodes with only one cpu

Diving deep / IO Latency

- IOPS on rbd volumes is great, more than 50k IOPS is easily achieved
  - with 4k random and 16k random writes
- network latency is ok between 0.08 - 0.09 ms
- commit lantencies are inconsistent and bad
  - between 0.1ms and 5ms

Optimizing Ceph

- Trying to tune Ceph for less jitter and better commit latencies
- range reduced to between 0.1ms and 2ms
- still way to big range
- found an old mailing list thread that talks about issues with OSDs on NVMe Storage with a size greater 1 TB
  - something about OSD Processes being single threaded
  - This was fixed in the Reef Release
  - We were still on Quincy

- Split NVMe drives into multiple OSDs with 1 TB size
  - 3 OSDs per drive
- commit latency reduces to between 0.1ms and 1ms
  - this is acceptable for our usecase

Openstack

- OpenStack upgrades are a pain, but less than in the past
- 29 Projects
- 297 VMs
- 3.7 TB RAM Used
- over 1000 vCPUs allocated

Ceph

- runs smooth
- upgrades are a piece of cake
- expanded from 60 TB RAW storage to +100 TB
  - 32 TB per Node
- reduced avg commit latency to 0.1 ms
  - 95th percentile around 0.2 ms
- In-/Egress around 250 MBps - 1.5 Gbps
- Cluster IO up to 150 Gbps
- 10k - 25k IOPS on average

The first incident

- On 2024-12-31 one of the Ceph nodes shows issues with multiple drives
- Cluster works fine, we push troubleshooting after new year
  - performance degradation, but no outage
- The node had a broken NVMe-backplane
- Technician with replacement arrives 2 days later
- Recovery runs with 50Gbps
  - restore done in under 4h!

The second incident

- End of July 2025; a node randomly drops from the network
- Still only performance degradation, no outage

- The NIC was slightly dislogded
- fans vibrate -> network goes down

Going Forward

- Ceph NVME-of as backend for Nova
- Ceph RGW as backend for Swift
- Maybe CephFS as backend for Manila
- Maybe local storage for high performance Databases

THANK YOU

To everyone that makes Ceph awesome.

Questions?

Let’s talk later