СЛЕРМ

Southbridge

Что такое Ceph

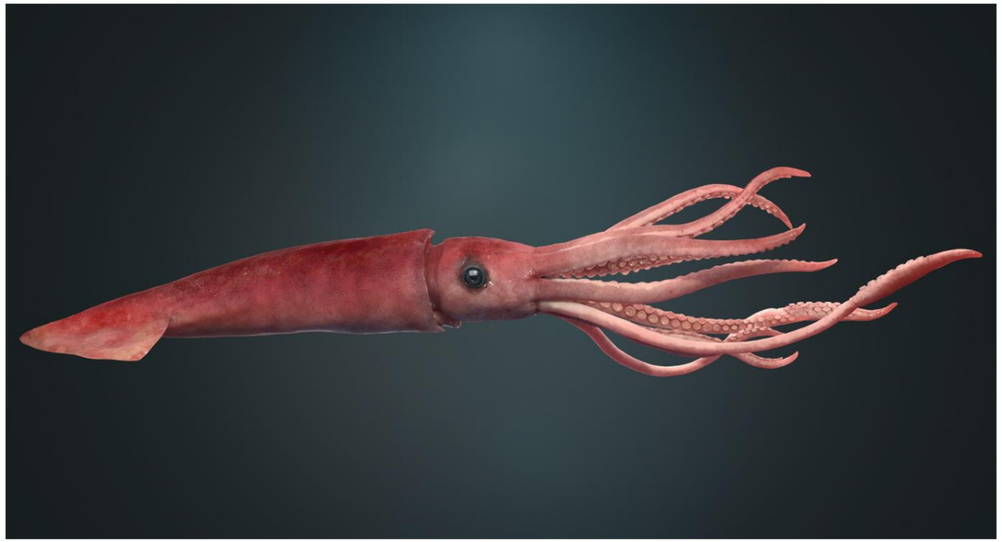
Виталий Филиппов

Виталий Филиппов

Разработчик-эксперт в компании CUSTIS, линуксоид. Занимаюсь разработкой на разных языках от node.js до C++, сильно упоролся по Ceph-у. :)

Автор статьи «Производительность Ceph»

Что такое Ceph и чем он не является



1. Перед использованием любой технологии надо понимать её ограничения.
   Ceph — не исключение.

2. Назван Ceph в честь головоногих (лат. Cephalopoda) — авторская метафора его симметричной архитектуры

3. Ceph — программная СХД (система хранения данных)
   Английский термин: SDS (Software Defined Storage)

4. Открытая и свободная система (GPL, Free Software)

5. Есть поддержка от RedHat и SuSE, если вдруг хочется много заплатить

Программная СХД

Вкратце это означает следующее:

1. Вы берёте обычные диски, обычные сервера
2. Соединяете их обычной сетью. Почти обычной — надо чуть быстрее гигабита: 10/25/40/100 Gbit/s. Это не проблема — 10 GbE дешёвый.
3. Получаете большое единое хранилище, устойчивое к вылетам отдельных дисков, или серверов, или стоек, или даже датацентров
4. У Ceph нет единой точки отказа by design. Все компоненты кластерные
5. Но, конечно, если вы потеряете больше дисков, чем у вас реплик данных — вы потеряете данные

Объектный, файловый, блочный протоколы доступа

1. Ceph поддерживает все 3 основных вида протоколов (*вида* протоколов, т. к. сами протоколы, кроме объектного S3, свои)

2. Объектный — самый простой. Крупные объекты типа веб-статики, 2 операции: GET и PUT (целиком, случайный доступ не нужен). Стандарт: S3 (аналог сервиса Amazon).

3. Файловый — самый сложный. Локально монтируемая ФС. Должна реализовывать семантику POSIX (атомарные переименования и т. п.), иначе ПО глючит. CephFS — чуть ли не единственная честная кластерная ФС (read-write с множества машин без единой точки отказа).

4. Блочный — хранилище дисков виртуалок для кластера виртуализации, стоящего рядом (в случае с Ceph — вероятно, на основе KVM).

Очень-очень много дополнительных фич

1. Снапшоты нескольких видов, Copy-On-Write клоны
2. Гибкое управление доменом отказа (что может умереть: диск/сервер/стойка/датацентр), схемой избыточности (N копий либо Erasure Code-ы — аналог RAID5+ с произвольным числом дисков данных и чётности)
3. Вынос метаданных на SSD, cache tiering (вынос горячих данных на SSD)
4. Синхронизация между кластерами/ДЦ (rgw multisite, mirroring)
5. Сжатие, шифрование, контрольные суммы, фоновая проверка данных
6. Несколько видов оркестрации, мониторинга, интеграция со всем, чем можно

Чему противопоставляется

1. В основном, классическим СХД (SAN, Storage Area Network)
2. Дорогой вендорский шкаф. IBM / Dell EMC / HPE / NetApp / Huawei и т.п.
3. В основе отказоустойчивости SAS диски. Основное отличие SAS дисков от SATA — наличие контактов второго Link'a, т. е. multipath — возможность подключения 1 диска к 2 контроллерам
4. Как правило, FC или iSCSI
5. Двукратное резервирование всех компонентов
6. Клоны-снапшоты и т. п. тоже у всех есть, возможностей тоже много

Ceph — серебряная пуля?

Всё то же самое, но бесплатно на дешёвом железе? Становись хостером и руби бабло?

Но нет, есть нюанс — ПРОИЗВОДИТЕЛЬНОСТЬ.

1. Честная кластеризация — высокий overhead.

2. Пример: любая серверная SSD даст вам 10000+ iops в 1 поток. Ceph по RBD не даст и 3000 — нет в мире железа, с которым он бы дал 3000.

3. Есть и другие архитектурные ограничения, но они не так критичны.

Чем Ceph не является

1. НЕ быстрое хранилище (1 локальная NVMe рвёт весь кластер, как Тузик грелку)

2. НЕ замена бэкапов (при всей отказоустойчивости — в Ceph встречаются баги, повреждающие данные. Помните про Cloudmouse)

3. НЕ система для слабого железа (ARM одноплатники — в мусорку, списанные HDD — туда же)

4. НЕ система для установки в облако или вообще на виртуалки. Только Bare-metal!

5. НЕ георасределённый кластер, НЕ система для дешёвой геосинхронизации данных

Когда же всё-таки применять Ceph

1. У вас очень много данных (хотя бы сотни терабайт), то есть действительно нужно масштабируемое хранилище

2. Вам важна надёжность хранения. Ceph не теряет данные. Без помощи админа. Почти всегда… :) — кроме ситуаций с багами.

3. Производительность не очень важна. Вы накинете дисков, а не будете пытаться выжать 100% из имеющихся.

4. Аналогично про ёмкость. 20-30% запаса для вас норма.

5. Вам нужно объектное (S3, RGW) или файловое хранилище (CephFS).

6. Под RBD — в эпоху SSD скорее нет, чем да. Зачем греть воздух?

Домашнее задание: подумайте о практических ситуациях, в которых бы применили и в которых бы НЕ применяли Ceph