Ceph In UnitedStack

朱荣泽

2013. 09. 06

UnitedStack

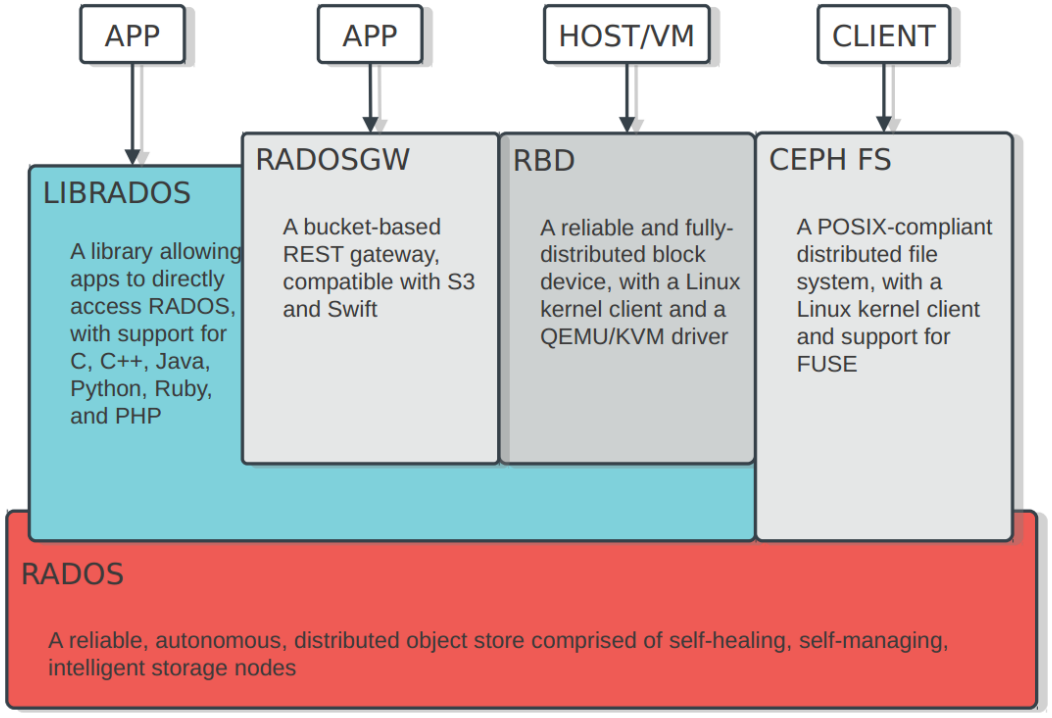
Contents

1. Ceph的介绍
2. Ceph的背景
3. Ceph的架构
4. Ceph的优点
5. Ceph的测试
6. Ceph的部署
7. Ceph与OpenStack整合

1. Ceph的介绍

Ceph是统一存储，支持三种接口：

- **Objects**：原生的API，兼容Swift和S3的API
- **Block**：支持精简配置、快照、克隆
- **File**：强一致，支持快照



Ceph是分布式存储，它的优点是：

- **高性能**：数据分布均衡，并行化度高。对于objects storage和block storage,不需要元数据服务器。
- **高可靠性**：没有单点故障，多数据副本，自动管理，自动修复。
- **高扩展性**：使用普通x86服务器，支持10~1000台服务器，支持TB到PB级的扩展。

2. Ceph的背景

Ceph的母公司：

目前Inktank公司掌控Ceph的开发，但Ceph是开源的，遵循LGPL协议。Inktank还积极整合Ceph和其他云计算和大数据平台，目前Ceph支持OpenStack、CloudStack、OpenNebula、Hadoop等。

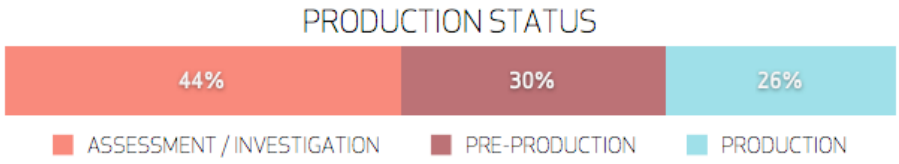
Ceph的社区：

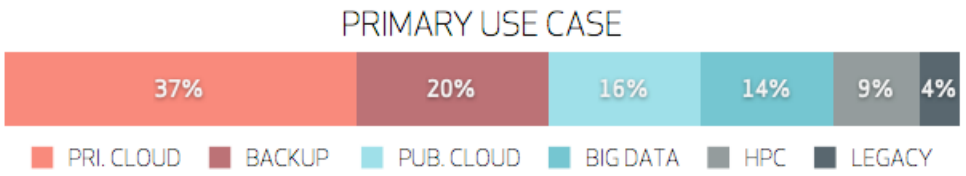
完善的社区设施和发布流程

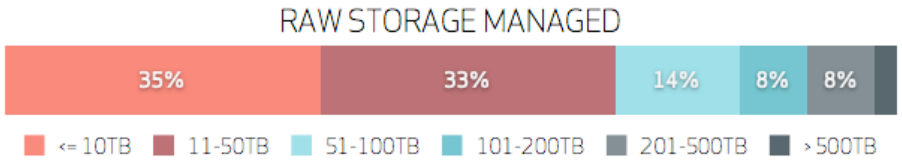
http://www.ustack.com/blog/ceph-distributed-block-storage/#2_Ceph

Ceph的用户调查(2013.03)：

http://ceph.com/community/results-from-the-ceph-census/







UnitedStack

2. Ceph的背景


Alcatel·Lucent


CALXEDA
Power Your Tomorrow


CANONICAL


Bloomberg


Walmart


HUAWEI


DELL
The power to do more


eNovance
Cloud & Managed Services Provider


hastexo!


Deutsche
Telekom


at&t


ebay


MIRANTIS


PENGUIN COMPUTING
EXPERTS IN HIGH PERFORMANCE COMPUTING SOLUTIONS


Piston
PISTON CLOUD COMPUTING


YouScribe


filoo
we make your net work.


OAK RIDGE
National Laboratory


Netelligent


unity


Zetta


CITRIX


SUSE


DreamHost


orange™


golden frog


UAB
THE UNIVERSITY OF
ALABAMA AT BIRMINGHAM


UnitedStack

3. Ceph的架构

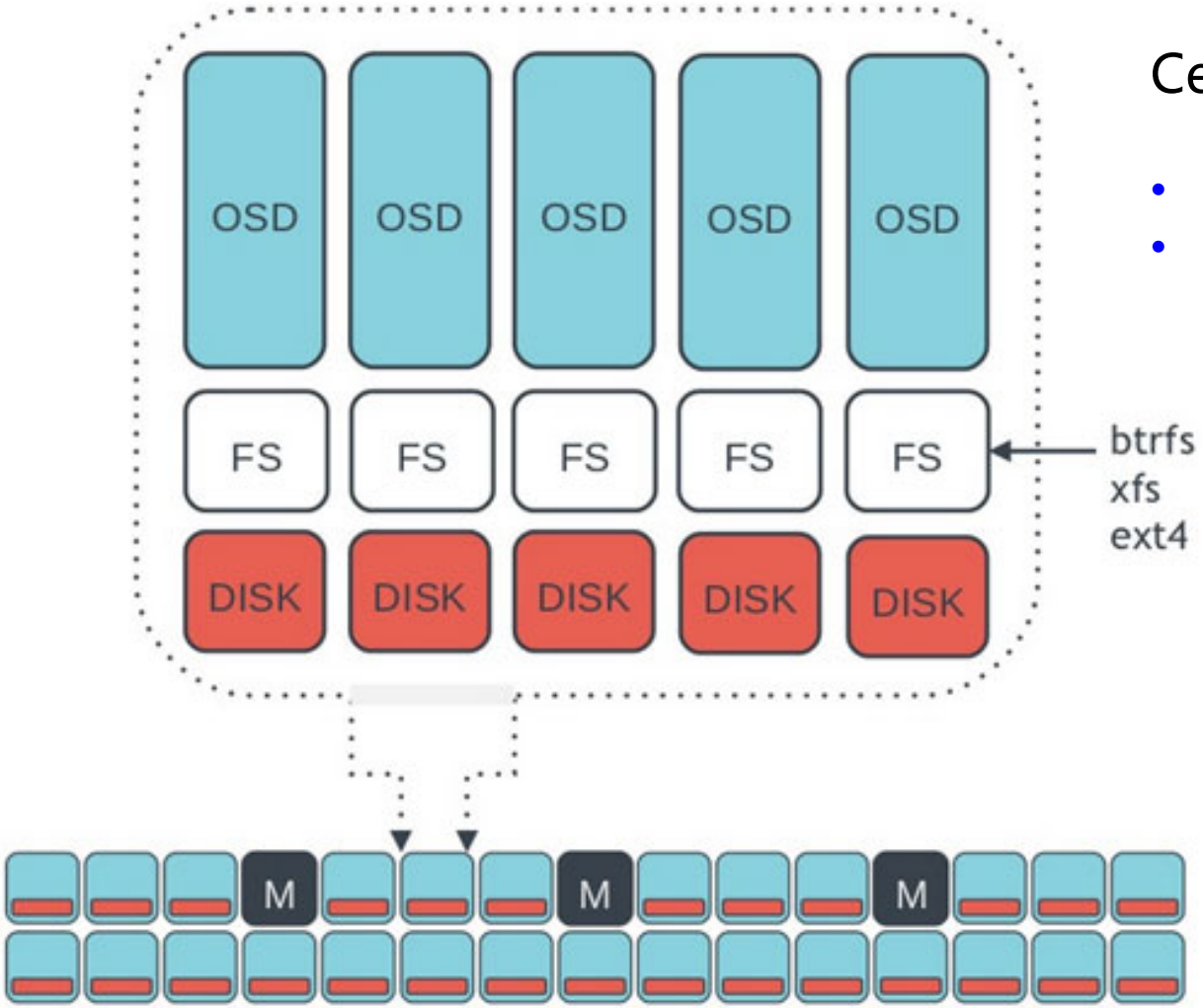
3.1 组件

3.2 映射

3.3 强一致性

3.4 容错性

3.1 Ceph的架构 > 组件



Ceph的底层是RADOS，由两个组件构成：

- OSDs：Object Storage Device提供存储资源
- Monitors: 管理整个集群的全局状态

A reliable, autonomous, distributed object store.

UnitedStack

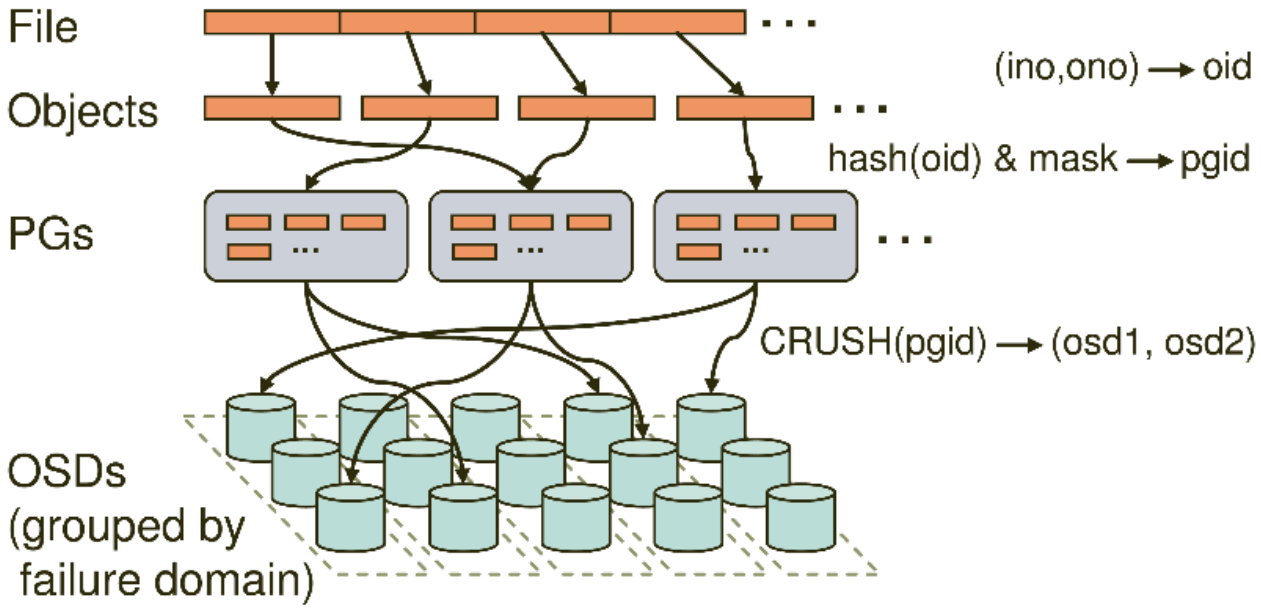
3.2 Ceph的架构 > 映射

数据映射(Data Placement)的方式决定了存储系统的性能和扩展性。映射由四个因素决定：

- CRUSH算法：一种伪随机算法。
- OSD MAP：包含当前所有Pool的状态和所有OSD的状态。
- CRUSH MAP：包含当前磁盘、服务器、机架的层级结构。
- CRUSH Rules：数据映射的策略。

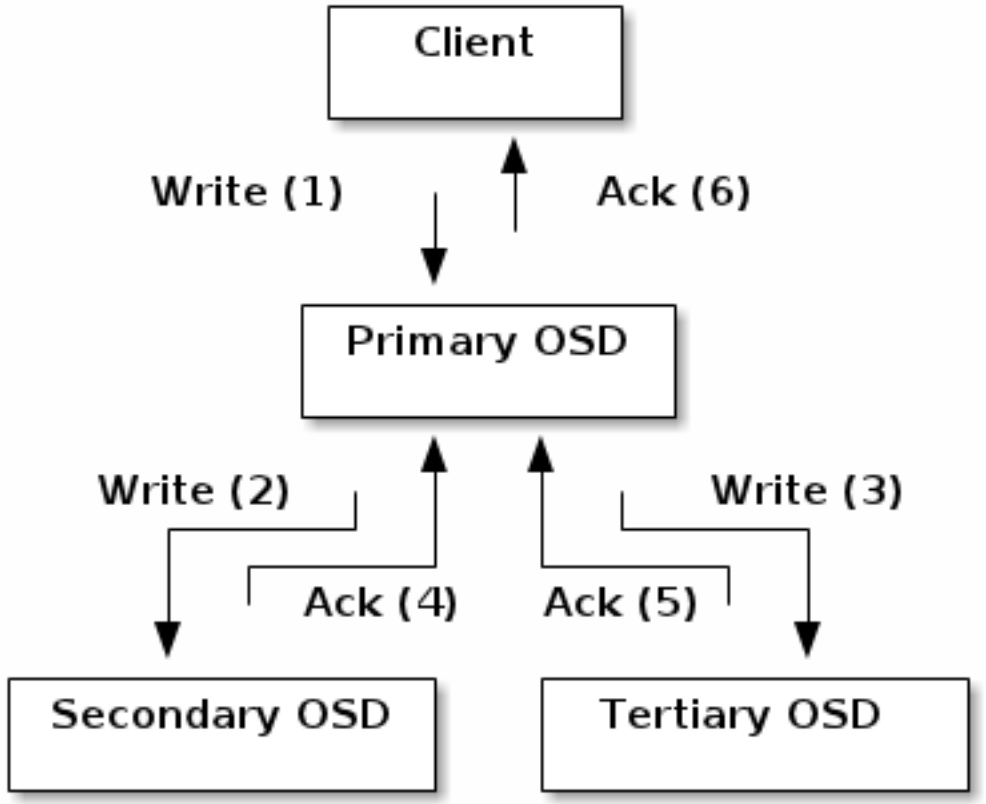
优点：

- 减少metadata的数量，降低管理开销
- 均衡OSD的负载，提高并行度
- 分隔故障域，提高数据可靠性



3.3 Ceph的架构 > 强一致性

- Ceph的读写操作采用Primary-Replica模型，Client只向Object所对应OSD set的Primary发起读写请求，这保证了数据的强一致性。
- 由于每个Object都只有一个Primary OSD，因此对Object的更新都是顺序的，不存在同步问题。
- 当Primary收到Object的写请求时，它负责把数据发送给其他Replicas，只要这个数据被保存在所有的OSD上时，Primary才应答Object的写请求，这保证了副本的一致性。



3.4 Ceph的架构 > 容错性

在分布式系统中，常见的故障有网络中断、掉电、服务器宕机、硬盘故障等，Ceph能够容忍这些故障，并进行自动修复，保证数据的可靠性和系统可用性。

- **Monitors维护着Ceph的全局状态**，它是Ceph管家，是最重要的组件。Monitors的功能和zookeeper类似，它们使用 quorum和Paxos算法去建立全局状态的共识。Monitors cluster容忍 N/2 - 1 个montior失效。
- **OSDs可以进行自动修复**，而且是并行修复。

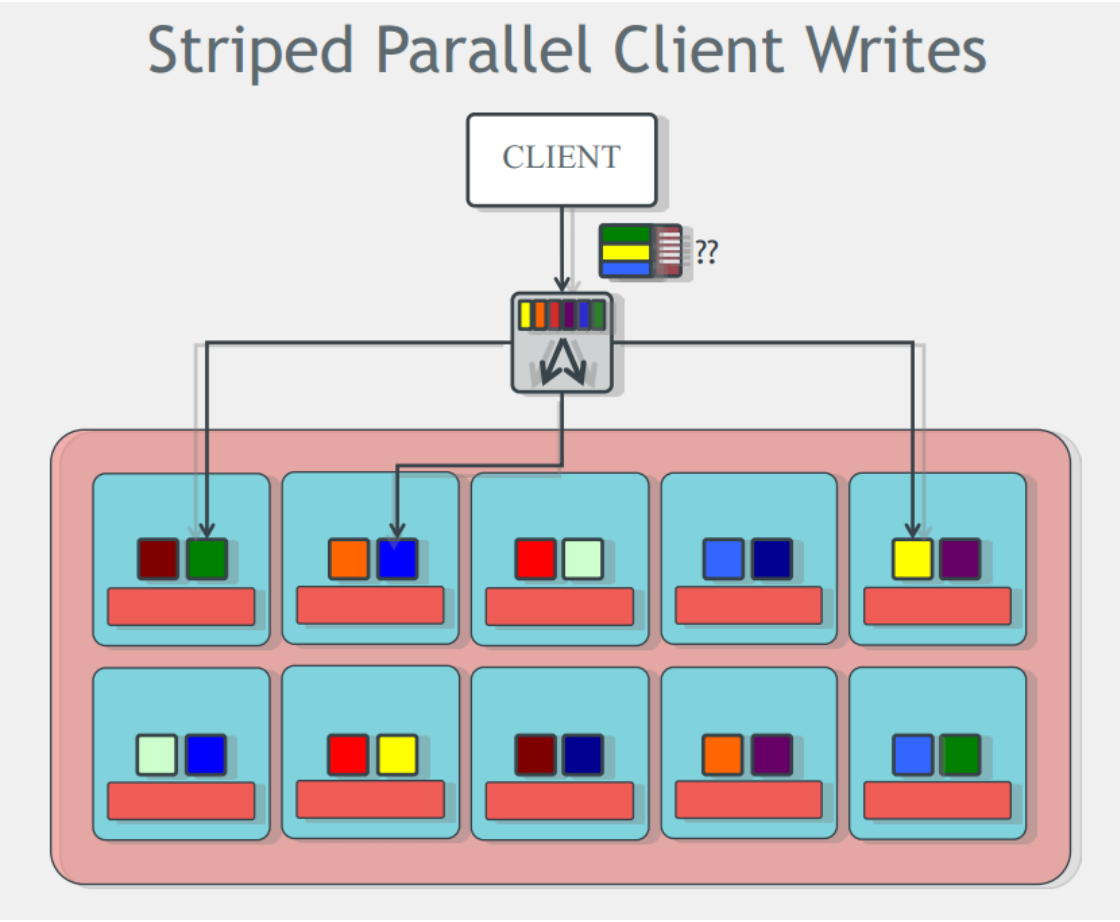
4. Ceph的优点

4.1 高性能

4.2 高可靠性

4.3 高扩展性

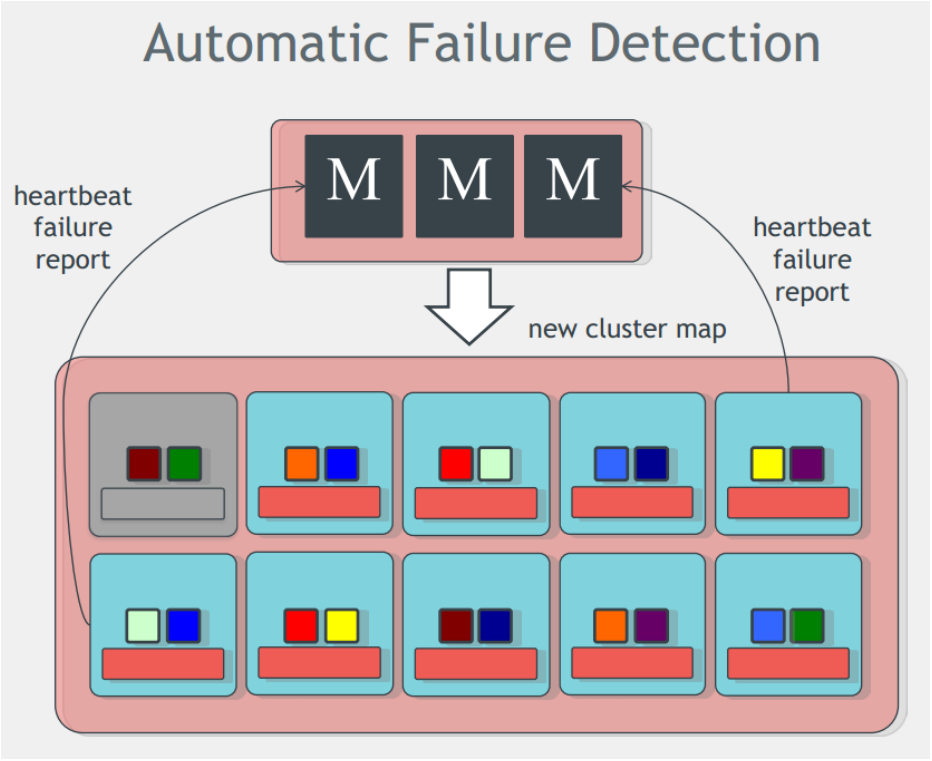
4.1 Ceph的优势 > 高性能

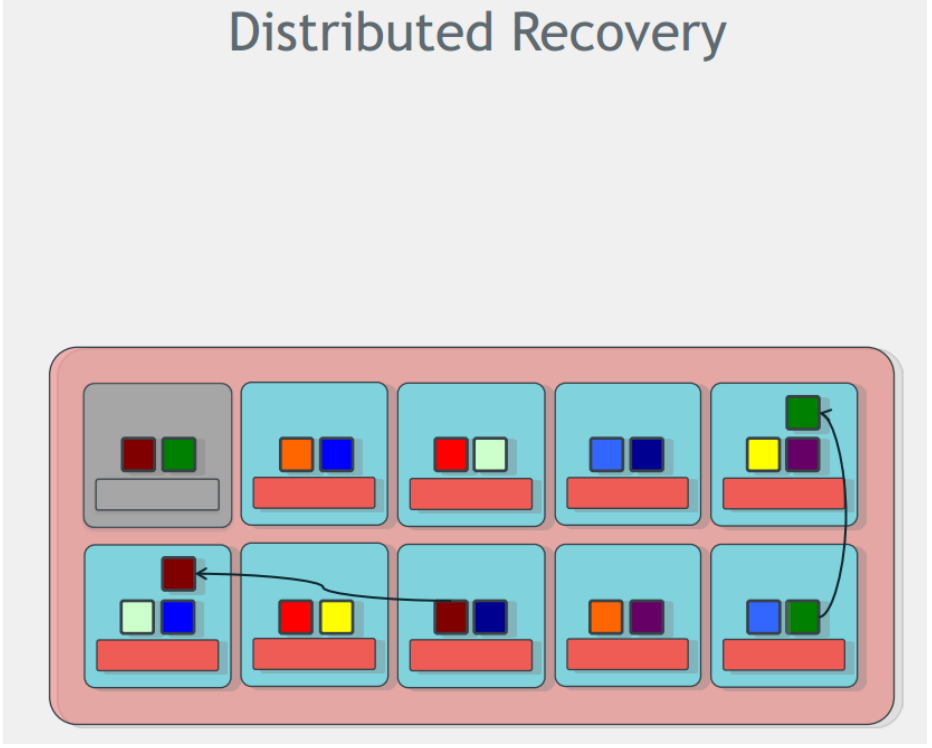


- Client和Server直接通信，不需要代理和转发
- 多个OSD带来的高并发度。objects是分布在所有OSD上。
- 负载均衡。每个OSD都有权重值(现在以容量为权重)。
- client不需要负责副本的复制，而是由primary来负责，这就降低了client的网络消耗。

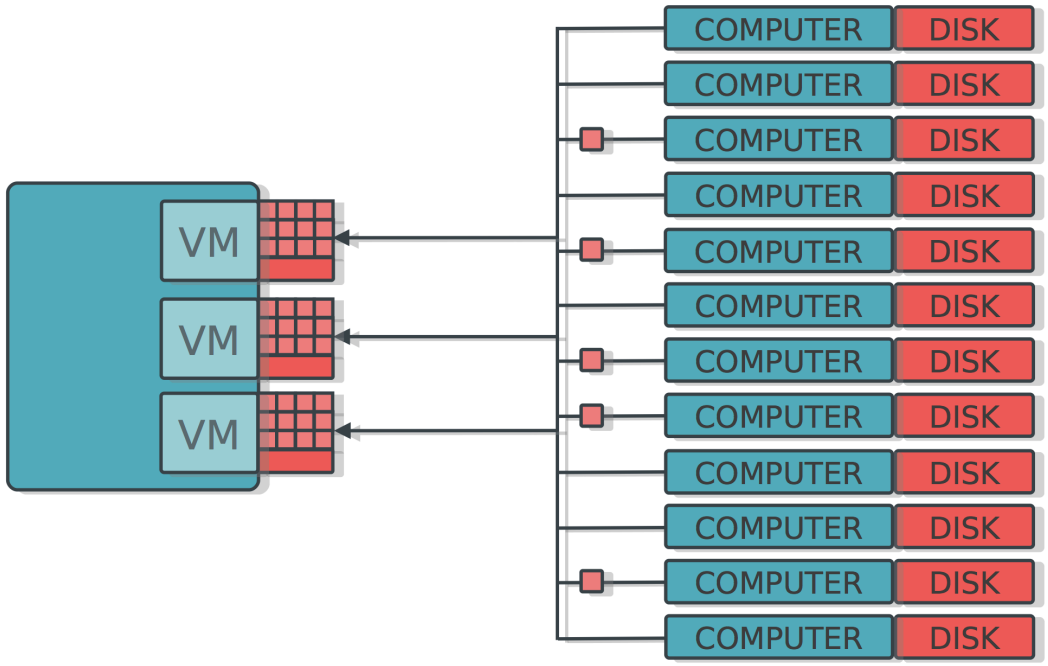
4.3 Ceph的优势 > 高可靠性

- 数据复制。可配置的per-pool副本策略和故障域布局，支持强一致性。
- 没有单点故障。可以忍受许多种故障场景；防止脑裂；单个组件可以滚动升级并在线替换。
- 所有故障的检测和自动恢复。恢复不需要人工介入，在恢复期间，可以保持正常的数据访问。
- 并行恢复。并行的恢复机制极大的降低了数据恢复时间，提供了数据的可靠性。





4.3 Ceph的优势 > 高扩展性



- 高度并行。没有单个中心控制组件。所有负载都能动态的划分到各个服务器上。把更多的功能放到OSD上，让OSD更智能。
- 独立。每个object的操作都只有一个已知的OSD负责。由该OSD负责这个object的数据完整性。
- 自管理。容易扩展、升级、替换。当组件发生故障时，自动进行数据的重新复制。当组件发生变化时(添加/删除)，自动进行数据的重分布。

5. Ceph的测试

5.1 测试环境

5.2 IOPS

5.3 吞吐率

5.4 写惩罚

5.4 结论

5.1 Ceph的测试 > 测试环境

测试工具：

- fio, bs=4K, ioengine=libaio, iodepth=32, numjobs=16
- dd, bs=512M,oflag=direct

测试服务器:

- 单台服务器
- 117GB内存
- 双路 E5-2650,共16核
- 24 * 2TB 硬盘

操作系统：ubuntu 13.04 (3.8.0-19 kernel)

Ceph: Cuttlefish 0.61版，副本数为2，块大小是1M

注意：

因为使用的是AWS上的虚拟机，所以它(Xen)挂载的磁盘都是设置了Cache的。因此下面测试的数据并不能真实反应物理磁盘的真实性能，仅供与RAID10进行对比。

5.2 Ceph的测试 > IOPS

磁盘数	随机写			随机读		
	Ceph	RAID10	性能比	Ceph	RAID10	性能比
24	1075	3772	28%	6045	4679	129%
12	665	1633	40%	2939	4340	67%
6	413	832	49%	909	1445	62%
4	328	559	58%	666	815	81%
2	120	273	43%	319	503	63%

5.3 Ceph的测试 > 吞吐率

磁盘数	顺序写(MB/s)			顺序读(MB/s)		
	Ceph	RAID10	性能比	Ceph	RAID10	性能比
24	299	897	33%	617	1843	33%
12	212	703	30%	445	1126	39%
6	81	308	26%	233	709	32%
4	67	284	23%	170	469	36%
2	34	153	22%	90	240	37%

5.4 Ceph的测试 > 写惩罚

- 在24块磁盘上创建Ceph集群，创建一个pool，副本数为2。在该pool上创建一个image,它的block size=64KB。
- 使用fio对这个image进行顺序写，fio的参数中bs=64KB。使用iostat观察数据的读写。
- 在后60秒中，fio对/dev/rbd1一共写入了 3667.0002MB的数据，Ceph往24块硬盘一共写入了 16084.6078MB的数据，，从24块硬盘中读取了 288.3773MB的数据。
- 对image写 1 MB = CEPH写 4.39 MB + 读 0.08 MB

Device:	rrqm/s	wrqm/s	r/s	w/s	rkB/s	wkB/s	avgrq-sz	avgqu-sz	await	r_await	w_await	svctm
xvdap1	0.00	13.48	1.13	11.02	6.13	201.20	34.13	0.04	3.18	7.59	2.73	1.09
xvdb	0.00	3.13	16.13	407.25	319.60	12005.88	58.22	0.48	1.14	6.45	0.93	0.34
xvdc	0.00	1.83	18.77	338.37	495.48	10003.83	58.80	0.43	1.19	5.80	0.94	0.34
xvdd	0.00	2.00	13.83	312.32	282.47	9178.73	58.02	0.38	1.17	5.58	0.98	0.36
xvde	0.00	3.35	11.25	440.18	51.73	12860.68	57.21	0.48	1.07	8.33	0.89	0.36
xvdf	0.00	2.83	7.65	328.65	34.73	9532.02	56.89	0.32	0.96	7.76	0.80	0.33
xvdg	0.00	3.03	10.73	393.65	80.73	11566.04	57.60	0.42	1.04	7.52	0.86	0.35
xvdh	0.00	3.15	13.13	455.33	122.60	13388.45	57.68	0.52	1.12	7.65	0.93	0.34
xvdi	0.00	3.65	20.07	504.07	413.73	14810.95	58.09	0.55	1.05	6.00	0.85	0.34
xvdj	0.00	2.77	9.22	385.45	41.48	11227.71	57.11	0.39	0.98	7.25	0.83	0.33
xvdk	0.00	2.92	11.25	399.43	51.47	11659.13	57.03	0.41	1.00	7.56	0.82	0.37
xvdl	0.00	2.03	8.22	330.35	37.80	9589.18	56.87	0.35	1.04	7.90	0.87	0.35
xvdm	0.00	2.83	19.92	495.62	417.60	14534.39	58.01	0.60	1.17	7.04	0.93	0.35
xvdn	0.00	2.25	9.27	366.73	42.47	10690.13	57.09	0.39	1.04	7.55	0.87	0.34
xvdo	0.00	2.02	6.92	280.05	31.67	8128.16	56.87	0.28	0.98	7.59	0.82	0.35
xvdp	0.00	2.60	12.00	468.75	53.87	13687.14	57.16	0.51	1.05	8.17	0.87	0.36
xvdq	0.00	2.72	15.58	364.70	316.33	10708.18	57.98	0.41	1.07	5.34	0.89	0.34
xvdr	0.00	2.57	17.23	421.95	320.27	12418.78	58.01	0.51	1.16	5.92	0.97	0.35
xvds	0.00	2.52	16.03	351.70	317.55	10333.96	57.93	0.45	1.24	5.92	1.02	0.36
xvdt	0.07	3.23	23.47	479.30	543.27	14138.79	58.41	0.60	1.19	6.33	0.94	0.36
xvdu	0.00	2.30	7.98	341.43	36.13	9953.21	57.18	0.34	0.98	7.08	0.83	0.33
xvdv	0.00	2.52	8.70	345.82	40.13	10089.62	57.15	0.36	1.03	8.17	0.85	0.35
xvdw	0.00	2.80	15.45	404.43	279.20	11837.48	57.71	0.46	1.11	6.37	0.91	0.34
xvdx	0.00	2.33	15.78	384.77	317.53	11289.79	57.96	0.44	1.11	6.15	0.90	0.33
xvdy	0.00	2.48	14.63	371.22	273.80	10878.41	57.81	0.44	1.14	5.75	0.96	0.34
rbd1	0.00	0.00	0.00	977.87	0.00	62583.47	128.00	31.91	32.64	0.00	32.64	1.02

UnitedStack

5.5 Ceph的测试 > 结论

- 在单机情况下，RBD的性能不如传统的RAID。这是因为RBD的I/O路径长：
  Librbd -> networking -> OSD -> FileSystem -> Disk

- 每个Client的写请求下发到OSD之后，会产生2~3个写操作：
  - 把写操作记录到OSD的Journal文件上。
  - 把写操作更新到Object对应的文件上。
  - 把写操作记录到PG Log文件上。

- Ceph的优势在于它的扩展性，它的性能随着磁盘数量线性增长，因此在多机的情况下，RBD理论的IOPS和吞吐率会高于单机的RAID。但是Ceph的性能还是受限于网络的带宽。

- Ceph是一款出色的分布式统一存储系统，使用Ceph能够降低硬件成本和运维成本。

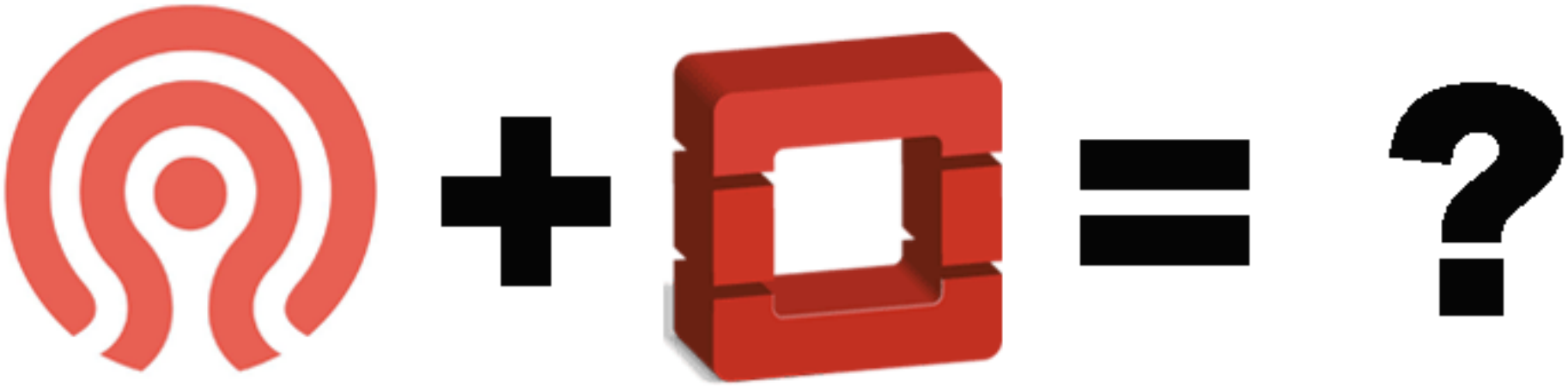
6. Ceph的部署

- 两个组件:OSD和Monitor
- 在每块硬盘上都跑OSD，部分节点上运行Monitor
- 部署工具：mkcephfs、ceph-deploy、puppet
- https://github.com/enovance/puppet-ceph

手工部署

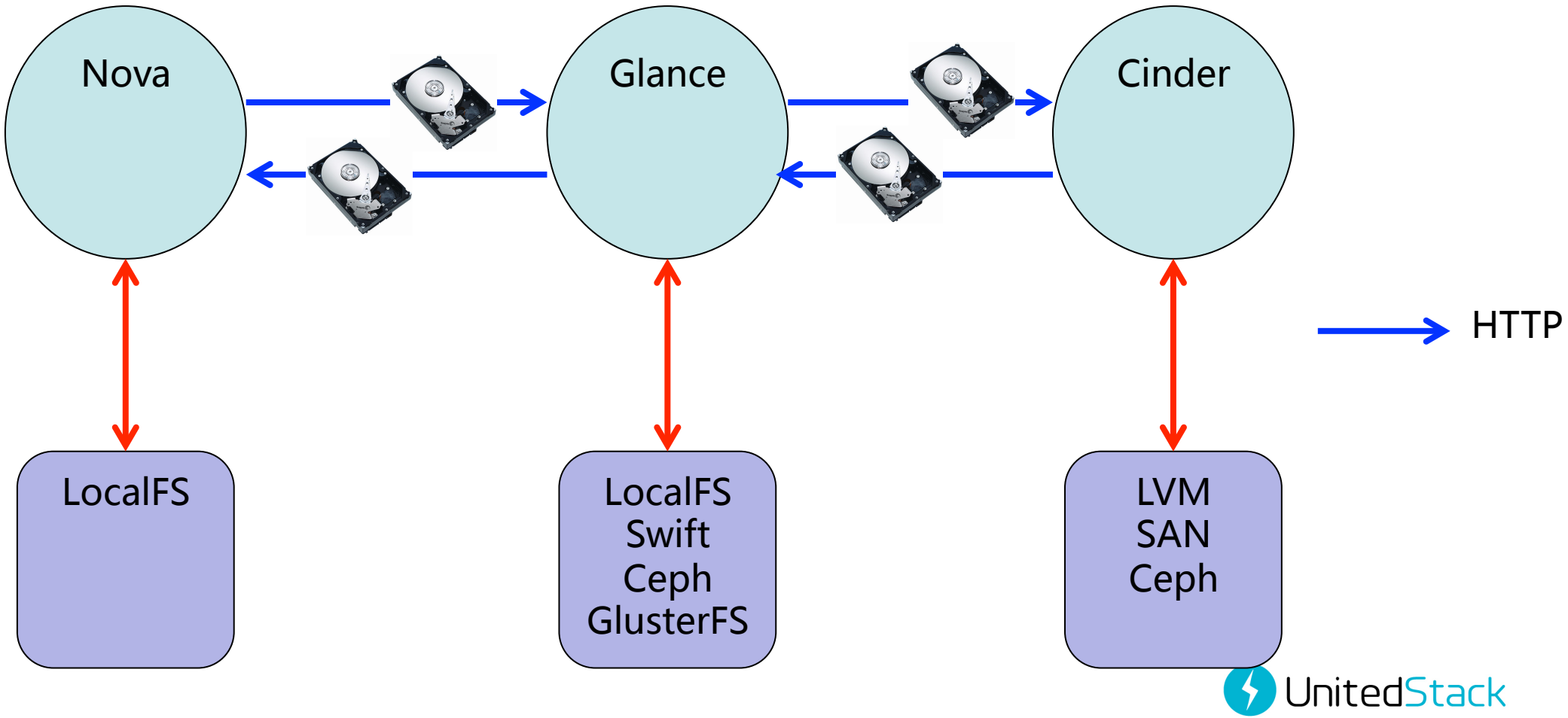
- 0 环境要求
- 1 安装依赖包
- 2 安装Ceph
- 3 生成Ceph集群所需的配置文件
- 4 创建/启动第1个monitor
- 5 收集keyrings
- 6 创建OSD

6. Ceph与OpenStack整合



6. Ceph与OpenStack整合

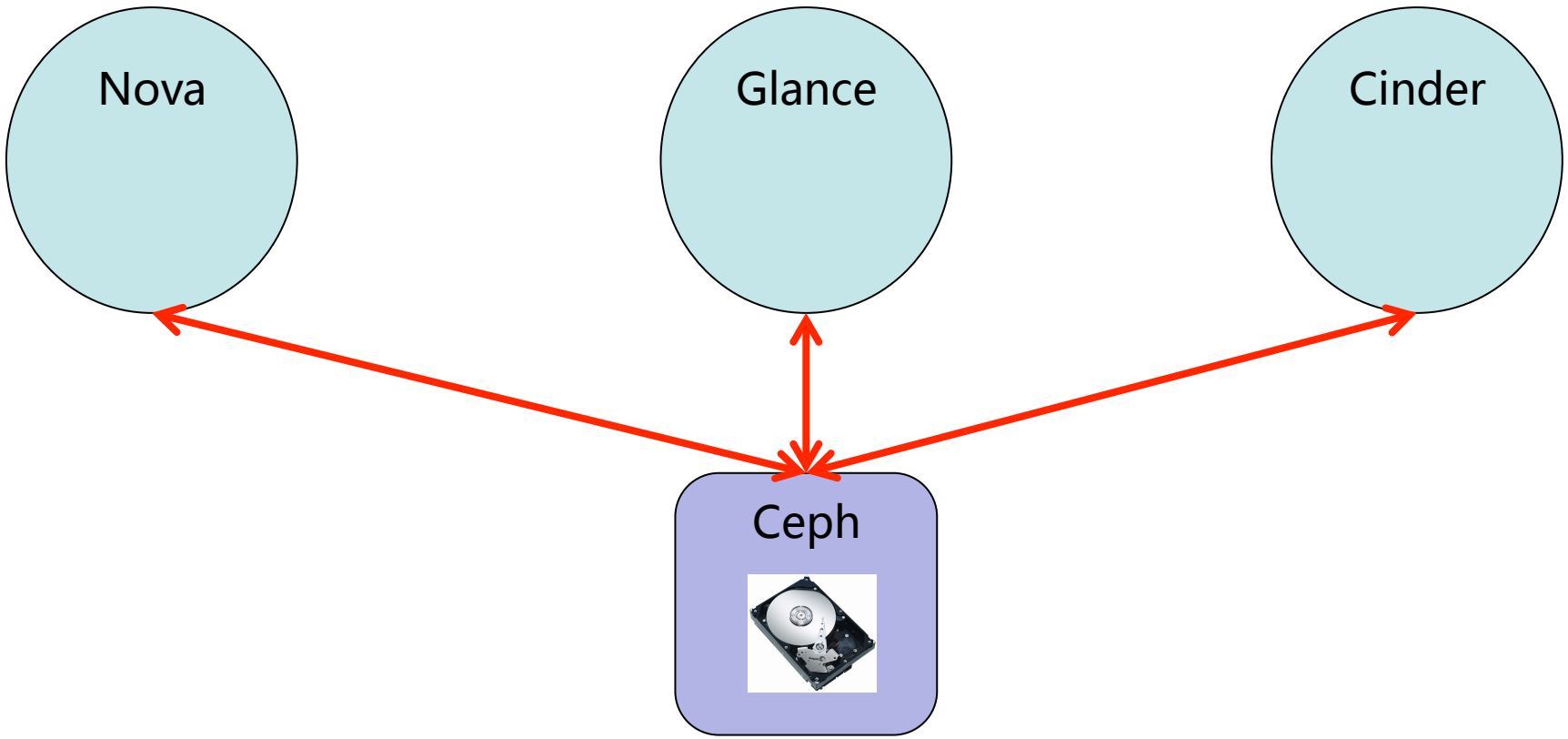
Nova, Glance, Cinder之间有数据传输，而且管理多个后端存储会很复杂



UnitedStack

6. Ceph与OpenStack整合

Nova, Glance, Cinder之间没有数据传输，快速创建虚拟机，只需要管理一个统一存储



Create an instance = clone an image (copy on write)

UnitedStack

Q&A

UnitedStack

THX

UnitedStack