ceph

Ceph中国社区 | IT大咖说 知识共享平台

CEPHALOCON APAC 2018

THE FUTURE OF STORAGE

22-23 March 2018 | BEIJING

企业级Ceph的机遇与挑战

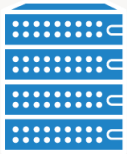
Red Hat &中国铁路信息技术中心

企业业务转型趋势

开发模式	应用架构	部署与封装	应用基础架构	存储
瀑布式	一体化	裸机	数据中心	纵向扩展
敏捷	N层	虚拟服务	托管	横向扩展
DevOps	微服务	容器	混合云	软件定义的存储

现代企业对存储的诉求

38%的IT决策人表示，存储能力不足是每周工作中最主要的三个痛点之一

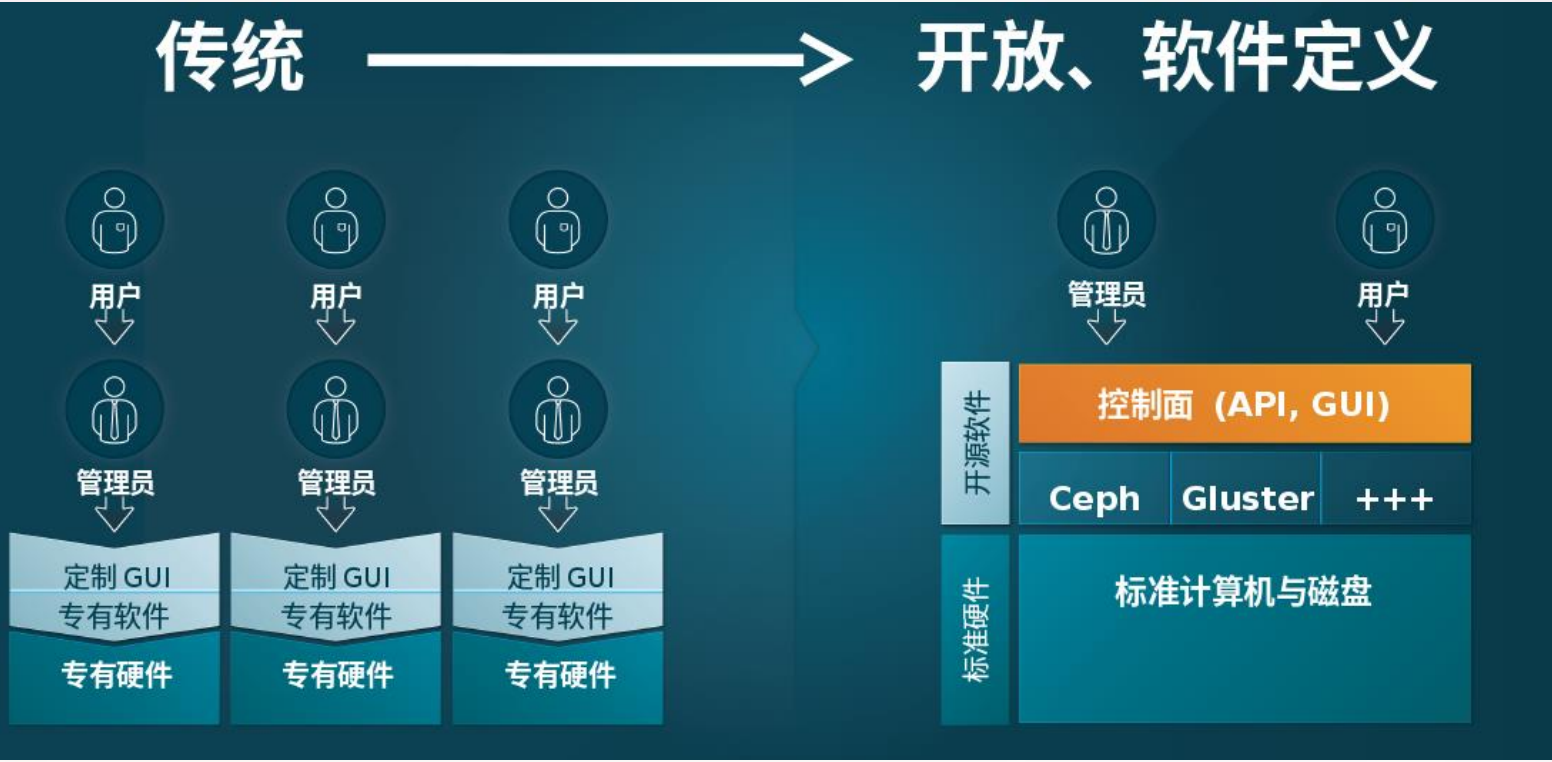


70%的IT决策人承认，其企业内当前的存储无法应对新型工作负载

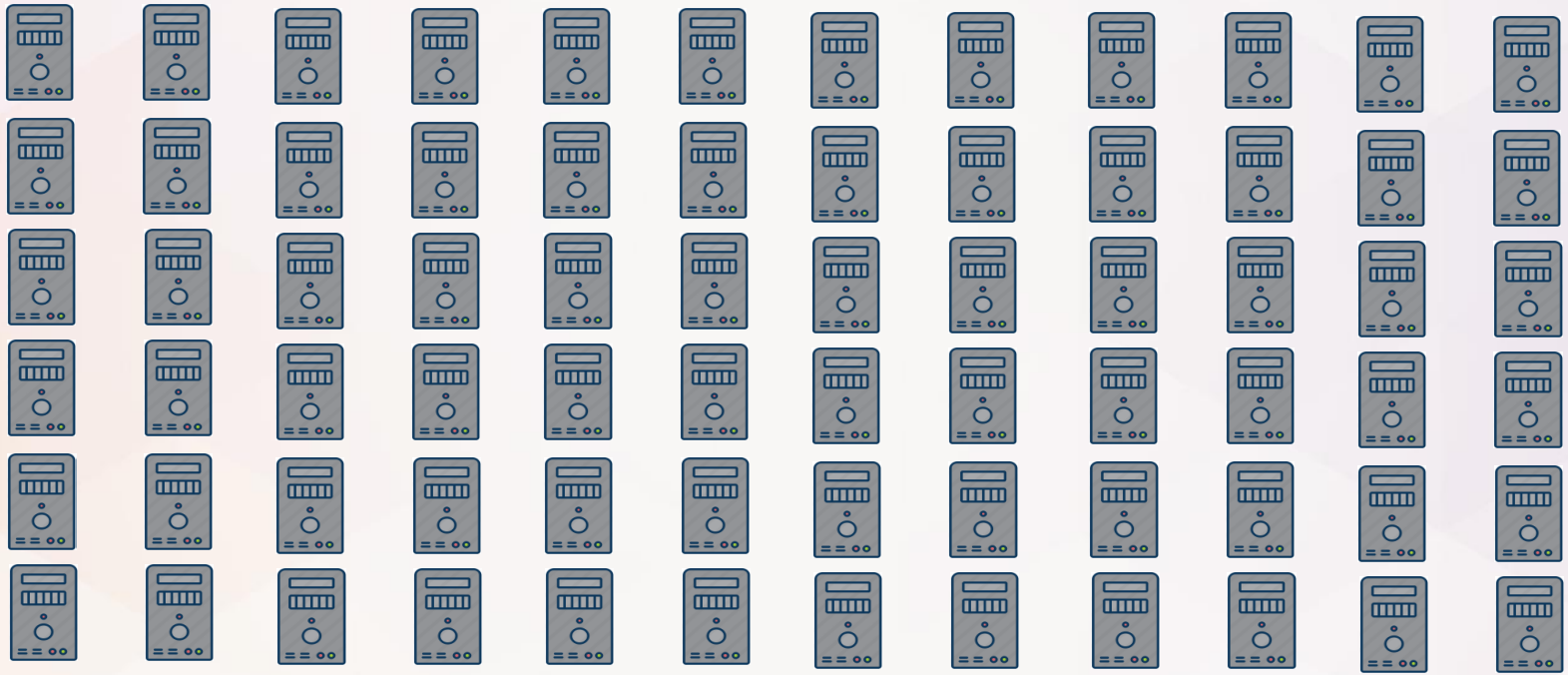
98%的IT决策人认为，更敏捷的存储解决方案对其企业有益

VansonBourne

存储的演进


传统
开放、软件定义
用户
用户
用户
管理员
管理员
管理员
定制 GUI
专有软件
专有硬件
定制 GUI
专有软件
专有硬件
定制 GUI
专有软件
专有硬件
管理员
用户
开源软件
控制面（API, GUI）
Ceph
Gluster
+++
标准硬件
标准计算机与磁盘

分布式软件定义存储



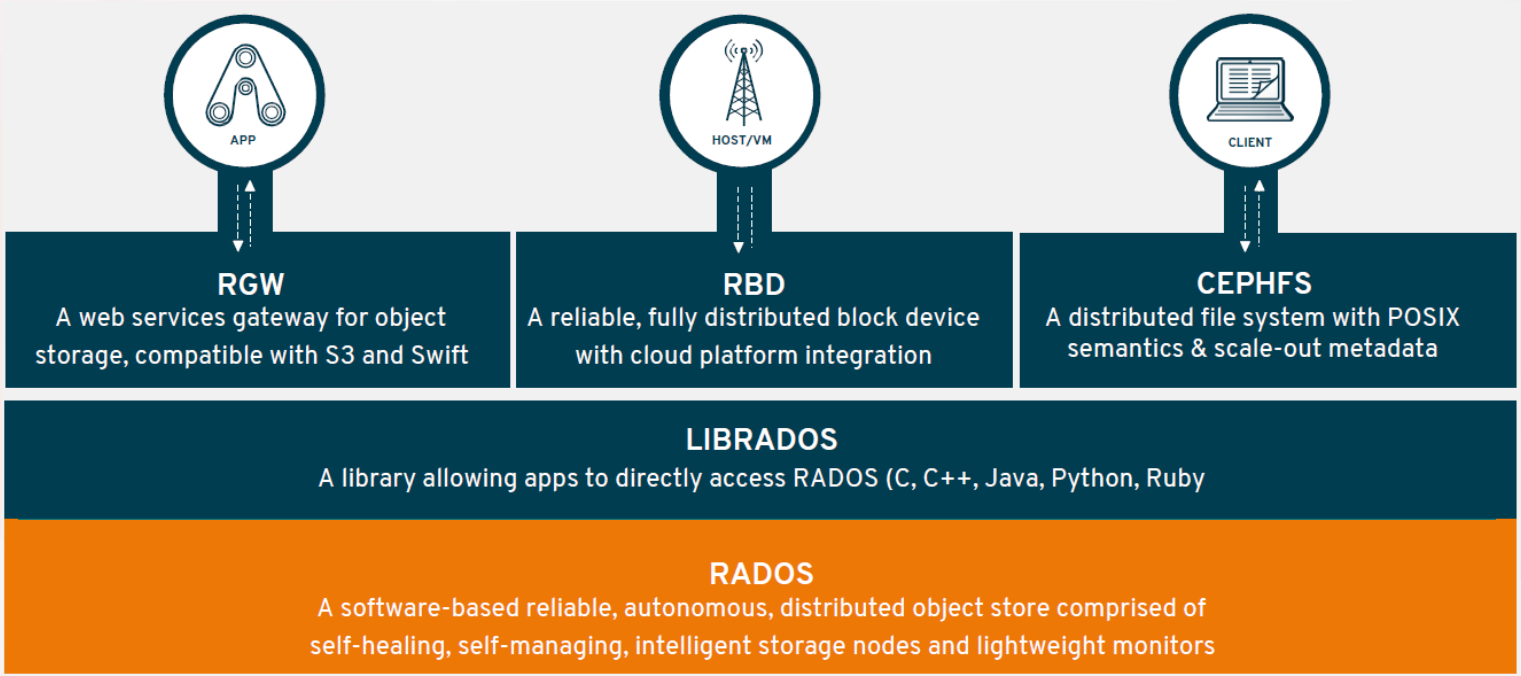
红帽存储的使命

打造统一、开放的软件定义存储产品，面向新型工作负载提供数据服务，从而加速企业向现代IT基础架构过渡。

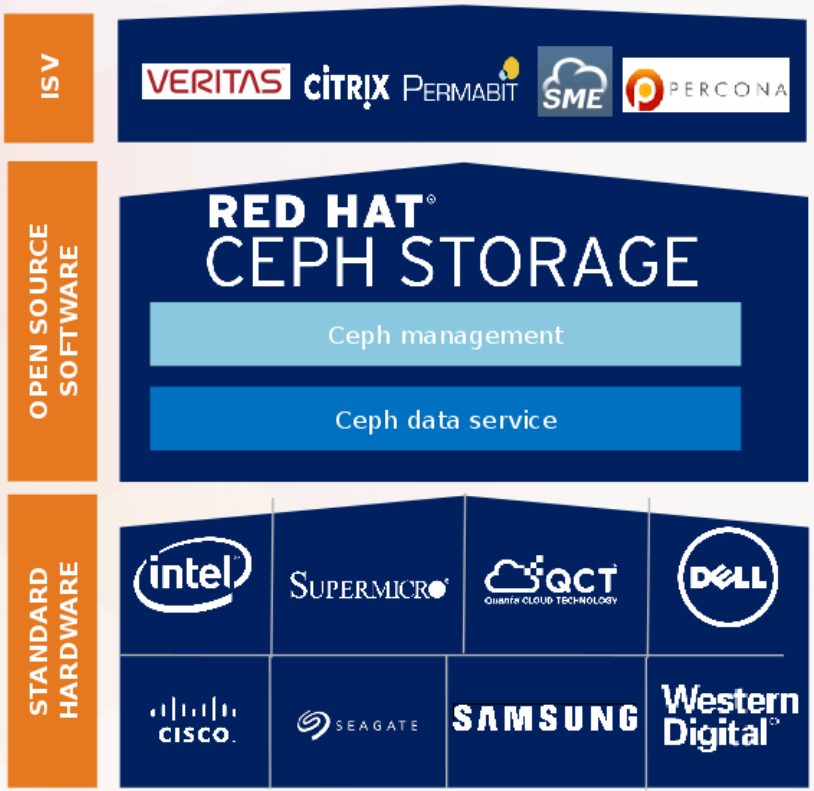
红帽存储

被Gartner列入首个分布式文件系统和对象存储魔力象限的***远见者。***

Ceph概览

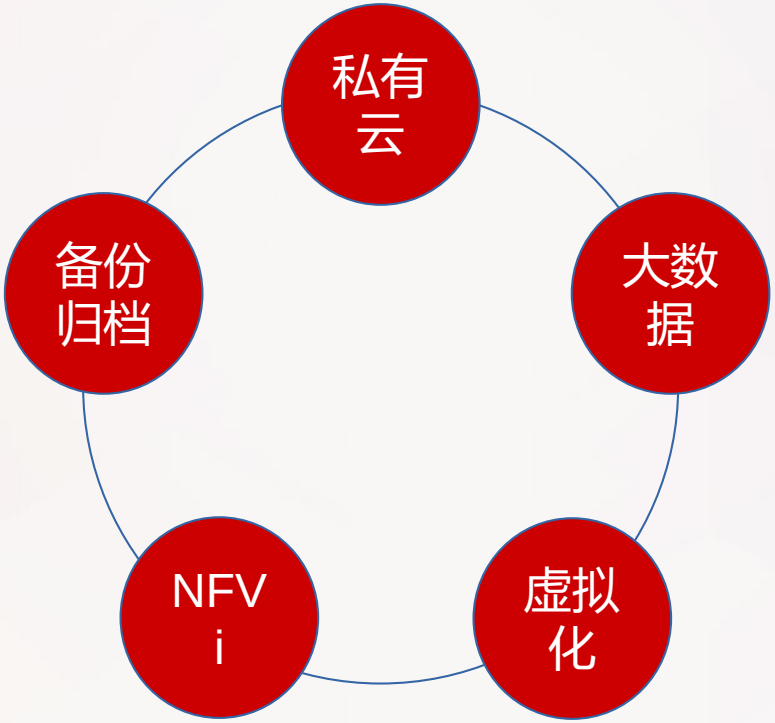

APP
HOST/VM
CLIENT
RGW
A web services gateway for object storage, compatible with S3 and Swift
RBD
A reliable, fully distributed block device with cloud platform integration
CEPHFS
A distributed file system with POSIX semantics & scale-out metadata
LIBRADOS
A library allowing apps to directly access RADOS (C, C++, Java, Python, Ruby
RADOS
A software-based reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes and lightweight monitors

红帽企业级Ceph存储



- 基于通用硬件的开源统一分布式软件定义存储，所有组件都可水平扩展
- 无单点故障，数据自愈、自管理
- 标准接口，全 API ，完善的生态
- 红帽提供企业版订阅和专业化的咨询服务

红帽企业级Ceph五大应用场景



企业级Ceph面临的业务挑战

1. 识别真正的横向扩展存储需求
2. 设计要满足目标工作负载的IO模式
3. 存储访问方式的选择
4. 容量性能规划
5. 确定故障域及风险容忍度
6. 选择数据保护方法

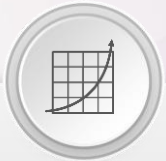
红帽是如何应对挑战的

弹性分配

廉价存储

数据访问方式

消除数据孤岛



滚动升级

1

红帽是如何应对挑战的

工作负载对IO的诉求:

- 吞吐量？ IOPS？ 价格？

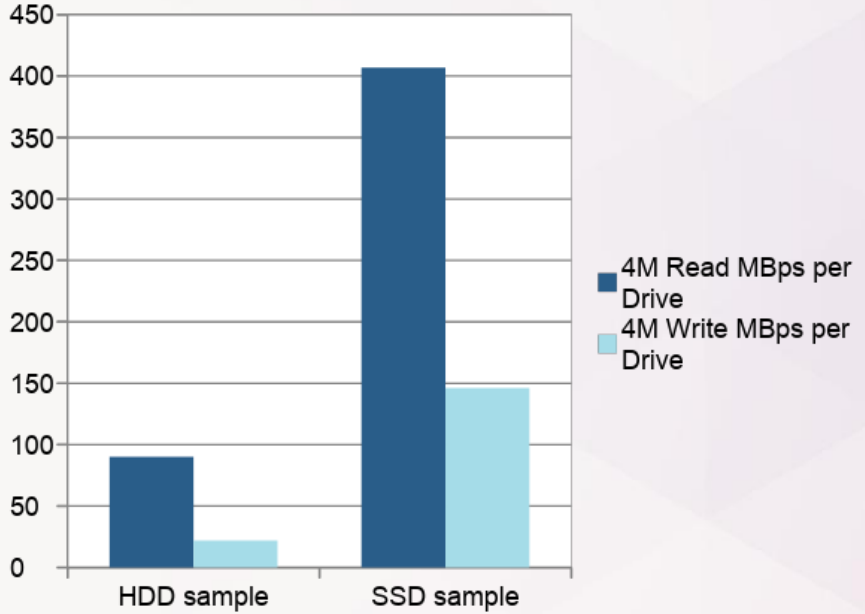
工作负载IO模式:

- IO Size大还是小？
- 顺序IO还是随机IO？
- 读还是写？读写比例？
- 对延迟的要求？

红帽是如何应对挑战的

一些相关的度量指标:

- MBps
- $/MBps
- MBps/provisioned-TB
- Watts/MBps
- MTTR (self-heal from server failure)

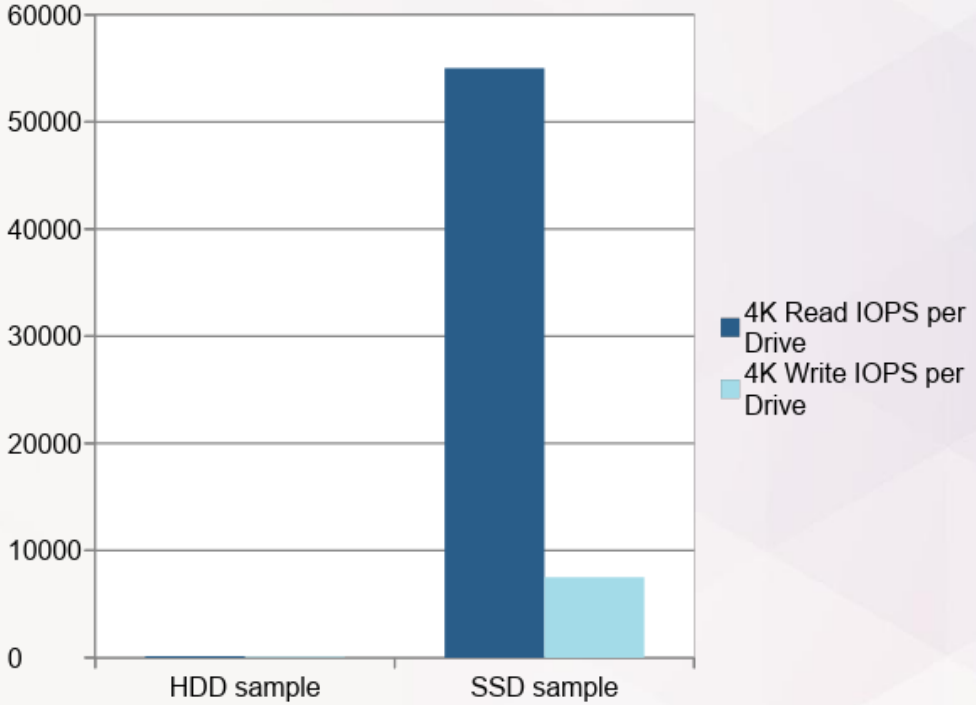


2

红帽是如何应对挑战的

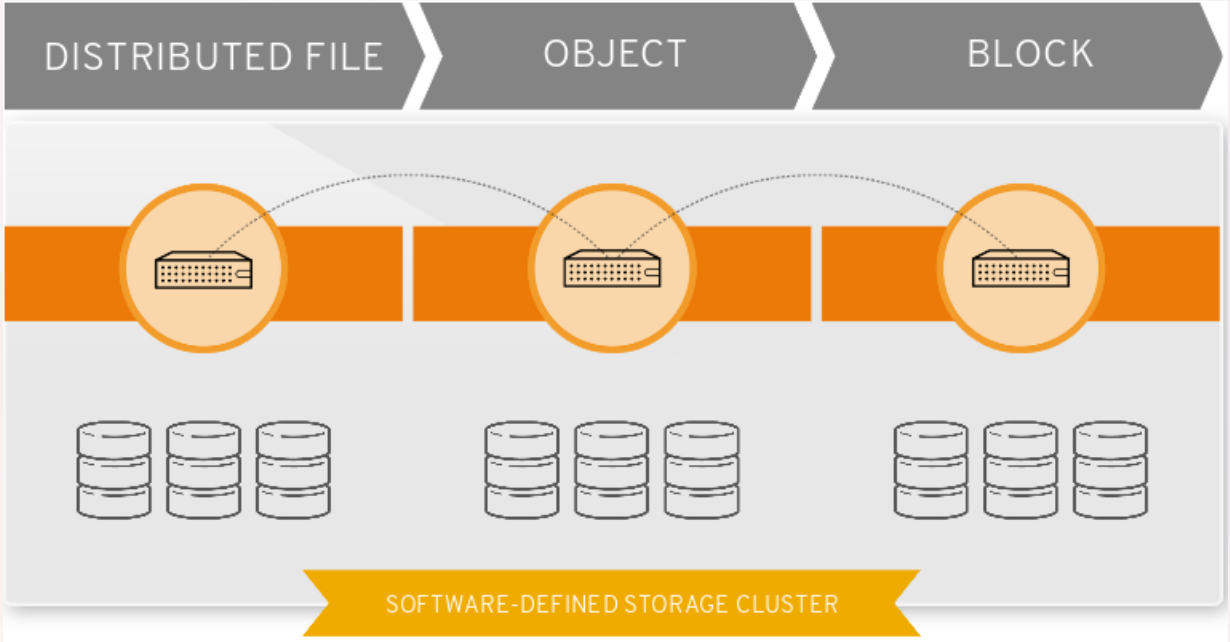
一些相关的度量指标:

- MySQL Sysbench requests/sec
- IOPS (4K, 16K random)
- $/IOP
- IOPS/provisioned-GB
- Watts/IOP



2

红帽是如何应对挑战的


DISTRIBUTED FILE
OBJECT
BLOCK
SOFTWARE-DEFINED STORAGE CLUSTER

红帽是如何应对挑战的

	S		M	L
	64TB	256TB	1PB +	2PB+
IOPS Optimized	• 1-2x P3700s per sled, OR • 4-8x S3710s per server			
Throughput Optimized	• 12-16x 3.5”-bay standard servers		• 24-36x 3.5” -bay dense servers	• 24-36x 3.5” -bay dense servers
Cost-Capacity Optimized				• 60-76x 3.5” -bay ultra-dense servers

红帽是如何应对挑战的

	S		M	L
	64TB	256TB +	1PB +	2PB+
IOPS Optimized	• Ceph block (RBD) • Intel® P3700s w/ co-located write journals, OR • Intel® S3710s w/ Intel® P3700 write journals • Multiple OSDs per flash drive • 10 Xeon® cores per P3700; 4 per S3710 • 2x or 3x replication (with backup)			
Throughput Optimized	• Ceph block or object (RBD or RGW) • HDDs w/ P3700 or S3710 write journals • 1 Xeon® core per 2 HDDs • Single OSD per HDD • 10GbE -> 40GbE with >12 HDD per chassis • 3x replication			
Cost-Capacity Optimized			• Ceph object (RGW) • HDD drives with no SSD journal • 1 Xeon® core per 2 HDDs • Single OSD per HDD • Erasure-coded	

红帽是如何应对挑战的

NODE AND CLUSTER SIZING PRINCIPLES:

- Ceph node recovery: Smaller cluster = larger impact to workload performance during recovery
- Ceph OSD server reserve capacity: Smaller cluster = greater % of reserve capacity on each node allocated to accommodate failed node

GUIDELINES:

Minimum supported cluster size: 3 nodes

Minimum recommended Ceph cluster: 10 nodes

红帽是如何应对挑战的

REPLICATION

- Multiple full data copies stored on different servers (2x or 3x replication)

ERASURE CODING (analogous to network RAID)

- Data encoded into k chunks with m parity chunks. (which are spread onto different volumes on different servers) Can tolerate m disk failures without data loss.

FOR EXAMPLE:

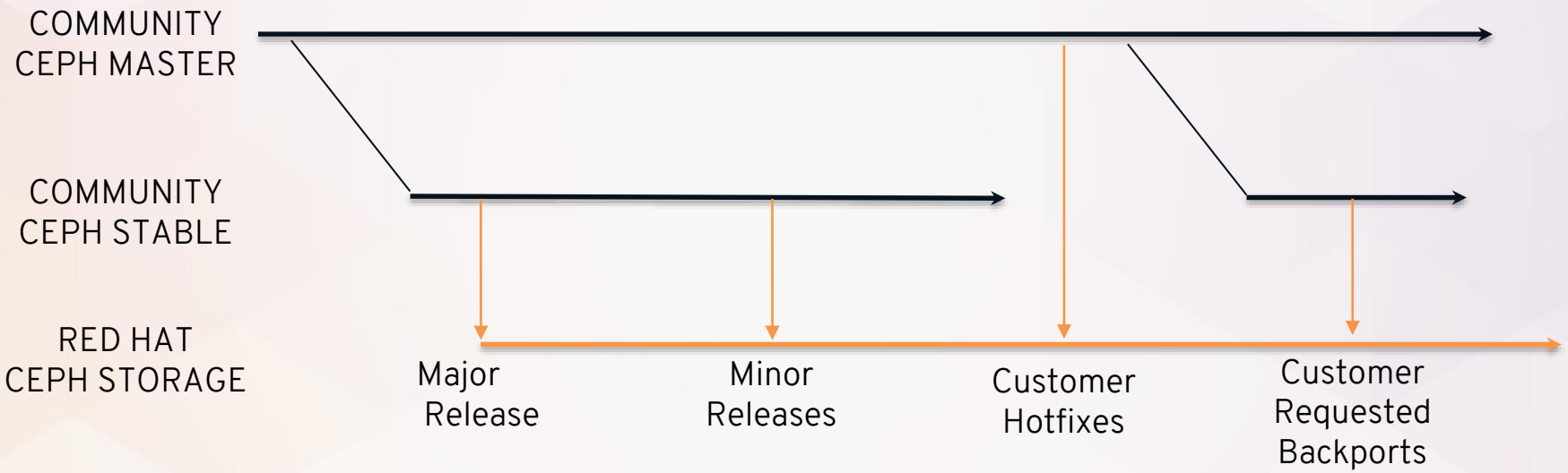
8+3 k+m

(One of the BIGGEST choices affecting purchase price in the entire solution!)

红帽Ceph -不仅仅是横向扩展

IOPS Optimized	Throughput Optimized	Cost / Capacity Optimized
NVMe SSD in SLED chassis	SSD, HDD in standard / dense chassis	HDD in dense / ultra-dense chassis
High IOPS / GB Smaller, random IO Read / write mix	High MB/s throughput Large, sequential IO Read / write mix	Low cost / GB Sequential IO Write mostly
Use Case: MySQL	Use Case: Rich Media	Use Case: Active Archives

红帽企业级Ceph的构建方法



红帽Ceph行业案例 - 中国铁路

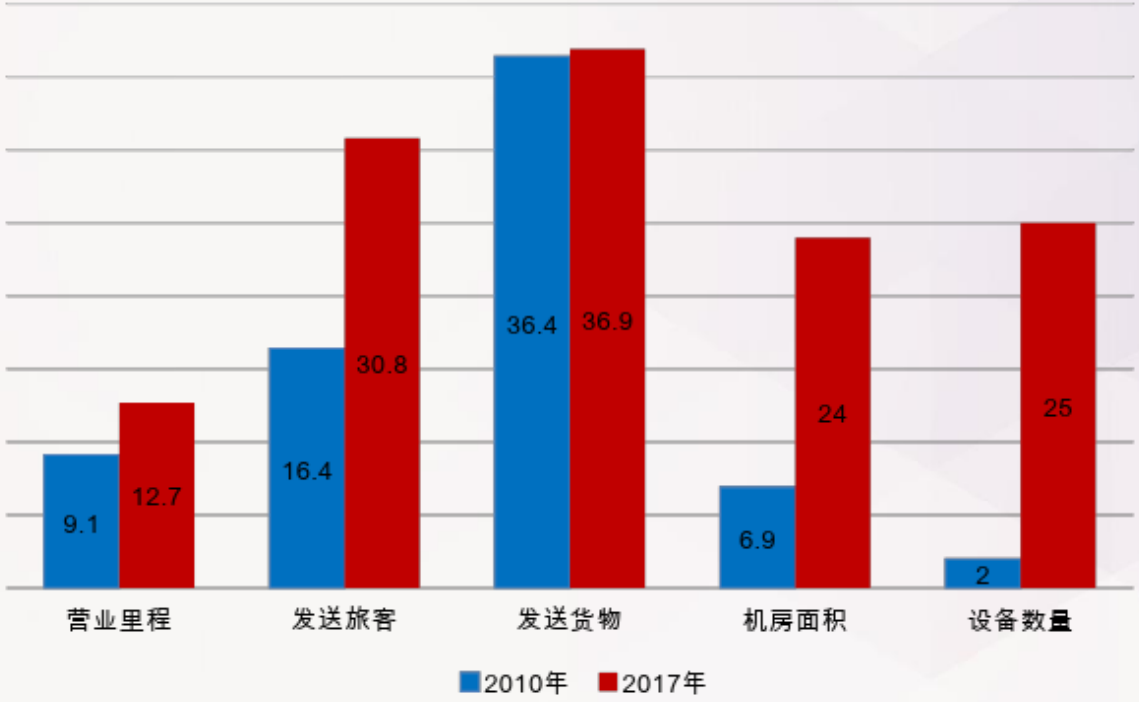
中国铁路总公司
是国家经济命脉型企业

主要业务
客货运输服务

业务特征
规模大，覆盖广，无间断

企业发展目标
向世界一流现代物流企业转型

铁路业务概况



中国铁路路网规划

红帽Ceph行业案例 - 中国铁路

Ceph中国社区 | IT大咖说 知识共享平台

铁信云需求

高效

- 支撑巨量客货运服务
- 适应互联网 + 和数字化转型
- 支撑铁路管理创新、业务创新、应用创新

便捷

- 集中统一，标准化建设，降低系统复杂度
- 资源灵活调整，业务快速上线
- 自动化运维管理

安全

- 满足信息系统安全等级保护要求
- 安全可控，不受特定厂商技术绑定
- 保障铁路业务安全有序运转

绿色

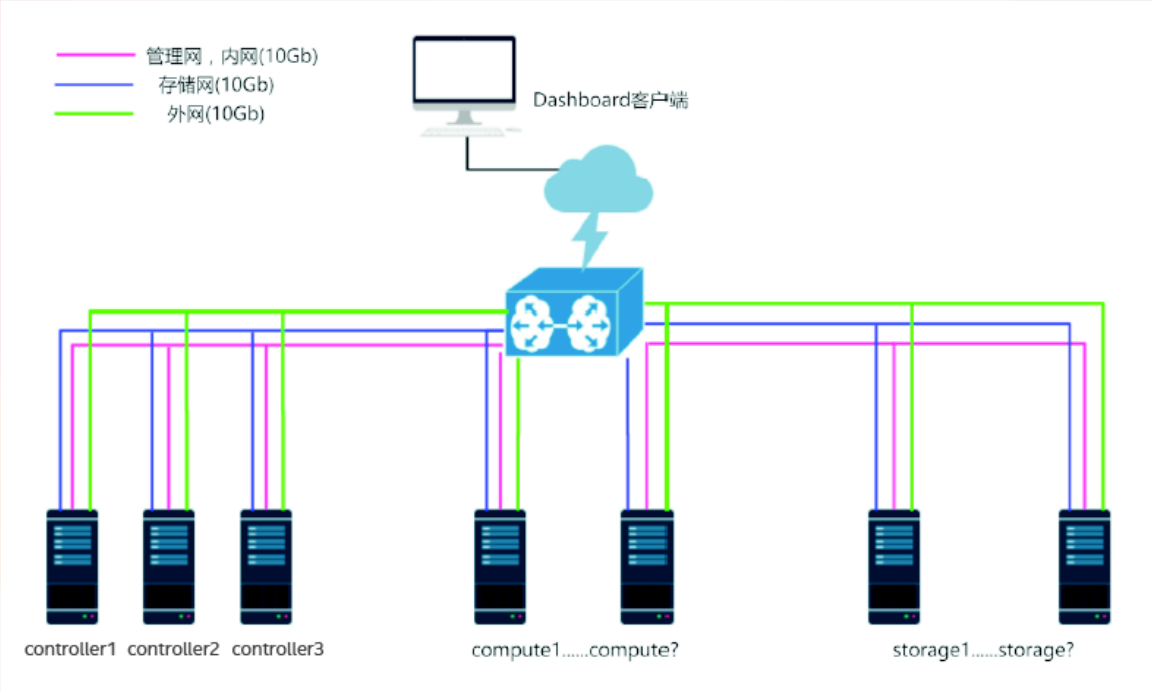
- 避免重复投资，减少资源浪费
- 降低建设成本和运维成本
- 降低数据中心能耗

红帽Ceph行业案例 - 中国铁路

铁信云逻辑架构

红帽Ceph行业案例 - 中国铁路

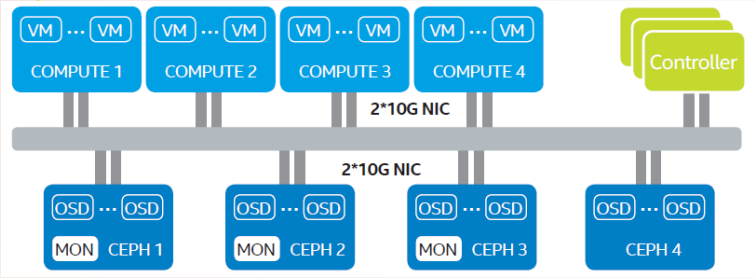
铁信云部署架构



红帽Ceph行业案例 - 中国铁路

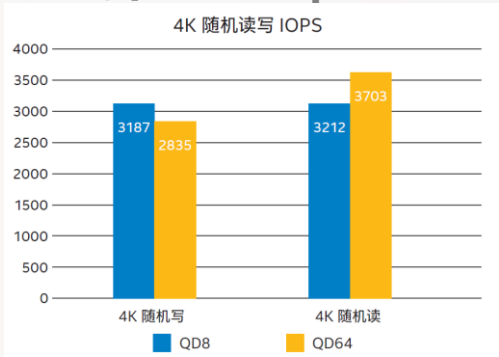
铁信云为什么选择红帽Ceph

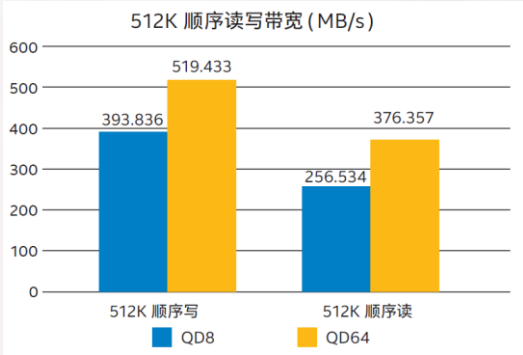
1 小规模性能测试下Ceph的表现

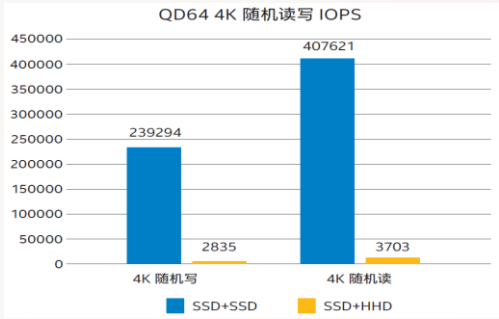


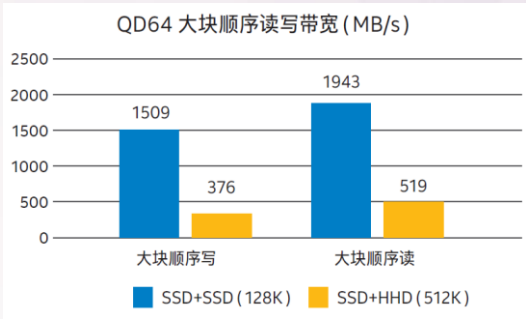
1. SSD盘+HDD盘部署：每台存储节点一块Intel P3700 800G SSD闪存硬盘加 8 块 4T 7200 转 HDD硬盘，其中SSD闪存盘用作Ceph的日志存储，HDD盘用作数据存储。

2. SSD卡+SSD盘全闪存部署：每台存储节点一块Intel Optane P4800X 375G SSD闪存卡加 Intel S3520 800G SSD的全闪存配置，其中Intel Optane SSD作日志存储，Intel S3520 SSD作数据存储。









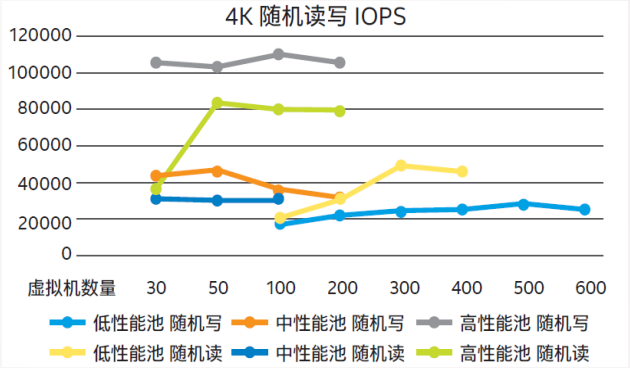
结论：HDD作为数据存储场景时机械盘性能已充分发挥，SSD作为数据存储场景时性能强劲

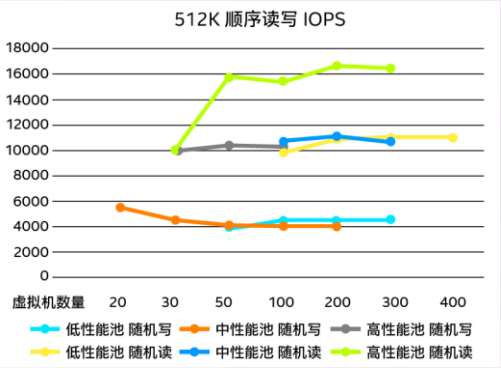
红帽Ceph行业案例 - 中国铁路

铁信云为什么选择红帽Ceph

2 大规模部署情况下Ceph的表现

Pool性能类型	Journal盘类型	存储节点数量(SAS HDD)	单节点OSD数量	Journal盘数量	Journal盘与OSD比	OSD总数
低	1.2T SAS HDD	27	10	未单独配置	N/A	270
中	480G SATA SSD	21	9	3	1:3	189
高	3.2T PCIE SSD	48	8	1	1:8	384



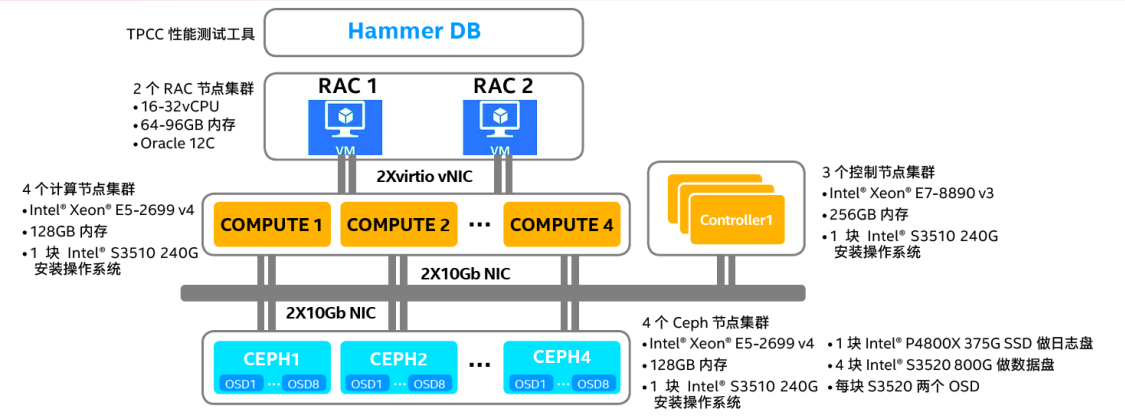


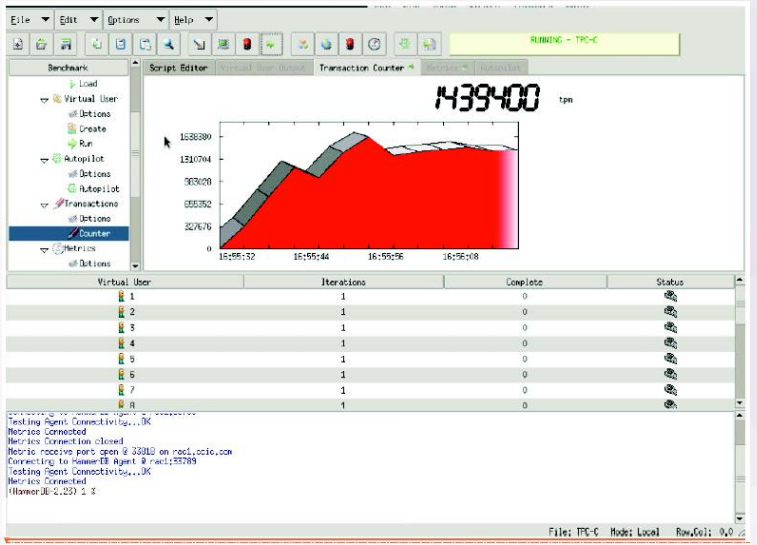
结论：存储池在VM并发量持续增加时，基本都能较好的保持整体最大IOPS，
能够满足大规模生产环境的业务需求

红帽Ceph行业案例 - 中国铁路

铁信云为什么选择红帽Ceph

3 关键业务应用下Ceph的表现





结论：经过硬件的配合（全闪配置）及软件的优化，Ceph可以满足关键业务（Oracle RAC）的性能诉求。

红帽Ceph行业案例 - 中国铁路

铁信云Ceph生产集群使用现状

多Ceph集群

主要承载业务：

- Web服务器
- 应用服务器
- 测试数据库

生产环境规模

- 星光外网：864个OSD
- 星光内网：795个OSD
- 酒仙桥内网：396个OSD

使用效果：

Ceph生产集群整体运行一年多，运行良好，没有影响生产业务，能够满足生产环境的需求。

Thank You!