如何评估Kubernetes持久化存储方案

关于焱融科技

焱融云 YAN RONG

2017年获得信雅达数千万元投资

SUNYARD

我们的理念：

致力于打造一流的软件定义存储和超融合产品

通过先进的互联网技术助力企业IT转型

我们的基因融合了：

国际级IT企业 + 一线公有云企业 = 企业级服务理念 + 最领先的互联网技术

IBM HUAWEI vmware 阿里云 金山云 WWW.KSYUN.COM 百度云 cloud.baidu.com

焱融分布式存储 YRCloudFile

利用x86服务器，搭建高性能、高可靠、可扩展的存储集群，提供文件、对象访问接口，支持K8S等容器编排平台

访问接口

- 文件存储，提供POSIX接口访问，对接各种文件访问应用
- 对象存储，S3标准接口，支持EC，适用于海量数据存储场景

高性能

- SSD + HDD，冷热数据自动分层
- 支持RDMA，性能比传统TCP连接方式提升80%

容器支持

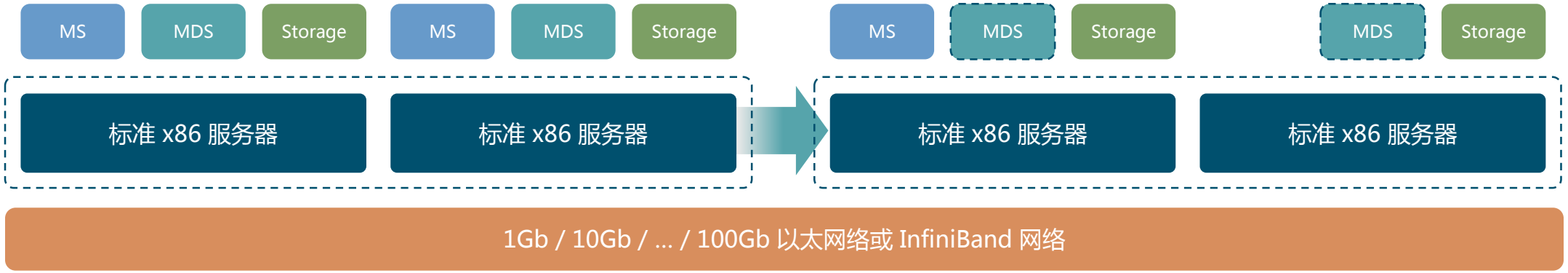
- 支持Kubernetes、Docker Swarm、Mesos等容器主流编排框架
- FlexVolume、CSI接口，动态、静态创建PV
- 支持容器存储QoS、Quota等高级特性

YRCloudFile 架构

YRCloudFile 最小规模 2 台服务器，可水平扩展

组件和架构：

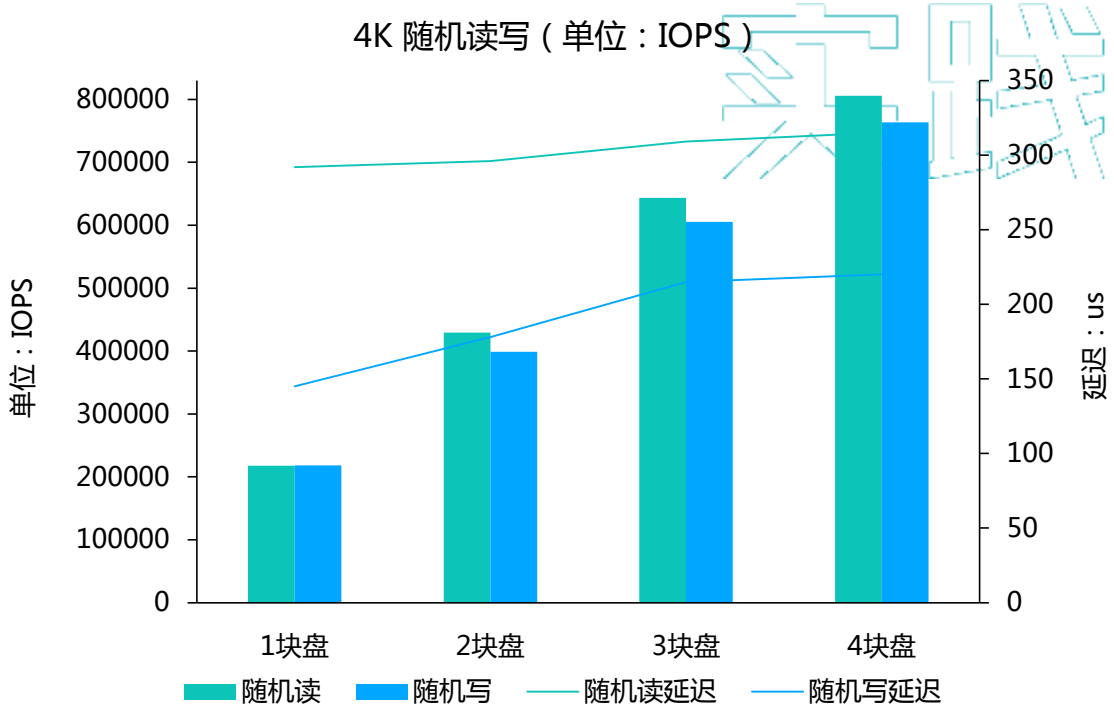
- MS：集群管理服务，通常运行在2-3台服务器上
- MDS：元数据服务，需要 SSD 磁盘支持，主要负责数据定位等工作，根据集群文件数量可随时水平扩展
- Storage：数据存储服务，可运行在 SAS / SATA / SSD / NVMe 磁盘上，存储实际数据
- 集群在部署、扩容时自动将MDS和Storage进行配对，用于进行副本复制
- 客户端通过以太网络或 InfiniBand 访问

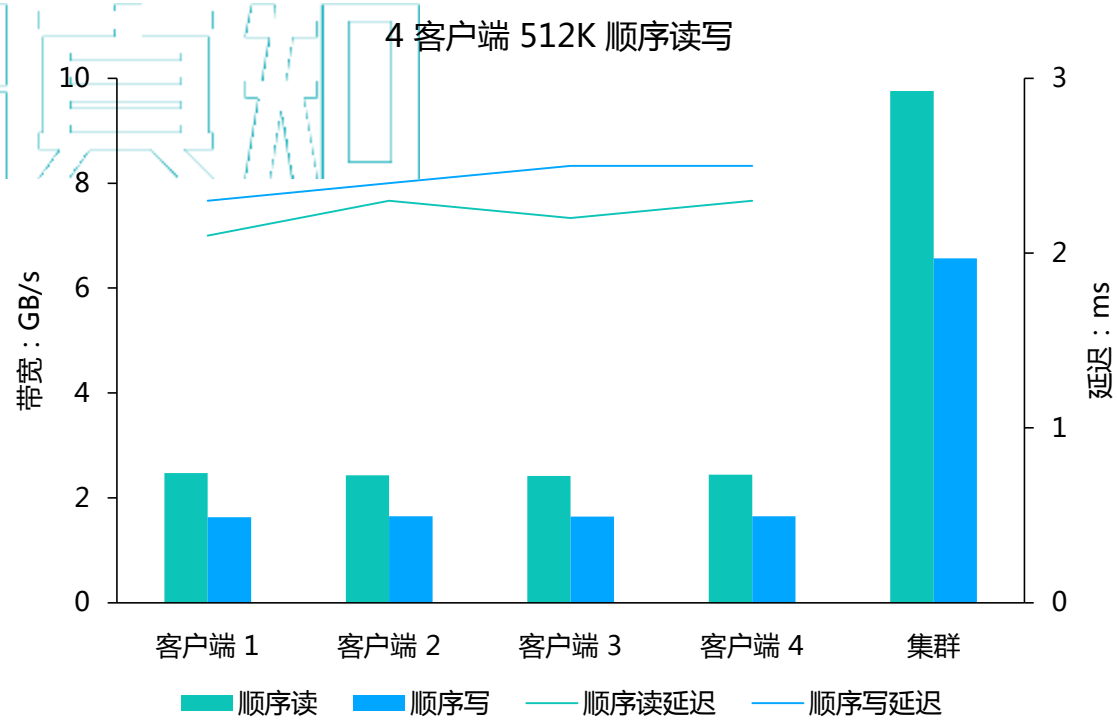


高性能 —— 为 GPU 提供强劲动力

两台服务器，四块 NVMe 磁盘，100 Gb RDMA 网络

- 随机读写（小文件）高 IOPS，顺序读写高带宽。将磁盘、网络性能发挥到极致，满足大/小文件混合读写需求
- IOPS 达80万，读带宽 10 GB/s，写带宽 6.5 GB/s
- 为 GPU 集群提供高性能数据支持， 加速 ML/DL 过程



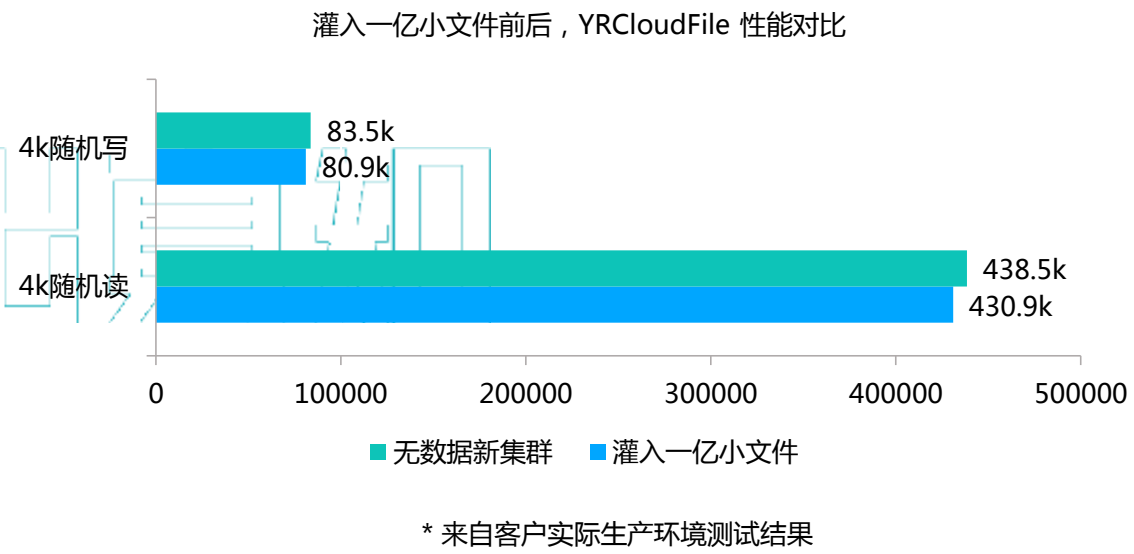


YRCloudFile 高扩展性 + 海量文件支持

机器学习必须基于海量数据，是 AI 业界的统一共识，扩展性和海量文件支持至关重要

YRCloudFile：

- 存储支持 1024 个节点，元数据支持 256 个节点
- 支持 10000+ 以太网客户端，2000+ RDMA 客户端
- 每 400GB 元数据空间，可支持 1 亿文件
- 单集群可支持千亿级别文件规模
- 元数据分布使用动态子树算法，避免数据访问热点

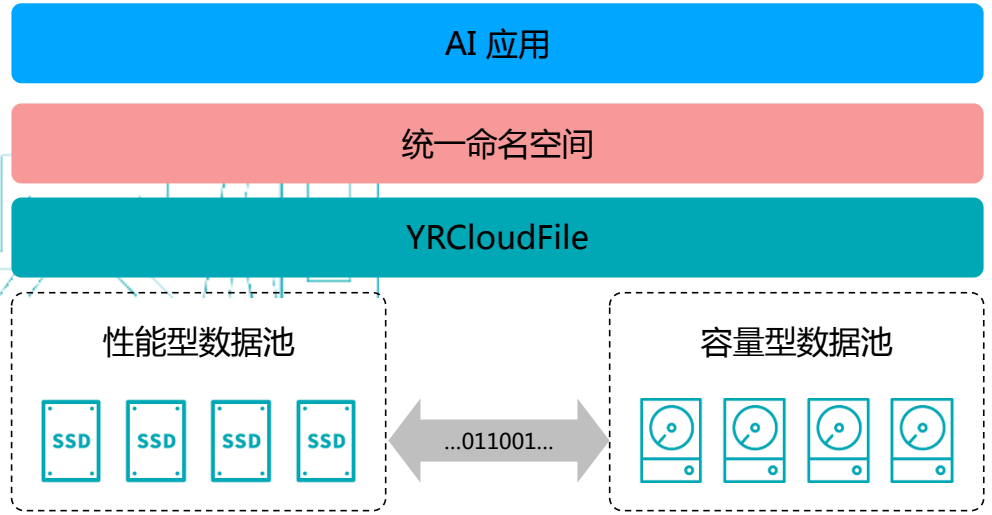


* 来自客户实际生产环境测试结果

YRCloudFile 统一命名空间数据治理

YRCloudFile：

- 集群采用基于以太网的 RDMA 技术进行数据传输
- 高性能数据池采用 NVMe / SSD 磁盘作为存储介质，通过副本方式进行数据可靠性保护
- 容量型数据池采用 SATA 磁盘作为存储介质，通过纠删码EC（Erasing Code）方式进行数据可靠性保护
- 性能型和容量型存储池位于同一命名空间管理下
- 可定义策略，在 YRCloudFile 集群内进行数据生命周期管理

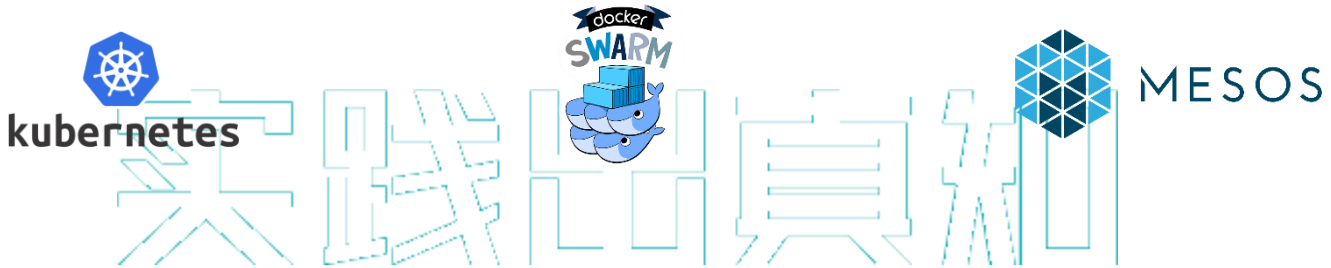


YRCloudFile vs GlusterFS

	YRCloudFile	GlusterFS
单客户端有效带宽	100% 有效带宽	1 / 副本数
交互式文件操作	**有元数据服务，体验好**	体验差，尤其是海量文件场景下
QoS	**支持**	不支持
扩展性	**好，扩容不影响性能，MDS 256节点，Storage 1024节点**	好，扩容期间对性能影响较大（100个节点以内）
数据访问网络和集群内部网络分离	支持	支持
Quota	支持	支持
接口	POSIX	POSIX / Fuse
文件Create/Stat性能	**比GlusterFS性能强10倍以上**	差
IOPS	**比GlusterFS性能强2倍以上**	一般
数据冗余	镜像	镜像/副本/纠删码
容器存储支持	支持	支持

支持多种容器编排框架

容器实例迁移时，能无感知地持续访问存储



Flex Volume 插件	✓		
CSI 插件	✓	✓	✓

焱融高性能容器存储特色功能

支持PV容量配额，保证容器应用的存储容量在可控范围

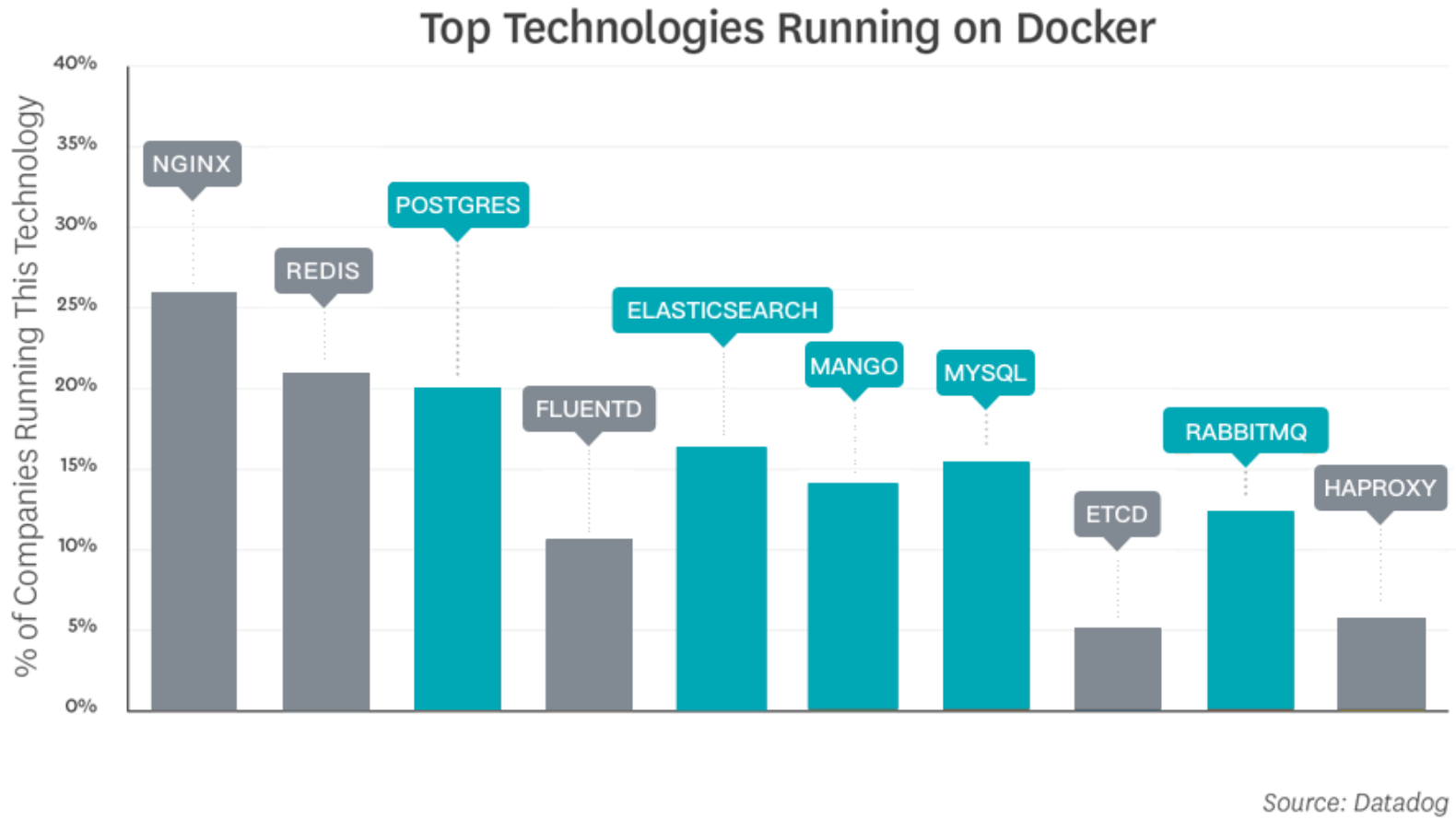
支持PV QoS，为容器提供持续稳定的性能支持，避免容器之间出现存储性能的抢占

支持RDMA，性能比传统TCP方式提升近一倍

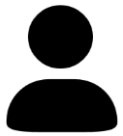
支持存储池划分，应对不同应用对存储特性的需求

为什么需要容器存储

有状态应用已经占容器应用的50%



存储成为容器应用面临的挑战



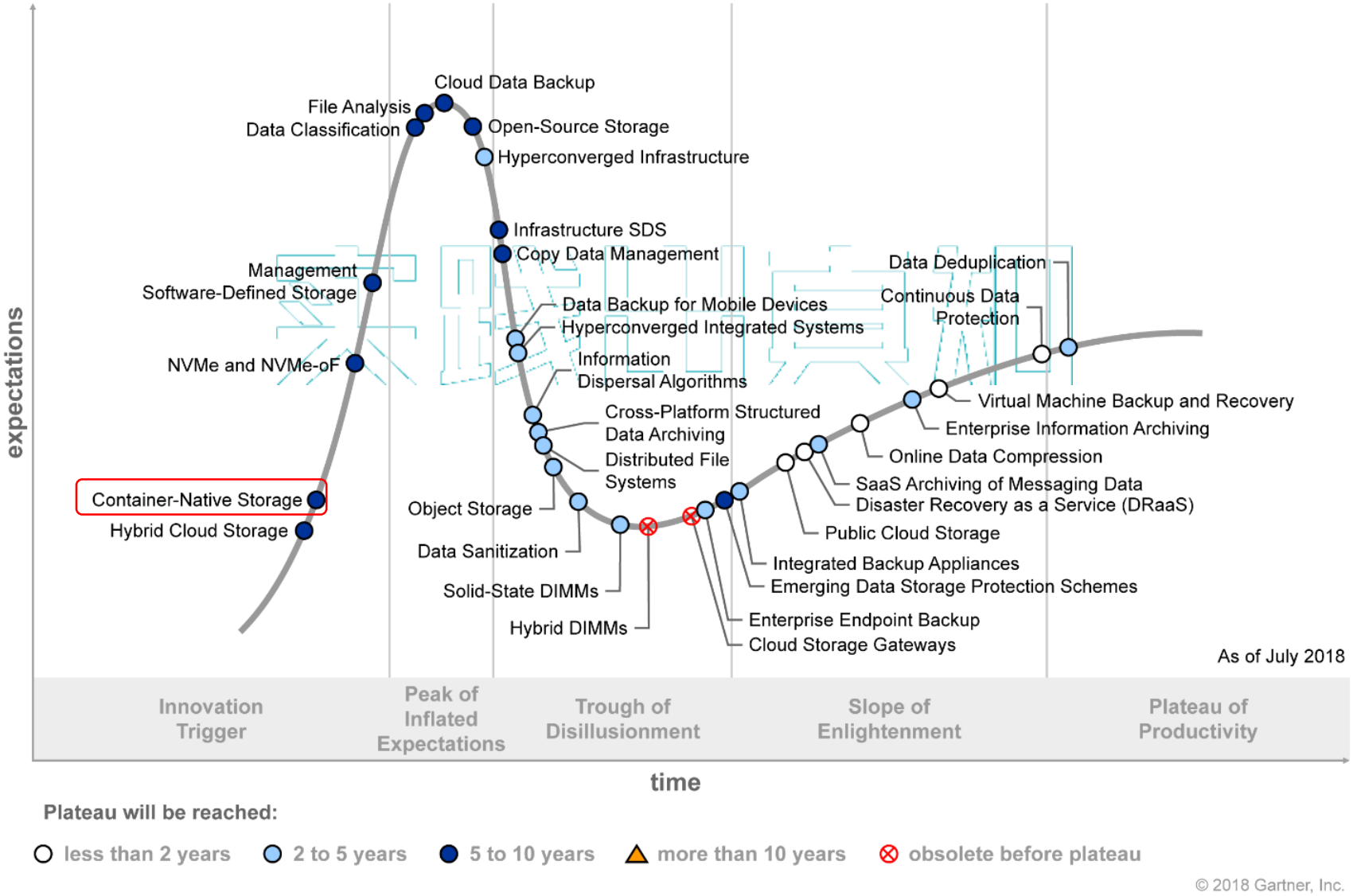
很多技术人员认为，容器存储是部署容器应用需要面临的问题，这些问题阻碍了容器在生产环境中大规模应用

The new stack, Multicloud Now a Chief Driver for Containers

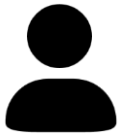
传统存储产品完全不是为容器时代设计，缺少与Kubernetes等主流编排平台的整合，挂载存储过程复杂，无法适应上层敏捷的应用需求，且成本高昂

An I&O Leader's Guide to Storage for Containerized Workloads, Gartner

Container-Native Storage首次出现在Gartner Storage技术成熟度曲线



不同视角看存储



用户

- 磁盘
- 目录

存储介质

- HDD
- SSD

产品形式

- DAS
- NAS
- SAN
- SDS

访问接口

- Block
- File
- Object

Kubernetes如何给存储定义和分类？

相关概念：

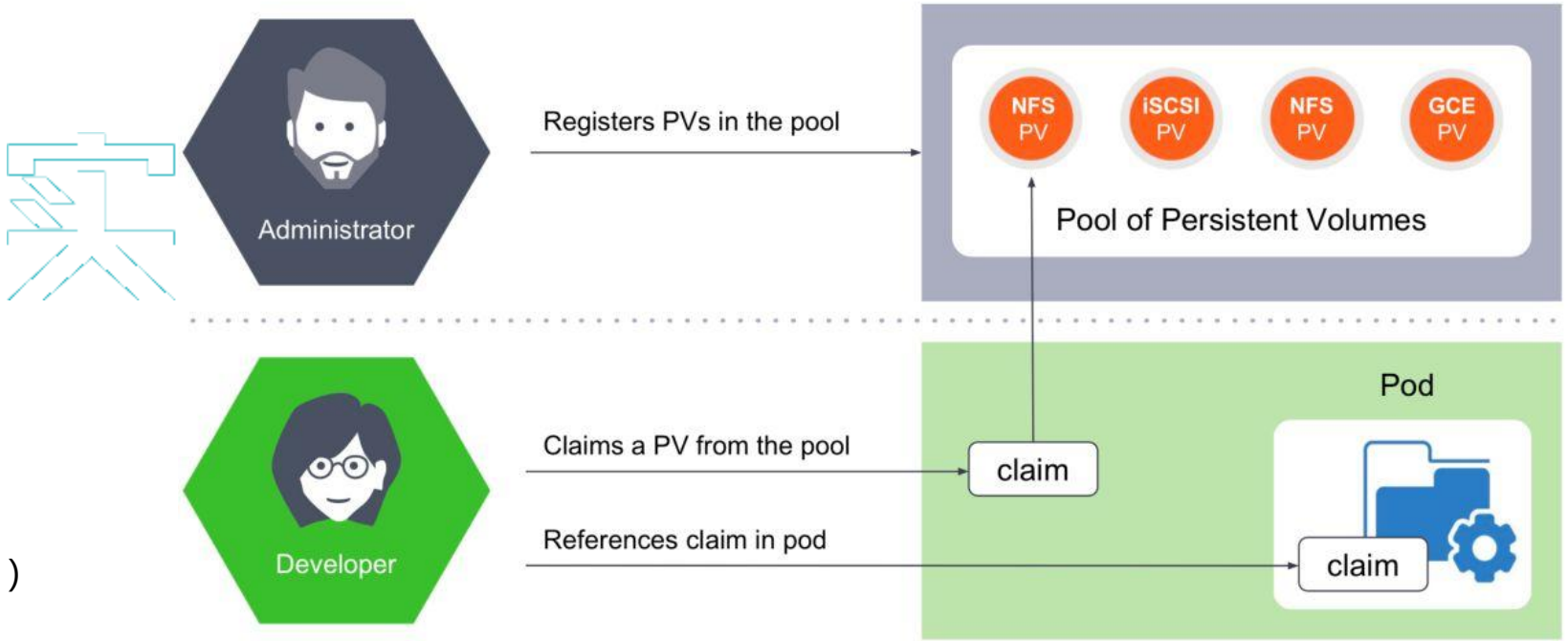
- PV（Persistent Volume）
- PVC（Persistent Volume Claim）

创建PV的方式：

- 静态PV
- 动态PV

创建静态PV的过程：

- 在K8S存储池中注册PV（创建PV）
- 从存储池中声明申请PV（创建PVC）
- 在Pod中引用PV声明（PVC）

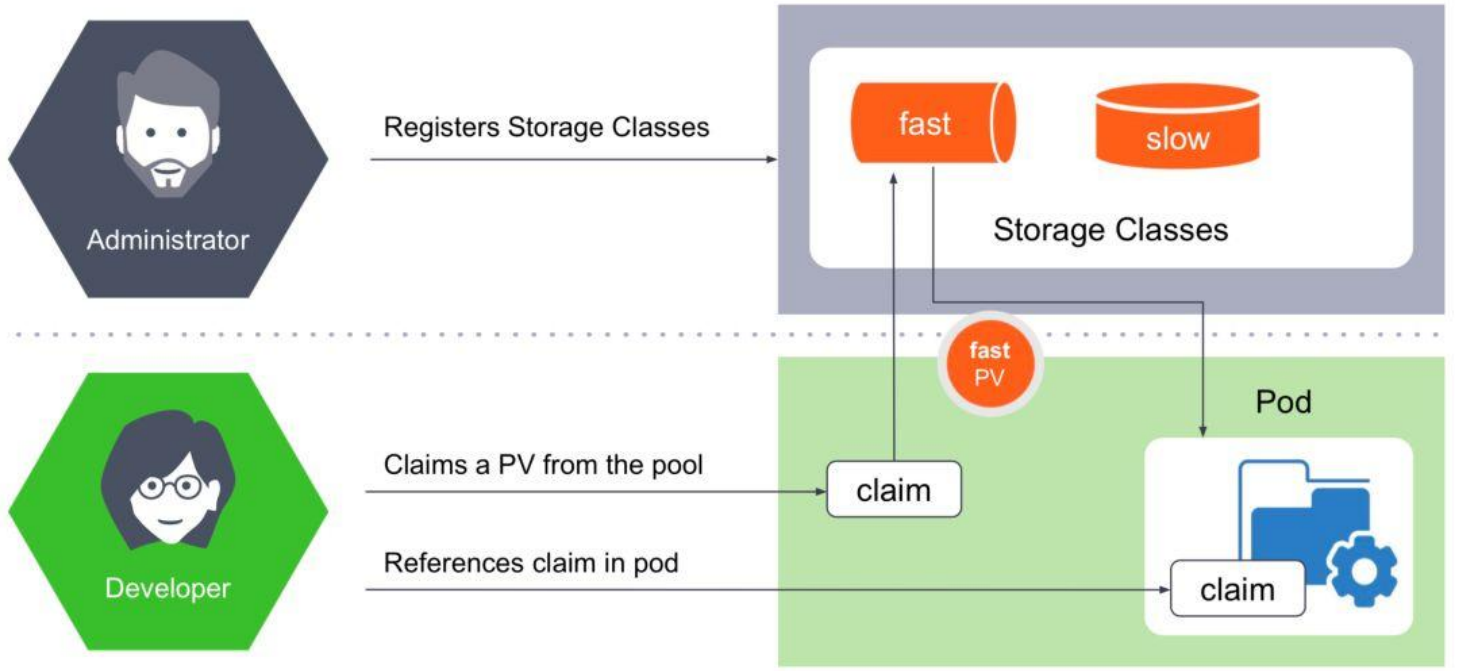


创建静态PV流程

Kubernetes中使用动态PV

创建动态PV的过程：

- 注册StorageClass（一次性操作）
- 直接从池中声明需要使用PV（创建PVC）
- 在Pod中引用PV声明（PVC）



创建动态PV流程

Kubernetes中PV的读写模式

PV支持三种读写：

- ReadWriteOnce（RWO）：PV只能被一个pod进行读写访问
- ReadOnlyMany（ROX）：PV可以被多个pod以read-only方式挂载
- ReadWriteMany（RWX）：PV可以被多个pod以read-write方式挂载

https://kubernetes.io/docs/concepts/storage/persistent-volumes/

PV读写模式与存储访问方式的关系：

- 块存储通常只支持RWO，例如AWS EBS、Azure Disk，有些块存储产品能支持ROX，例如GCEPersistentDisk、RBD、ScaleIO
- 文件存储（分布式文件系统），支持RWO/ROX/RWX三种模式，例如CephFS、GlusterFS和AzureFile
- 对象存储可通过Http RestAPI访问，不需要借助PV/PVC

Volume Plugin	ReadWriteOnce	ReadOnlyMany	ReadWriteMany
AWSElasticBlockStore	✓	-	-
AzureFile	✓	✓	✓
AzureDisk	✓	-	-
CephFS	✓	✓	✓
Cinder	✓	-	-
FC	✓	✓	-
Flexvolume	✓	✓	depends on the driver
Flocker	✓	-	-
GCEPersistentDisk	✓	✓	-
Glusterfs	✓	✓	✓
HostPath	✓	-	-
iSCSI	✓	✓	-
Quobyte	✓	✓	✓
NFS	✓	✓	✓
RBD	✓	✓	-
VsphereVolume	✓	-	- (works when pods are collocated)
PortworxVolume	✓	-	✓
ScaleIO	✓	✓	-
StorageOS	✓	-	-

业务类型对存储选型至关重要

业务类型	典型应用	场景特点	适用的存储
配置	集群配置、应用配置	需要并发访问，需要支持ROX/RWX读写模式，从而使不同集群或不同节点都能访问同样的配置文件。	分布式文件存储
日志	ElasticSearch	同样需要RWX读写模式，高吞吐，可能会产生大量小文件。日志分析场景会产生大量并发读操作。	分布式文件存储
数据库/消息队列/大数据	Kafka，MySQL，Cassandra，PostgreSQL，HDFS	对底层存储的要求就是高IOPS，低延迟。底层存储最好有数据冗余机制，上层应用就可以避免复杂的故障和恢复处理。	高性能分布式文件存储和高性能分布式块存储
备份	数据的备份或者数据库的备份	高吞吐，数据量大，低成本	文件存储和对象存储最优

市场上的容器存储产品和方案

对于容器场景，主要集中在5种方案：

- 分布式块存储，包括开源的Ceph、Sheepdog，商业产品EMC Scale IO、VMware vSAN。分布式块缺少对RWX读写模式的支持，限制了分布式块在容器存储中的应用。由于不支持RWX，因此在使用过程中，还会面临其他诸多问题。
- 分布式文件存储，包括开源的GlusterFS、CephFS、Lustre、MooseFS/LizardFS，商业产品EMC Isilon、IBM GPFS。性能是分布式文件存储是否适合容器存储的关键。
- Local-Disk，节点故障后，数据无法使用，可用性差。
- 传统NAS，性能受限于NAS机头，性能比较差，例如nfs。
- SAN存储，依赖于FC/iscsi协议的方案，例如openebs

市场上开源容器方案对比

存储产品	优点	缺点
Ceph	• 功能丰富 • 社区活跃 • 用户广泛	• 性能不稳定，延迟不稳定 • 代码复杂，代码成熟度参差不一 • 运维难度大
Sheepdog	• 代码简洁 • 功能丰富 • 运维简单	• 有数据不一致风险 • 社区已死
GlusterFS	• 代码模块清晰 • 社区较活跃 • 运维便捷	• 性能一般，小文件、海量文件性能差
CephFS	• Ceph的文件接口	• 不够成熟，尤其是多MDS方案
Lustre	• 性能好	• 需要使用集中式存储 • 运维难度较大
MooseFS/LizardFS	• 运维简单 • 功能丰富	• 小文件性能很差 • 社区不活跃

存储方案的评估策略

存储核心需求是稳定、可靠、可用，通过以下几个方面进行评估

可靠性

可靠性是指数据不丢失的概率，通常情况下，可通过计算得出几个9的数据可靠性，或给出最多允许故障盘/节点个数。评估方式就是拔盘、掉节点，只要数据不损坏，说明可靠性没问题。

可用性

数据可用性和数据可靠性很容易被混淆，可用性指的是数据是否在线。比如存储集群断电，这段时间数据是不在线，但是数据没有丢失，集群恢复正常后，数据可以正常访问。评估可用性的主要方式是拔服务器电源，再有查看存储的部署组件是否有单点故障的可能。

一致性

数据一致性是最难评估的一项，因为大部分场景用户不知道程序写了哪些数据，写到了哪里。该如何评估数据一致性呢？普通的测试工具可以采用fio开启crc校验选项，最好的测试工具就是数据库。如果发生了数据不一致的情况，数据库要么起不来，要么表数据不对。

块存储性能评估策略

存储的性能测试很有讲究，块存储和文件存储的侧重点也不一样

fio/iozone是两个典型的块存储测试工具，重点测试IOPS，延迟和带宽，以fio为例，测试命令如下：

```
fio -filename=/dev/sdc -iodepth=${iodepth} -direct=1 -bs=${bs} -size=100%
--rw=${iotype} -thread -time_based -runtime=600 -ioengine=${ioengine}
-group_reporting -name=fioTest
```

参数：iodepth，bs，rw和ioengine

- 测试IOPS，iodepth=32/64/128，bs=4k/8k，rw=randread/randwrite，ioengine=libaio
- 测试延迟，iodepth=1，bs=4k/8k，rw=randread/randwrite，ioengine=sync
- 测试带宽，iodepth=32/64/128，bs=512k/1m，rw=read/write，ioengine=libaio

文件存储性能评估策略

fio/vdbench/mdtest是测试文件系统常用的工具，fio/vdbench用来评估IOPS，延迟和带宽，mdtest评估文件系统元数据性能。

以fio为例，测试命令如下，与块存储的测试参数有一个很大区别，就是ioengine都是用的sync，用numjobs替换iodepth：

```
fio -filename=/mnt/yrfs/fio.test -iodepth=1 -direct=1 -bs=${bs} -size=500G
--rw=${iotype} -numjobs=${numjobs} -time_based -runtime=600 -ioengine=sync
-group_reporting -name=fioTest
```

- 测试IOPS，iodepth=32/64/128，bs=4k/8k，rw=randread/randwrite
- 测试延迟，bs=4k/8k，rw=randread/randwrite，numjobs=1
- 测试带宽，bs=512k/1m，rw=read/write，numjobs=32/64

mdtest是专门针对文件系统元数据性能的测试工具，主要测试指标是creation和stat，需要采用mpirun并发测试：

```
mpirun --allow-run-as-root -mca btl_openib_allow_ib 1 -host yanrong-
node0:${slots},yanrong-node1:${slots},yanrong-node2:${slots} -np ${num_procs}
mdtest -C -T -d /mnt/yrfs/mdtest -i 1 -I ${files_per_dir} -z 2 -b 8 -L -F -r -u
```

性能测试的其它场景

存储性能测试不仅仅测试集群正常场景下的指标，还要包含其他场景：

- 存储容量在70%以上或者文件数量上亿的性能指标
- 节点/磁盘故障后的性能指标
- 扩容过程时的性能指标

容器存储的特殊性

除了存储的核心功能（高可靠/高可用/高性能），对于容器存储，还需要几个额外的功能保证生产环境的稳定可用：

- Flexvolume/CSI接口的支持，动态/静态PV的支持
- 存储配额，对于Kubernetes的管理员来说，存储的配额是必须的，否则存储的使用空间会处于不可控状态
- 服务质量（QoS）
- 如何支持有状态Pod跨节点快速迁移
- 多数据中心容灾
- Kubernetes平台PV的数量会比虚拟机云平台的卷多数十倍，如何定位PV数据热点
- PV Resize
- 如何跟Kubernetes的监控平台关联
- 如何对Kubernetes镜像库提供存储支持
- PV是否能提供监控、告警
- PV内部数据分布状况

总结

Kubernetes持久化存储方案的重点在存储和容器支持上。因此首要考虑存储的核心功能和容器的场景支持。

- 存储的三大核心，高可靠，高可用和高性能
- 业务场景，选择分布式文件存储
- 扩展性，存储能横向扩展，应对业务增长需求
- 可维护性，存储的运维难度不亚于存储的开发，选择运维便捷存储产品
- 成本

焱融云
YAN RONG

THANK YOU

谢谢观看