白皮书

面向 Ceph 部署的英特尔解决方案

基于常见 Ceph 用例的英特尔® 组件基本配置指南。

Anjaneya (Reddy) Chagam
英特尔公司

Dan Ferber
英特尔公司

David J. Leone
英特尔公司

Orlando Moreno
英特尔公司

Yaguang Wang
英特尔公司

Yuan (Jack) Zhang
英特尔公司

Jian Zhang
英特尔公司

Yi Zou
英特尔公司

Mark W. Henderson
英特尔公司

简介

并非所有 Ceph 存储解决方案都是相同的，了解工作负载和容量要求对于设计 Ceph 解决方案至关重要。Ceph 可帮助企业通过统一的分布式集群提供对象存储、块存储或文件系统存储。在设计流程中，这些集群解决方案针对每项要求都进行了优化。该设计流程的首要因素包括 IOPS 或带宽要求、存储容量需求以及架构和组件选择，确保这些因素的合理性有助于完美平衡性能和成本，如图 1 所示。

不同类型的工作负载需要不同的存储基础设施方案。例如，关系数据库管理系统（RDBMS）工作负载在订单提交事务中需要 IOPS 和时延优化的存储，并避免受到使用限制，而对象归档可能需要容量优化。举例来说，视频流需要连续的、流传输带宽优化的解决方案。它不同于您在备份中可能使用的带宽优化解决方案，因为视频传输需要连续进行。

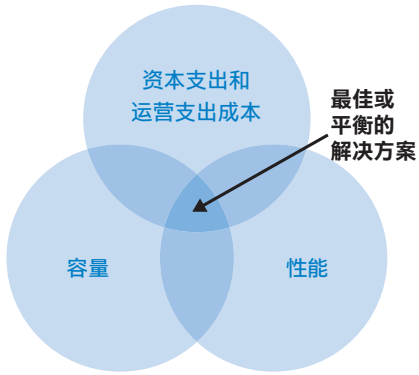


图 1. 不同的存储工作负载和容量需要平衡各种因素，如组件选择、集群组织和采用的 Ceph 参数等。

目录

本白皮书的预期用途

本白皮书旨在指导英特尔客户选择合适的英特尔架构解决方案。

本白皮书应该用于补充而非替代英特尔存储合作伙伴已经发布的参考架构和解决方案。在实际使用中，较为严谨和全面的英特尔解决方案参考架构（SRA）应始终优先于任何指南。

如需了解英特尔解决方案参考架构，请访问：www.intel.cn/content/www/cn/zh/storage/storage-products.html

如需了解英特尔合作伙伴参考架构，请访问 ISV 和/或 OEM 存储网页。

有关本白皮书的反馈、建议和修正，请发送至 mark.w.hendreson@intel.com

通常，我们可按照三种主要类型对存储工作负载分类：

集群优化条件	属性	使用示例
IOPS 优化	• 最高 IOPS • 最低的每 IOPS 成本 • 通常 3 个副本 • 单个节点少于或等于 10% 的集群（出于容错考虑）	• RDBMS 虚拟化 • 通常为块存储
带宽优化（即吞吐量）	• 最高的带宽 • 最低的每指定带宽单位成本 • 最高的每 BTU 带宽 • 最高的每瓦带宽 • 提高读取吞吐量通常需要 3 个副本 • 单个节点少于或等于 10% 的集群（出于容错考虑）	• 块或对象存储 • 视频流
容量优化	• 最低的每 TB 成本 • 最低的每 TB BTU • 最低的每 TB 瓦数 • 最高的每 TB 密度 • 实现最高可用容量通常需要纠删码 • 单个节点少于或等于 15% 的集群（出于容错考虑）	• 通常为对象存储 • 冷/归档存储

表 1. Ceph 集群优化条件

除工作负载特性描述外，在直接影响成本和性能的进一步集群定义中需要确定所需的存储容量。根据总体经验法则，在各种 TB 和 IOPS 优化存储要求中，配备十几块 2.5 英寸固态盘（SAS/SATA）或 4 块 NVMe 固态盘的服务器可能具有较出色的每美元 IOPS。另一方面，配备 60 块或更多 3.5 英寸机械硬盘的 PB 级容量优化节点目前可提供最佳的每美元资本支出。然而，不断进步的闪存存储技术（具有更低的故障率和更长的摊销周期）在运营支出方面正对机械硬盘构成挑战。两者之间的产品需要节点保持适当平衡：具有较少容量但具有较高的 CPU 和网络吞吐量，适当混合使用固态盘和机械硬盘，确保特定工作负载实现最低的资本支出和运营支出。

英特尔® 硬件和软件配置为 Ceph 提供支持

Ceph 是一款基于冗余服务器的高度可扩展的开源存储产品，可在单一存储集群上提供对象、块和文件系统存储。Ceph 可在具有裸机或虚拟化配置的大容量英特尔架构硬件上运行。Ceph 之所以广受欢迎，是因为其支持根据工作负载要求从不同存储厂商中灵活选择硬件配置，而且通过块、文件和对象接口的统一访问支持广泛的应用。Ceph 支持灵活使用上游 Linux 内核驱动程序和用户模式 QEMU/libvirt 接口来利用块存储。它可扩展至数百台存储服务器和数千个存储客户端。

Ceph 社区

有关 Ceph 的更多信息，请访问 http://ceph.com。许多商业和教育机构使用 Ceph，如需了解其中部分机构，请参见他们在 Ceph Day 上的演讲：http://www.slideshare.net/Inktank_Ceph。如需了解 Ceph 使用 OpenStack 的信息，请访问 http://www.openstack.org/surveys/。英特尔存储解决方案方面的信息发布在 http://www.intel.cn/content/www/cn/zh/storage/storage-products.html 上。

英特尔的 Ceph 开发人员为上游 Ceph 社区贡献了大量基于性能的增强功能和特性。事实上，英特尔在这方面的贡献通常仅次于 Redhat。英特尔还在全球举办了多场 Ceph Day 会议，并在 2015 年面向 Ceph 开发人员主办了首届社区面对面黑客马拉松，聚焦性能这一重要主题。

本白皮书描述了当前最常见的 Ceph 块和对象用例。对于每个用例，本文阐述了面向 Ceph 存储集群的典型英特尔架构硬件配置，并提及了该配置的任何公开性能信息。有时，还可考虑具有不同特征的第二种配置。也就是说，本文及本文中的配置并非参考架构，并不保证能够实现特定的性能水平，只是基于我们及社区经验的指导性配置。如果需要固定的性能水平或性价比水平，您可从提供 Ceph 解决方案的众多英特尔合作伙伴中选择一个进行合作。您也可通过访问 http://www.intel.cn/content/www/cn/zh/storage/storage-products.html 或相应的 Ceph 解决方案提供商网站查看详细的参考架构。

未包含的元素

- 请注意，相关指导并非针对客户端节点，因为它们差异较大
- Ceph 文件（Cephfs）的用例和配置

Ceph 池类型及与硬件和用例的关系

Ceph 具有三类池：1）复制池，2）纠删码池，3）高速缓存层池。池类型的选择与指导性硬件配置无关，简介中提及了复制和纠删码存储类型。然而，性能数据的确取决于所用池的类型设置。这里的块用例较多显示了复制池的数据，对象用例使用纠删码池，这两种池适合各自的存储类型。

高速缓存层池尚未广泛使用，因为其性能仍在优化中，因此，这些指导性配置没有关于高速缓存层的数据。同样，本文也未介绍文件用例，这些用例尚处于 Ceph 生产使用的早期阶段，不像块和对象用例一样已经非常普及。

涵盖的 Ceph 用例

本文介绍了部分最常见的块和对象用例。更多详情请访问：http://pad.Ceph.com/p/hack-athon_2015-08

在阅读本文和回顾我们与您讨论的用例、配置及性能要素时，您可思考如下两种基本的高级 Ceph 存储节点硬件配置：

- 标准 Ceph 配置。采用英特尔® 至强® 处理器 D 或英特尔® 至强® 处理器 E5 节点，采用英特尔® 固态盘数据中心级产品家族 PCIe 设备闪存存储（SATA 或 PCIe NVM 固态盘）作为日志存储，机械硬盘用作数据驱动器。可添加英特尔® 高速缓存加速软件（CAS）来提升性能。
- 高性能 Ceph 配置。采用英特尔®至强® E5 节点，采用英特尔闪存 NVM/PCIe 固态盘作为日志存储，英特尔 SATA 固态盘用作数据驱动器。可添加英特尔® CAS 来提升性能。

复制和纠删码池等 Ceph 存储池类型分层部署在这些硬件配置上，这些将在用例中进行讨论。

块用例

最常见的块用例是：

- 虚拟桌面托管：VDI 或虚拟桌面映像，或类似应用
- 数据库或类似应用
- 通用：表示面向应用的通用块存储

下文描述了 VDI、数据库和通用块模式的 I/O 特性。

对象用例

最常见的对象用例是：

- 数字视频录制（DVR）或类似应用
- 视频点播（VOD）或类似应用
- 备份或类似应用

下文描述了 DVR、VOD 和备份的 I/O 特性。

名称	用途	池配置/复制	集群前置条件
通用	通用原始综合性能	3 个副本	60% 满
VDI	虚拟桌面环境	3 个副本	60% 满
VirtInfr	虚拟客户端	3 个副本/快照	60% 满
OLTP	RDBMS 性能	3 个副本	60% 满
NoSQL	Cassandra 等	3 个副本	60% 满

表 2. DVR、VOD 和备份的 I/O 特性

英特尔硬件和软件组件

这些指导性系统采用了下列英特尔硬件：

- 英特尔® 至强® 处理器家族
- 英特尔® 固态盘
- 英特尔® 千兆以太网服务器适配器

这些指导性系统采用了下列英特尔软件：

- 英特尔® 存储加速库
- 英特尔® 高速缓存加速软件

本文还描述了英特尔提供的下列社区工具：

- 虚拟存储管理器
- Ce-Tune，基于用户界面的 Ceph 基准测试和分析工具
- 用于基于对象的基准测试的 COSBENCH

以下示例只是众多实施特定解决方案的方式之一，结果可能有所不同。您首选的系统提供商可提供英特尔架构解决方案。

通用指南

以下内容适用于我们讨论的所有用例。

Ceph 版本

每个 Ceph 版本通常可提供增强的性能、可靠性和功能。我们推荐部署长期支持（LTS）版本。在撰写本文时，最新的 LTF 版本为 “hammer” 或 0.94 版。我们预计后续的硬件（CPU、闪存固态盘/NVMe）和软件（Ceph、ISA-L、SPDK 等）版本将持续改进。

Ceph 和 Linux 操作系统调优参数

英特尔的 Ceph 开发团队维护着一份 Ceph 调优指南。该指南中的建议普遍适用于所有用例，除非用例描述中另有说明。参见附录 A 查看各种调优的概述，完整文档需在签署保密协议的情况下查看。

NUMA 配置

下图显示了单个 CEPH 节点 NUMA 配置。CPU 支持可在 BIOS 内启用的超线程功能。逻辑核心编号分别为 0-5、12-17（CPU 0）和 6-11、18-23（CPU 1）。

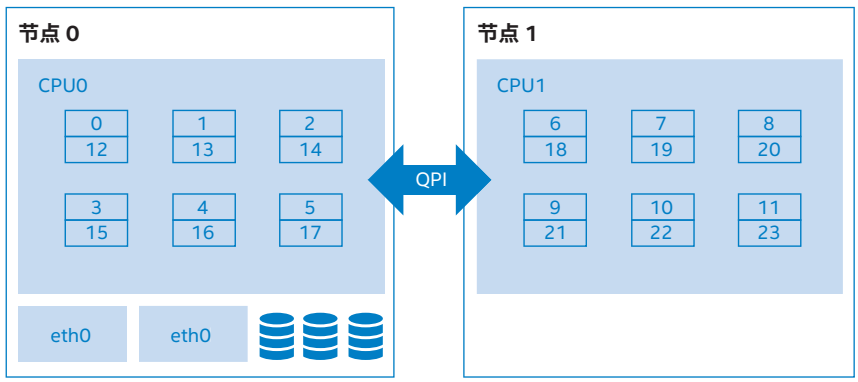


图 2. 单个 CEPH 节点 NUMA 配置

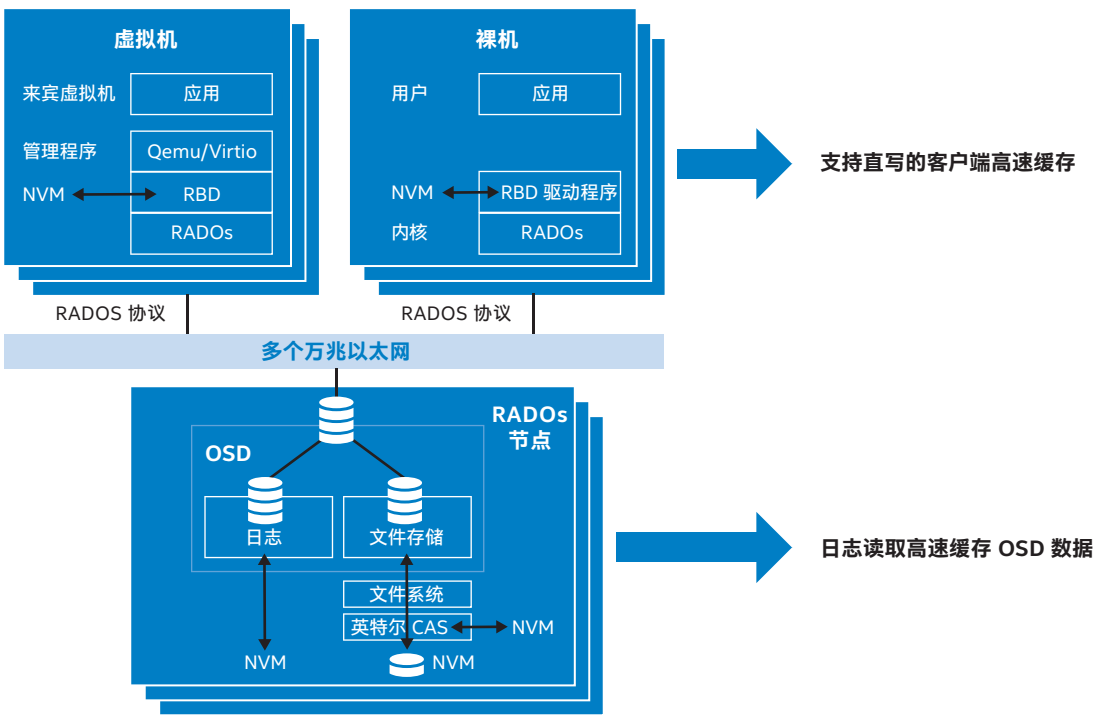


图 3. 显示 Ceph 基础设施使用固态盘/NVM 的顶部视图

提供存储连接功能的网卡和 HBA 连接至 NUMA 节点 0，后者是在逻辑核心 0-5 和 12-17 上运行的进程的最佳配置。

为优化 CEPH 进程的硬件访问速度，Ceph 启动脚本需进行修改以迫使 Ceph-osd 进程在 CPU0 上运行。详情请见附录 B，但该操作通常使用下列命令实施：

```
setaffinity= "numactl --mem-bind=0 --cpunodebind=0"
```

该操作还可释放 CPU1 上的资源，让它可以运行 Velocix 中间件服务（在真实 cDVR 配置下运行），并确保配置更符合现实需求。

英特尔固态盘/NVM 技术：Ceph 概述

本节介绍通用固态盘/NVM 技术，以及兼顾性能和成本的存储节点配置建议。

面向 Ceph 的固态盘使用场景

独立的固态盘池

将不同服务器/节点的固态盘部署至一个与机械硬盘池分离的 OSD 固态盘池中。该固态盘池将专用于高性能应用。

多个性能/容量池可共存于同一 Ceph 集群中，图 5 列出了三种推荐配置。

三种存储节点配置

良好/标准配置

固态盘用作日志和高速缓存驱动器，机械硬盘用于存储 OSD 数据，比率如下：

强烈建议使用 PCIe*/ NVMe 固态盘实现高性能和低时延，因为：1）NVMe 技术针对非易失性内存（NVM）/固态盘进行了全新的设计和优化，而 PCIe 让存储更靠近 CPU 以减少时延，2）NVMe 正在推动数据中心从 SATA/SAS 转向 PCIe 接口。

1. **PCIe/NVMe** 固态盘：机械硬盘，1:12，英特尔 PCIe/NVMe P3700 用作日志驱动器
2. **SATA** 固态盘：机械硬盘，1:4，英特尔 SATA S3700 用作日志驱动器

协同使用英特尔 CAS 与基于 hint 的 I/O 分类、分配和优先级划分技术，以及英特尔® 差异化存储服务（DSS），可对能提升性能的存储元素进行识别和高速缓存。由于 Ceph 具有分布式性质，高速缓存和日志可位于同一固态盘上。

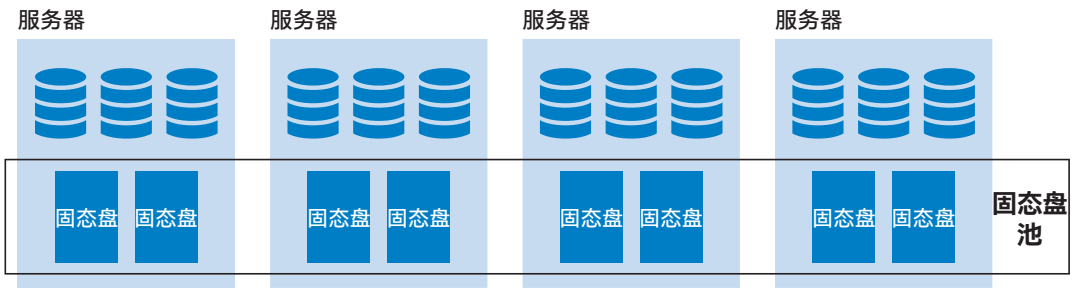


图 4. 独立的固态盘池

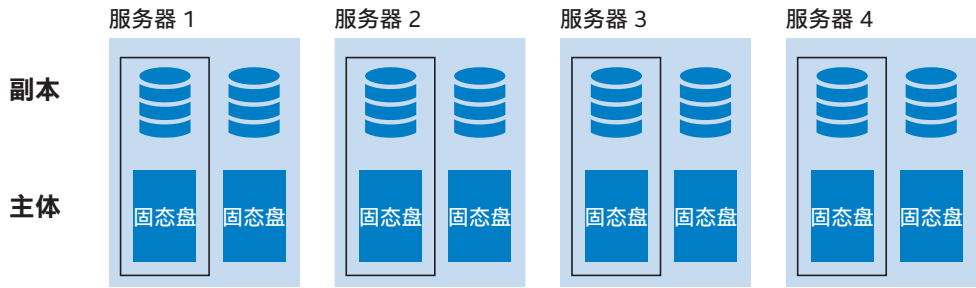


图 5. 推荐配置

CEPH 存储节点 — 良好	
CPU	英特尔® 至强® 处理器 E5-2650v3
网卡	10GbE
驱动器	1 个 1.6 TB P3700 12 个 4 TB 机械硬盘（比率为 1:12）（P3700 用作日志和高速缓存驱动器）
软件	英特尔 CAS RSTe/MD4.3（可选）

表 3. 良好配置

CEPH 存储节点 — 更好	
CPU	英特尔® 至强® 处理器 E5-2690
内存	128 GB
网卡	双 10GbE
驱动器	1 个 800 GB P3700 4 个 1.6 TB S3510 （P3700 用作日志和高速缓存驱动器）
软件	英特尔 CAS

表 4. 更好配置

CEPH 存储节点 — 最好	
CPU	英特尔®至强® 处理器 E5-2699v3
内存	>=128 GB
网卡	2个 40GbE 4 个双 10GbE
驱动器	4-6 个 2 TB P3700

表 5. 最好配置

该配置面向具有高吞吐量性能的高容量存储。

请注意，在 Ceph 配置中，故障域是整个存储节点。因此，具有单个固态盘（基于 SATA 或 PCIe）并不表示 Ceph 系统会发生单点故障。硬件系统具有冗余性。这与非横向扩展系统形成了鲜明对比，故障域通常包含在单个非横向扩展系统中。横向扩展故障域的优势在于 RAS、可扩展性、站点级复制、EC 和节点升级时的持续运行。

高级配置：

在该配置中，PCIe/NVMe 固态盘用作日志驱动器，大容量低成本 SATA 固态盘用作 OSD 数据驱动器。对于需要较高性能、尤其是需要较高 IOPS 和 SLA 且具有中等存储容量要求的用例/应用，该配置最为经济高效。

最佳性能配置：

该配置全部使用 NVMe/PCIe 固态盘，是需要最高性能和低时延的用例/应用的最佳性能存储解决方案。

示例：4 - 6 个 P3700 2 TB 固态盘用作 OSD。

固态盘选择指南

企业固态盘对比客户端固态盘

固态盘并非全部一样，同一固态盘厂商也可能有不同的固态盘生产线。一般而言，客户端固态盘和企业固态盘在性能、可靠性和耐用性方面具有不同的测试要求和规格。例如，客户端固态盘具有每天 8 小时的使用规定和较低的数据集成要求，而企业固态盘需要每天 24 小时全天候使用，具有端到端数据集成和电源保护等功能。更多详情参见行业标准 JEDC 的固态盘可靠性和耐用性要求。

每日整盘写入次数

不同于机械硬盘，固态盘是一种消耗性资源，只能写入较多但有限的次数。固态盘具有一定的大小，其使用时的性能特征与其保修和性能预期一致。术语 “每日整盘写入次数”（DWPD）是测量固态盘耐用性的行业标准/规格，每日可写入的数据量取决于 JEDC 工作负载（4K 块大小随机写入）、DWPD 和固态盘可持续正常运行的年数。

例如，英特尔 P3700 800 GB 的 DWPD 值为 10，它可持续运行 5 年，这将支持主机在 5 年内每日写入共 8 TB 数据。

下方表 6 列出了典型的 DWPD 值。相关数值仅供参考，强烈建议在生产前评估/测量耐用性，在生产中通过 SMART 指标监控固态盘耐用性。例如，可通过 SMART 属性 E2、E3 和 E4，以及 E9 和 E8 等实时/生产寿命/耐用性指标对英特尔固态盘进行离线耐用性评估。

英特尔数据中心级固态盘

根据性能、耐用性和成本，英特尔数据中心级固态盘可分为三类。

英特尔 PCIe/NVMe 固态盘的型号以 “P” 开头，如 P3700。英特尔 SATA 固态盘的型号以 “S” 开头，如 S3700。两种规格的固态盘均提供所有三种耐用性级别。

参见英特尔数据中心级固态盘产品家族，了解有关英特尔固态盘的更多信息，包括英特尔傲腾和 3D NAND 技术。

固态盘 SMART 属性

表 7 和 8 中的 SMART 属性用于监控固态盘运行状况，如耐用性指标、错误日志、主机读取/写入、NAND 读取/写入等，可通过英特尔® 虚拟存储管理器（VSM）等外部管理软件管理/监控这些属性。

固态盘类型	每日整盘写入次数（DWPD）	保修（年数）	备注
高耐用性（P3700 或 S37xx）	10	5	高密集性随机写入
中等耐用性（P3600 或 S36xx）	3	5	平衡的读取与写入
标准耐用性（P3500 或 S35xx）	0.3	5	高密集性读取工作负载

表 6. 英特尔数据中心级固态盘

字节	字节数	详细信息
0	1	临界警告：若需要，设置如下位数，标记各种警告源 位 0：可用备用容量低于阈值 位 1：温度超过阈值 位 2：由于过多的介质或内部错误而降低可靠性 位 3：介质处于只读模式 位 4：易失性内存备份系统已失效（如 PCI 电容器测试故障） 位 5-7：预留
1	2	温度：整个设备的当前温度（开氏度）
3	1	可用备用容量：包含剩余备用容量的标准百分比（0 到 100%）
4	1	可用备用容量阈值
5	1	使用百分比估计
21	16	读取的数据单位
48	16	写入的数据单位
64	16	主机读取命令
80	16	主机写入命令
96	16	控制器忙时
112	16	供电周期
128	16	通电时间
144	16	不安全关机
160	16	介质错误
176	16	错误信息日志条目数量

表 7. 示例，NVMe 运行状况日志表（日志页标识符 02h）

字节	字节数	属性	描述
0	1	AB（程序故障次数）	原始值：显示程序故障总数。 标准值：开始为 100，显示容许程序故障的剩余百分比。
3	1	标准值	
5	6	当前原始值	
12	1	AC（擦除故障次数）	原始值：显示擦除故障总数。 标准值：开始为 100，显示容许擦除故障的剩余百分比。 原始值： 1-0 字节：最小擦除周期 3-2 字节：最大擦除周期 5-4 字节：平均擦除周期 标准值：从 100 缩减至 0。
15	1	标准值	
17	6	当前原始值	
24	1	AD（平均擦写次数）	
27	1	标准值	
29	6	当前原始值	
36	1	B8（端到端错误检测次数）	原始值：按硬件报告检测和纠正的端到端错误的数量。 标准值：始终为 100。 原始值：显示接口发生循环冗余检验（CRC）错误的总数。 标准值：始终为 100。 原始值：测量固态盘发现的损耗（从重置工作负载计时器属性 E4h 起），并且表示为最大额定周期的百分比。将原始值除以 1024，获得具有三个小数点的百分比。 标准值：始终为 100。
39	1	标准值	
41	6	当前原始值	
48	1	C7（CRC 错误数量）	
51	1	标准值	
53	6	当前原始值	
60	1	E2（定时工作负载，介质损耗）	
63	1	标准值	
65	6	当前原始值	
72	1	E3（定时工作负载，主机读取百分比）	原始值：显示 I/O 操作中读取操作所占的百分比（从重置工作负载计时器属性 E4h 起）。 报告为从 0 到 100 的整数百分比。 标准值：始终为 100。
75	1	标准值	
77	6	当前原始值	
84	1	E4（定时工作负载，计时器）	原始值：测量经过的时间（从启动此工作负载计时器起经过的分钟数）。 标准值：始终为 100。 原始值：报告控制状态百分比和事件数量。 0 字节：控制状态报告为整数百分比。 1-4 字节：控制事件数量。 激活热量控制的次数。重启时保存。 5 字节：预留。 标准值：始终为 100。
87	1	标准值	
89	6	当前原始值	
96	1	EA（热量控制状态）	

表 8. 示例，英特尔 NVMe/PCIe 固态盘的其他 SMART 属性

Ceph 块存储：虚拟桌面托管用例 1

该用例的焦点是 4K 块随机 IOPS。在模拟虚拟机托管 IOPS 用例中，不同客户使用不同比例的读取和写入操作。同样，每位客户应为每个虚拟机用户提供不同的 IOPS。在 2015 年春季温哥华 Open Stack 峰会上的演讲中，CERN 表示他们为每个虚拟机用户提供的默认 IOPS 为 100。[1]

该用例基本上能以最低成本实现 IOPS 目标。这与一些数据库或定制 IOPS 用例形成了鲜明对比，这类用例希望实现最高的 IOPS，但 IOPS 越高，成本也越高。如需了解最高的 4K 随机 IOPS，请参见我们的数据库用例及其中的指导性配置。

我们发现，虚拟机托管 IOPS 的指导性系统可实现 100% 4K 随机读取以及 100% 随机写入的性能。在远程客户端节点上的 FIO I/O 测试中，虚拟机也实现了这种写入和读取性能。

我们假设 Ceph 可实现每虚拟机 100 IOPS，并据此推测该指导性配置在保持 IOPS 性能的情况下可支持多少虚拟机。

首先，我们将说明指导性硬件配置，然后展示性能结果。

指导性硬件配置

（有关软件配置信息，请参见常规软件配置和调优部分）。

Ceph 存储集群

- 最少 4 台服务器
- 每台服务器 10 个 OSD
- 每台服务器 2 个英特尔 SATA 固态盘
  - 每个固态盘 5 个 OSD 日志分区
- OSD 主机上的监视器（最少 3 台监视器）
- 架顶式交换机

网络：架顶式

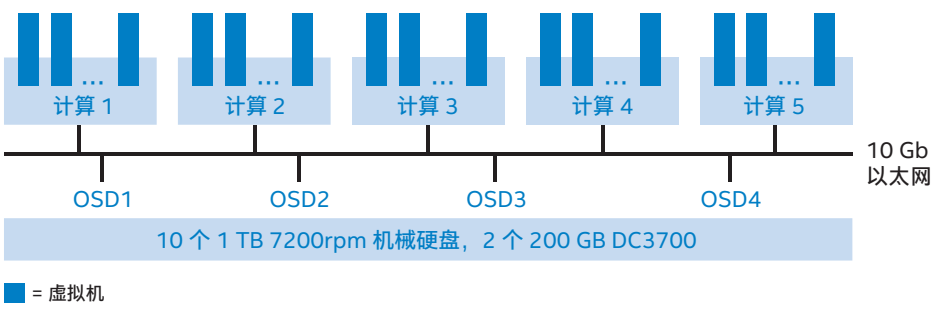
- 1 台 10GbE 公共（面向公共网络客户端）交换机
- 1 台可选 1GbE 交换机 - 管理
- 1 个可选 10GbE 专有（面向集群网络集群）网络 – 集群数据
- 可选IPMI交换机

每个 Ceph 存储节点

- 处理器
  - 1 个英特尔® 至强® 处理器 E3-1200 v2 CPU
- 内存
  - 16 GB（若在节点上还运行 Ceph 监视器，则为 32GB）
- 机械硬盘
  - JBOD 磁盘控制器（8 个端口）+ 芯片组 SATA（4 个端口）
  - 数据磁盘：10 个 1 TB 企业级 SATA 3.5 英寸
- 日志固态盘
  - 2 个 200 GB 英特尔 DC S3700
- 网络
  - 1 个面向公共网络数据的 10GbE 英特尔® 82599ES 端口
  - 1 个面向集群网络数据的可选 10GbE 英特尔® 82599ES 端口
  - 可选 1GbE 英特尔® 82574L 管理端口
- 管理
  - 可选 IPMI 端口

英特尔针对 Ceph Firefly* 版本测试了上述配置。在测试之后，英特尔陆续推出了多个 Ceph 版本，以及 PCIe 固态盘和英特尔® 至强® 处理器 D 产品家族。预计后续版本和技术将能提升性能，该文档的更新版本将反映相关信息。

目前测试的配置如下图所示：



- 10GbE 网络
- 面向 Ceph 集群的英特尔® 至强® 处理器 E3 服务器
  - 16 GB 内存（每个节点）
  - 10 个 1 TB SATA 机械硬盘通过 LSI9205 HBA（JBOD）用于存储数据，每个硬盘分成一个 OSD 守护进程分区
  - 2 个 SSD 用作日志驱动器，直接连接 SATA 控制器，每个 OSD 使用 20 GB（3、3、4）
- 5 节点客户端集群
  - 虚拟机主机系统

图 6. 测试配置

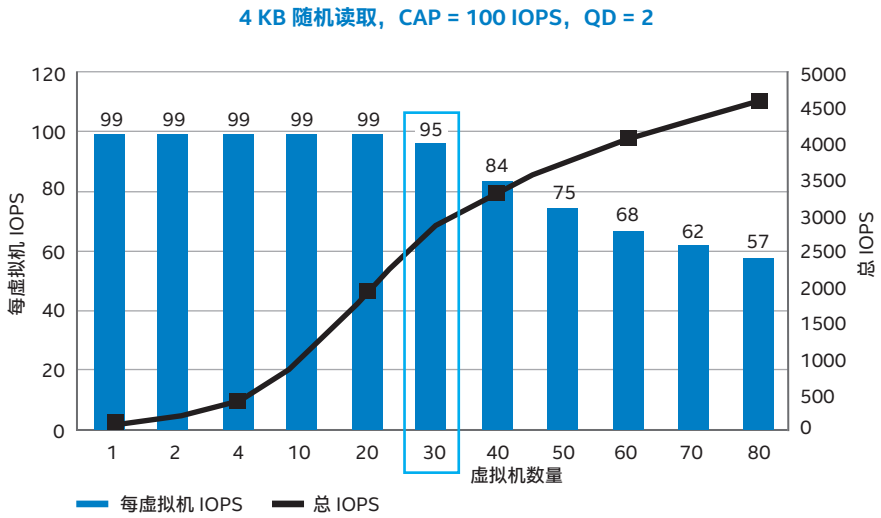


图 7. 40 个机械硬盘阵列 4K 随机读取：共 4500 IOPS，每机械硬盘 112 IOPS

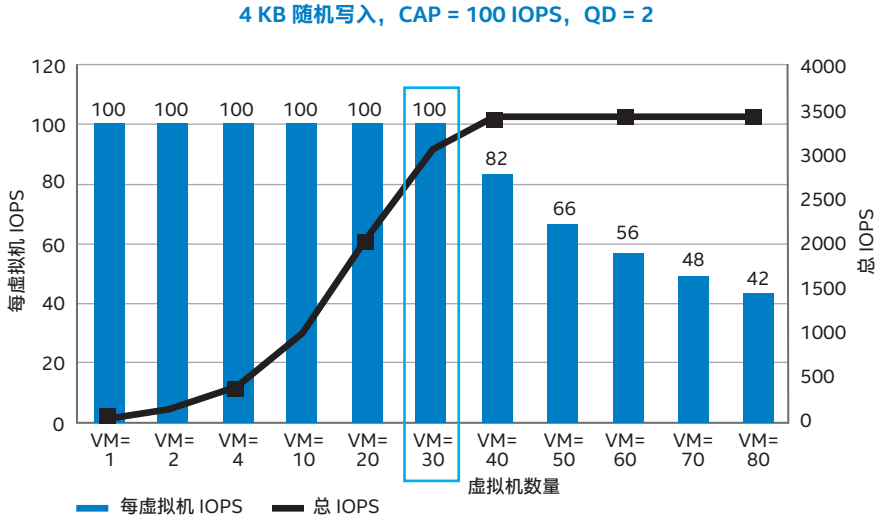


图 8. 40 个机械硬盘阵列 4K 随机读取：共 3358 IOPS，112 IOPS（用于 2 个副本），每机械硬盘 84 IOPS

如需计算日志可使用大约多少空间，请访问：

http://docs.Ceph.com/docs/v0.94/rados/configuration/osd-config-ref/

Ceph 将日志存储在 Ceph OSD 守护进程数据所在的同一磁盘上，且未进行性能优化。进行性能优化的 Ceph 实现可使用单独的磁盘存储日志数据（如固态盘具有高性能日志存储功能）。

Ceph 的默认 osd 日志大小为 0，您需要在 ceph 配置文件中对此进行设置。日志大小应为文件存储最大同步间隔和预期吞吐量的乘积乘以 2 所得的值，这样可以在提交一个日志的同时确保另一个日志继续记录数据。

osd 日志大小 = {2 *（预期吞吐量* 文件存储最大同步间隔）}

预期吞吐量应包括预期磁盘吞吐量（即持续的数据传输速率），以及网络吞吐量。例如，7200 RPM 磁盘可能具有大约 100 MB/s 的速率。最小的磁盘和网络吞吐量应为合理的预期吞吐量。一些用户开始仅使用 10GB 日志大小。例如：

osd 日志大小 = 10000

默认情况下同步间隔为 5 秒。因此我们希望日志能够容纳磁盘在同步间隔之间可容纳的所有数据量的两倍（对于当前数据和以前的同步间隔）。假设 SATA 机械硬盘可实现最高 100 MB/s 的速率，则容量仅为 1 GB。显然，使用闪存介质时计算结果会有所变化。鉴于闪存的速度较高，日志大小的限制因素最可能是系统其余部分支持这一速度的能力，而非闪存盘。（网络、处理器、内存或其他限制因素）。

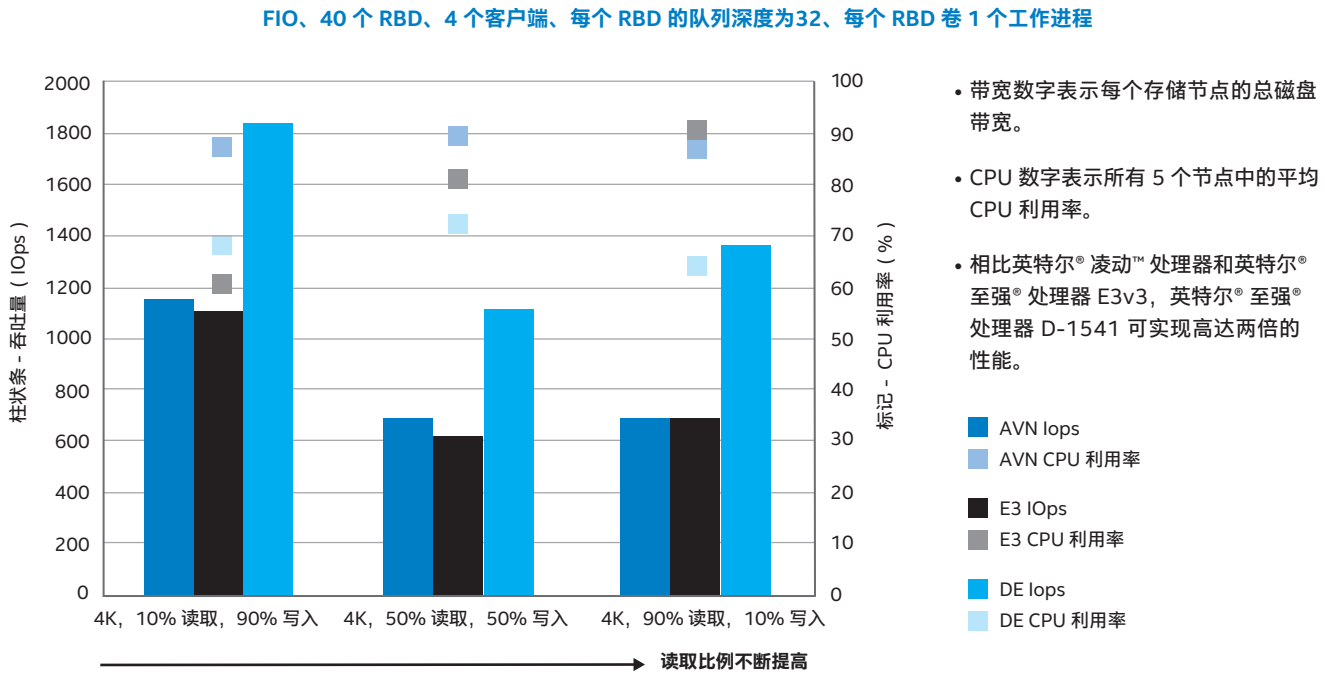


图 9. 英特尔® 至强® 处理器 D 的性能

用例 2 英特尔 PCIe 固态盘而非 SATA 固态盘用于存储 Ceph 日志

对于 Ceph 日志，一般的建议是使用固态盘，此处的固态盘基于 SATA，与机械硬盘的比率一般为 1:5，每个节点一般使用两个 SATA 固态盘（存储日志）和 10 个机械硬盘。目前，该比率主要是为了适应常见的 12 和 24 个驱动器托架系统。随着闪存不断发展，该比率也可能变化，以反映实际的性能要求与常见封装件之间的关系。

作为行业领导者，英特尔发布了速度更高的全新闪存技术 NVMe。该技术相当于在 OSD PCIe 总线上部署固态盘。相比传统的 SATA 固态盘，NVMe 取消了多个中间 I/O 功能，因而大幅降低了时延；此外，NVMe 在 PCIe 总线上具有更高带宽，而且由于尺寸大于 2.5 英寸，所以具有更高的潜在容量。NVMe 闪存设备具有更高的 IOPS 和更低的时延，因而在支持特定底层内存技术方面优于 2-4 个 SATA 固态盘。

对于全 PCIe 固态盘配置，参见我们的数据库用例。

何时考虑英特尔® 至强® 处理器 D 而非英特尔® 至强® E3 处理器

在固态盘用于存储日志和机械硬盘用作辅助存储的标准用例中，全新英特尔® 至强® 处理器 D 是每台存储服务器中替代英特尔® 至强® E3 处理器的理想选择。

正如下图所示，英特尔® 至强® 处理器 D-1541 相比英特尔® 至强® 处理器 E3v3 可实现高达 2.5 倍的性能提升。集成英特尔 D 有助于以更低的成本提供出色性能。

有关英特尔® 至强® 处理器 D 图示中使用的配置和调优参数，参见附录 B。

用例 3 Ceph 块存储：SQL 数据库和高 IOPS

在本节中，我们将概述支持低时延、高吞吐量数据库工作负载的 Ceph 集群的最佳配置。尽管任何 Ceph 集群配置均可提供任何应用使用的块存储，根据预期的主要工作负载，还需考虑一些配置因素。例如，归档工作负载可能仅要求网络带宽足以支持大型块数据传输，而 VDI 工作负载可能具有较高的事务处理速率或较小的块大小，但小范围内的时延一致性可能并不是同样重要。

关系数据库工作负载（通常称为 OLTP）通常具有较小的传输大小（每 I/O 4k-8k），本质上具有较高的随机性。在读取和写入操作中，读取操作通常偏多，70% 或 80% 的事务为读取。对于关系数据库工作负载，存储事务时延非常重要，因为每个数据库事务可能包含多个序列化存储事务。

为确保这些工作负载实现最佳性能和最低时延，需要利用全闪存，特别是基于 PCIe 的闪存设备（采用 NVMe 架构）。为实现最高的闪存利用率，多内核数

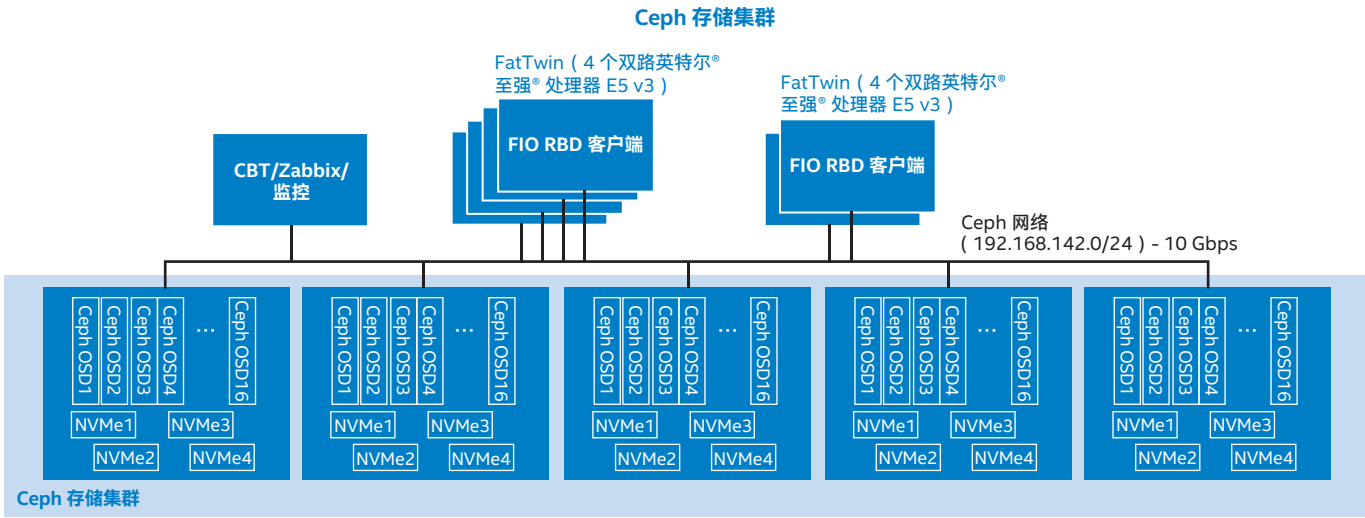


图 10. 高密度 5U 全闪存 Ceph 集群和客户端

工作负载模式	最大 IOPS	时延
4K 100% 随机读取（4.8 TB 数据集）	1.15 M	1 毫秒
4K 100% 随机写入（4.8 TB 数据集）	200 K	3 毫秒
4K 70%/30% 读取/写入 OLTP 混合模式（4.8 TB 数据集）	452 K	3 毫秒

表 9. 在所描述的配置中，可实现下列最高性能

CPU 被用于提升存储节点内的并行性和性能。这是因为 Ceph OSD 包含许多活跃线程，系统中的内核数越多，越多的进程可同步运行，从而提升事务吞吐量。

该集群包括五个 1U 全闪存系统，每个系统配备两个英特尔® 至强® E5-2699 v3 18 核 CPU，总可用核心数量为 54（启用超线程，操作系统利用 36 个 CPU）、128 GB DDR4 以及四个英特尔 P3700 800 GB NVMe 闪存设备。每个节点具有四个 10GbE 网络端口，在该配置中，单个 10GbE 链路用于公共和专有 Ceph 网络。连接单个 10GbE 节点的目标工作负载（小块事务）能够实现超过 100 万随机 IOPS，从而节省交换机端口。需要时可利用多个端口提升连续吞吐量。安装的软件为 CentOS 7.1 发行版 Linux 和 Ceph 版本 0.94.3（Hammer 版）。

高密度 5 节点集群仅占用 5U 机架空间，包含 16TB 原始容量。每节点使用多达 10 个 2.5 英寸 SFF NVMe 插槽，使用 P3700 800 GB NVMe 可安装多达 40TB 闪存，或在使用更高容量闪存设备时可安装更高容量的闪存。该集群可动态进行扩展，能够并发使用 Ceph 的自动数据分发方案。Ceph 还支持其他高级特性，如精简配置、快照、高速缓存分层和 OpenStack 集成。

该工作负载使用的一个配置策略是对 NVMe 设备实施 “多重分区”。在该场景中，NVMe 设备分区为四个 OSD 设备，每个设备具有四个日志分区。如此，Ceph 可在 4 个物理设备上每节点运行 16 个 OSD。通过使用多个 OSD 分区，减少了单个 OSD 进程中的锁争用情况，从而降低了所有队列深度的时延，大幅提高了最高吞吐量。尽管这会使得 Ceph 将多个 OSD 的数据存储在同一物理设备上，默认 crush 映射（将副本分发至同一节点之外）可保持 Ceph 中的数据耐用性。若需要，可创建 “设备” 的额外 Crush 映射级别。

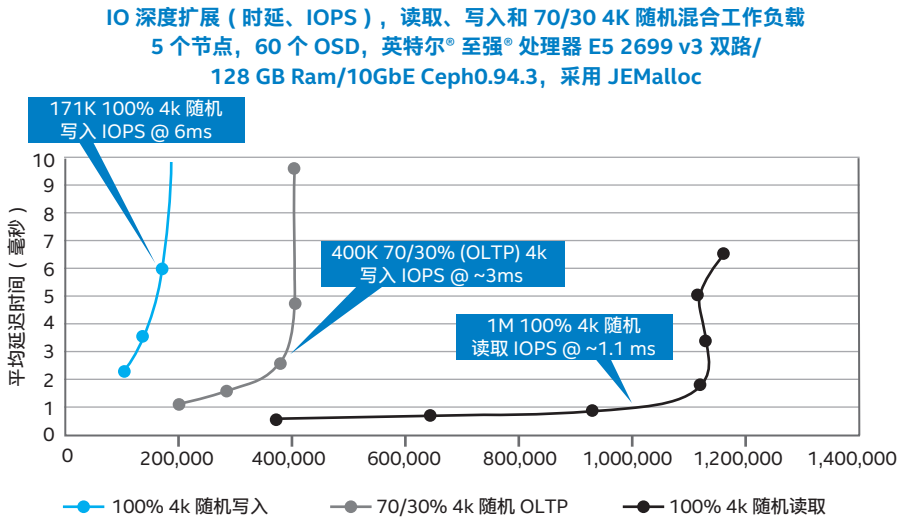


图 11. IO 深度扩展

双副本池用于容纳 RBD（块设备卷）。使用两个而非三个数据副本的原因是闪存和 NVMe 可大幅提升耐用性和可靠性。较小的容量和较高的速度也缩短了重构时间，所有这些因素共同降低了发生故障时暴露数据的风险。相比机械旋转硬盘，闪存设备发生故障的可能性明显更低，P3700 DC 设备的耐用性超过每日 10 次全盘写入。物理硬件配置也更为强悍，因为存储路径中仅从 CPU 到设备进行了 PCIe 连接，不同于 SAS/ SATA 设备，后者必须先使用 PCIe HBA、再使用 SAS/SATA 线缆进行连接，需要多个接口。

除了最大事务吞吐量之外，了解 “有用性能” 也很重要，即工作负载所需的特定时延下的最大性能。下图展示了对三个混合工作负载的 IOPS 性能与时延进行的比较。

如图 11 所示，在 70%/30% 混合工作负载中拥有 4K 随机读写能力的 “数据库” 混合工作负载的性能可达到 400,000 IOPS，平均时延为 3 毫秒。在较低的个位数时延下随机读写少量数据的能力对于性能关键型数据库工作负载非常重要。

总而言之，本节介绍了一种低时延 Ceph 集群的硬件和软件配置，这种集群可提供适用于关系数据库的高级事务性能。高级特征包括：数据中心级英特尔 DC P3700 NVMe 设备、高内核数量的英特尔至强 E5 2699v3 处理器和高密度 1U 服务器。Ceph 软件配置为在单个 NVMe 设备上使用多个分区，以提高吞吐量并降低时延。最终您将获得一个可扩展的 5U 存储集群，能够在 1 毫秒的时延下支持超过 100 万次的读取 IOPS 操作、多达 20 万次的写入 IOPS 操作，在 4 毫秒的时延下提供超过 40 万次 IOPS 70%/30% 混合读写性能。

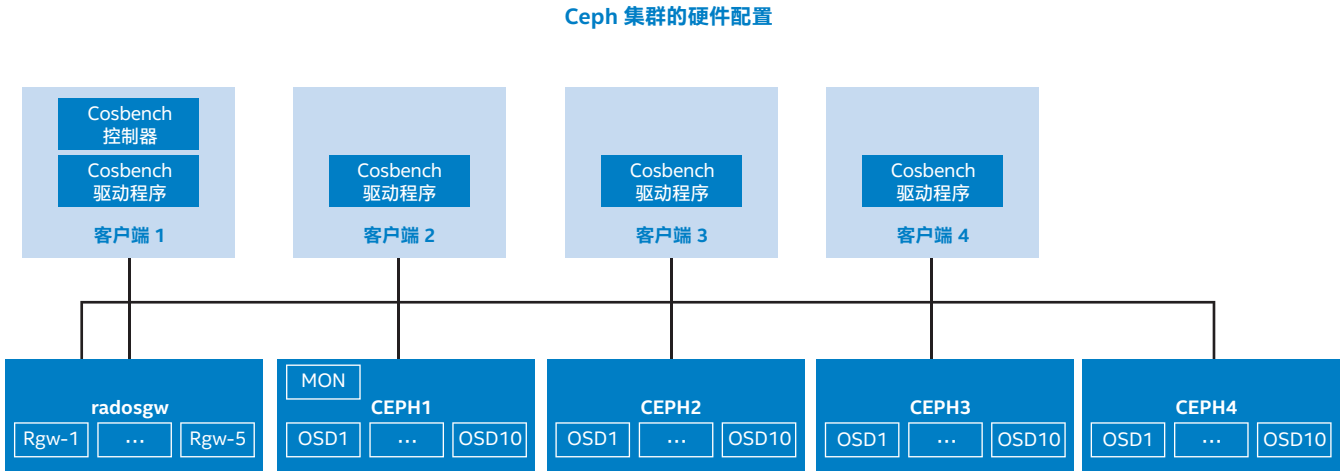


图 12. 硬件配置

示例 4 Ceph 对象存储

在本节中，我们将概述 Ceph 对象存储的优化配置，并对一个典型的 Ceph 对象集群进行简要的性能概述。对象存储是一种不同于传统文件系统（如 NFS）或块设备系统（如 iSCSI）的新兴技术。Amazon S3 和 Openstack* swift 是众所周知的对象存储解决方案。在本节中，我们使用 COSbench 作为基准测试工具。COSBench 由英特尔开发和维护，专用于通过多对象接口（包括 OpenStack* Swift、Amplidata v2.3、2.5 和 3.1 以及定制适配器）测量云对象存储服务的性能。

Ceph 对象网关（Rados 网关，RGW）是一个构建于 librados 之上的对象存储接口，为应用提供了一个连接 Ceph 存储集群的 RESTful 网关。提到 “Ceph 集群的对象接口性能” 时，实际上有两种执行性能测试的方法：a. 采用一个或多个 radosgw 节点作为访问 Ceph 集群的网关；b. 直接通过 Librados 库放置/获取对象。我们在下面的部分中采用第一种方法进行性能测试。测试涵盖两种不同场景：小型对象（128 KB）和大型对象（10 MB）。

图 12 描述了 Ceph 集群的硬件配置。

集群中有四个 Ceph-osd 节点和一个 radosgw 节点。每个 Ceph-osd 节点都有 10 块 Seagate 3.5 英寸 3 TB 7200rpm 机械硬盘作为 OSD 设备，2 块英特尔 S3500 400 GB 2.5 英寸固态盘作为日志设备。radosgw 节点将占用更多的处理器资源，因此这一节点配备了 2 个英特尔至强处理器 E5-2699 v3 @ 2.30 GHz 和 64 GB DDR3 内存。

对于客户端，我们使用 4 个节点作为 COSBench 驱动程序，其中一个节点也充当 COSBench 控制器。所有四个 COSBench 驱动程序将生成一个随机布局对象请求并发送到 radosgw 节点。所有节点都通过 10GbE 英特尔® 82599ES 端口连接。

机械硬盘配置	Readahead 2048, writecache on
HAPROXY 配置	Listen on 5 ports: 7850 - 7854
CEPH 优化调试	mount omap of each osd to a SSD partition turn down all debug log in Ceph.conf [global] mon_pg_warn_max_per_osd = 1500 ms_dispatch_throttle_bytes = 1048576000 objecter_inflight_op_bytes = 1048576000 objecter_inflight_ops = 10240 throttler_perf_counter = false [osd] osd_op_threads = 20 filestore_queue_max_ops = 500 filestore_queue_max_bytes = 1048576000 filestore_queue_committing_max_ops = 500 filestore_queue_committing_max_bytes = 1048576000 journal_max_write_entries = 1000 journal_queue_max_ops = 3000 journal_max_write_bytes = 1048576000 journal_queue_max_bytes = 1048576000 filestore_max_sync_interval = 10 filestore_merge_threshold = 20 filestore_split_multiple = 2 osd_enable_op_tracker = false filestore_wbthrottle_enable = false

表 10. Ceph 集群系统配置

工作负载模式	对象大小	容器
小型对象	128 KB	随机（1,100）
大型对象	10 MB	随机（1,100）
工作负载模式（续）	**对象**	**工作进程**
小型对象（续）	随机（1,100）	320
大型对象（续）	随机（1,100）	320

表 11. 工作负载模式

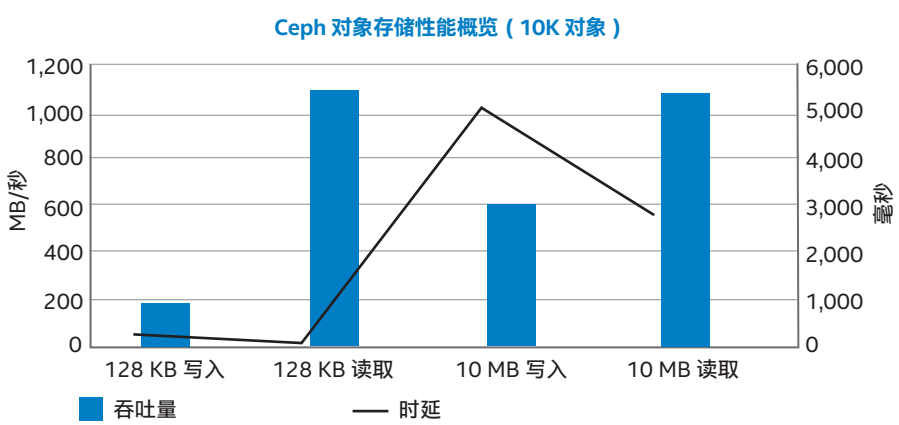


图 13. Ceph 集群对象性能概览

对于 Ceph 集群配置，以下性能基于最新版本 Infernalis（9.2.0），副本规模为 2。详细配置如表 10 和表 11 所示。

图 13 显示了按照小型/大型对象进行测试的 Ceph 集群对象性能，以及优化的 Ceph 调试。

我们可以看到，小型对象或大型对象的读取性能都会达到 radosgw 节点 NIC 最大吞吐量。不过，小型对象（128k 写入）写入性能有一些问题，怀疑是由 RadosGW 实施引起，需要进一步调查。

示例 5 云数字视频录制（云 DVR）

云 DVR 和美国法律

由于版权法的演变，美国 DVR 市场仅对个人所有和单一副本的广播内容开放。起初，公平使用法律只适用于在家用 VCR 上录制节目，但现在这一法律已应用于数字视频录像机中。然而，随着云计算和横向扩展存储的出现，它正在成为一种越来越受欢迎的架构，支持人们通过互联网存储和播放电视节目。人们可以通过任何互联网连接随时随地灵活地下载个人消费记录。根据具体的法律和国家规定，在云中存储内容时，内容可以作为共享副本或私有副本。美国数字版权管理和公平使用法律规定，在云 DVR 上记录内容的每个人都有一个单独的副本，不能出于数据保护或其他任何原因复制这个副本。这严格来讲包括为提高存储效率而对相同内容进行的任何重复数据删除。此外，DVR（云或单机）上的大部分内容（约 85%）都未观看过。因此，您需要有一个以写入为主导的流对象存储云环境，就像归档、备份或数字视频监控一样。

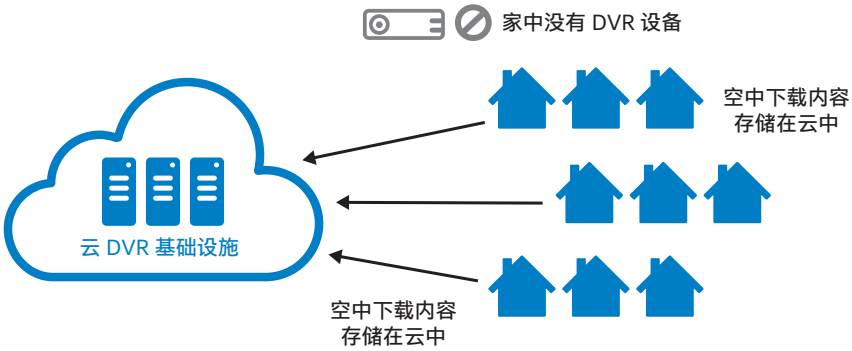


图 14. 云 DVR 基础设施

操作类型	比特率	对象百分比
8-14 秒用户段读取	4 MB	80
	2.5 MB	10
	1.9 MB	5
	1.25 MB	5
8-14 秒用户段写入	4 MB	25
	2.5 MB	25
	1.9 MB	25
	1.25 MB	25

表 12. 用户读写对象大小和百分比的典型组合

随着内容的复制，这个单独的副本对空间回收有其他影响。（例如，“我只想付费保存 DVR 中最后一周的节目。”）由于存在仅限一个副本的限制，若要从对象存储中删除内容，必须对通常存储为几秒内容的每个记录段进行重写。因此，这并不会影响进出云端的带宽，但由于对象必须在作为可用空间返回之前进行重写，所以对云 DVR 存储系统影响很大。

云中存储 DVR 内容提供以下优势：

- 降低存储成本，因为拥有独特外部视图的不同帐户持有人之间（美国之外）可以在内部共享内容。
- 云 DVR 可以充分利用存储创新，如擦除编码来实现空间节省。
- 当内容存储在云中时，能够从大地理区域内的任何设备访问内容。
- 通过数据中心实践提供额外的数据保护，因为内容可以在云中的多个服务器或数据中心之间共享。家庭 DVR 设备可能会发生故障，导致内容丢失。
- 通过有针对性的广告等捆绑服务，为服务提供商提供额外的收入机会。

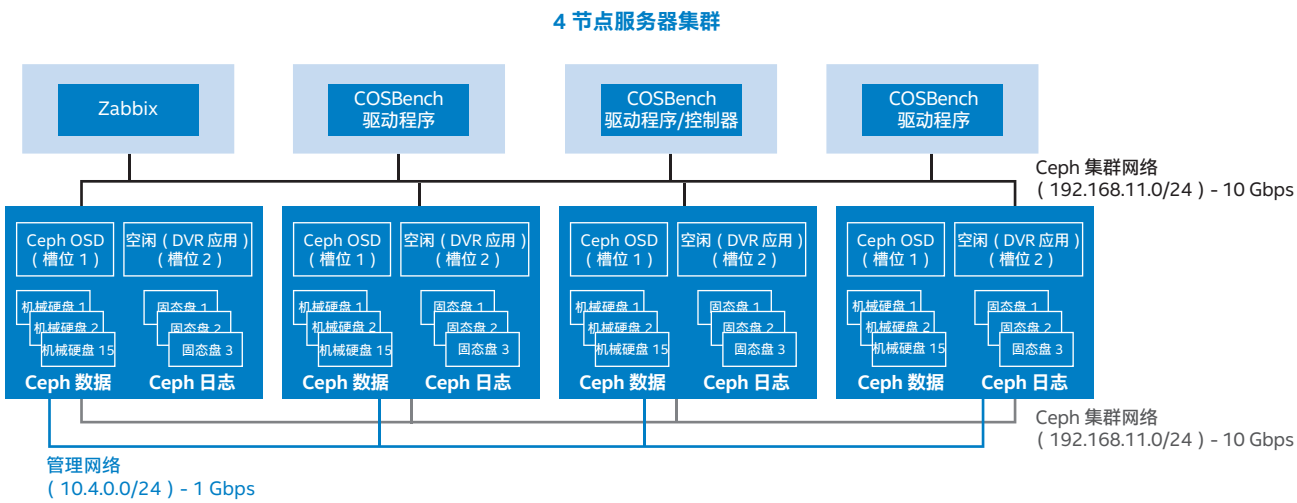


图 15. 用于测试 DVR 场景的 4 节点服务器集群

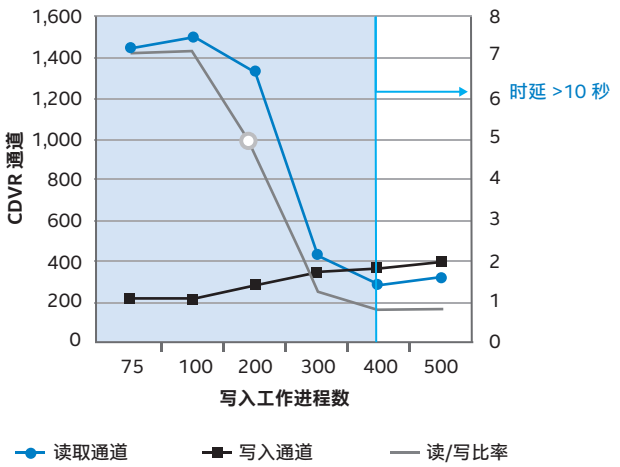


图 16. 100 个读取器

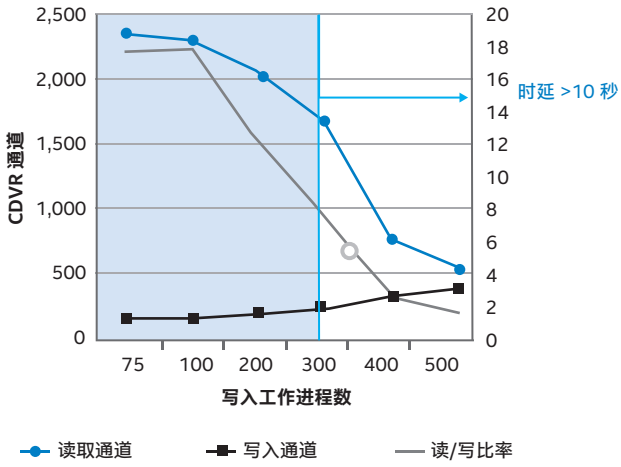


图 17. 200 个读取器

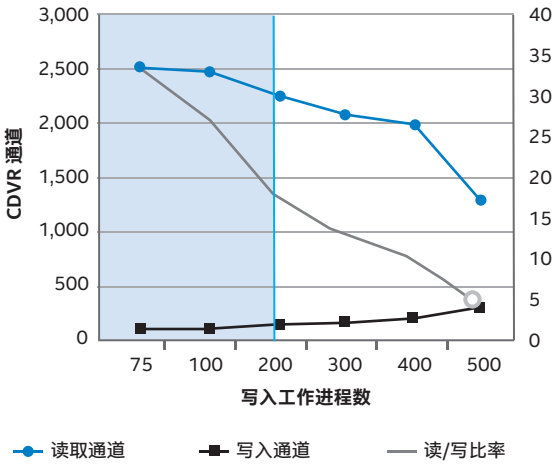


图 18. 300 个读取器

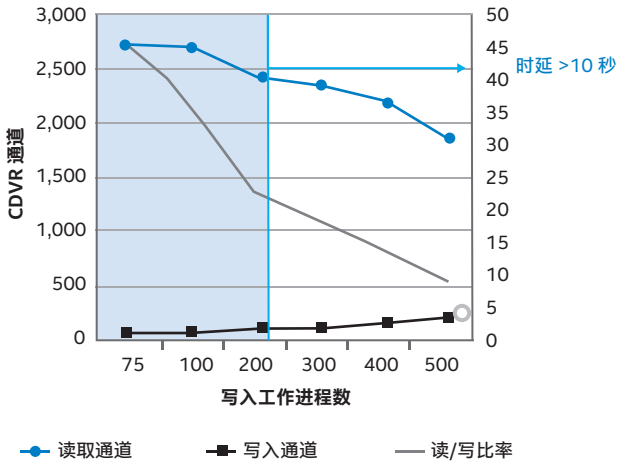


图 19. 400 个读取器

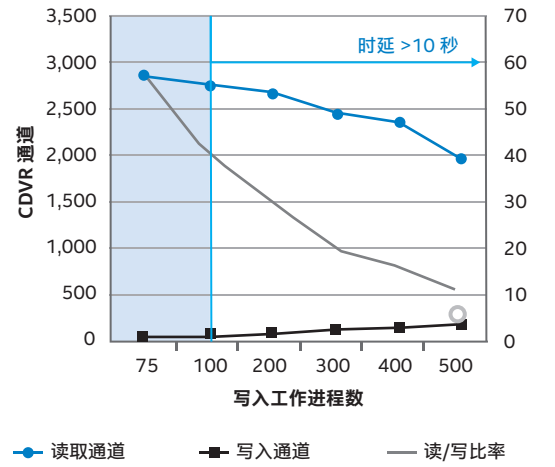


图 20. 500 个读取器

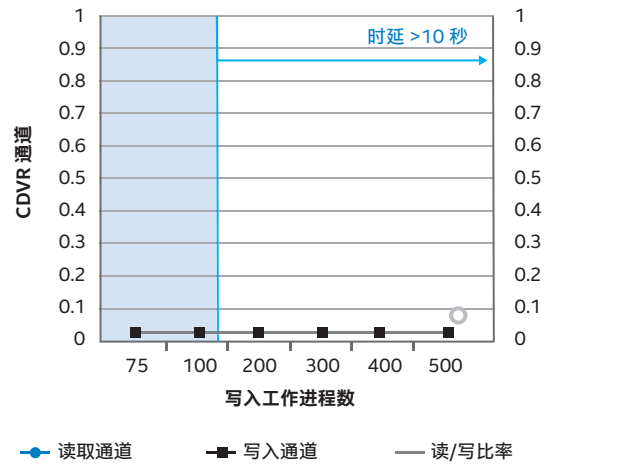


图 21. 600 个读取器

Ceph 云 DVR 实验室测试

Ceph 可提供用于存储云 DVR 内容的横向扩展存储，正成为一种越来越受欢迎的选择。Ceph 提供擦除编码存储，其中包含优化的 ISA-L EC 插件，用于加速读写 DVR 内容的擦除编码。

私人副本用例是一种写入操作密集型工作负载，因为内容需要按照用户进行存储，即使对于其他用户是相同的空中下载内容。共享副本用例是一种读取操作密集型工作负载，因为内容是共享的，所以写入通常比读取少。媒体内容在云中基于片段进行存储，通常为 8-14 秒的播放时间。

以下四节点服务器群集用于测试各种 DVR 场景。这是一种超融合配置，只有一个槽位用于运行 Ceph 存储服务。每个双路存储节点包含英特尔至强处理器 E5-2620 v3 @ 2.40 GHz、24 核（支持超线程技术）、96 GB、高速缓存 15360 KB、8 通道 SATA 6 Gb/秒 SAS 半高 600 MBps PCIe 3.0 x8、DCS3700 400 GB 固态盘。Ceph 日志存储在固态盘上。一个固态盘用于存储 5 个机械硬盘日志。COSBench 用于模拟使用三台服务器的用户段读写操作。

图 16-21 显示了一个不同的用户读取、写入和时延组合，用于实现共享副本和私人副本云 DVR 用例的最佳读写操作组合。

示例 6 备份/存档

备份/归档简介

Ceph 对象存储与擦除编码池相结合可以高效地存档和备份对象。从小对象（如图片和图像）到数 GB 超大对象的各种对象都在备份范围内。擦除编码是一种数据保护方法，可防范节点、机架和数据中心故障，从而有效地保护数据。下图概述了选取原始对象、将其编码为片段并在整个集群中分发编码片段来实现数据保护目标的编码操作。同样，解码过程通过选取分片和解码来创建原始对象，从而重构原始对象（图 22）。

英特尔加速擦除编码算法的优势

Intel EC Acceleration 在 Ceph 中作为标准版本的一部分提供，只需添加库，便可享受更快的吞吐量。擦除编码（EC）算法经过配置，可提供与三重 RAID 数据冗余相同或更出色的数据耐用性，同时还可节省高达 50% 的存储。计算基于英特尔对 320 块硬盘（共 960TB 原始容量，无单点故障）较之 3 路 RAID 设置的可用容量的内部测量；EC 使用可配置方案，因此数量可能不同，但相同点是，对于每 14 块硬盘，您可以存储 10 块硬盘的数据，因此 14n/10n = 1.4x，而典型三重复制为 3x。特性和优势可能需要支持的系统和第三方硬件、软件或服务得以激活。产品性能因系统配置而有所差异。详情请咨询您的系统提供商。有关更多信息，请访问：http://www.intel.cn/content/www/cn/zh/benchmarks/intel-product-performance.html

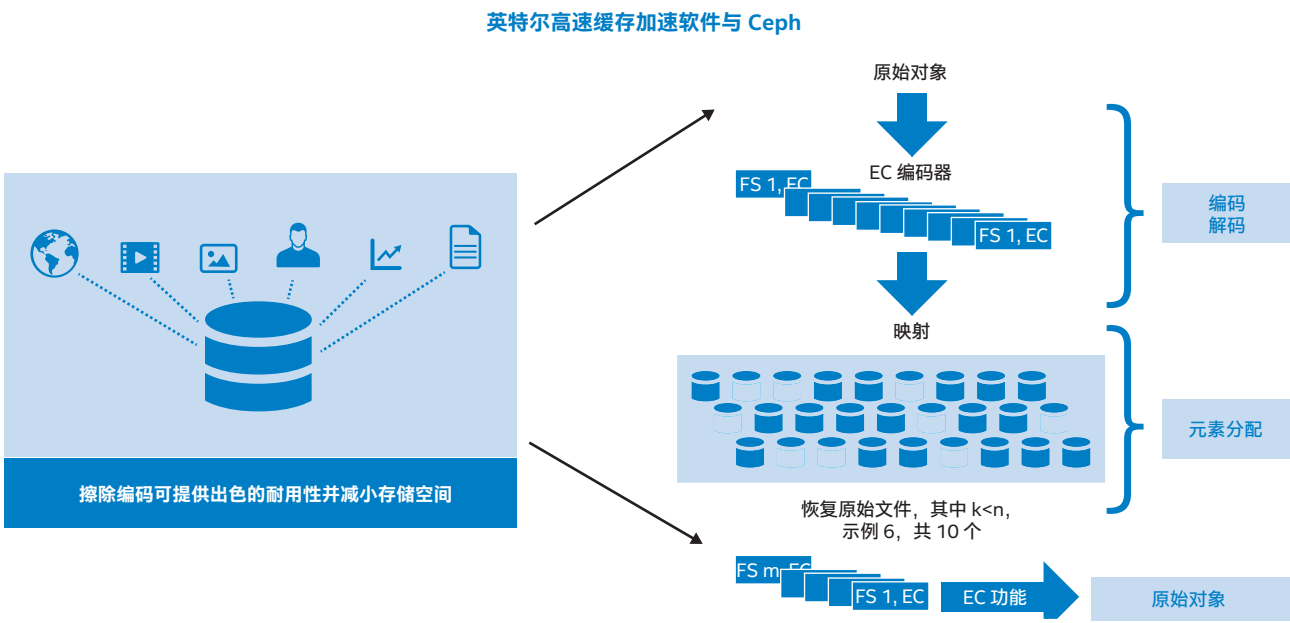


图 22. 高速缓存加速软件

示例 7：英特尔高速缓存加速软件与 Ceph

如果由于成本或容量而无法迁移到 100% 固态盘存储（大规模 Ceph 部署就可能出现这种情况），英特尔® 高速缓存加速软件（CAS）可提供一款性能中等、成本较低的替代方案。

英特尔 CAS 是一款面向 Windows* 的文件级解决方案和面向 Linux* 软件缓存解决方案的块级解决方案。与英特尔® DC 系列 PCIe NVMe 固态盘配合使用时，英特尔 CAS 可为您现有的机械硬盘备份存储提供出色的性能改进，且所需的费用仅为用固态盘完全替换机械硬盘所需费用的一小部分。

英特尔 CAS for Linux 作为可加载内核模块和用户空间管理工具加以实现。这是一款真正的 “插件” 解决方案，不需要更改您的操作系统、应用或硬件基础设施。

英特尔 CAS 将最热门的数据存在快速固态盘上，以实现高吞吐量和低时延交付。

过去，和大多数其他缓存解决方案一样，英特尔 CAS 使用最近最少使用（LRU）驱逐算法来缓存所有块存储访问，以将最热门的数据保存在缓存中，并且逐出最旧的未使用数据。然而，Intel CAS v3.0 for Linux 提供了一种称为 I/O 分类的新功能，该功能基于英特尔实验室发明的英特尔® 差异化存储服务（DSS）技术。这种新的 I/O 分类功能支持以比传统 LRU 缓存算法更精细的粒度进行 I/O 分类、优先级排序和选择性分配，让用户能够比以往更加灵活地控制缓存的内容和缓存中存留的内容。最终，这将提升缓存性能（从而提高吞吐量和降低时延），同时获得降低缓存容量需求的附加优势。

英特尔 CAS 公开了以下 I/O 类，每个类可以针对缓存进行启用/禁用，并通过一个用户可编辑的配置文件获得自己的清除优先级。参见表 13。

在过去一年中，英特尔与雅虎合作改进了 Ceph 在 Yahoo Mail*、Flickr* 和 Tumblr* 中的性能。Yahoo 的典型工作负载使用 Ceph 对象存储实现，将 XFS 作为基本文件系统并使用 8+3 擦除编码（EC 8+3）（不是 3 份副本，可最大限度提高机械硬盘的容量可用性）来存储 Yahoo Mail*、Flickr* 和 Tumblr* 附件、照片和视频。

问题（积少成多）

在这一实现中，单个 Ceph 存储节点的文件系统通常包含数亿个小文件。检索这些文件之中的一个文件需要收集 8+3 擦除编码片段，将它们重组成原始对象。为了获取每个片段，必须首先通过 XFS inode 进行跟踪，在托管相应 XFS 文件系统的磁盘上找到该文件。由于磁盘上有数亿个文件，所以在加载小（少量块文件）文件之前，必须通过 4-6 个 inode 块进行跟踪。这会直接导致比加载小文件本身长 6 倍的时延以及相当于整体潜力六分之一的吞吐量。更糟糕的是，重建对象的整体时延取决于检索 11 个 EC 片段中前 8 个片段的最长（长尾）时延。

CAS DSS IO 类
未分类
元数据 (Superblock、GroupDesc、BlockBitmap、InodeBitmap、Inode、IndirectBlk、Directory、Journal、Extent、Xattr)
<=4KiB
<=16KiB
<=64KiB
<=256KiB
<=1MiB
<=4MiB
<=16MiB
<=64MiB
<=256MiB
<=1GiB
>1GiB
O_DIRECT
其他

表 13. 用户可编辑配置文件

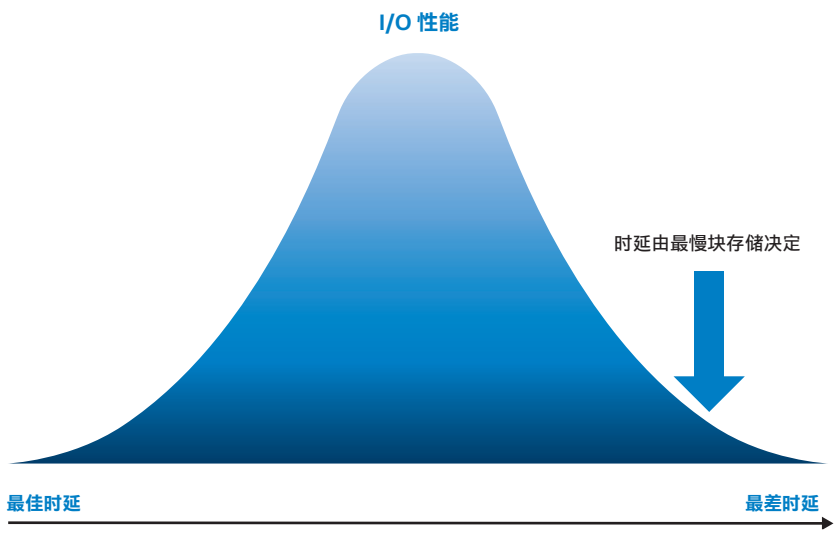


图 23. I/O 性能

雅虎与客户签订的服务级别协议规定，雅虎会超额提供 3 倍容量，以满足性能预期。这些并不是小数目。当 Flickr 需要 1 PB 存储时，雅虎必须提供 3 PB（全部并行运行）才能达到所需的吞吐量水平。

解决方案：借助采用英特尔® 差异化存储服务技术的英特尔高速缓存加速软件，另辟蹊径

英特尔为雅虎设计的 CEPH 解决方案可将 XFS 文件系统的所有元数据移动到使用英特尔® 差异化存储服务（DSS）技术和英特尔® CAS 的高速缓存中。检索文件时重复使用 inode，而文件本身的使用可能不太频繁，缓存起来可能不太有效。主要的性能改进是通过缓存文件系统元数据来实现的。这就是英特尔 CAS I/O 分类发挥作用的地方。英特尔 CAS 允许用户在不缓存文件数据的同时缓存元数据，从而以最小的缓存实现最佳性能。

通过缓存文件系统元数据，所有 inode 访问都以 PCIe NVMe 吞吐量和时延提供，从而实现吞吐量的总体改进和时延降低。

配置详细信息：

硬件配置：

- **服务器：** HP ProLiant DL180 G6 ySPEC 39.5
- **CPU：** 2 个英特尔® 至强® 处理器 X5650 2.67 GHz（启用超线程，共 12 个核心，24 个线程）
- **芯片组：** 英特尔® 5520 IOH-36D B3（Tylersburg）
- **内存：** 48 GB 1333 MHz DDR3
  - 12 个 4 GB PC3-10600 DDR3-1333 ECC Registered CL9 2Rx4
- **机械硬盘：** 10 块 8 TB 7200 RPM SATA 机械硬盘
- **固态盘：** 1 块 1.6 TB 英特尔 P3600 固态盘
- **网卡：**
  - 2 个 HP NC362i/英特尔 82576 Gigabit
  - 2 个英特尔® 82599EB 10Gbe
- **操作系统：**
  - RHEL 6.5，kernel 3.10.0-123.4.4.el7

Ceph 配置：

- Ceph Giant v87.1
- 1 个管理节点
- 2 个监控节点
- 8 个 OSD 节点，每个拥有 10 块 8 TB 企业级 SATA 机械硬盘和 1 块 1.6 TB 英特尔 DC P3600 固态盘。
  - 注：固态盘用于日志记录和缓存。硬盘进行了分区，其中 1.5 TB 的分区用于缓存，10个 10 GB 的分区用于日志记录（每个 OSD 一个 10 GB 分区）。
- EC 8+3

基准测试：

为了评估性能，我们利用 rest-bench 对 GET 和 PUT 性能进行了采样。我们在不同的集群数据加载水平（满负荷度）下采集带缓存和不带缓存的样本。基准测试流程如下：

1. 在下面的每个步骤之间清除页面和磁盘缓存
2. 在禁用缓存的情况下填充集群 10%（PUT）
3. 在启用缓存的情况下填充集群 10%（PUT_cache）
4. 在启用缓存的情况下读取（GET）测试（GET_cache）
5. 在禁用缓存的情况下读取测试（GET）
6. 在集群满 50% 和满 70% 的情况下利用 GET 测试点进行重复测试。
7. 将非缓存性能与缓存性能进行对比。

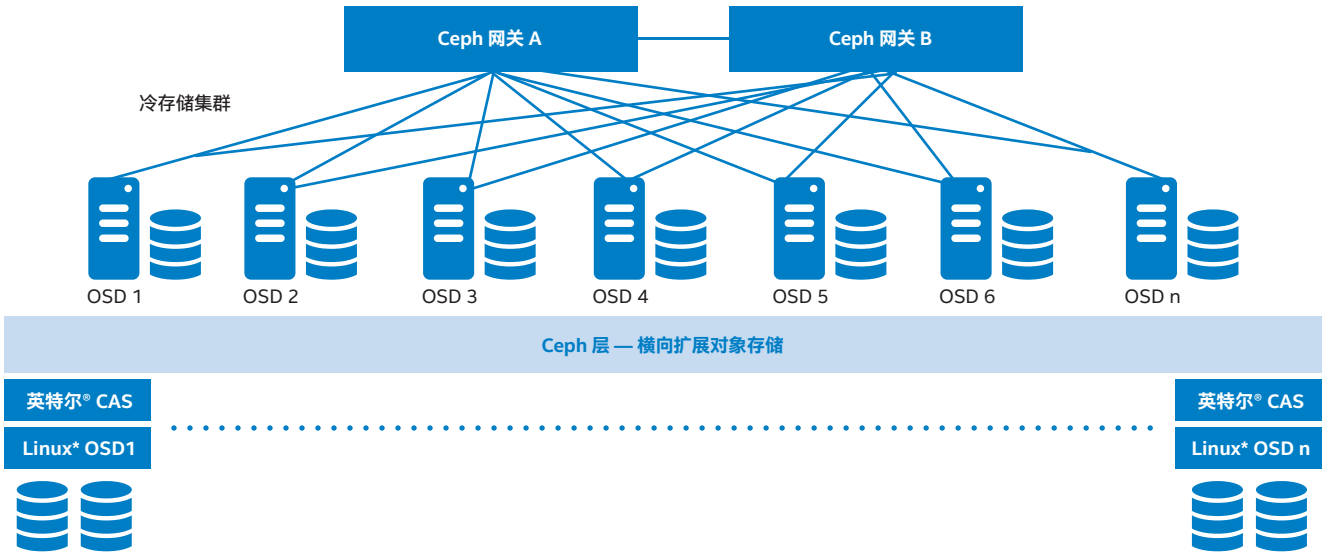


图 24. Ceph 配置

结果：

通过使用支持 I/O 分类的英特尔 CAS 软件结合一块只占后备存储大小 2% 的英特尔 DC P3600 固态盘高速缓存，雅虎实现了：

- 200% 的GET 吞吐量增长
- 50% 的GET 时延降低
- 100% 的PUT 吞吐量增长
- 30% 的PUT 时延降低
- 满足 SLA 性能要求所需的集群数量减半（由于 GET 吞吐量增长和时延降低，雅虎现在只需先前所需的超额供应量的一半便可达到 SLA 级别吞吐量，因此所需的机械硬盘和服务器数量将减少）。

从图 25 中可以看出，随着请求数量的增加，使用采用英特尔 DSS 的英特尔 CAS 可降低时延并提高利用率。（RPS = 每秒请求数）随着利用率的增长，CAS 解决方案不会出现随机访问应用中机械硬盘的磁头寻道时间越来越长而导致开销增加的问题。

在写入方面，结果更加引人注目（图 26），因为所有机械硬盘解决方案30% 的时间都会超时。

这一解决方案可带来实际的资本支出节省（超额供应）、运营成本节省（降低功耗、空间和散热成本），以及改进的可扩展性计划（性能和可预测性）。

总之，如果您正在寻求一款性能中等的低成本 Ceph 性能加法器，请考虑采用英特尔® 高速缓存加速软件、英特尔® 差异化存储服务（DSS）技术和英特尔 PCIe NVMe 固态盘。

有关更多信息，请联系您的英特尔代表或访问：http://www.intel.cn/content/www/cn/zh/software/intel-cache-acceleration-software-performance.html

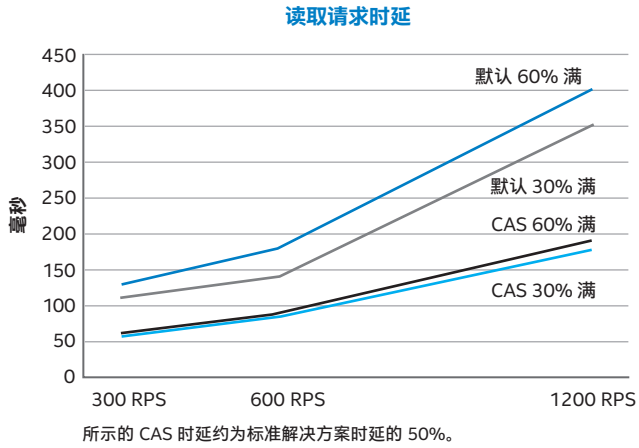


图 25. 读取请求时延

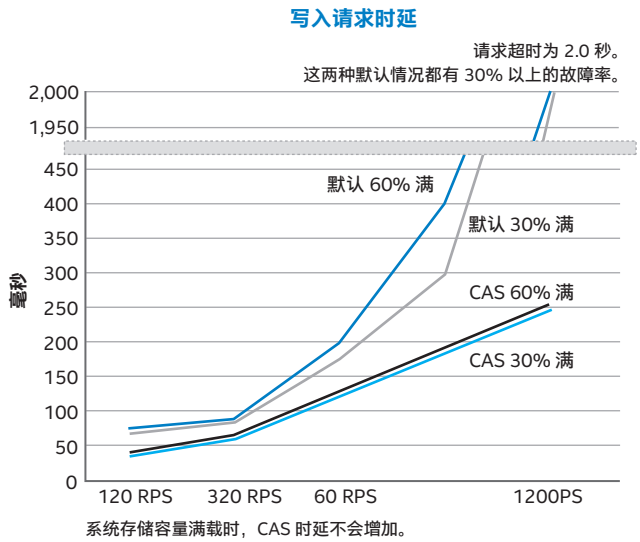


图 26. 写入请求时延

差异化存储服务：后续步骤

在上一节对雅虎使用的大规模 Ceph 对象存储的讨论中，我们已经展示了应用来自英特尔® 高速缓存加速软件和英特尔 PCIe NVMe 固态盘的英特尔® 差异化存储服务（DSS）技术的潜力。此外，Ceph 能够提供数据块级和文件级存储服务，这使其成为最具吸引力的存储服务解决方案之一。

然而，使用 Ceph 的应用的 I/O 特性可能与 Ceph 存储节点实际看到的特性非常不同，例如雅虎的例子。同样，如果使用用于虚拟机（VM）块存储的 Ceph RADOS 数据块设备（RBD），一个虚拟机内一款应用的 I/O 特性可能与 Ceph 存储节点实际看到的特性非常不同，另外也可能与另一个虚拟机内的相同应用非常不同。传统存储缓存可能无法很好地运行，因为即使将高容量固态盘作为高速缓存设备，对应用的 I/O 也不了解。与此同时，几乎所有的云服务提供商（CSP）正在寻求更好的方式（1）实现服务级别对象（SLO），并（2）创造服务级别协议（SLA）的差异化。前者可让云服务提供商最大限度降低违反客户 SLA 的风险，后者可让云服务提供商为客户构建灵活的定价模型。

英特尔实验室正在引领云存储服务的研究，帮助充分发挥 Ceph 的多功能性。更具体地说，英特尔实验室正在积极研究 QEMU 中的 Ceph RADOS 块设备（RBD）。借助 DSS I/O 提示机制，来自任何特定虚拟机的各 I/O 都可以传送自己的 I/O 类信息。每虚拟机 I/O 提示机制支持基于每台虚拟机对存储缓存策略进行更高级的调整，甚至基于每个虚拟机主机对每个虚拟机进行调整。这可以让 Ceph 通过基于每个虚拟机 I/O 提示机制调优的 DSS 缓存策略来充分利用高性能的英特尔® PCIe NVMe 固态盘，同时在无需了解实际应用的情况下提供存储服务。另一方面，在雅虎这样的用例中，英特尔实验室也在积极研究基于原生对象的 I/O 提示。例如，客户可能喜欢更快速地访问来自相同对象存储的图片和文档，其中 Ceph 通过 Ceph RADOW Gateway（RGW）提供服务。此外，适用于 Ceph 的 DSS I/O 提示，无论来自数据块还是对象存储服务，都是云服务提供商的宝贵资产，可支持他们根据自己定义的 SLA 差异化优势获得更智能的存储缓存策略。英特尔实验室也在积极利用云 DSS 来寻找在这方面帮助 Ceph 用户的创新方法。

SLA 是总体协议，SLO 是提供服务的特定性能指标。

CeTune：基于用户界面的 Ceph 基准测试工具

随着将大量数据存储到云并进行检索的格局不断增长和扩展，以轻松高效的方式评估、分析和调优云存储解决方案的需求也在增长。英特尔在生产环境中观察到，客户在实现最佳性能方面仍然面临许多挑战，包括如何解决瓶颈问题，从众多（500+）参数中确定最佳调优旋钮，并处理频繁发布之间的意外性能回归。为了简化这一工作，英特尔开发了开源 CeTune，Ceph 基准测试、分析和调优工具。

CeTune 框架包含 5 个不同的组件：

1. Deployer，可以在几分钟内轻松部署一个 Ceph 集群。
2. Benchmarker 生成明确定义的用例，并自动评估带各种可插拔工作负载的 RBD、对象和 CephFS 性能。
3. 分析器通过单一界面监控各个方面的性能（系统特征数据、工作负载吞吐量和时延），并通过基于 Lttng 和 Zipkin 的通用可视化 GUI 揭示 Ceph 软件堆栈时延。
4. Tuner，动态注入参数并比较性能，以确定最佳调优旋钮。
5. Visualizer，在基于 Web 的性能门户上自动呈现数据。

借助 CeTune，我们可以评估每个主要版本的 Ceph 性能，识别性能瓶颈，并在短时间内发布可供用户/开发人员参考的结果（图 27 和图 28）。

CeTune 可通过以下网址下载：
https://github.com/01org/CeTune

用于管理 Ceph 集群的虚拟存储管理器（VSM）

2015 年温哥华 OpenStack 峰会调查显示，44% 的 Openstack 采用者正在使用 Ceph 作为块存储选项（图 29）。

但对于存储管理员和操作员来说，Ceph 集群的操作仍然复杂繁琐。为了减少采用障碍并加速基于 Ceph 的解决方案的实施，我们开发了虚拟存储管理器（VSM）来弥补操作差距。虚拟存储管理器（VSM）是一款 Ceph 管理工具，于 2014 年 11 月的OpenStack 巴黎峰会上作为开源项目发布。

虚拟存储管理器（VSM）包含两个主要组件：控制器和代理。

- VSM 控制器在专用服务器或服务器实例上运行，并通过 VSM 代理管理 Ceph 集群。此外，如果用户希望为 OpenStack 提供存储池资源，VSM 控制器负责连接到 OpenStack 集群。
- VSM 代理在每台 Ceph 服务器上运行，它接受来自控制器的请求，并将服务器配置和状态信息转发到 VSM 控制器。

图 27. CeTune 生成的性能报告

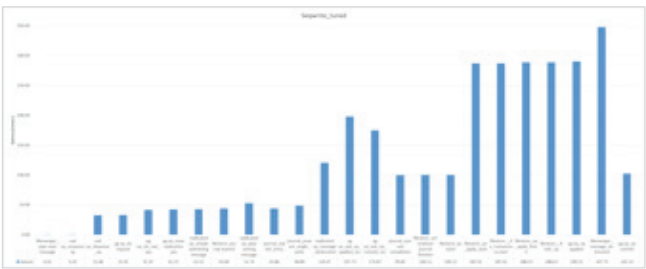


图 28. CeTune 生成的时延故障报告

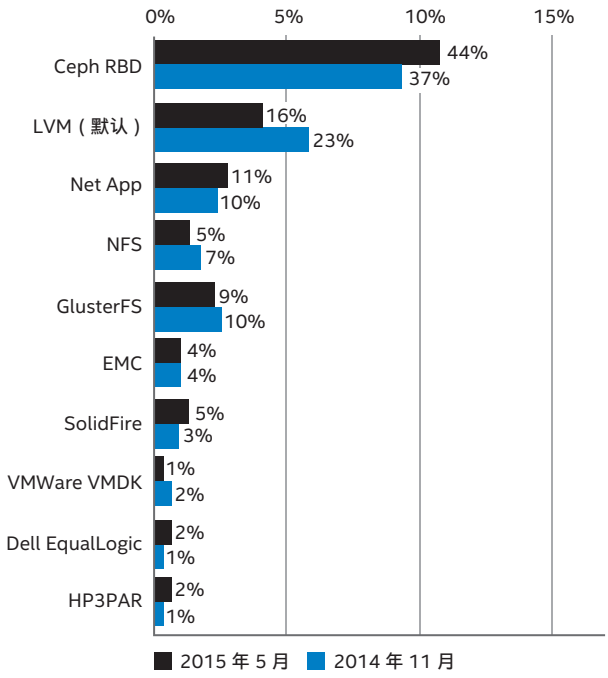


图 29. 块存储驱动程序：生产

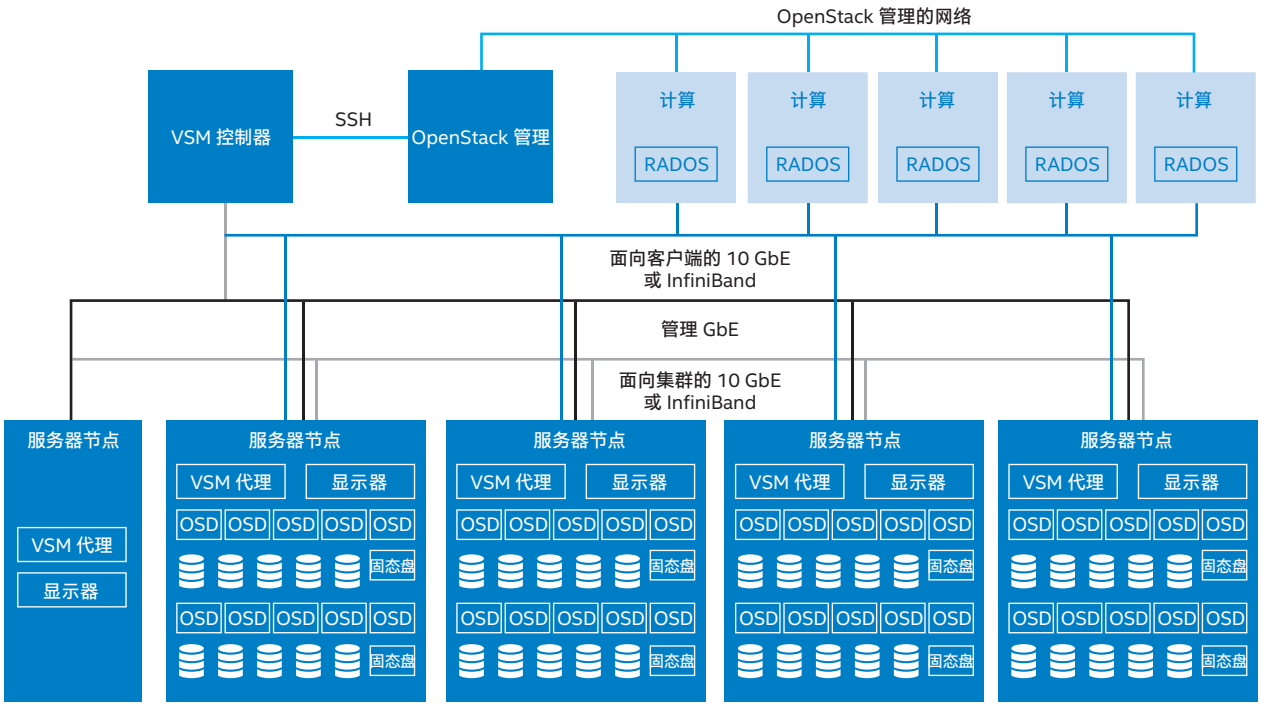


图 30. VSM 控制器

为支持 Ceph 操作，VSM 提供了一些新功能，这些功能大致分为四个方面：

1. 部署

VSM 支持从头开始创建 Ceph 集群，它需要清单文件来描述集群拓扑。VSM 还可以通过导入现有 Ceph 集群对其进行管理。

VSM 提供一个自动安装程序，以帮助早期用户使用一个命令行部署 VSM。

2. 管理和维护

VSM 提供帮助存储管理员进行日常运维的功能。

日常运维的一个常见任务是更换故障磁盘或服务器。VSM 可以协助和简化这些操作并减小影响。

此外，VSM 还可以将不同的存储介质分类到存储组中，然后用来创建不同的存储池，以满足不同的业务需求。VSM 可以创建和管理复制池、擦除编码池和高速缓存池，还支持分配共有配额，并允许一个池利用多个存储组，这意味着支持先前提到的主副本模式。

对于长时间运行的系统而言，维护是一个重要方面，VSM 可将自身升级到较新版本，此外它还支持将 Ceph 升级到指定版本。

3. 监控

若要了解集群状态，监控集群运行状态是至关重要的一步。除集群整体运行状态之外，VSM 还会监控容量利用率，包括集群级别、存储组级别和 OSD 级别的容量利用率。

VSM 还会监控其他资源，如服务器、设备和池。对于设备而言，VSM 还会尝试检索磁盘 SMART 日志，以了解设备状态。

除资源状态以外，VSM 还会监控整体集群性能，如 IOPS、带宽、时延以及不同时间的 CPU 利用率。

4. 集成

Ceph 经常与 OpenStack 云平台进行集成，VSM 具有向 OpenStack 提供池的功能。通过它，OpenStack 可以在具有不同特征的池上创建虚拟机实例来满足不同的业务需求。

为了便于将 VSM 与第三方工具和工作流进行集成，VSM 提供了一套 REST API 和 CLI 工具。

VSM 项目主页位于 https://01.org/zh/virtual-storage-manager，代码库位于 https://github.com/01org/virtual-storage-manager，二进制包可从 https://github.com/01org/virtual-storage-manager/releases 下载。社区位于 http://vsm-dis-cuss.33411.n7.nabble.com/。

图 31. VSM 项目主页

当前工作

英特尔将继续更新 Ceph 相关开发，请获取位于 https://soco.intel.com/docs/DOC-2146636 的本文的最新版本，获取最新的性能和配置信息。

结论

云工作负载和成本推动了对横向扩展存储解决方案的需求。Ceph 是一款热门开源存储软件，已出现许多有意义的生产部署案例。备份、归档、虚拟块和流媒体工作负载在当前的 Ceph 生产部署中占主要地位。以上部分介绍的基于英特尔® 架构和闪存的参考解决方案概述了最终客户如何部署优化的 Ceph 配置，通过最佳基础设施实现性能、时延和空间效率目标。

附录 A：推荐的调优参数

完整的当前调优的参数文档 *Ceph Cookbook：配置指南*仅供签署英特尔保密协议的用户使用。介绍如下。

简介

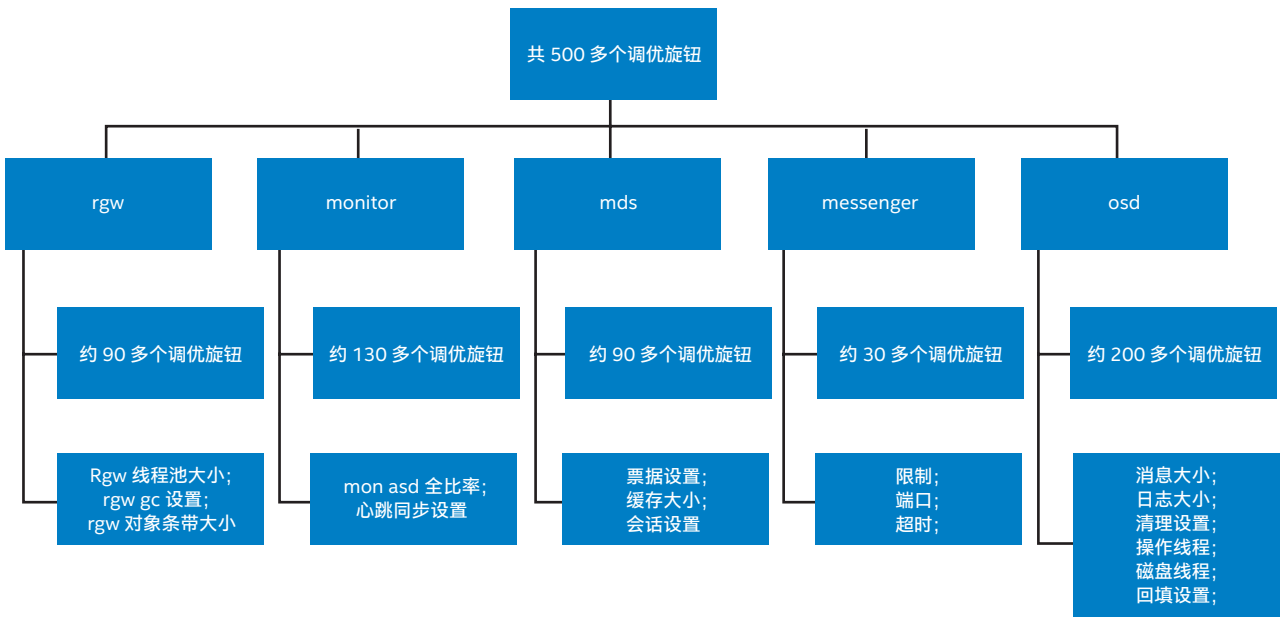
除了传统的企业级存储技术，许多组织现在拥有性能和价格要求各异的存储需求。OpenStack 支持对象和块存储，并提供了许多部署选项，支持用户根据用例进行选择。

作为存储技术的未来发展趋势，Ceph 是一种大规模可扩展且开源的软件定义存储系统，在商用硬件上运行。Ceph 专为在自我管理、自我修复的单一软件平台上提供对象块和文件系统存储而开发，没有单点故障。由于 Ceph 具备高度可扩展、软件定义的存储架构，它是传统存储系统的理想替代方案，并且为云计算环境中的对象和块存储提供了功能强大的存储解决方案。

Ceph 共有 500 多个调优旋钮。本文介绍了我们在测试中发现的有关 Ceph 性能调优的最知名方法。其中一些调优方法可大幅改进性能。

本文假设读者对 Linux 操作系统和云存储基础设施有基本的了解。

为完成本文，引用了许多互联网资源。由于资源和知识的限制，当前版本肯定有很多待改进之处或者错误。我们将不断更新文档，添加新的知识或发现，欢迎提出意见。有关更多信息，请访问 https://github.com/01org/CeTune。在这里，您可以找到联系信息、邮件列表链接、WIKI 和问答等。



附录 B：英特尔至强处理器 D 和 Ceph 配置及调优

Ceph 性能配置

Ceph 版本：0.94.3 “Hammer” 与 JEMalloc

（OSD 数量 * 100 ）/尺寸 = PG 数量

50 个预填充 25 GB RBD 卷映射到池，每个客户端 10 个

5:1 OSD/日志比例

10 GB 日志大小

FIO 基准配置

版本：2.2.9

I/O 引擎：libRBD

Direct：是

队列深度：

4 KB 随机和 1M 顺序 I/O 为 32，

32 M 顺序 I/O 为 8

Numjobs：1

缓冲时间：30 秒

运行时间：300 秒

重填缓冲区：1

失效：0

工作负载	对象大小	模式
小型工作负载	4KB	90% 写入 10% 读取 50% 写入 50% 读取 10% 写入 90% 读取
中型工作负载	1MB	90% 写入 10% 读取 50% 写入 50% 读取 10% 写入 90% 读取
大型工作负载	32MB	90% 写入 10% 读取 50% 写入 50% 读取 10% 写入 90% 读取

Ceph 配置（续）

```
[global]
auth_cluster_required = cephx
auth_service_required = cephx
auth_client_required = cephx
filestore_xattr_use_omap = true

debug_default = 0
debug_lockdep = 0/0
debug_context = 0/0
debug_crush = 0/0
debug_buffer = 0/0
debug_timer = 0/0
debug_filer = 0/0
debug_objecter = 0/0
debug_rados = 0/0
debug_rbd = 0/0
debug_journaler = 0/0
debug_objectcatcher = 0/0
debug_client = 0/0
debug_osd = 0/0
debug_optracker = 0/0
debug_objclass = 0/0
debug_filestore = 0/0
debug_journal = 0/0
debug_ms = 0/0
debug_monc = 0/0
debug_tp = 0/0
filestore_op_threads = 8
filestore_max_inline_xattr_size = 254
filestore_max_inline_xattrs = 6
filestore_queue_max_ops = 500
filestore_queue_committing_max_ops = 5000
filestore_merge_threshold = 40
filestore_split_multiple = 10
Journal_max_write_entries = 1000
Journal_queue_max_ops = 3000
Journal_max_write_bytes = 1048576000
osd_mkfs_options_xfs = -f -I size=2048
osd_mount_options_xfs = noatime,largeio,nobar
rier,inode64,allocsize=8M
ods_op_threads = 32
osd_journal_size = 10000
filestore_queue_max_bytes = 1048576000
filestore_queue_committing_max_bytes =
1048576000
journal_queue_max_bytes = 1048576000
filestore_max_sync_interval = 10
filestore_journal_parallel = true
```

Linux 启动器调优

映射 LUN 配置

```
# echo 1024 > /sys/block/$DEVICE/queue/nr_requests
# blockdev –setra 32768 /dev/sd{}
# echo 128 > /sys/block/sd{}/device/queue_depth
# echo noop > /sys/block/sd{}/queue/scheduler
```

iSCSI 设置：更改 /etc/iscsi/iscsid.conf

```
Node.session.cmds_max=2048
Node.session.queue_depth=128
```

网络配置

TCP 设置：推荐更改 /etc/systemctl.conf

```
Net.ipv4.tcp_rmem= 10000000 10000000 10000000
Net.ipv4.tcp_wmem= 10000000 10000000 10000000
Net.ipv4.tcp_mem= 10000000 10000000 10000000
Net.core.rmem_default=524287
Net.core.wmem_default=524287
Net.core.rmem_max=524287
Net.core.wmem_max=524287
Net.core.netdev_max_backlog=300000
```

i40e 驱动程序设置

```
# service irqbalance stop
# ./scripts/set_irq_affinity <net_dev>
```

Linux LIO 驱动程序调优

iSCSI 目标门户组的参数值

```
FirstBurstLength=65536
MaxBurstLength=262144
MaxRecvDatSegmentLength=8192
ImmediateDataYes
InitialR2T=Yes
Default_cmdsn_depth=128
```

CPU	核心数量	8 个核心，16 个线程	8 个核心，8 个线程	4 个核心，8 个线程
	CPU 名称	英特尔® 至强® 处理器 D 1541	英特尔® 凌动™ 处理器 C2750	英特尔® 至强® 处理器 E3v3-1265L
	频率	2.1 GHz	2.4 GHz	2.5 GHz
内存	规格	DDR4 2400 MT/秒	DDR3 1600 MHz	DDR3 1600 MHz
	大小	32 GB，2 个内存通道 每通道 2 个 1 GB DIMM	16 GB，2 个内存通道 每通道 2 个 8 GB DIMM	32 GB，2 个内存通道 每通道 2 个 16 GB DIMM
存储后端	驱动器配置	4 TB WD SATA 64 MB 高速缓存	3 TB WD SATA 64 MB 高速缓存	3 TB WD SATA 64 MB 高速缓存
	RPM	7200	7200	7200
	硬盘数量	20 个用于存储的 OSD， 4 个用于日志记录的固态盘	10 个用于存储的 OSD， 2 个用于日志记录的固态盘	10 个用于存储的 OSD， 2 个用于日志记录的固态盘
网络	带宽	20GbE，MTU 9000	20GbE，MTU 9000	10GbE，MTU 9000
操作系统	发行版	Ubuntu Server 14.04.2	Ubuntu Server 14.04.2	Ubuntu Server 14.04.2
	内核	3.16.0-30-通用内核	3.16.0-30-通用内核	3.16.0-30-通用内核

Ceph 存储节点配置

[1] 参见：https://www.openstack.org/summit/vancouver-2015/summit-videos/presentation/Cph-at-cern-a-year-in-the-life-of-a-petabyte-scale-block-storage-service

0416/SM/HBD/PDF 334220-001CN