LightOS 与 Ceph 在 块存储应用的比较

August 2019

目录

导言

Ceph平台是云数据中心中常用的分布式存储系统，它在同一个包中提供块存储、对象存储和文件系统功能。虽然最初设计时，机械硬盘驱动器是流行的形式存储媒体，今天，它也使用与 NVMe SSD 作为底层媒体。

Ceph是一个复杂的系统，它提供了许多特性，但也很难设置、管理和控制。它使用自己的自定义 RADOS 协议在 Ceph 集群中的节点之间进行通信。应用机器 客户端 包括 RADOS 块设备 (设备驱动程序和对对象和文件系统访问的附加支持。

Lightbits 的 LightOS 是一种即将出现的集群存储系统，它 只关注块存储， 使用 由 Lightbits 发明并标准化 NVME/TCP 具有 创新 性 的 分类存储协议。 LightOS 提供了与 Ceph 相同的一些功能。
这包括：

- 分布式操作
- 数据保护和擦除编码的高可用性
- 在出现跨服务器集群的失败负载平衡时
- API 驱动的操作

与C eph 不同， LightOS 不需要客戶端驱动程序。客户端使用 NVMe/TCP 驱动程序，这些驱动程序在 Linux 发行版中是标准的，不久也将用于其他操作系统和管理程序。

LightOS与Ceph在性能方面的比较

如果存储解决方案标准是可伸缩性、易用性、IOPS 、延迟、故障时的快速恢复以及 TCO ，则基于 Light OS 方案 会击败 基于 Ceph 的 方案 。在类似的配置中，我们量化了Ceph 和 Light OS 之间的性能和延迟差异。对于 Ceph ，我们使用 Ceph 开发人员发布的结果来保证最佳的测量结果。对于 LightOS 来说，这一结果来自于在 L ightbits Labs 进行的实验。

我们的测试结果表明，LightOS 与 Ceph 相比可以提供最多

- 14 倍的 IOPS 计算 机 核
- 87 倍 优于 平均延迟
- 305 倍 改进 99B 的尾部延迟

这些结果来自于只使用两个 L ightOS 服务器 共有 56 个核 ） ） ，而 Ceph 的 5 个服务器 相当于 318 个核

注: LightOS 和 Ceph 不提供相同的功能集，所以一定要选择最适合需要的存储解决方案。

Ceph 设置

对于Ceph Performance 结果，我们使用了以下已发布的结果：

- Part 1 : BlueStore (Default vs. Tuned) Performance Comparison
- Part 2: Ceph Block Storage Performance on All Flash Cluster with BlueStore backend..

基于Ceph 博客文章，五个 RHCs OSD 节点 服务器 使用了以下硬件和软件配置：

Chassis	Cisco UCS C220-M5SN Rack Server
CPU	2 x Intel® Xeon® Platinum 8180 28 core (56 HT cores) @ 2.50 GHz
Memory	196 GB
NIC	Cisco UCS VIC 1387 2 port x 40Gb
Storage	Ceph Data: 7x Intel® SSD DC P4500 4.0 TB Ceph Metadata (RocksDB/WAL) : 1x Intel® Optane™ SSD DC P4800X 375 GB Ceph Pool Placement Groups : 4096
Software Configuration	RHEL 7.6, Linux Kernel 3.10, RHCS 3.2 (12.2.8-52)

Ceph七个客戶端分别使用了以下硬件和软件配置：

Chassis	Cisco UCS B200 M4 Blade servers
CPU	2x Intel® Xeon® CPU E5-2640 v4 @ 2.40GHz
Memory	528 GB
NIC	Cisco UCS VIC 1387 2 port x 10Gb
Software Configuration	RHOSP 10, RHEL 7.6, Linux Kernel 3.10, Pbench-FIO 3.3

注

本节开始时链接到的博客文章进一步详细介绍了网络拓扑结构、确切的 Ceph 软件配置以及与 FIO 一起运行的工作负载。

Ceph 测试有 105 卷或 84 卷， Ceph 基准测试有 40GbE 网络。由于 Ceph 无法饱和网络，增加更多的网络带宽不会改善 Ceph 的结果。

LightOS 设置

虽然Ceph 安装程序需要五个 OSD O bject Storage Device 节点 服务器 才能达到最大性能，但我们只需要两个 Light OS 服务器，配置如下：

Chassis	SuperMicro
CPU	2x Intel Xeon Scalable Gold 5120 14c/28t
Memory	512 GB
NIC	2x Mellanox ConnectX-5 En 100Gbps
Storage	14x Intel P45104TiB SSDs
Configuration	LOS+LF: LightOS v1.1 with LightField™ hardware acceleration and compression/decompression enabled, with mirrored replication

我们将10 个客戶端连接到 L ightOS 服务器，每个服务器都使用标准的 NVMe/TCP 同时连接到两台 LightOS 服务器。由于服务器数量很少，我们没有使用本机 LightOS 集群，这需要三个或更多的服务器。相反，每个客戶端通过镜像复制将数据复制到两个 LightOS 服务器。

客戶端配置如下：

Chassis	SuperMicro
CPU	2x Intel Xeon E5-2603 v4 6c/6t
Memory	64 GB
NIC	2x Mellanox ConnectX-4 LX En DP 25Gbps
Software	Ubuntu 16.04 with Linux 4.14.47

每个客戶端配置了11 个 NVME/TC P 卷，与 C eph 设置中的 84 卷或 105 卷相比，总共有 110卷。

LightOS 设置使用了 100 GbE 网络和与 Ceph 不同的网络拓扑，但正如上面所指出的， Ce ph无法满足用于 Ceph 基准测试的 40 GbE 网络设置。

在下面的表格中可以看到，Ceph 有相当于 318 Skylake 2.2GHz 的核心，而 LightOS 只有 56个核心。因此，我们给出了每 2.2GHz 核心归一化的 IOPS 结果。

存储系统	服务器数	CPUs L服务器	核数LCPU	核的速度[GHz]	归一化 2.2GHz 核的数量
Ceph	5	2	28	2.5	318
LightOS	2	2	14	2.2	56

试验结果

我们的实验首先比较了以下每种配置所提供的 IOPS/CORE 、平均延迟和 99% 尾延迟：

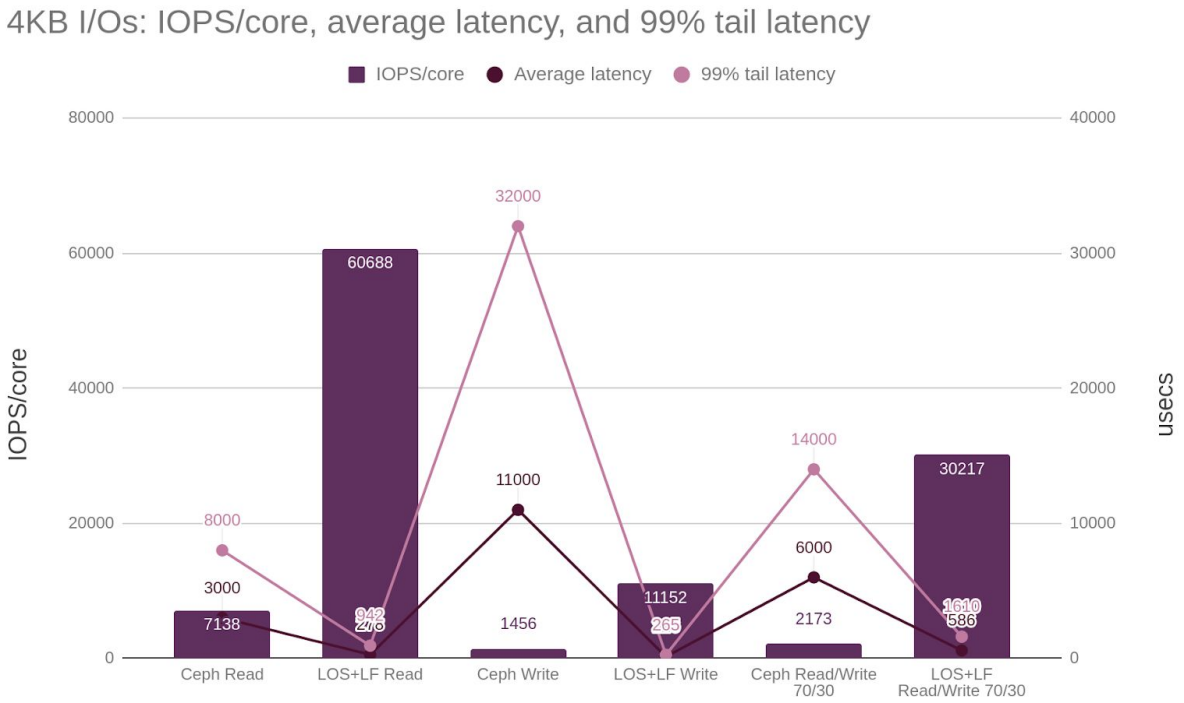
配置选项	系统	
	Ceph	LightOS 配LightField硬件加速卡
透明压缩/解压	NA	✔
多重数据复制保护服务器失效	✔	✔
图标记	Ceph	LOS+LF

采纳Ceph 博客中的方法之后，我们评估了纯读取工作负载、纯写工作负载和 70/30 读 写混合工作负载的每个配置。

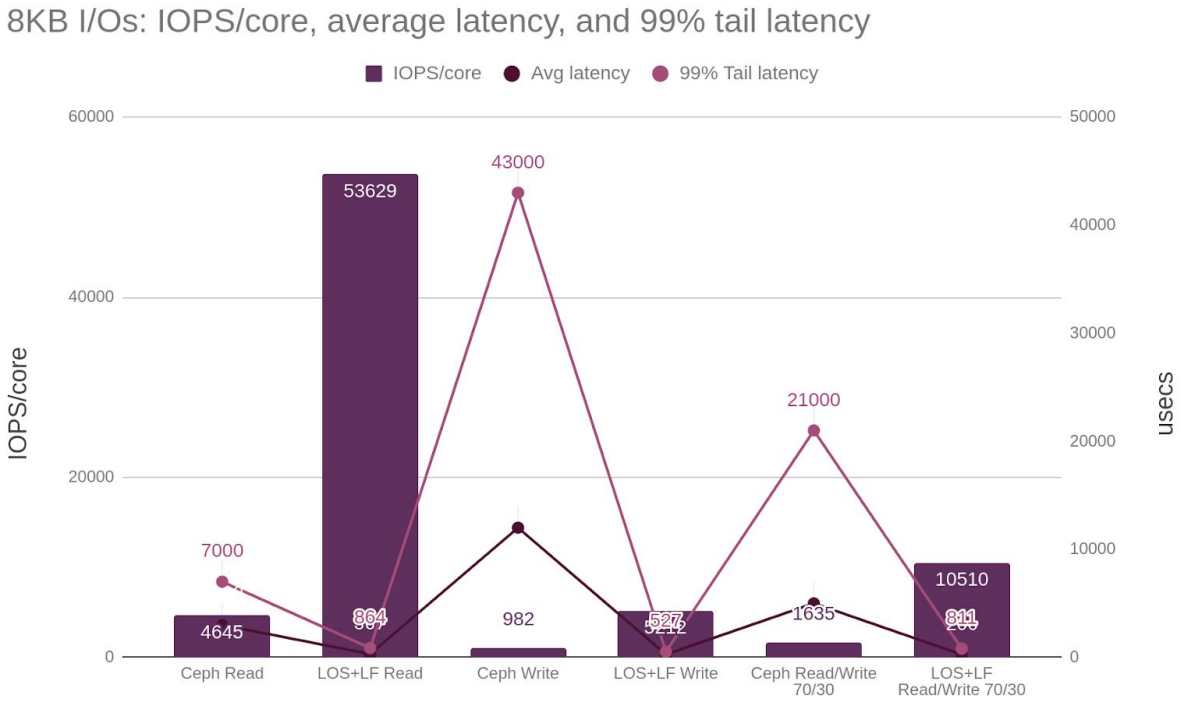
如以下图表所示

- **已读工作负载：** **LightField Acceleration** 的 **LightOS** 在 **11x** 更好（较低）平均延迟和 **9x** 更好的尾部延迟方面提供了 8.5 倍的较好的 IOPS/ 内核。
- **写入工作负载**： LightOS 和 LightField 在 87x 更好的平均延迟和 120x 更好的尾部延迟方面提供了 8 倍的较好的 IOPS/ 内核。
- **混合工作负载**： LightOS 和 LightField 在 10 倍的平均延迟和 9x 更好的尾部延迟方面提供了 214x 更好的 IOPS/ 内核。

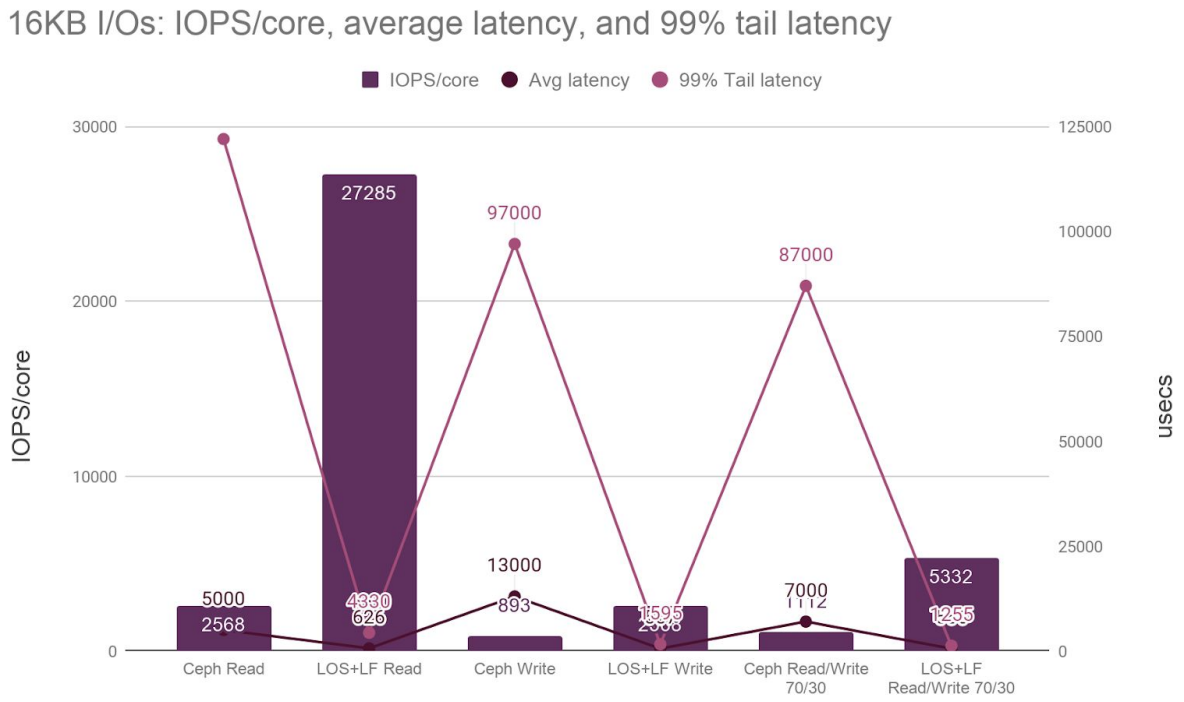
显然，对于工作负载主要为4KB IO 的主数据 块存储工作负载来说，具有硬件加速的 LightOS在整个主板上比 C eph 好得多。

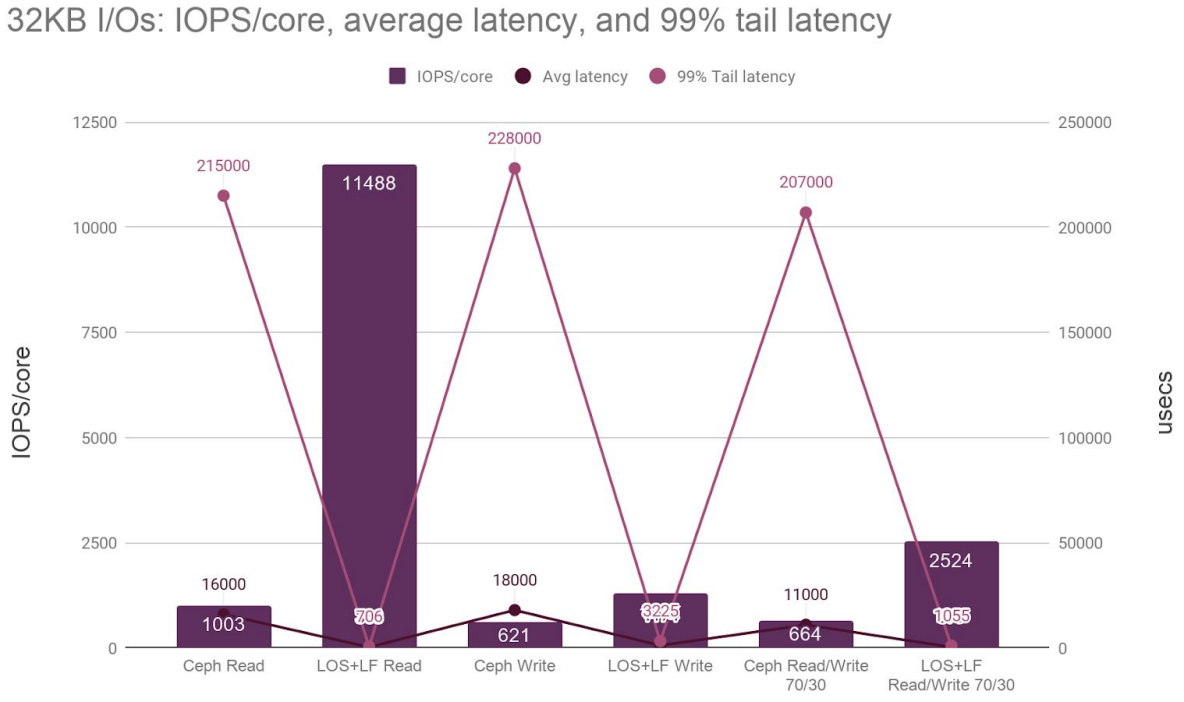


对于数据 量 大一点的 I/O 会 如何 呢 ？ 对于 8k IO/s ，具有硬件加速的 Light OS 提供了高达 12倍的高 IOPS/ 核心，平均延迟高达 44 倍，尾部延迟提高了 82 倍。



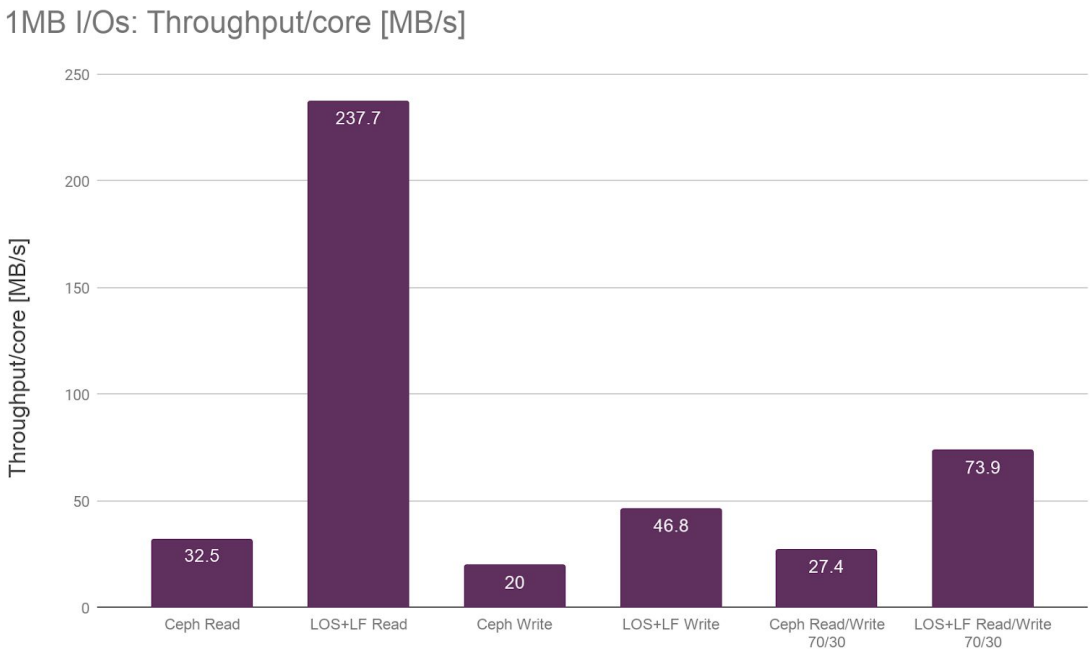
对于16k IO/s 和 32k IO/s 来说，故事是相似的，因为 L ightOS 更全面。 LightOS 提供了最多11倍更好的 IOPS/core，44 倍更好的平均延迟， 305 倍更好的尾延迟。





通常主块存储 I/O 相对较小，典型的 I/O 大小为 4K 、 8K 、 16K 或 32K 。但是，有时存在 I/O更大、大小高达 1MB 的工作负载。那 会 发生 什么?

如下图所示，每个核心的标准化、有硬件加速和不加硬件加速的灯塔比Ceph 高出 7 倍的吞吐量。



在故障情况下的LightOS与Ceph的对比

稳态性能和延迟是非常重要的。尽管在此评估中，我们没有比较LightOS 和 C eph 的性能和延迟，但 有 证据表明， LightOS 可以从服务器（例如 SSD 故障）和整个服务器故障中恢复，但比 C eph 更快，同时 对运行工作负载的影响为最小。从故障中恢复并更快地返回到稳定状态意味着系统易受另一故障影响的时间窗减小。在故障情况下对 C eph 和 LightOS 进行详细的定量评估仍然是未来的工作。

总结及结论

如果需要为主要存储工作负载构建云级块存储服务，则应仔细考虑性能、延迟和TCO 需求。当使用较慢的媒体 如硬盘驱动器和 SATA SSD) 时，Ceph 可能非常有用。

但是，当您的介质是NVMe SSD时，您应该使用完全能够利用NVMe速度的存储解决方案。如我们所显示的，仅使用两台总共56个内核的LightOS服务器，而Ceph使用五台318个内核的服务器，LightOS可以提供最高14倍的IOPS /内核，87倍的平均延迟和305倍的99％尾部延迟。

About Lightbits Labs™

当今的存储方法是为企业设计的，无法满足不断发展的云规模基础架构要求。例如，SAN因缺乏性能和控制能力而广为人知。从规模上讲，直接连接的固态硬盘（DAS）变得过于复杂，无法顺利运行，成本也很高，并且固态硬盘利用率低下。

云规模的基础架构需要对存储和计算进行分解，顶级云巨头从低效的直接连接SSD架构过渡到低延迟共享NVMe闪存架构证明了这一点。

与其他NVMe-oF方法不同，Lightbits NVMe / TCP节省成本的解决方案将存储和计算分开，而不会影响网络基础架构或数据中心客戶端。Lightbits团队成员是NVMe标准的主要贡献者以及NVMe over Fabrics（NVMe-oF）的发起者之一。

现在，Lightbits正在制定新的NVMe / TCP标准。作为该领域的开拓者，Lightbits解决方案已经在行业领先的云数据中心中成功进行了测试。

该公司的共享NVMe架构提供了有效而强大的分解。如此平稳的过渡过程，您的应用团队甚至都不会注意到这一变化。他们现在可以以比本地SSD更好的尾部延迟发狂！

www.lightbitslab.com ✉ info@lightbitslabs.com

US Office
1830 The Alameda,
San Jose, CA 95126, USA

Israel Offices
17 Atir Yeda Street,
Kfar Saba, Israel 4464313

3 Habankim Street,
Haifa, Israel 3326115