腾讯云私有云对象存储服务
技术白皮书

腾讯云

目录

概述

简介

数据正以前所未有的速度增长，「云」提供了一种基础架构服务，解决大量数据和计算的问题。要想跟上业务发展的快速步伐，有效管理 TB ~ EB 级数据，您需要以灵活、经济高效的方式在云端存储数据。

腾讯云私有云存储（Cloud Storage on Private，后文简称 CSP）旨在帮助客户降低企业存储数据成本，通过高效、灵活、自动化的方式，管理呈指数级增长的业务数据。腾讯云私有云存储提供了大规模、可扩展的持久化分布式存储平台，管理分布式计算机群集上的数据，并为对象级别存储提供接口，让您可以集中精力运行主要业务。

CSP 对象存储服务是在腾讯多年的海量数据存储的经验之上，结合开源存储项目生态与自研组件服务，对外提供的可靠、安全、易用的海量对象存储平台。您可以按需部署对象存储服务以实现企业的海量文件存储的需求，例如：文档、图片、视频等非结构化文件的存储。使用 HTTP RESTful API 作为基础接口，可以支持原生云计算应用、批量计算分析、归档备份以及内容分发等应用场景。

特点

对象存储是云存储中一种无层次结构的数据存储方法，其不使用目录树结构，各个单独的数据（对象）单元存在于存储池中的同一级别。每个对象都有唯一的识别名称，可用于列出检索操作，另外每个对象还包含对应的元数据信息。其对比传统文件存储方式，具有如下特点：

- 数据作为单独的对象进行存储
- 数据并不放置在目录层次结构中，而是存在于平面地址空间内
- 应用通过唯一地址来识别每个单独的数据对象
- 专为使用 HTTP RESTful API 在应用级别进行访问而设计

优势

CSP 对象存储提供对象存储能力，接口完全兼容 AWS S3 协议与 OpenStack Swift 协议，通信基于 HTTP/HTTPS 网络协议，并使用 REST(REpresentational State Transfer)风格，协议简单清晰无状态，易于访问，支持多语言的 SDK。

CSP 对象存储提供了用户隔离，不同的用户之间是资源隔离的，资源所有者可以决定

是否允许其它用户访问。

CSP 对象存储提供了“存储桶（Bucket）”与“对象（Object）”两个存储级别，存储桶可以理解为文件系统中的目录，存储桶的命名全局唯一，一些通用配置作用在存储桶级别。对象可以理解为文件系统中的文件。对象存储舍弃了传统文件系统中复杂的读写语义与目录结构，可以存储海量的文件，提高文件检索与存储性能。

CSP 对象存储具备如下关键优势：

- 扩展性：参考业界成熟的分布式架构，从所需的最小实例开始，根据负载和存储要求逐步扩展存储规模，业务平滑扩容
- 持久性：支持多节点、机柜、机房的分布式管理，实现了数据多副本或纠删码冗余，极大地提升了数据可靠性
- 经济性：使用兼容性软硬件平台，避免专用设备开销，显著降低每 GB 的部署和管理成本，以应对高速增长的数据存储
- 安全性：可以被部署硬件独占集群，并实现网络层面的软件安全隔离，数据支持 SSL 传输加密和 SSE 服务端加密
- 整合性：与虚拟化平台、容器服务、大数据计算等服务紧密集成，为整个云平台提供完善的存储支撑，并提供友好的 S3 API 兼容性
- 多租户：支持多租户模式，提供丰富的权限管理方式，最大限度的满足企业环境资源分配需求

产品特性

特性类型	特性	描述
管理页面	集群概览	提供资源使用情况、集群运行状态基本信息展现面板
	存储池管理	支持多存储策略，每个存储池可以配置不同副本策略，解决不同业务场景需求
	主机管理	提供添加、删除主机功能，通过可视化界面实现集群扩缩容；支持主机 CPU，内存，磁盘状态监控，实时掌握主机运行状态
	磁盘管理	磁盘健康状态实时监控，支持在线换盘
	监控告警	支持多级事件告警，系统主动向注册邮箱推送告警信息，支持管理页面查询告警事件
	用户管理	基于角色的用户权限管理，实现资源视图隔离
	日志管理	支持系统运行日志、用户操作日志检索功能，用于异常跟踪、系统排障和用户行为审计
容灾恢复	多副本冗余	支持 2 副本及以上的副本策略，针对不同业务设置不同

		等级的数据可靠性
	EC 纠删码	相比于副本方式，纠删码采用计算时间换取存储空间的方式，只需更少的存储空间，来保证数据可靠性
	容灾策略	支持主机、机架、机房纬度容灾，解决不同业务对不同安全等级要求
	多数据中心	支持跨区域复制，多数据中心数据同步
	恢复控制	当节点从异常状态恢复后，存储系统默认实现数据迁移以及数据重平衡，同时提供恢复控制（Recovery QoS）特性，让数据在恢复过程中，对业务正常读写影响降到最低
	集群巡检	定期集群状态检查，提前发现系统潜在异常
	硬件热插拔	支持主机、磁盘热插拔，实现在线扩缩容、异常硬件剔除
开放协议	S3 接口	兼容 Amazon S3 (Amazon Simple Storage Service) 对象存储接口
高级特性	读写缓存	支持配置 Cache，提升文件读取性能
	在线扩容	在线扩容，现有业务无感知
	在线缩容	在线扩容，现有业务无感知
	REST API	支持 REST API，便于二次开发对接现有系统
	多存储介质	支持 NVMe SSD，SATA SSD、SAS HDD、SATA HDD 多种存储硬件，提供更多的性能、容量、成本选择

应用场景介绍

内容托管与分发

CSP 对象存储可以提供高可靠性和高可用性的数据存储服务，将传统存储设施中的数据转移到对象存储集群中，实现高性价比的、根据流量要求扩展的内容托管分发解决方案。随着数据和业务的增长，可以通过合理预留资源以及简便的扩容处理高峰流量，并可以作为 CDN 服务的源站存储，提供分发服务。

大数据分析

CSP 对象存储支持作为大数据的原始存储，适合政务、财务、医疗、图片、音视频以及日志等文件。集成腾讯云大数据处理能力，可以通过 Hadoop 文件系统中标准的 S3N

协议访问，作为大数据计算的存储后端，提供高吞吐、高并发的读写能力，同时保障数据存储的持久性和经济性。

备份与灾难恢复

CSP 对象存储提供高持久性、高安全性的存储引擎，可用于备份和归档关键数据，并提供在灾难时的恢复解决方案，实现数据保护。不仅实现跨节点、机架和机房容灾，针对多地多集群部署的场景，还能通过跨区域复制的能力实现跨数据中心容灾。

技术规格与功能

整体架构

CSP 对象存储从上到下依次分为协议层、接入层、存储层，各层的功能如下：

层次	功能
协议层	协议层提供 AWS S3 对象存储协议
接入层	接入层负责将协议层的请求转发到存储层
存储层	存储层则负责提供底层存储、数据索引等核心存储服务

主要的模块如下：

Data Access：负责对外提供 API 接口，解析 HTTP 请求。该模块无状态，可以根据用户流量的需求，实现分布式部署。

Index Storage：负责存储元数据。

Data Storage Device：模块负责管理磁盘存储，保障数据可靠性、安全性。

Cluster Monitor：模块负责集群元信息的一致性的存储，例如集群拓扑结构，存储结点状态等。

Cluster Manager: 模块提供 WEB 控制台，通过控制台查看集群和服务的状态，管理存储集群。

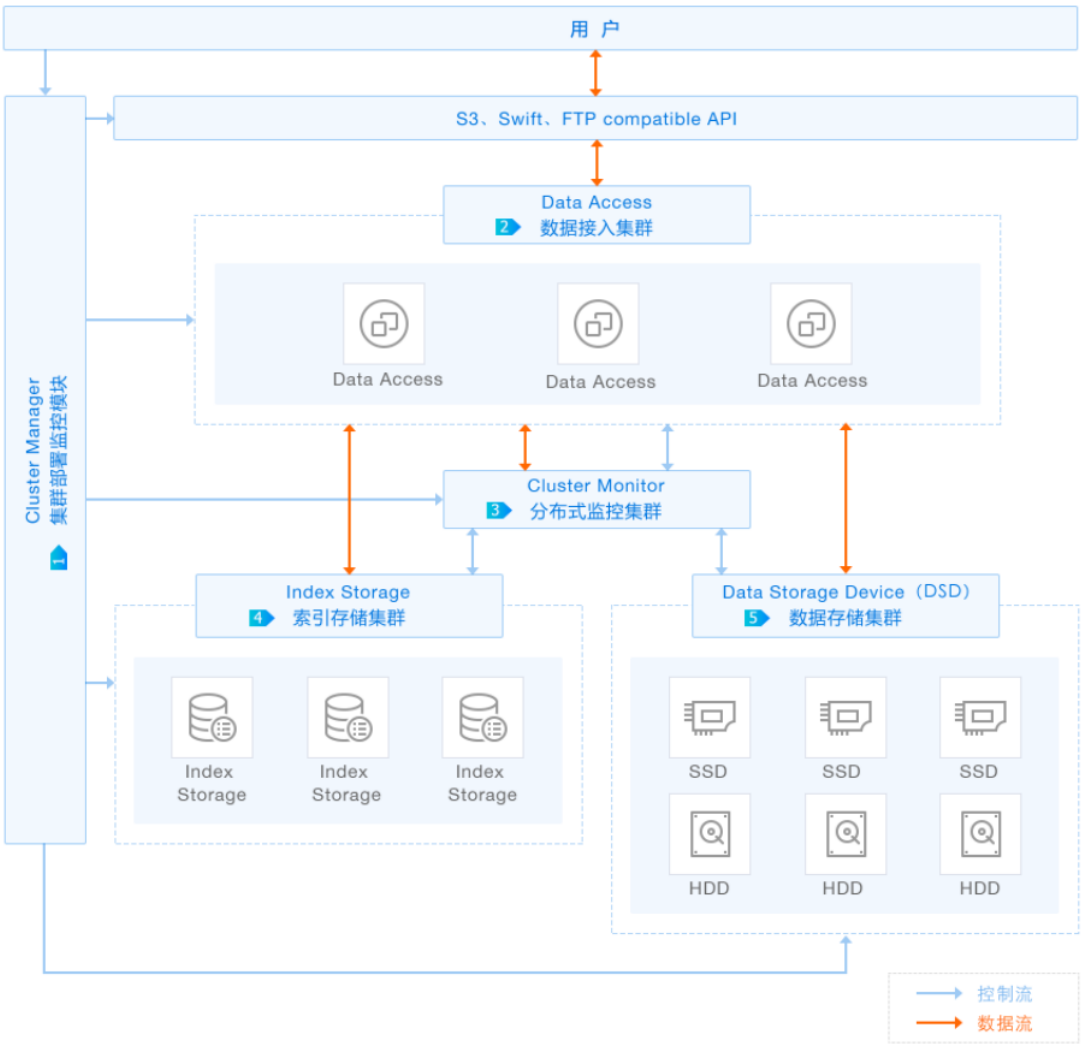


图 1：CSP 对象存储架构图

关键业务流程

写请求

数据写入可以分为如下步骤：

1. 请求接入：使用特定语言的开发套件，与对象存储服务的接入节点建立连接，并向其发送数据
2. 数据分片：接入节点根据收到的数据，采取最优的分片策略，将一个较大的对象分为多个小文件
3. 数据路由：在接入节点完成分片后，根据这些信息，计算分片对应存储的机器节点，并将数据发送到对应的存储节点
4. 数据存储：存储节点在收到数据后，将数据按照策略安全的保存为多副本或者纠删码

读请求

数据读取操作与写入类似，只是数据流的流向不同。

1. 请求接入：客户端与对象存储服务的节点建立连接，客户端向访问的节点请求数据
2. 数据路由：接入节点根据路由规则寻址到存储节点，读取相应的数据片
3. 数据修复：如果某些数据片损坏，存储节点将根据数据的存储策略进行相应的修复操作
4. 数据聚合：接入节点将数据片聚合为完整的数据，发送给客户端

数据冗余策略

CSP 对象存储支持多种副本策略，包括主流的副本冗余以及适应于冷数据的纠删码策略。

对于 3 副本策略，对于每一个对象，存储系统都会将其按照用户要求的故障域隔离级别，在多台机器上保存该对象。在机器宕机或者其它不可恢复故障时，CSP 对象存储自动地将对象迁移到其它节点上，确保数据的完整性。3 副本的冗余策略可以提供较高性能的读写能力，容忍磁盘损坏或者机器宕机等多种故障，保证数据可靠性。

CSP 对象存储同时支持纠删码的冗余策略，EC（Erasure Code）算法实现数据冗余存储，确保硬件失效时的数据可靠性和可用性。纠删码（EC）技术主要是对数据分片进行分组，每个分组有数据块和校验块组成，其中校验块即为产生的部分冗余数据。如果数据的一部分损坏或丢失，对象存储服务能够利用冗余的数据重建并修复损坏数据。该策略数据不仅具有较高的可靠性，而且存储空间利用率非常高（相比多副本模式），是可靠性和经济性平衡的最佳选择。

对于用户上传的数据，CSP 对象存储的接入节点在将数据切分为数据片的过程中，会将连续的 N 个数据片划分为一个 EC 块，并利用纠删码技术对 EC 块进行计算，生成 M 个校验数据片。每个 EC 块的数据片和校验数据片，将存储在存储集群上不同的存储节点上，确保其可靠性。只要每个 EC 组损坏的数据片数量不超过 M，CSP 对象存储的存储结点都能利用 EC 组的其它数据片将损坏数据片修复。

关键技术

无单点架构设计，PB 级存储能力

CSP 对象存储系统的每个模块都是水平拓展的，没有故障单点，可以提供 PB 级别的存储能力。在任意组件发生故障的时候，系统会自动修复数据。CSP 存储系统基于分布式哈希算法来路由存储请求，客户端可以访问存储结点，简化存储路径，让 CSP 存储系统具有优异的性能和海量的存储能力。

多种副本策略，优化成本与性能

CSP 对象存储系统支持多种副本策略，您可以根据特定的业务场景和性能要求选择合适的副本策略。3 副本存储适合于较高性能的场景，通过 3 副本来保证系统的高可用性、高可靠性以及高吞吐能力。纠删码副本存储则使用更少的存储的空间来提供高可靠性的存储能力，使用 1.3 倍的冗余达到与多副本一样的可靠性，适用于较冷数据的存储场景。

海量文件存储与检索能力

CSP 对象存储针对对象存储的读写场景，提供了 key-based 存储模型以及存储桶、对象的存储单位概念，简化了大量文件系统的不必要属性与语义，同时不同于传统文件系统的 inode 设计，CSP 对象存储独立数据与元数据的存储，让整个系统拥有海量文件的存储能力，不受 inode 的个数限制。

CSP 对象存储提供了简单实用的一致性模型"read-after-write"，只要确认一个文件写成功，后续的所有读取操作都可以访问该文件，同时因为元数据的独立存储，检索路径短，与传统文件系统的多次系统调用相比，性能提升明显，让海量小文件存储没有瓶颈。

最小集群

若采用 3 副本方式，存储集群至少需要 3 台服务器（Data Access、Index Storage、Data Storage Device、Cluster Monitor 可以复用部署），加上集群管理模块需要 1 台独立服务器进行部署，仅仅需要 4 台普通 x86 机器，就能实现 CSP 对象存储集群部署。

说明：最小集群仅仅可以用作测试，不能用作生产环境。

各模块推荐机器配置如下表：

模块名称	机型配置	备注
Data Access	6 核 32G 内存，1TB SATA 磁盘	可以根据业务情况进行水平扩容
Index Storage	6 核 32G 内存，4TB SATA 磁盘	数据索引模块，推荐使用 SSD 盘
Data Storage Device	6 核 32G 内存，12*4TB SATA 磁盘	数据存储模块，大容量 SATA 盘
Cluster Monitor	6 核 32G 内存，1TB SATA 磁盘	集群监控，计算，内存消耗较大
Cluster Manager	4 核 4G 内存，1TB SATA 磁盘	建议独立部署，与底层存储业务隔离

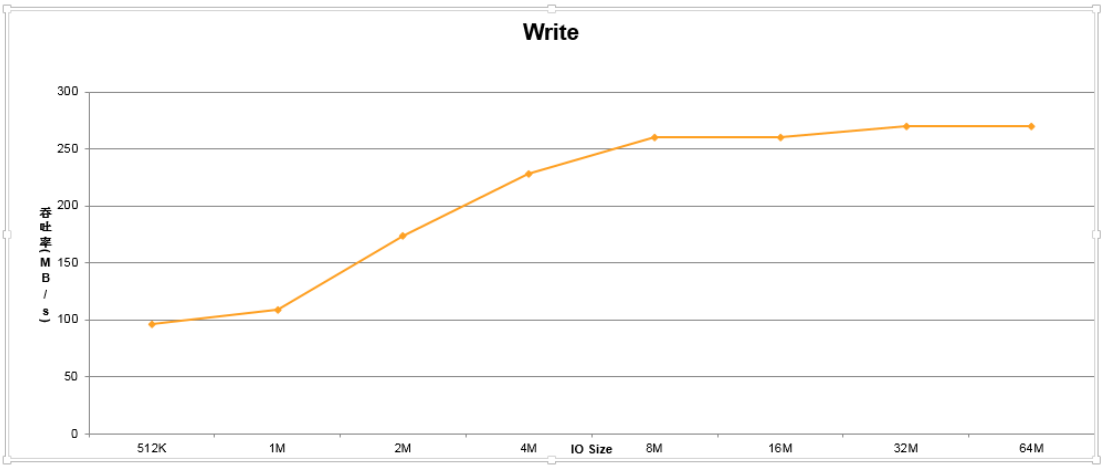
考虑到 Data Access、Index Storage、Data Storage Device、Cluster Monitor 可以复用部署，这个模块使用 3 台 6 核 12G 内存，12*4TB SATA 磁盘的服务器，即可实现裸数据 48TB，3 副本的对象存储集群。

最小集群测试

软硬件环境配置：

3 台 6 核 32G 内存、12*4TB SATA 磁盘、10Gb 网卡的服务器，3 台作为存储集群 Index Storage、Data Storage Device、Cluster Monitor 组件部署，1 台用于部署 Data Access，存储池采用 3 副本。

写吞吐率测试

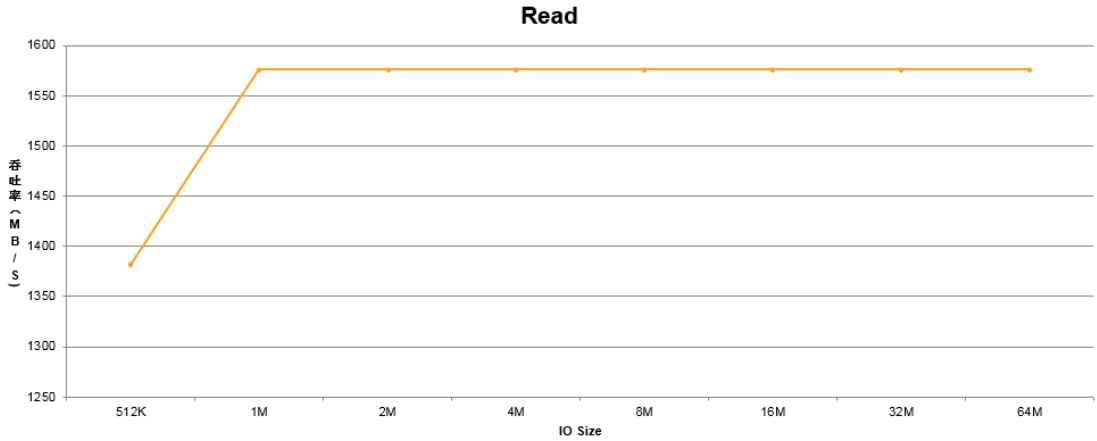


IO Size	吞吐率（MB/s）	并发数	Bucket 数
512K	96	1024	32
1M	109	512	32
2M	174	512	32
4M	228	512	32
8M	260	256	32
16M	260	256	32
32M	270	256	32
64M	270	256	32

IO Size 超过 8M，并发度为 256 时，写吞吐率性能达到 270MB/s 的上限

写入采用三副本模式，网络带宽只能利用三分之一

读吞吐率测试



读性能表现优异，达到网卡上限：

512K IO　　读取吞吐率 1.3GB/s

>=1MB IO　读取吞吐率达到网络上限，为 1.5GB/s