JWanFS
广域网分布式云存储

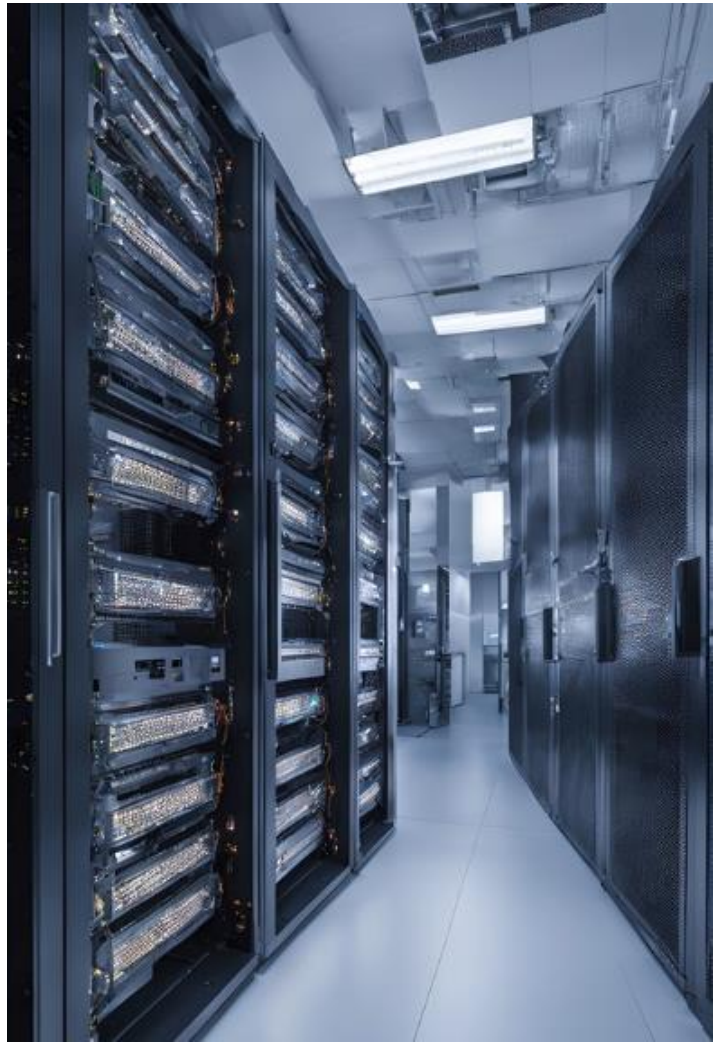
隗立畅

weilc@ihep.ac.cn

高能物理研究所计算中心

金钱猫科技股份有限公司

2024/8



1 背景

2 系统概述及架构

3 客户端

4 性能分析

5 Ocloud——JWanFS应用实例

PART ONE

01

背景

Background

1.1 异地存储痛点问题

客户存储节点位于全国各地不同数据中心，或不同公有云的云空间，数据需跨异地和异构系统传输和访问。

数据需要在各节点间进行流转，在其中一个节点写入数据，要求其它异地节点可以立刻读取到数据。

各地存储节点容量不同，大少不一，无法灵活调度，以满足海量大数据跨异地灵活存储需求。

部分节点网络安全性要求高，对公网用户数据存储和访问仅提供单一链路单一端口，链路条件苛刻。

部分节点无公网地址，但又希望作为存储节点提供其他数据中心访问。

1.2 现有系统存在问题

Lustre

网络延迟敏感：

Lustre的架构设计对网络延迟非常敏感，异地环境中的高延迟会显著降低其性能。

集中式元数据管理：

MDS的集中式管理在异地环境中容易成为瓶颈，增加单点故障的风险。

GlusterFS

数据同步挑战：

去中心化架构在异地环境中面临数据同步的挑战，网络延迟和带宽限制会影响数据一致性和可用性。

小文件性能问题：

GlusterFS在处理小文件时性能较差，异地环境中的网络开销会加剧这一问题。

EOS

高性能网络需求：

EOS对网络性能的要求非常高，异地环境中的网络延迟和带宽限制会显著影响其性能。

数据局部性假设：

EOS的设计假设计算节点和存储节点在物理上接近，这在异地环境中难以实现。

网络端口需要完全打开

PART TWO

02

系统概述及架构

system overview and architecture

2.1 JWanFS优势

海量数据存储

先进的数据管理技术架构，有力支撑百万级文件、ZB 级别数据扩展，可承载海量数据

数据统一管理

提供Web控制台，用户可轻松创建和管理文件系统，实现多协议数据统一视图查看，省去本地搭建与运维成本

存储高可靠性

全方位安全措施，满足企业数据安全与合规要求。系统设计可用性达到99.99%，数据可靠性达到99.9999%

多协议访问

支持多种数据访问接口协议，包括NFS、FTP、S3、WebDAV、本地挂载等。支持Linux、MacOS与Windows客户端使用

公网安全传输

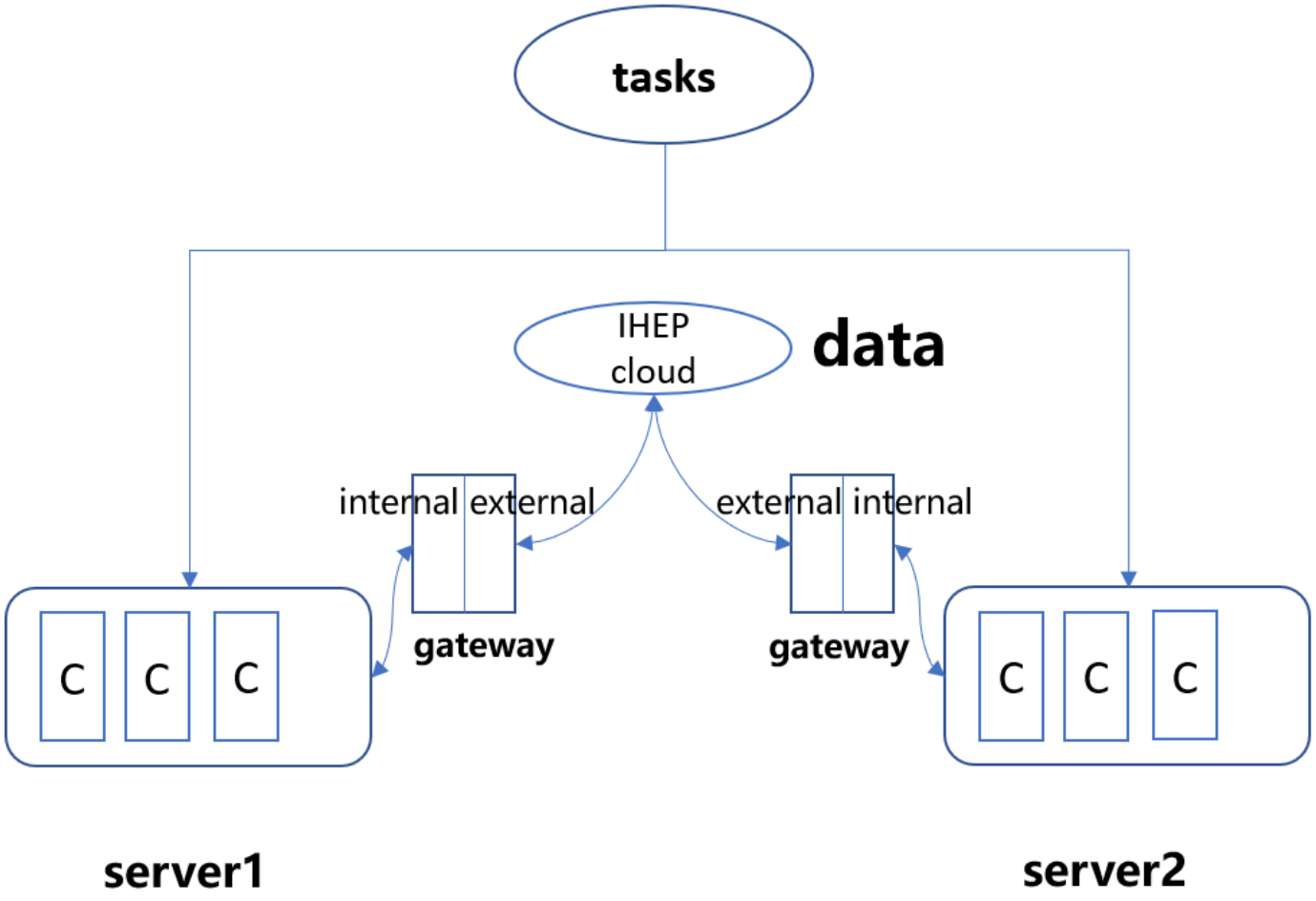
广域网节点间通过传输节点互联，数据在公网间传输只需要开放各传输节点，同时设置节点IP白名单，实现数据在公网上安全传输

异地数据实时同步

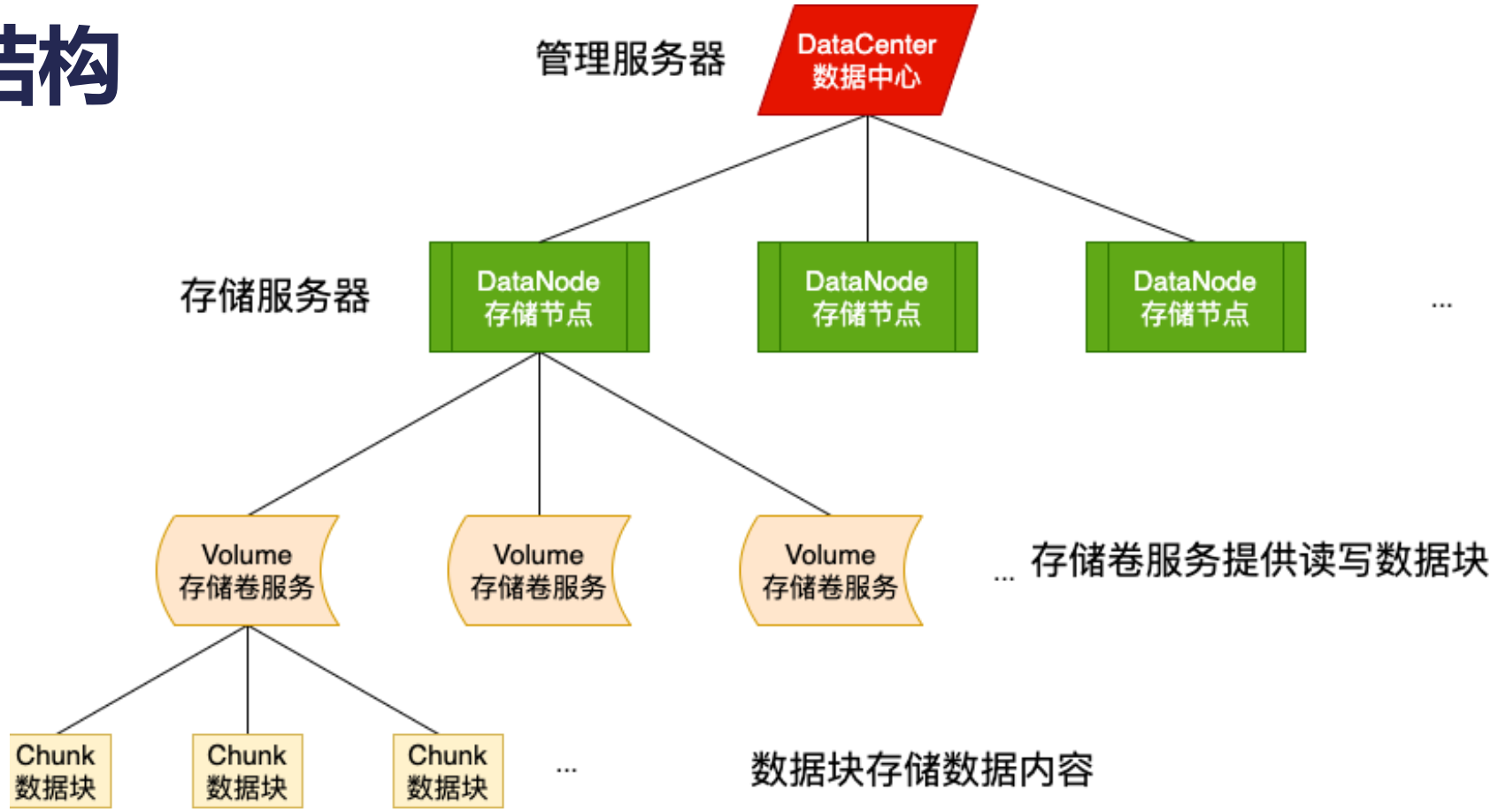
优秀的读写架构允许跨地域读写数据实时同步，当其中一节点写入数据时，另一异地节点可以立即读取，无需等待异步复制同步

JWanFS广域网分布式云存储，统一调度，统一分配，支持多种数据访问接口协议，统一视图数据查看，数据加密传输、加密存储，保证数据高可靠、高可用、高安全性

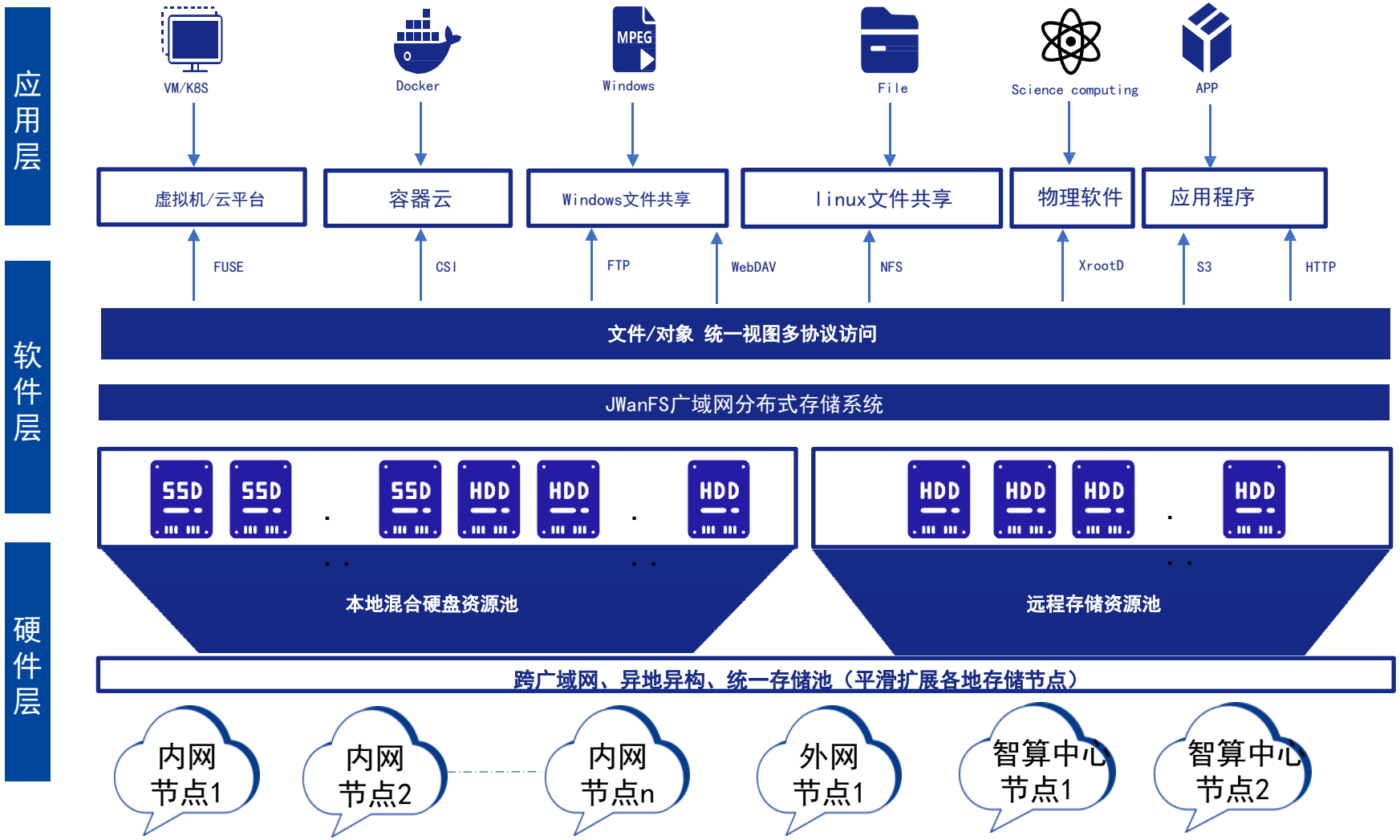
数据共享模式


tasks
IHEP
cloud
data
internal external
external internal
gateway
gateway
C C C
C C C
server1
server2

存储层级结构


管理服务器
DataCenter
数据中心
存储服务器
DataNode
存储节点
DataNode
存储节点
DataNode
存储节点
...
Volume
存储卷服务
Volume
存储卷服务
Volume
存储卷服务
... 存储卷服务提供读写数据块
Chunk
数据块
Chunk
数据块
Chunk
数据块
...
数据块存储数据内容

2.2 系统架构



多应用访问支持

通过文件网关提供对全部类型应用访问数据的解决方案。

统一视图查看

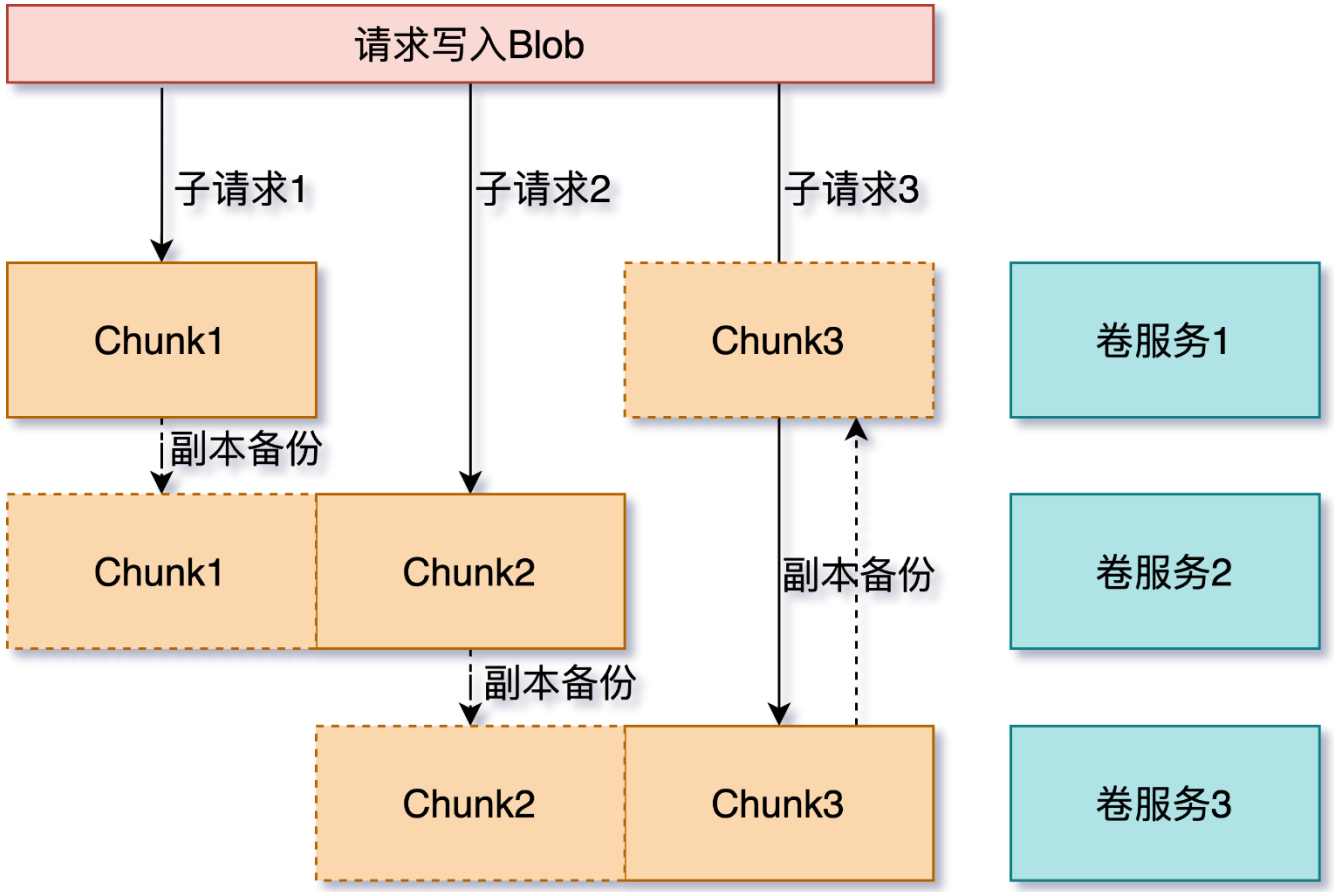
不同地域数据统一管理，多协议统一视图查看，打通应用隔阂。

分布式存储架构

全对称的分布式架构，容量和性能 可横向扩展。软硬件解耦合，支持利旧，扩容、替换成本更低。

2.2 系统架构

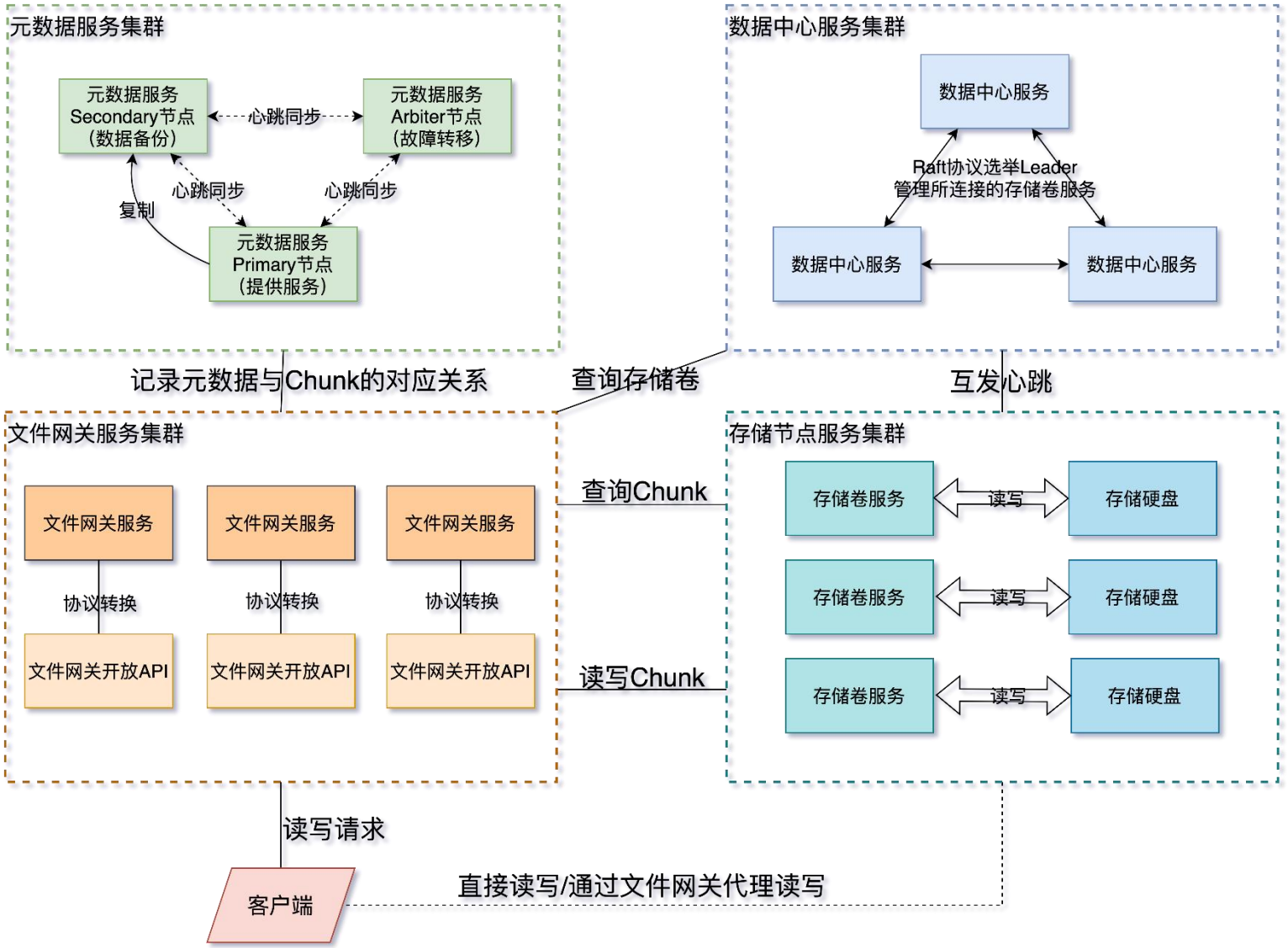
高效可靠的副本备份方案



2.2 系统架构

高可靠元数据服务

多协议兼容



卷服务管理

硬件读写交互

2.3 硬件部署要求

- **支持Intel/AMD、海光、鲲鹏、飞腾、龙芯等各类架构的异构服务器的混合部署**
- **服务端支持运行在各种Linux操作系统上，包括：CentOS、Ubuntu、OpenKylin、统信UOS、华为EulerOS等**

	标准配置
核心数	4核心以上
操作系统	Linux
内存信息	32G以上
磁盘信息	300G 系统盘 + 若干存储盘
网络带宽	50Mbps以上存在固定公网IP 可开放1-2个端口

	推荐配置
核心数	8核心以上
操作系统	Linux
内存信息	64G以上
磁盘信息	300G 系统盘 + 若干存储盘
网络带宽	50Mbps以上存在固定公网IP 可开放1-2个端口

	极简配置
核心数	2核心以上
操作系统	Linux
内存信息	16G以上
磁盘信息	100G 系统盘 + 若干存储盘
网络带宽	50Mbps以上存在固定公网IP 可开放1-2个端口

系统内存占用参考：
如果管理1TB存储空间，仅占用1GB内存
如果需要管理100TB的数据空间，推荐部署服务器内存在64GB以上

PART THREE

03

客 户 端

jcliuse

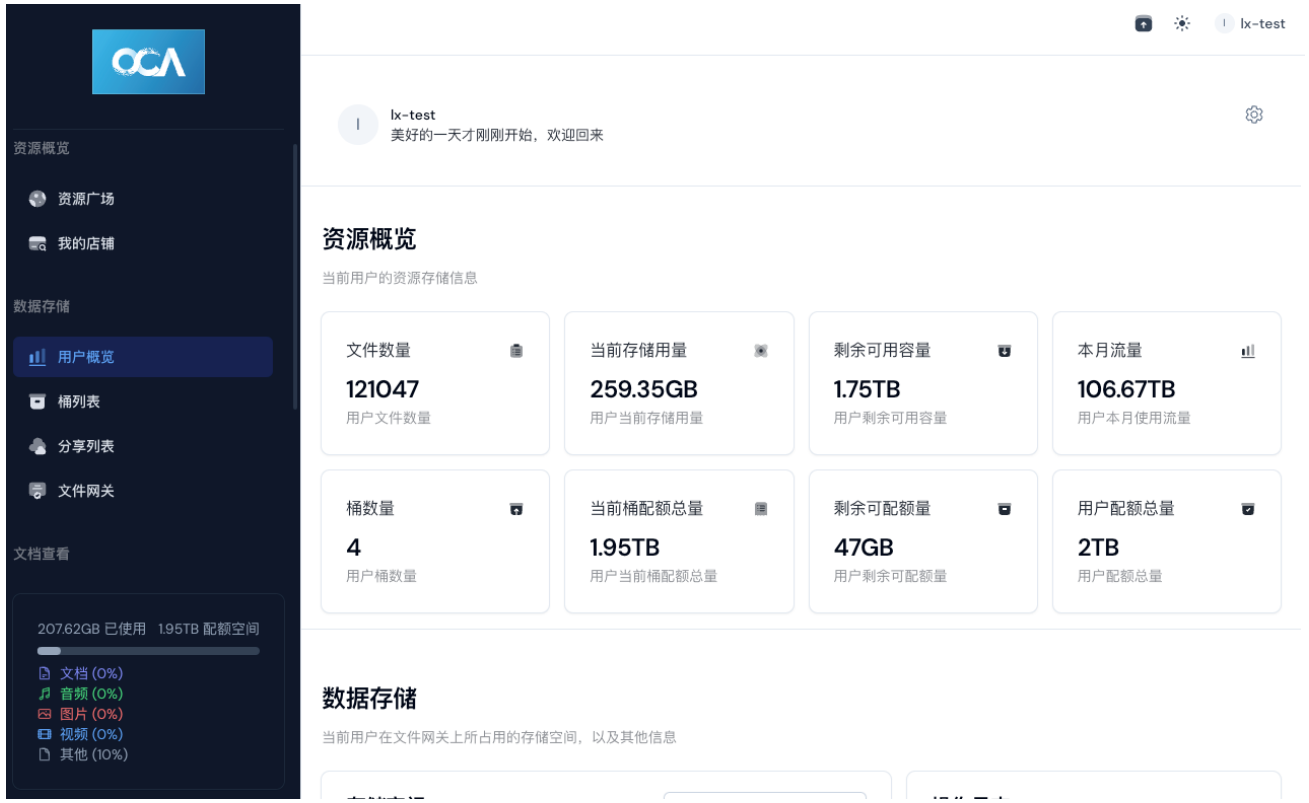
3.1 WEB客户端

用户登入

- 支持接入OAuth2.0协议统一认证
- 支持Carsi统一认证，接入全国773所高校机构
- 采用AcessKey/SecertKey鉴权

https://ocloud.ihep.ac.cn

3.1 WEB客户端



灵活的数据管理

- 创建、查看、删除 Bucket
- 用户可以自定义设置需要创建的 Bucket 数量，并支持根据需求修改配额
- 上传、查看、删除数据
- 用户可以存储的 数据没有数量限制
- 数据支持文本、多媒体、二进制等任何类型的数据

3.2 命令行

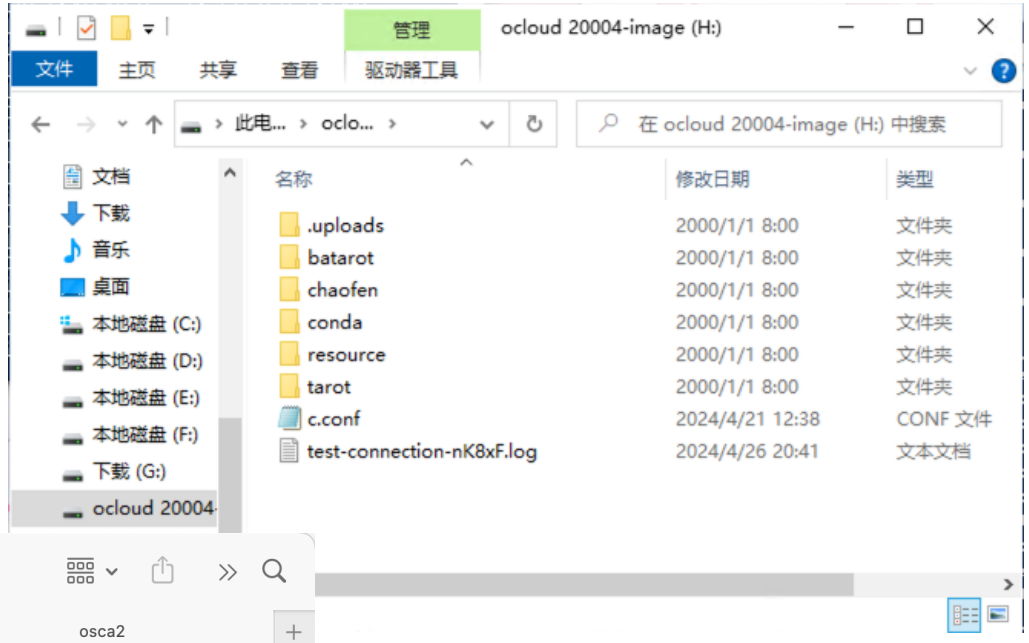
支持命令行进行授权以及各种文件操作

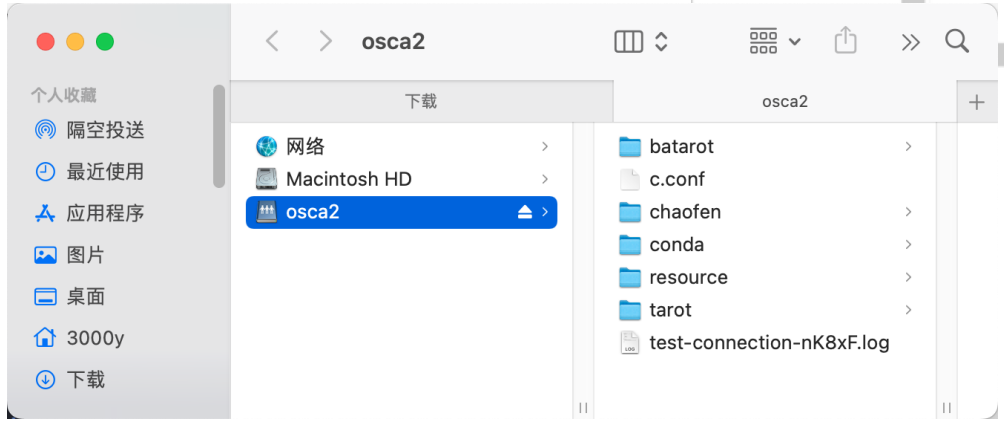
```
[root@ocloud ~]# jcli
子命令列表：
  命令      别名           作用
  auth                   认证服务
  config                 配置认证信息
  cp                     复制对象
  rm        delete,del   删除文件
  etag                   计算文件ETag
  expire                 获取AKSK过期时间
  get                    下载文件
  list      ls           查看文件列表
  mv                     移动对象
  put                    上传文件
  stat                   获取文件信息
  token                  创建临时AKSK
  update                 更新客户端版本

使用 help [子命令] 查看子命令详细信息
[root@ocloud ~]#
```

3.3 挂载访问

提供Linux、Windows、MacOS挂载解决方法
访问数据像访问本地磁盘一样轻松

```
[3000y@iZbp1ghimkhmv0ihg7x4ajZ osca]$ ll
total 23
drwxr-xr-x 1 3000y 3000y 4096 Apr 28 08:14 batarot
-rw-r--r-- 1 3000y 3000y 1328 Apr 21 12:38 c.conf
drwxr-xr-x 1 3000y 3000y 4096 Apr 21 12:43 chaofen
drwxr-xr-x 1 3000y 3000y 4096 May 11 14:54 conda
drwxr-xr-x 1 3000y 3000y 4096 Apr 26 20:16 resource
drwxr-xr-x 1 3000y 3000y 4096 Apr 26 10:47 tarot
-rw-r--r-- 1 3000y 3000y 1072 Apr 26 20:41 test-connection-nK8xF.log
[3000y@iZbp1ghimkhmv0ihg7x4ajZ osca]$
```



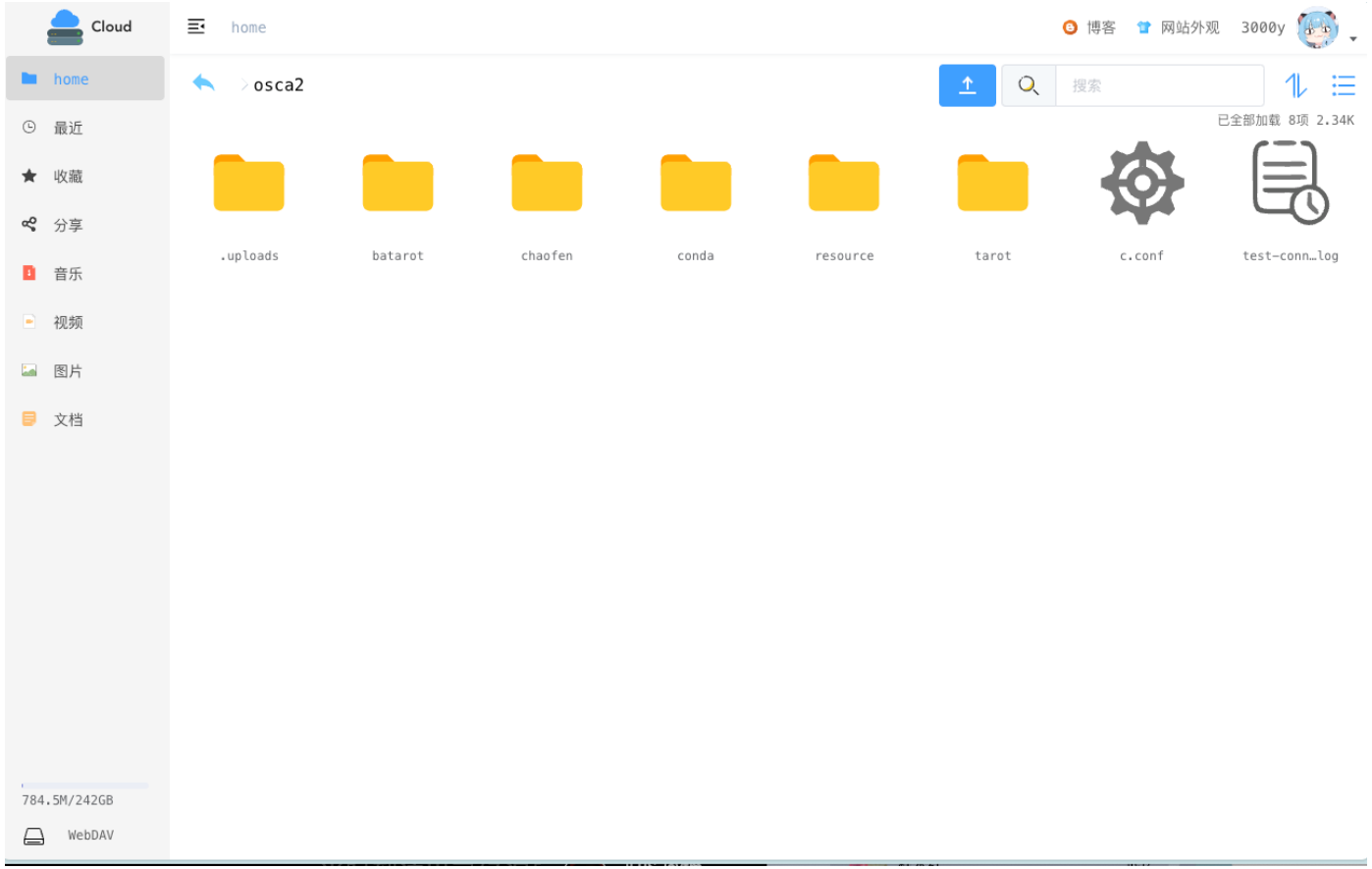


3.4 第三方客户端

第三方客户端使用开放API进行接入

- 用户可以在应用程序中调用开放API读写数据
- 提供SDK方便用户调用
- 支持桶对象管理
- 支持分块上传下载
- 支持断点续传
- ...

jmalcloud



PART FOUR

04

性　　能　　分　　析

Performance analysis

4.0 测试环境

以下测试全部在 c6420（64 CPU，128G RAM），CentOS 7 (Kernel 3.10.0) 系统进行（客户端）。

测试网关为OSCA-ocloud高能所数据中心双副本网关（ocloud.ihep.ac.cn:6100）
存储服务器为armeos01与armeos02。

- 元数据测试采用 **mdtest 4.0**，参数为-b 6 –z 4 –I 8
- 读写性能测试采用 **fio 3.7**，参数尽可能采用默认值，不针对系统、硬件等进行调优。各个测试任务中都加了refill_buffers 参数，让fio 生成的数据随机且没有规律，效果尽可能接近于实际业务场景中性能较差的情况

网络速率

	ocloud	c6420
Rec(MB/s)	1119	1118
Send(MB/s)	1120	1118

ocloud与c6420通过万兆交换机连接，传输速度达到**1G/s**以上。

硬盘速率(MB/s)：采用dd命令进行测试

		data01	data02	data03	data04	data05	data06
armeos 01存储服务器	write	262	151	151	151	214	267
	read	260	147	112	152	249	274
armeos 02存储服务器	write	242	151	151	266	249	154
	read	244	142	153	274	253	155

4.1 元数据性能(FUSE)

单位：OPS

对比项	JWanFS	Oceanstor pacific	EOS
目录创建	2578.072	496.717	90.410
目录信息	76812.309	229669.087	98.310
目录重命名	140.575	125.680	38.284
目录删除	5664.520	217.118	96.478
文件创建	2075.686	567.775	78.913
文件信息	51415.777	1508.539	99.657
文件读取	5706.691	1502.852	82.180
文件删除	5840.902	373.135	117.065
目录树创建	2435.896	414.418	101.798
目录树删除	5028.068	264.708	107.178

注：

- 华为oceanstor pacific 采用NFS协议挂载，硬件为SSD
- EOS系统元数据使用QuarkDB内存数据库，硬件为sata盘
- JWanFS系统硬件为sata存储盘+部分SSD缓存盘

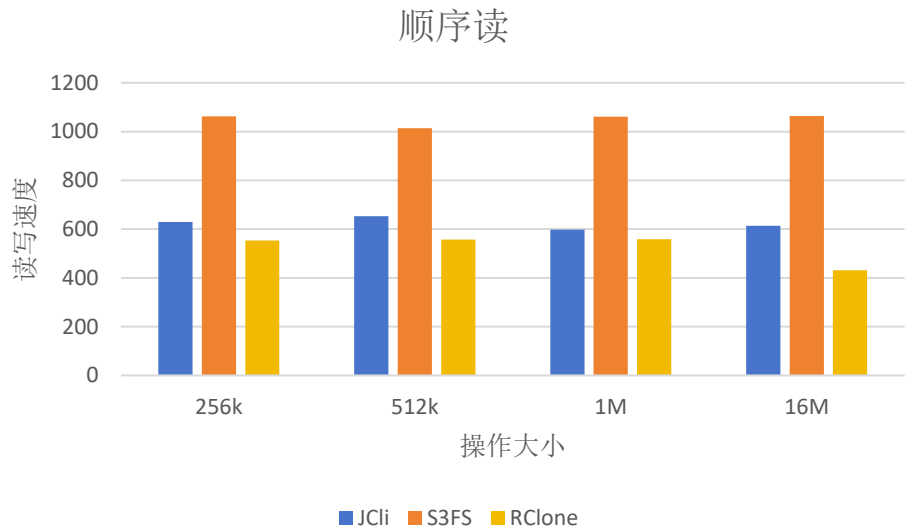
采用mdtest进行测试：

mdtest -d test -b 6 -z 4 -I 8

每层有6个目录，共4层，
每个目录有8个文件

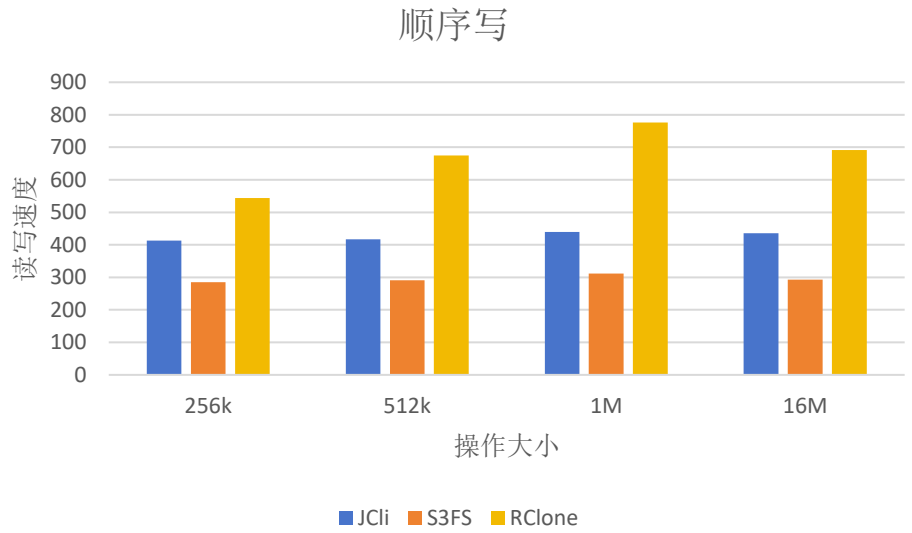
4.2 大文件读写性能

单位：MB/s　　注：操作大小为fio的bs参数，即数据块大小;文件size为4G



顺序读

操作大小	256k	512k	1M	16M
JCli	630	653	598	614
S3FS	**1063**	**1014**	**1061**	**1064**
RClone	553	557	558	431

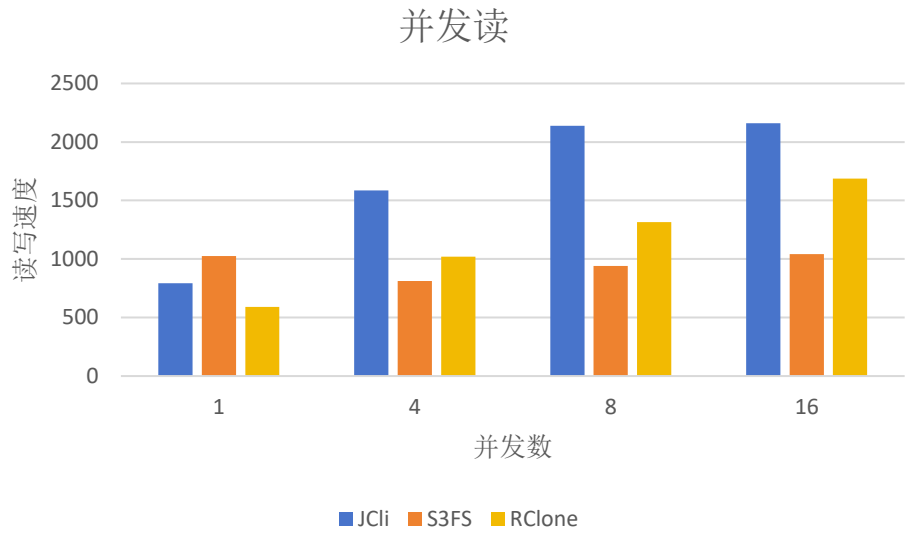


顺序写

操作大小	256k	512k	1M	16M
JCli	413	417	440	436
S3FS	285	291	312	293
RClone	**544**	**675**	**776**	**692**

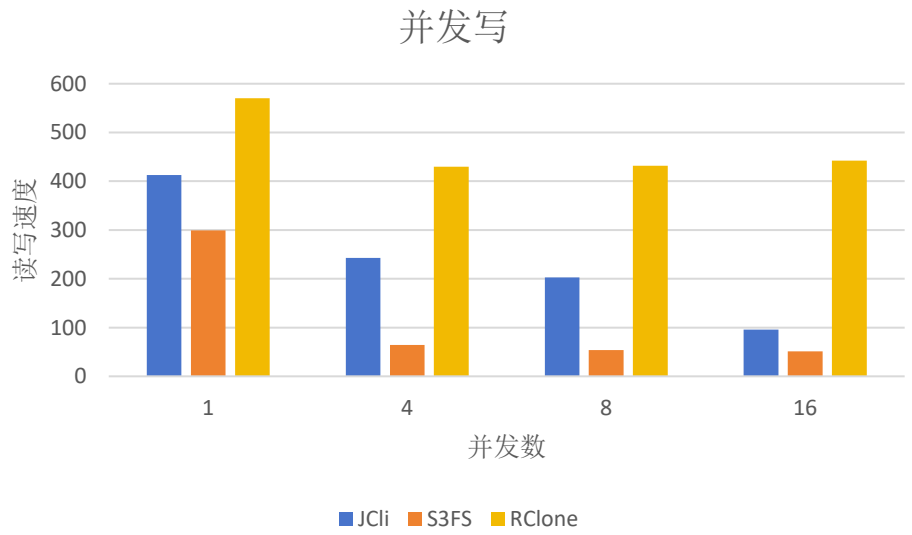
4.2 大文件读写性能

单位：MB/s



并发读

并发数	1	4	8	16
JCli	792	**1587**	**2139**	**2159**
S3FS	**1026**	813	940	1041
RClone	590	1021	1315	1687



并发写

并发数	1	4	8	16
JCli	413	243	203	96
S3FS	299	64	54	51
RClone	**570**	**430**	**432**	**442**

4.2 大文件读写性能

单位：MB/s

随机读

带内核缓存

操作大小	4k	16k	64k	256k
JCli	49.5	160	365	**494**
S3FS	**94.2**	**204**	**387**	462
RClone	22.6	91.1	155	269

去内核缓存

操作大小	4k	16k	64k	256k
JCli	**50.1**	**177**	**456**	**572**
S3FS	49.9	116	179	282
RClone	25.0	71.2	140	198

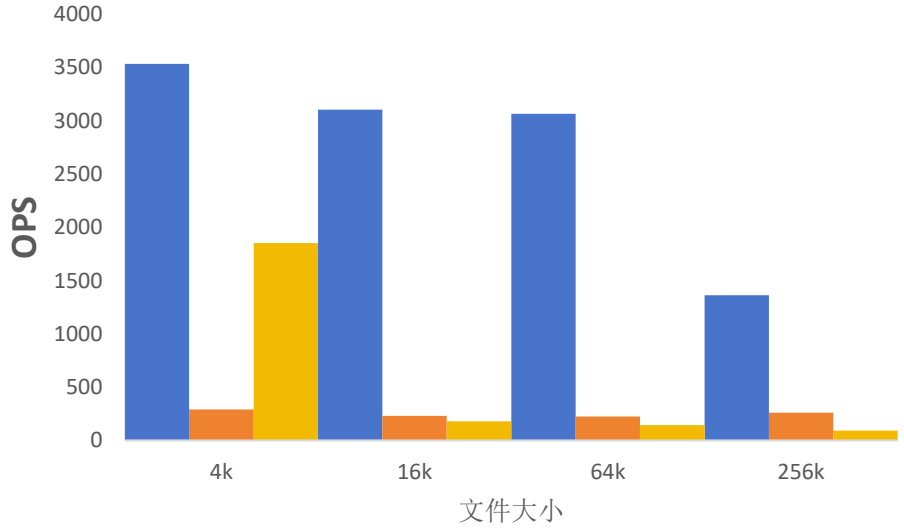
随机写

操作大小	4k	16k	64k	256k
JCli	0.18	1.87	1.09	1.95
S3FS	**3.54**	**16.3**	**90.3**	**221**
RClone	0.033	0.095	0.032	0.031

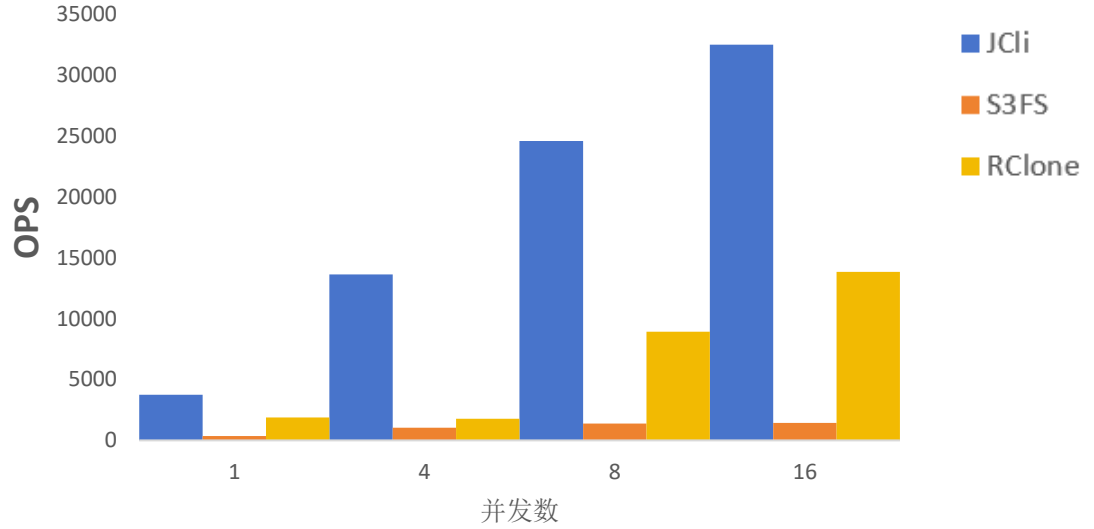
注：并发时测试文件大小为4K

4.3 小文件读写性能

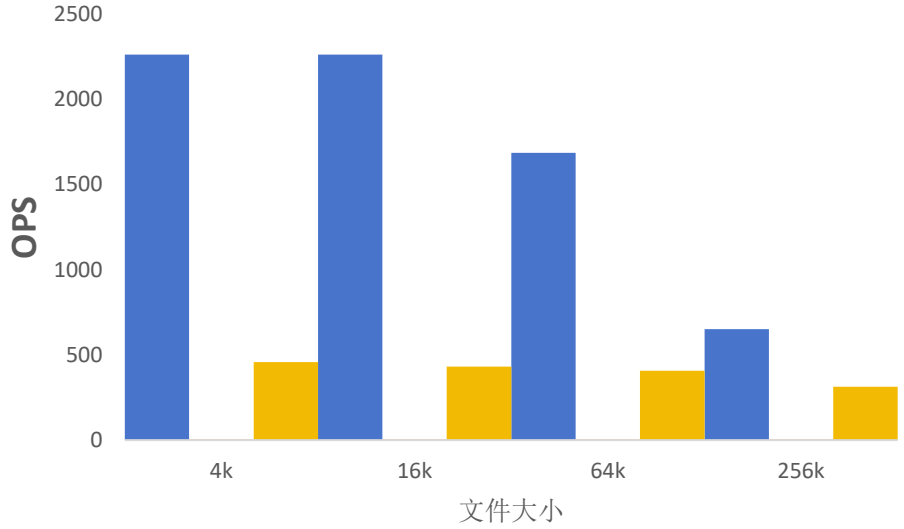
读



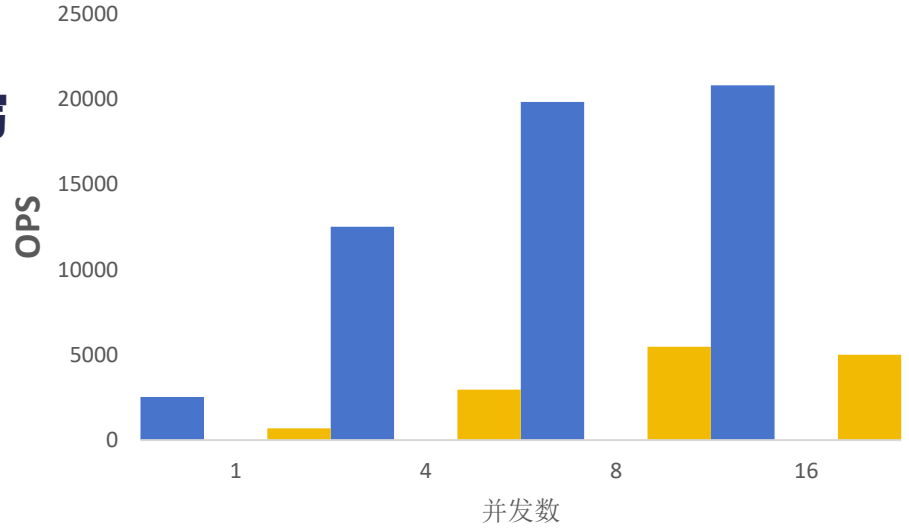
并发读



写



并发写



4.3 小文件读写性能

读

MB/s

文件大小	4k	16k	64k	256k
JCli	**14.5**	**50.9**	**201**	**358**
S3FS	1.1	3.7	14.7	68.5
RClone	7.5	2.9	9.3	24.2

OPS

文件大小	4k	16k	64k	256k
JCli	**3533**	**3105**	**3067**	**1364**
S3FS	289	231	224	261
RClone	1851	179	143	92

写

文件大小	4k	16k	64k	256k
JCli	**9.2**	**37.1**	**111**	**171**
S3FS	0.015	0.064	0.253	1.03
RClone	1.87	7.08	26.8	82.4

文件大小	4k	16k	64k	256k
JCli	**2262**	**2262**	**1686**	**652**
S3FS	4	4	3	3
RClone	458	432	408	314

4.3 小文件读写性能

并发小文件size：**4K**

并发读

MB/s

并发数	1	4	8	16
JCli	**15.3**	**53.2**	**96.1**	**127**
S3FS	1.4	4.08	5.42	5.63
RClone	7.4	6.9	34.9	54.1

OPS

并发数	1	4	8	16
JCli	**3745**	**13619**	**24601**	**32512**
S3FS	363	1044	1387	1441
RClone	1894	1766	8934	13849

并发写

并发数	1	4	8	16
JCli	**9.89**	**48.9**	**77.5**	**81.3**
S3FS	0.011	0.023	0.026	0.026
RClone	2.71	11.6	21.4	19.6

并发数	1	4	8	16
JCli	**2531**	**12518**	**19840**	**20812**
S3FS	2	5	6	6
RClone	693	2969	5478	5017

4.4 服务端性能指标

- 测试方案：本地挂载后使用FIO对文件系统进行测试
- 测试结果显示可以跑满整个网络带宽（测试数据为万兆）

	写入平均速度（MiB/s）				读取平均速度（MiB/s）			
并发数 \ 文件大小	1	10	50	100	1	10	50	100
4K	6.88	5.91	5.67	5.72	275	449	326	270
10M	1082	1066	1059	991	407	1035	1093	1100
100M	969	1016	1007	1008	597	1033	1102	1110

OSCA到成都超算测试：跑满带宽

```
-dsk/total- -net/total- ---paging-- ---system-- ------memory-usage----- ----swap---
 read  writ  recv  send  in   out  int  csw  used  buff  cach  free  used  free
  48k  260k 6796M 3489M  32k   20k 253k 148k 48.1G 7164k 63.4G 14.1G 4133M   60G
  92k  476k 6965M 3657M  76k   68k 422k 150k 47.8G 7164k 63.3G 14.6G 4133M   60G
 536k 1164k 6629M 3321M 264k   24k 302k 143k 48.1G 7164k 63.3G 14.3G 4133M   60G
  68k  432k 7179M 3700M  68k    0  223k 128k 48.1G 7164k 63.3G 14.3G 4133M   60G
  32k 7812k 7278M 3695M  32k    0  258k 144k 48.1G 7164k 63.3G 14.3G 4133M   60G
  76k  468k 6513M 3368M  76k   40k 312k 135k 48.1G 7164k 63.2G 14.3G 4133M   60G
 104k  188k 7237M 3746M 104k   28k 374k 138k 48.1G 7164k 63.2G 14.3G 4133M   60G
 336k   28M 7388M 3932M 336k    0  377k 156k 47.9G 7164k 63.2G 14.6G 4133M   60G
    0 6224k 5502M 2628M   0  108k 443k 143k 48.3G 7164k 63.0G 14.4G 4133M   60G
    0 1372k 7369M 3867M   0    0  229k 143k 48.1G 7164k 63.0G 14.5G 4133M   60G
    0  264k 6749M 3433M   0    0  259k 154k 48.2G 7164k 63.0G 14.4G 4133M   60G
  32k  564k 7108M 3685M  32k 8192B 397k 138k 48.0G 7164k 63.0G 14.6G 4133M   60G
 628k   10M 8106M 4269M 628k    0  294k 140k 47.6G 7164k 63.0G 15.0G 4133M   60G
 620k  296k 5992M 3004M 604k    0  291k 146k 48.2G 7164k 63.0G 14.4G 4133M   60G
```

PART FIVE

05

Ocloud——JWanFS在OSCA上的应用实例

Example of JWanFS in OSCA

5.1 Ocloud——JWanFS实践

项目背景

开放科学计算联盟（英文简称：OSCA）由国内高校、科研院所等科学计算资源拥有者、服务提供者、应用方及其他开放计算平台组成。

OSCA联盟云存储项目由OSCA联盟组织各成员共同参与贡献资源，由中国科学院高能物理研究所计算中心负责运营维护，提供公益性质的专用云存储服务，为高校和研究机构提供统一的跨站点数据管理和访问，从而推动科学研究和技术创新。

用户需求

海量数据存储

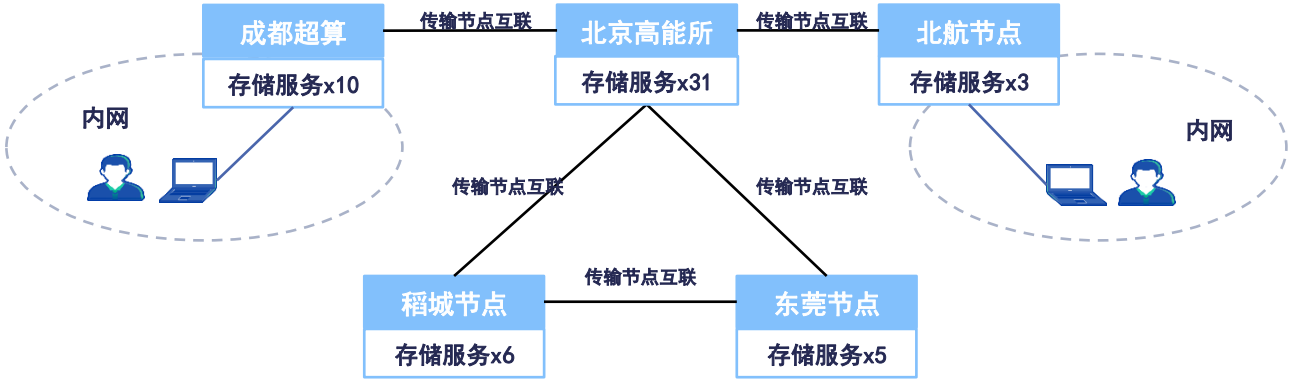
需要对OSCA联盟全部机构提供存储服务，存储海量数据，但部分存储节点可用存储容量不足，服务节点位于全国各地，需要打通全部服务节点，进行安全跨地域传输数据。

跨地域数据流转

服务节点位于全国各地，部分节点仅提供单一链路单一端口访问，部分节点无公网地址，但又希望作为存储节点提供其他数据中心访问。且数据需要在各节点间进行流转，在其中一个节点写入数据其他异地节点可以立刻读取到数据。

解决方案

本次配置5节点JWanFS服务，分别部署于北京、东莞、稻城、成都等地，提供高性能跨异地广域网存储，可用容量5PB。



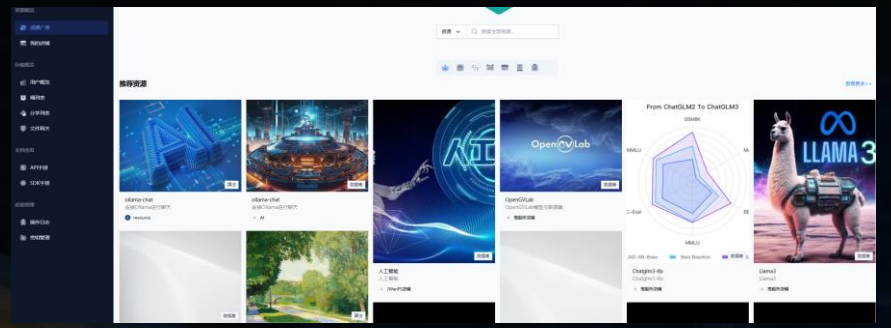
客户端价值

- ✓ **安全数据传输**：JWanFS各节点间通过传输节点互联，数据在公网间传输只需要开放各传输节点，且传输时采用TLS加密，保障数据安全。
- ✓ **打破存储极限**：JWanFS存储架构将各节点存储卷组合为大存储池，支持横向扩展，当一个节点存储容量不足时会自动调度其他节点的存储卷，打破存储极限，轻松存储海量数据。
- ✓ **异地数据实时同步**：JWanFS优秀的读写架构允许跨地域读写数据实时同步，当其中一节点写入数据时，另一异地节点可以立即读取，无需等待异步复制同步。

5.2 Ocloud

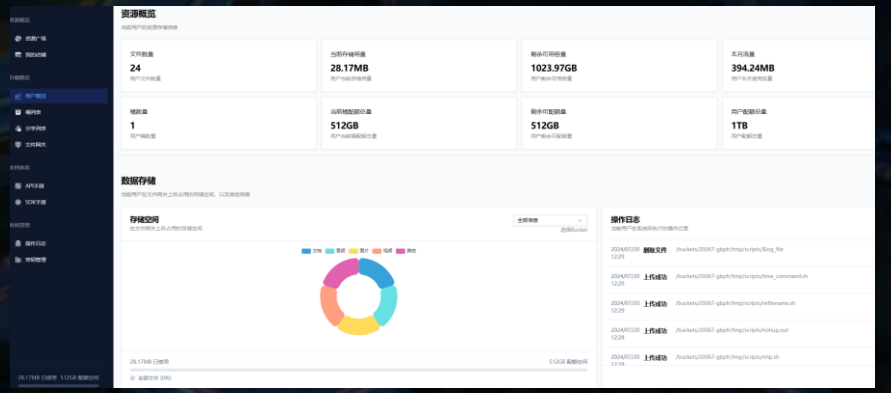
OSCA联盟云网址
https://ocloud.ihep.ac.cn

各类资源



Carsi登录无需审核！
国内各高校账号直接登录使用！

方便快捷的文件管理



Ocloud欢迎大家体验使用！
欢迎各联盟成员一起为云存储贡献资源

5.3 总结

- 支持海量存储、实时读写、高可靠性、高安全性，解决了广域网分布式云存储的部分难题
- 多数据中心、多平台、多协议、多架构
- 优秀的元数据性能
- Jcli客户端优秀的小文件性能、高并发性
- Ocloud联盟云项目

JWanFS未来要走开源路线，诚挚邀请大家一起贡献代码

Ocloud是联盟的一个平台，欢迎大家贡献资源、软件以及算法

欢迎大家的合作及对我们项目提出的任何建议！

感谢！

Thank You

高能所計算中心
IHEP Computing Center