ceph

Ceph中国社区 | IT大咖说 知识共享平台

CEPHALOCON APAC 2018
THE FUTURE OF STORAGE
22-23 March 2018 | BEIJING

统一存储新边界-UMStor的创新

最具CBA时代气息的存储系统

朱荣泽 - UMCloud

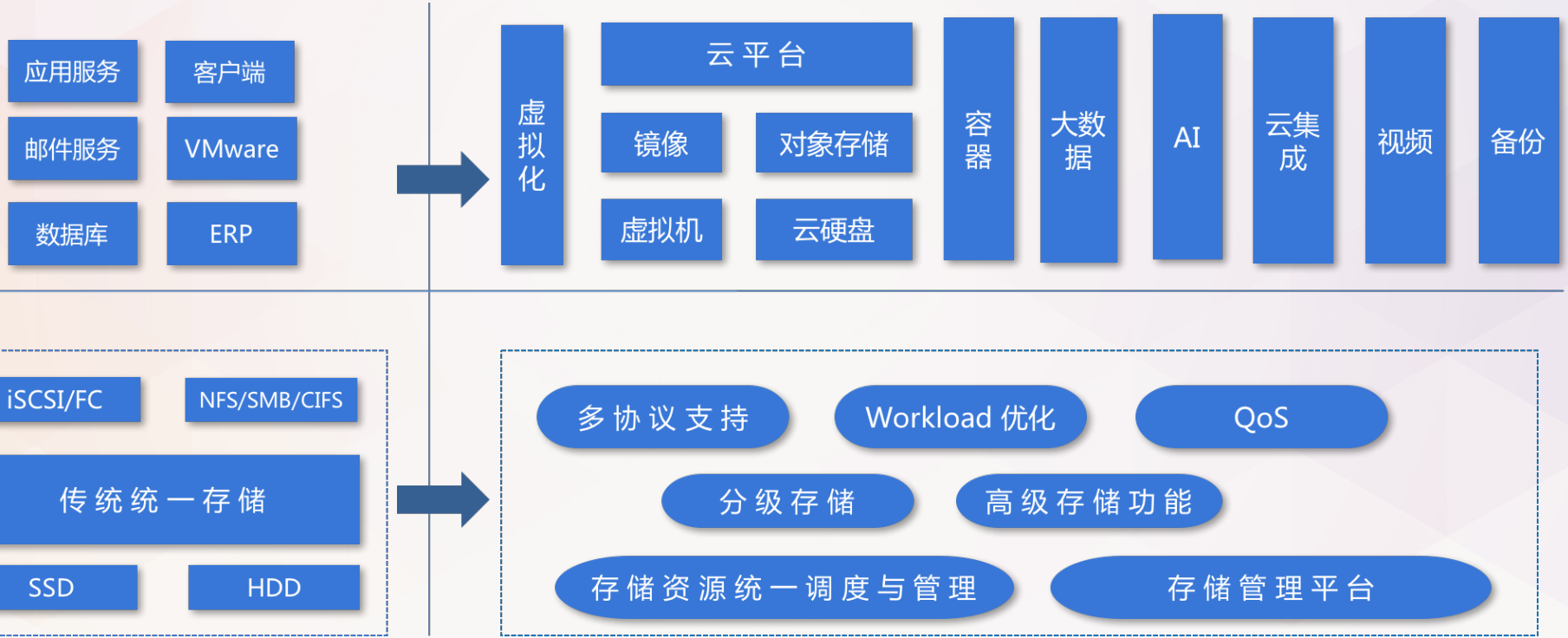
2018.03.23

OUTLINE

1. CBA时代对存储的要求
2. 数据冰山的挑战
3. 统一存储的新边界
4. UMStor产品架构
5. 超大规模部署 – 30PB云存储案例
6. 满足复杂场景 – 20PB数据湖案例
   I. 多协议支持
   II. 对象存储高级功能
   III. 新的大数据存储
   IV. 融合数据湖
7. AI计算引擎下沉

1. The Storage Requirements in the CBA Era
2. Data Iceberg Challenge
3. The New Boundary for Unified Storage
4. UMStor Product Architecture
5. Very Large Scale Deployment – 30PB Object Storage Case
6. Meet Complex Scenes – 20PB Data Lake Case
   I. Multi-Protocol Support
   II. Object Storage Advanced Features
   III. Another Big Data Storage
   IV. Fusion Data Lake
7. AI Compute Engine

CBA时代对存储的要求



数据冰山的挑战

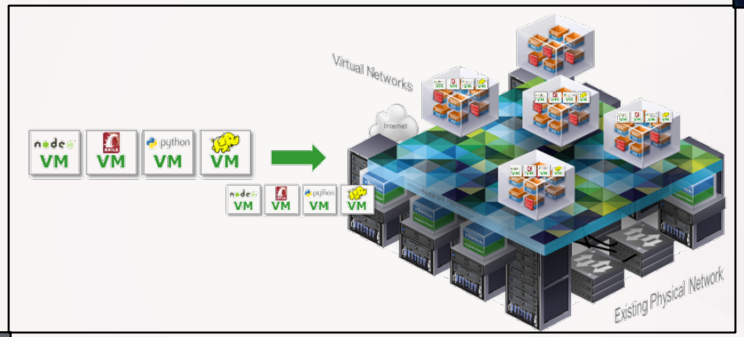
主存储
关键任务类应用

二级存储
云计算
人工智能
数据分析
数据归档
数据备份
文件共享

统一存储的新边界

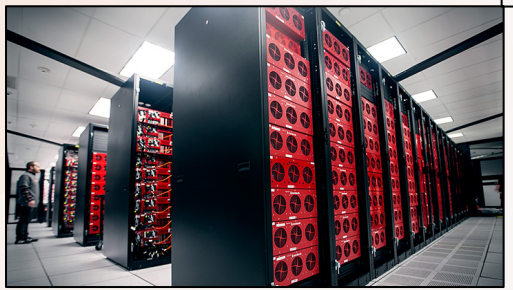
业务边界

利用AI技术，为业务赋能。



复杂边界

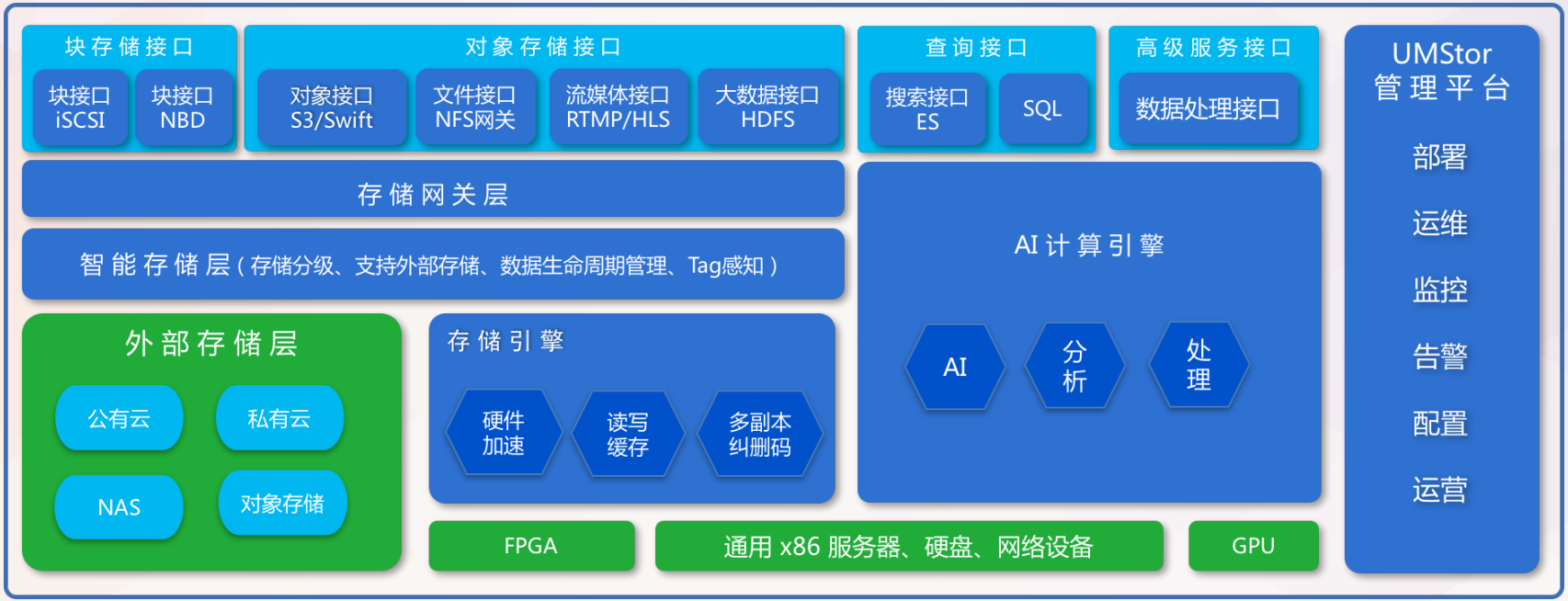
在CBA时代满足多种应用场景负载，轻松解解决数据挑战。



规模边界

10PB级存储规模成为常态。

UMStor产品架构


块存储接口
块接口 iSCSI
块接口 NBD
对象存储接口
对象接口 S3/Swift
文件接口 NFS网关
流媒体接口 RTMP/HLS
大数据接口 HDFS
查询接口
搜索接口 ES
SQL
高级服务接口
数据处理接口
UMStor 管理平台
部署
运维
监控
告警
配置
运营
存储网关层
智能存储层（存储分级、支持外部存储、数据生命周期管理、Tag感知）
AI计算引擎
AI
分析
处理
外部存储层
公有云
私有云
NAS
对象存储
存储引擎
硬件加速
读写缓存
多副本纠删码
FPGA
通用 x86 服务器、硬盘、网络设备
GPU

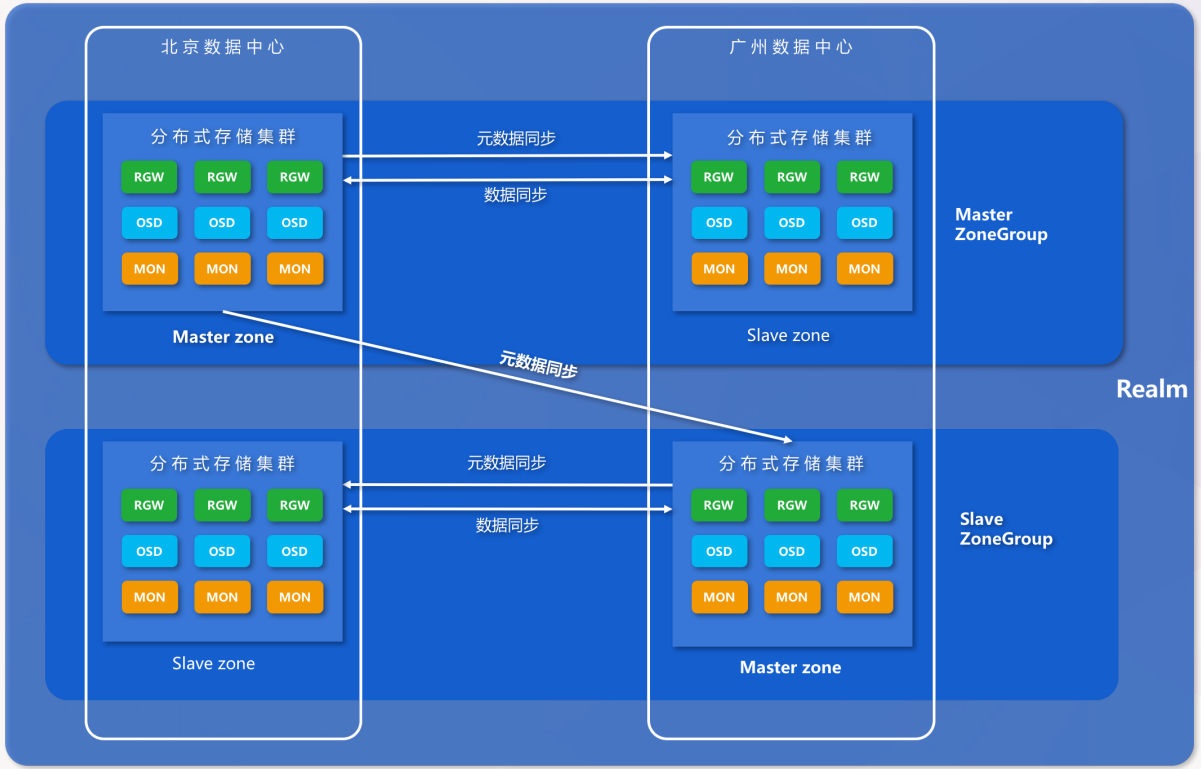
超大规模部署 – 30PB云存储案例

规模：

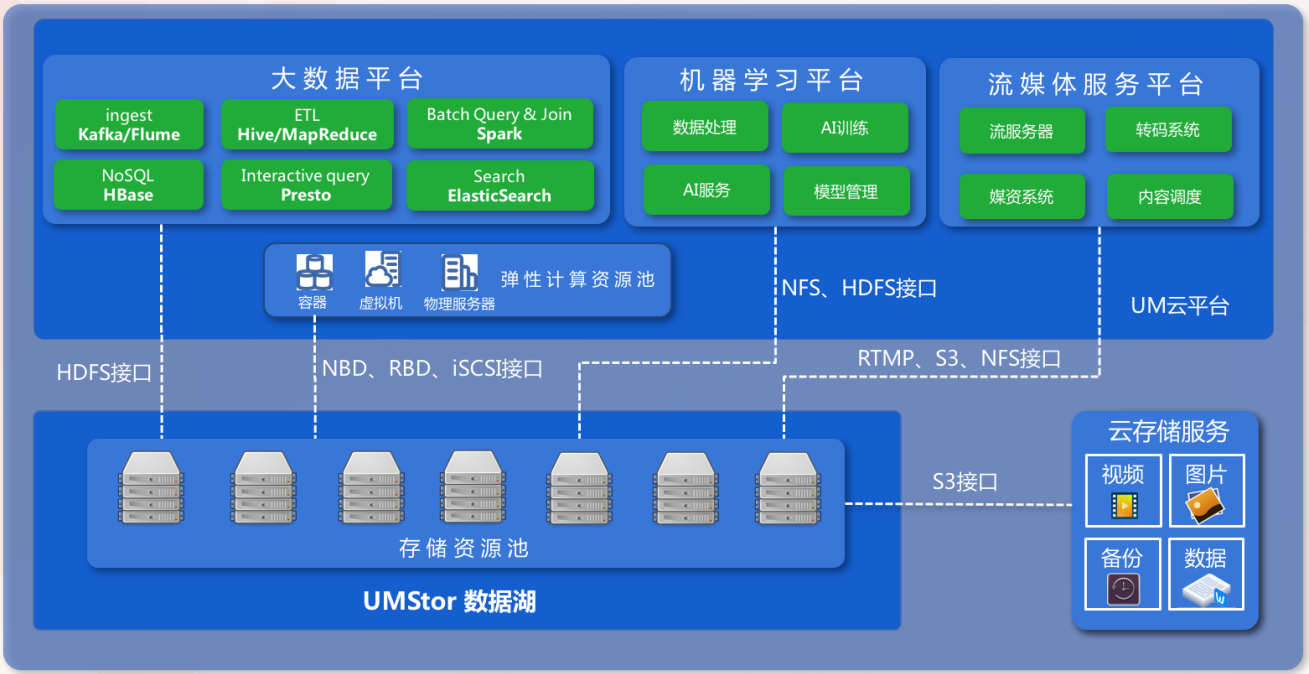
- 两地4个集群共520台服务器。
- 每个集群2376个OSD。
- HDD+SSD。
- 上线1年多。

如何大规模部署：

- 整体架构设计
- 多站点设计
- 网络架构规划
- 节点规划
- 负载均衡设计
- 硬件配比配置
- 软件配置
- CRUSH MAP设计
- 扩容方案
- 自动化交付



满足复杂场景 – 20PB数据湖案例



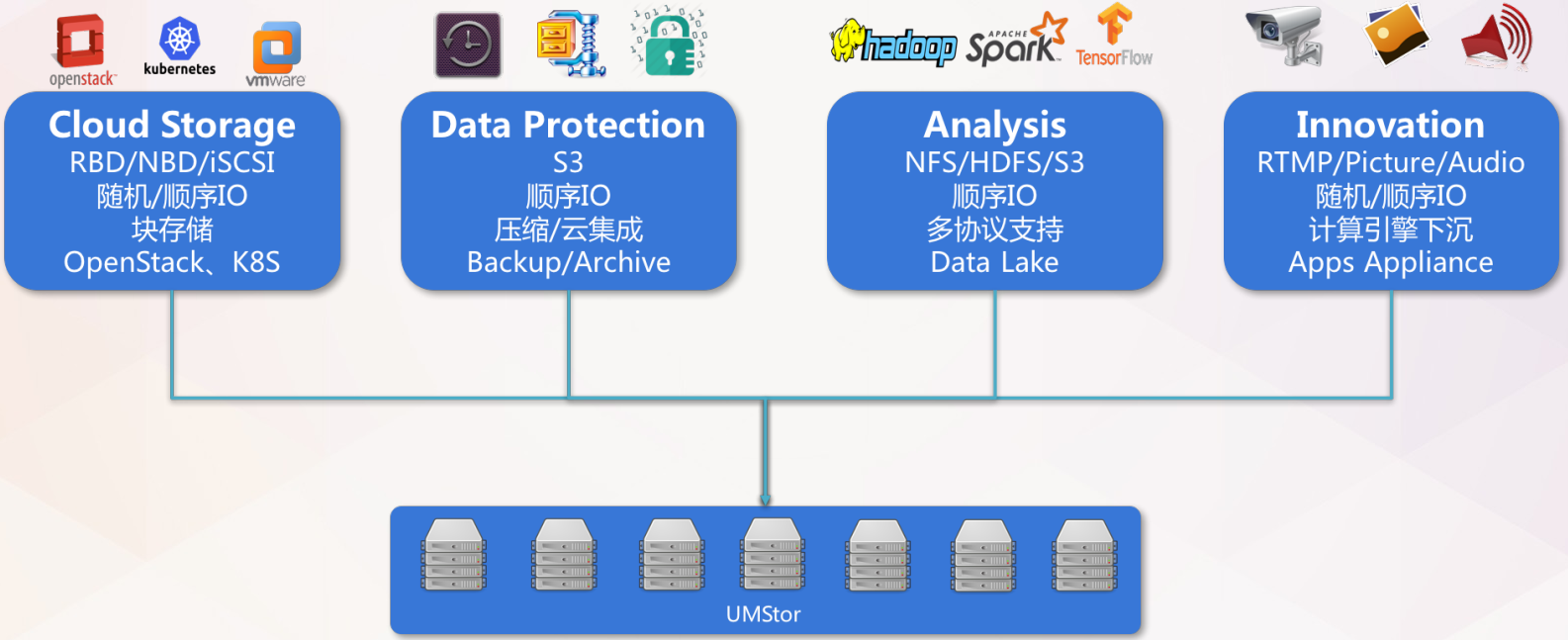
挑战：

- IT资源池两期共4000台服务器，用于虚拟机、裸金属服务器、分布式存储。
- 分布式存储需20PB，同时支持虚拟化平台、大数据业务、媒资业务，需支持多种数据接口。
- 需同时支持4000台虚拟机。

UMStor方案：

- **一站式方式**，提供块存储、对象存储、大数据接口、文件接口。
- **高性能**，同时满足4000台虚拟机的读写负载，还有大数据和媒资业务的负载。
- **弹性**，快速部署，快速扩容。

多协议支持


openstack
kubernetes
vmware
Cloud Storage
RBD/NBD/iSCSI
随机/顺序IO
块存储
OpenStack、K8S
Data Protection
S3
顺序IO
压缩/云集成
Backup/Archive
hadoop
APACHE Spark
TensorFlow
Analysis
NFS/HDFS/S3
顺序IO
多协议支持
Data Lake
Innovation
RTMP/Picture/Audio
随机/顺序IO
计算引擎下沉
Apps Appliance
UMStor

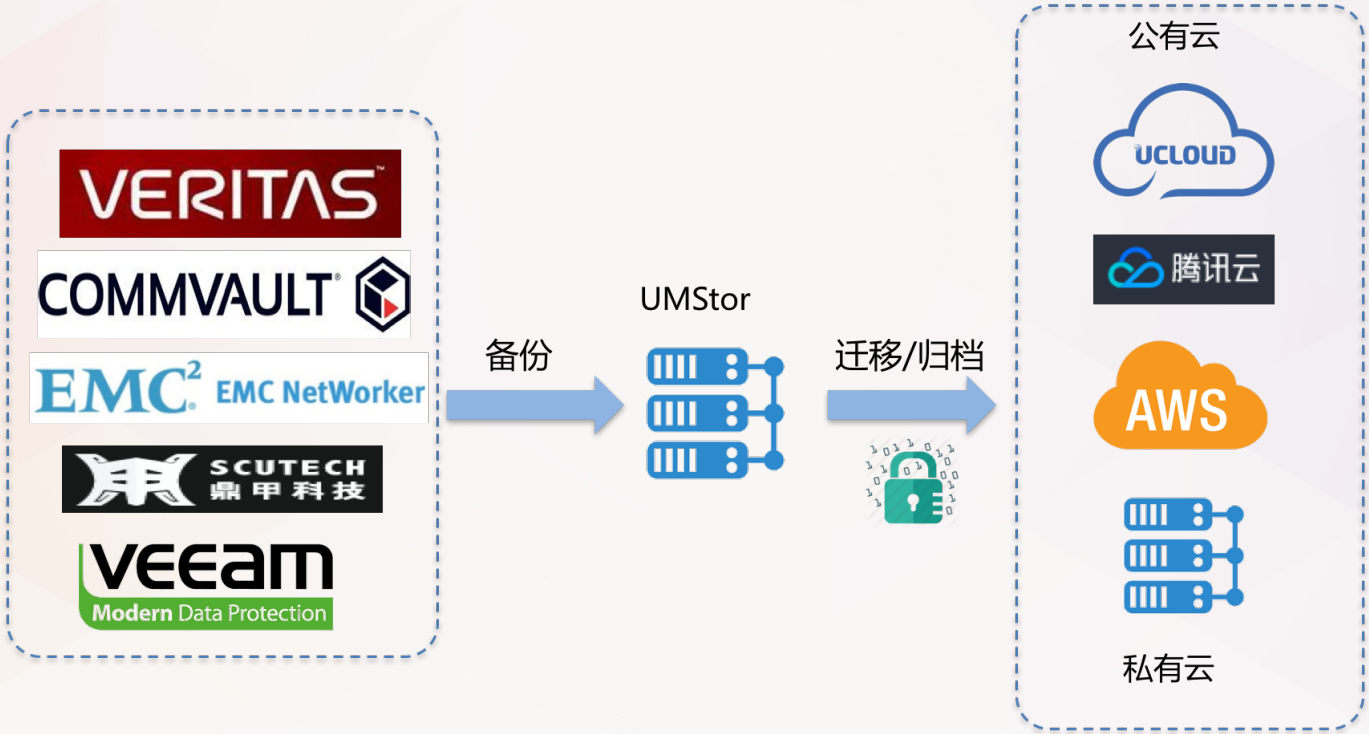
对象存储高级功能

ceph

Multi-Site 多站点多活 → Cloud Sync 云同步 → **Object Tagging ElasticSearch 自定义多标签+搜索** → **Cloud Tier 云集成** → **Storage Classes 存储类别**

Time	AWS S3 Key Feature	Time	AWS S3 Key Feature	Time	AWS S3 Key Feature
2006/3/13	Introducing Amazon S3	2010/7/14	Event notification(SNS notice RRS objec	2012/11/13	**Archiving Data to Glacier**
2006/6/14	Hierarchical listing of keys	2010/9/2	Support AWS Identity and Access Mana	2012/12/27	Lifecycle Management(versioning object su
2006/6/14	Virtual Hosting of buckets	2010/11/10	Upload via Multipart	2014/1/30	Signature Version 4
2006/10/2	**Bucket Logging**	2011/1/14	Support setting Get response header	2014/6/12	Server-side encryption(customer-provied e
2007/5/18	Signed requests for virtual hosted	2011/2/17	Static Website Hosting	2014/11/12	Server-side encryption support AWS Key M
2007/11/6	Support location constraints	2011/6/21	Server-side Copy support Multipart	2014/11/13	**Event Notification(SNS,SQS,AWS Lambd**
2007/12/18	Support DevPay	2011/8/3	Temporary token from AWS IAM	2015/3/24	**Cross-region replication**
2007/12/17	Support Upload via POST	2011/10/4	Server-side encryption	2015/7/28	Amazon CloudWatch Integration
2008/5/2	Server-side Copy	2011/12/7	Multi-Object Delete via single request	2015/9/1	AWS CloudTrail Integration
2008/12/31	Requester Pays bucket	2011/12/27	Object Expiration	2015/9/16	**Storage Classes(STANDARD_IA)**
2010/3/16	Object Versioning	2012/7/10	Multi-Factor Authentication(MFA)-prot	2016/4/19	Transfer Acceleration
2010/5/18	**Storage Classes(Reduced Redun**	2012/8/21	Cost Allocation Tagging	2016/11/29	**Object Tagging**
2010/7/6	Bucket policy	2012/8/31	Cross Origin Resource Sharing(CORS)	2016/11/29	Inventory
2010/6/9	Support AWS Management Conso	2012/10/4	Static Website Hosting(root domain sup	2016/11/29	Storage Class Analysis

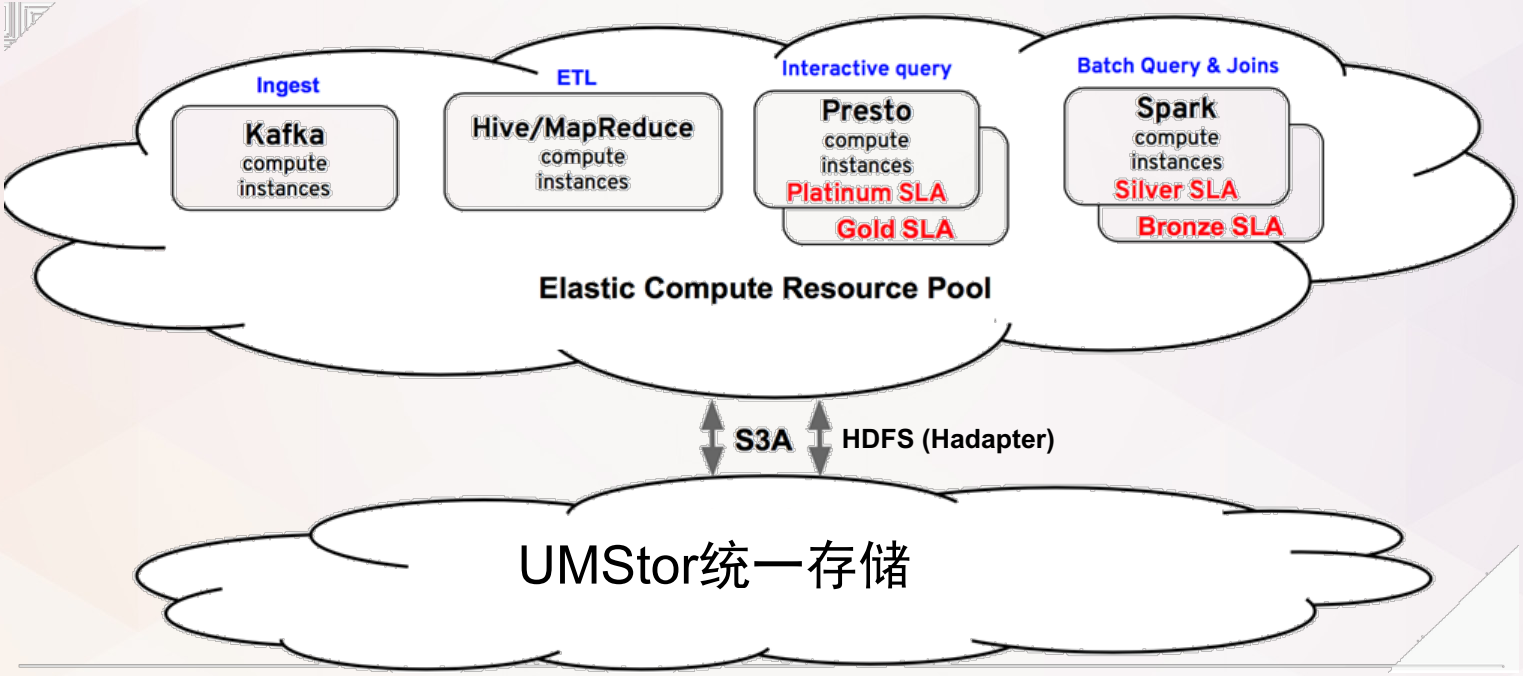
云集成 Cloud Tier



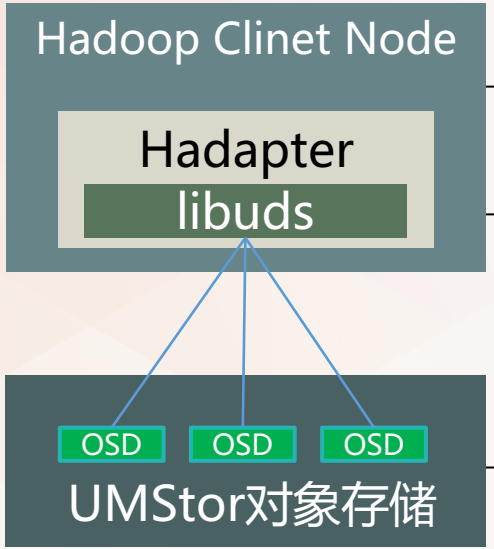
存储类别 Storage Classes

	标准存储	低频存储	归档存储
存储池介质	SSD/SATA	SATA	SATA
数据保护	三副本/纠删码	纠删码	纠删码
压缩	不使用	不使用	使用
延迟	低	一般	高
访问频度	高	低频	非常少
多存储池	支持	支持	支持
得盘率	30%	75%	75%~150%

新的大数据存储


Ingest
Kafka
compute
instances
ETL
Hive/MapReduce
compute
instances
Interactive query
Presto
compute
instances
Platinum SLA
Gold SLA
Batch Query & Joins
Spark
compute
instances
Silver SLA
Bronze SLA
Elastic Compute Resource Pool
S3A
HDFS (Hadapter)
UMStor统一存储

UMStor 大数据存储接口



1、在Hadoop应用节点上安装Hadapter jar包，不需要修改应用程序。
2、Hadoop fs client 发起读写请求，使用 url： uds://localhost/bucket1/test_obj 。
3、Hadapter 把Hadoop fs client的请求转给 libuds。

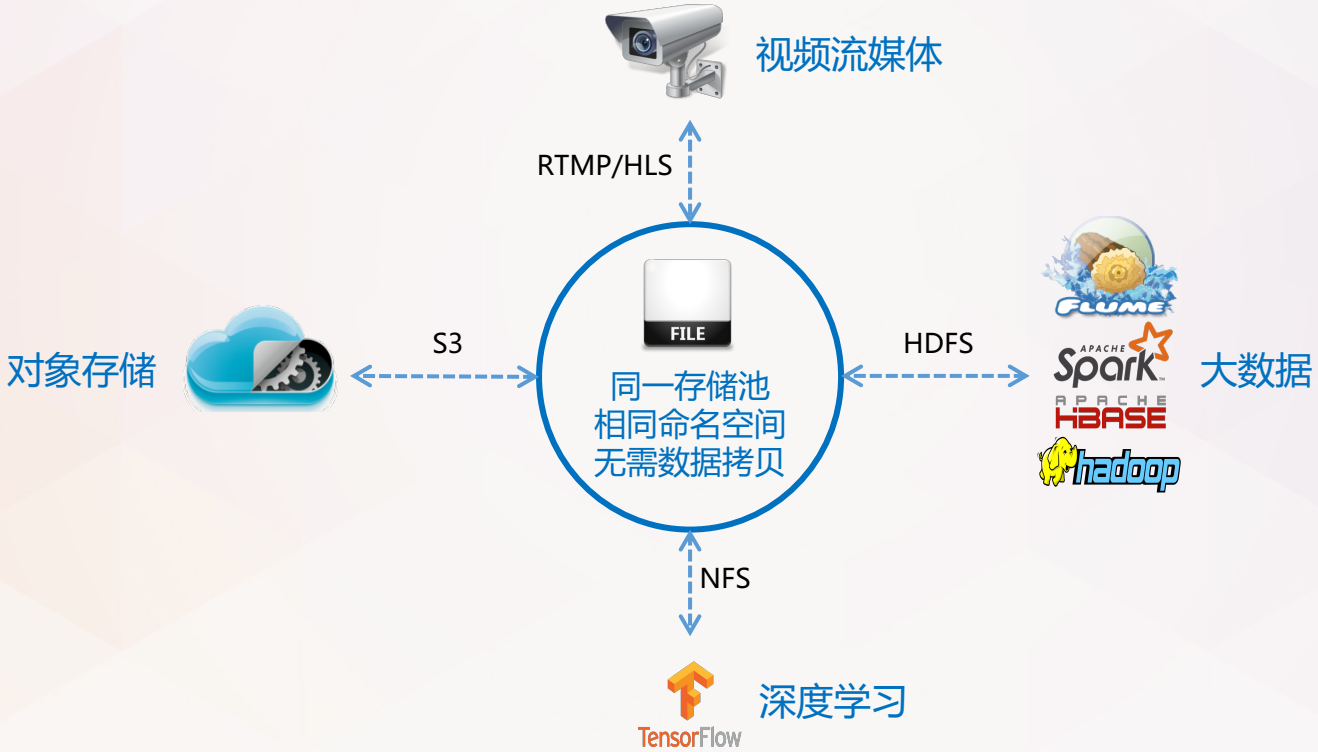
4、Libuds是UMStor高性能原生对象存储SDK，可以直接把请求转成“对象存储协议”请求，并发给对应的OSD，不需要绕过“对象存储网关”，可以打满存储服务器的网络带宽。

5、UMStor存储引擎收到请求，并进行处理。

UMStor 大数据存储

	HDFS	UMStor+Hadapter	UMStor+S3A
读写吞吐	高	高	低
MapReduce测试	3m2.40s	3m35.84s	6m10.69s
数据保护	三副本/纠删码(新)	三副本/纠删码	三副本/纠删码
数据分布 (海量文件)	NameNode	CRUSH	CRUSH
弹性	低	高	高
存储功能	少	多	多
规模与扩展性	一般	高	高

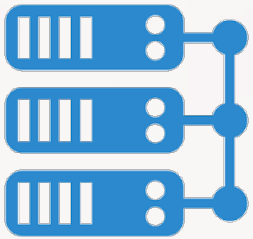
融合数据湖 – 解决数据孤岛



AI计算引擎下沉

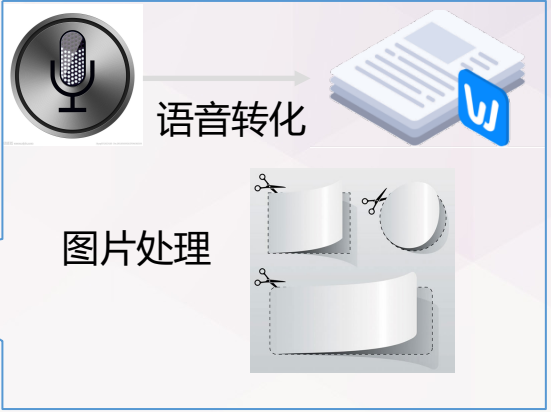
客户端服务器 客户端服务器 客户端服务器

Search Engine



UMStor

Apps offload Engine



核心客户案例

ceph

中国移动 China Mobile

咪咕

UnionPay 银联 中国银联 China UnionPay

百联集团 BAILIAN GROUP

北京金融资产交易所 BEIJING FINANCIAL ASSETS EXCHANGE

深圳证券通信有限公司 SHENZHEN SECURITIES COMMUNICATION CO.,LTD.

華林證券 CHINALIN SECURITIES

铜牛信息 TOPNEW INFO

中国邮政储蓄银行 POSTAL SAVINGS BANK OF CHINA

深圳航空 Shenzhen Airlines

中国盐业总公司 China National Salt Industry Corporation

国家电网 STATE GRID

国网信通北京国电通网络技术有限公司 BEIJING GUODIANTONG NETWORK TECHNOLOGY COMPANY LTD. STATE GRID INFORMATION & TELECOMMUNICATION CORPORATION

CHEMCHINA 中国化工集团电商平台

华大基因 BGI SHENZHEN

中国体育彩票 CHINA SPORTS LOTTERY