DTCC 2010-2022

第十三届中国数据库技术大会

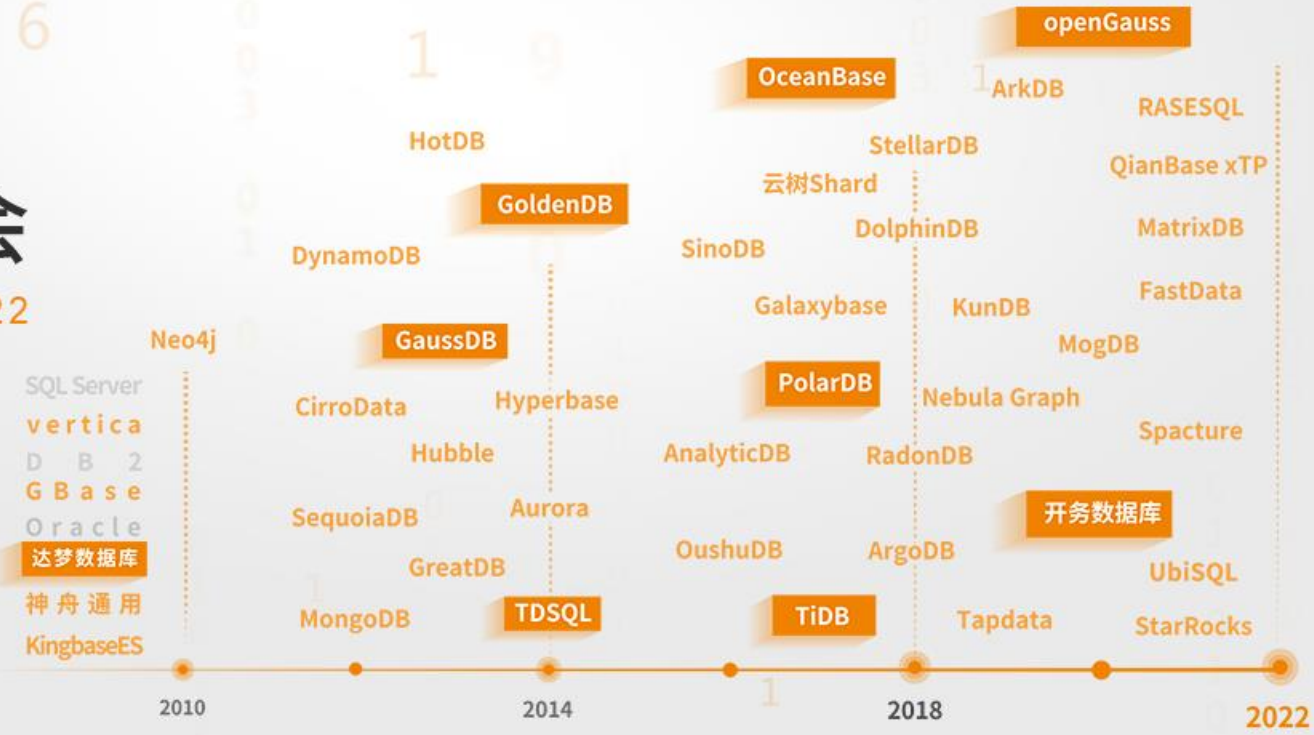
DATABASE TECHNOLOGY CONFERENCE CHINA 2022

数据智能 价值创新

IT168.com ChinaUnix.net ITPUB

线上直播 | 2022/12/14-16

数据来源：数据库产品上市商用时间



探究企业级数据存储高可靠与高效的实现方法
-数据与存储技术

成思敏
天翼数字生活科技有限公司

目录/CONTENTS

目录/CONTENTS

1、现代企业级存储综述-数据存储解析

数据存储就是围绕数据存、取、算三件事的一系列解决方案集合

存储

基本释义 详细释义

- 把钱或物等积存起来。
- 指积存的钱或物等。
- 把信息记录在电子设备(计算机)内,需要时可将资料从中取出。

数据属性

- A（attributable)-可溯源
- L（legible)-清晰
- C（contemporaneous)-同步
- O（original or true copy)-原始或真实复制
- A（accurate)-准确

数据存储

就是根据不同的应用环境通过采取合理、安全、有效的方式将数据保存到某些介质上并能保证有效的访问

包含两个方面的含义：它是数据临时或长期驻留的物理媒介；是保证数据完整安全存放的方式或行为。存储就是把这两个方面结合起来，向客户提供一套数据存放解决方案。

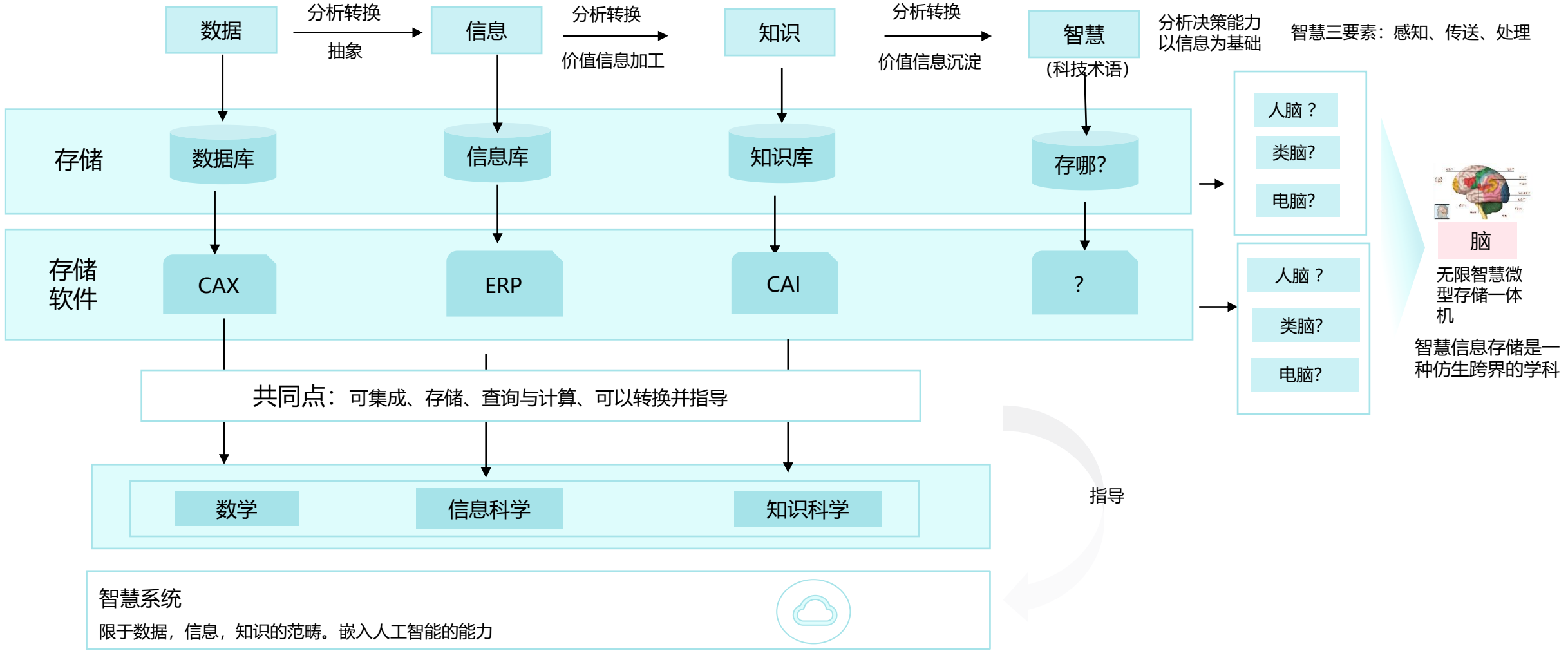
数据存储是一套方法与工具的方案集

数据存储内容

数据创建 | 数据冗余 | 数据分配 | 数据维护 | 数据读取 | 数据重构 | 数据服务

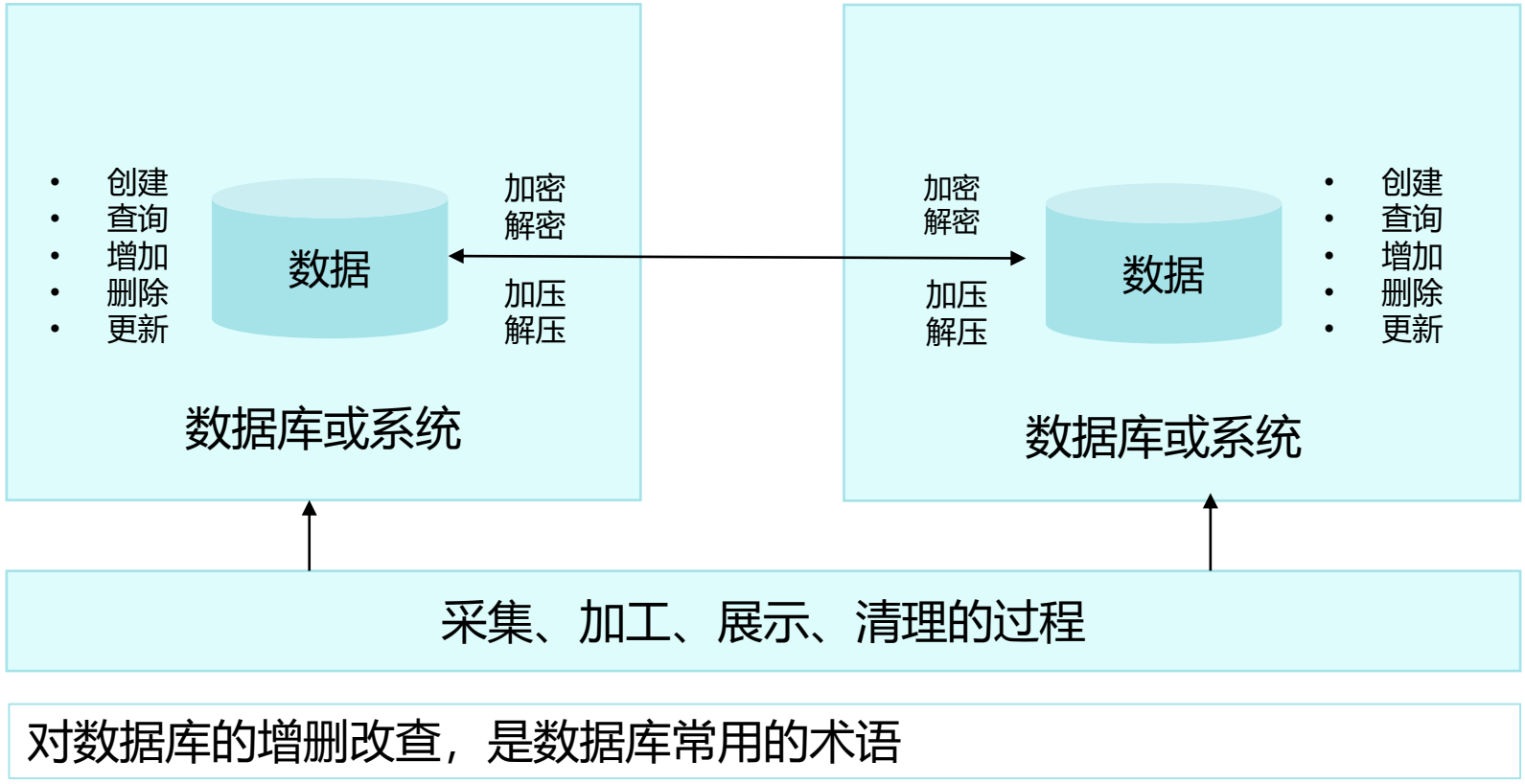
1、现代企业级存储综述-数据关联性与存储

数据存储为信息、知识、智慧提供基本获取性能力



1、现代企业级存储综述-对数据的操作

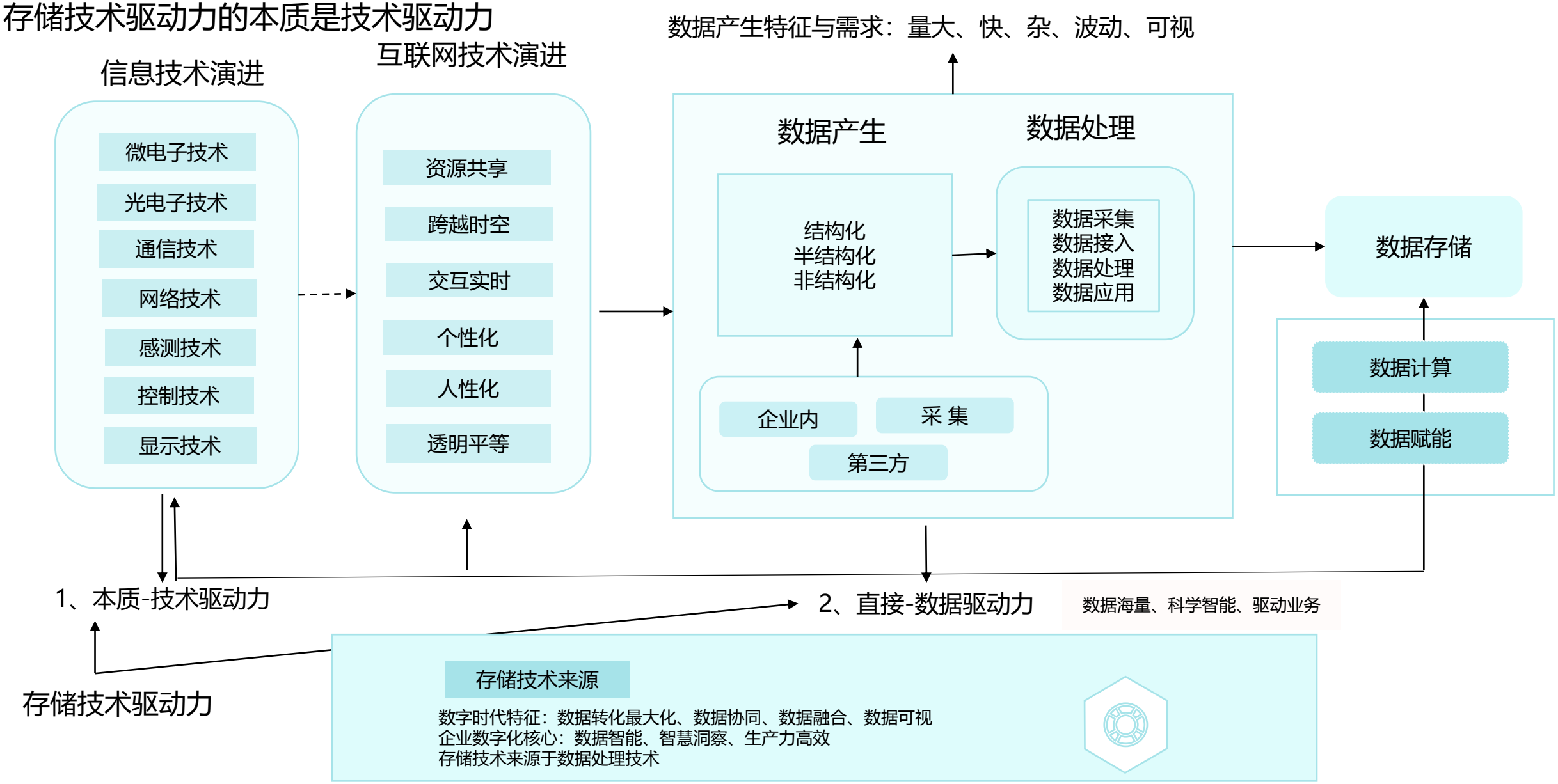
对数据操作具有动态性，属于数据加工的集合体



对数据库的增删改查，是数据库常用的术语

1、现代企业级存储综述-存储技术驱动力

存储技术驱动力的本质是技术驱动力



1、现代企业级存储综述-市场中的存储

政 策

数字技术 / 实体经济 → 融合 → 赋能 → 传统产业升级

新型生产要素 → 数据 → 存储技术创新

政策层面的支持是市场繁荣的前提

创新方向

- 存储芯片
- 软件定义存储
- 海量异构存储
- 智能存储
- 存储安全

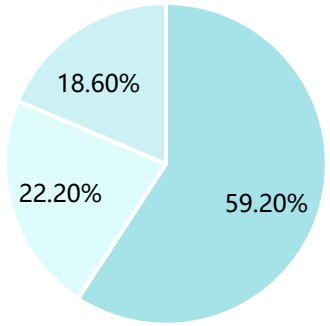
数据中心存储配套组件

- 高密度硬盘
- 混合硬盘
- 固态硬盘
- 冷数据存储
- 分层存储
- 集装箱储存
- 虚拟存储
- 高密度磁带存储
- 存储系统管理软件

云-存储产品

- 简单存储
- 块存储
- 网络存储
- 档案存储
- 并行文件存储
- 存储网关
- 灾难恢复方案

2020年Q1-Q3企业级存储市场份额

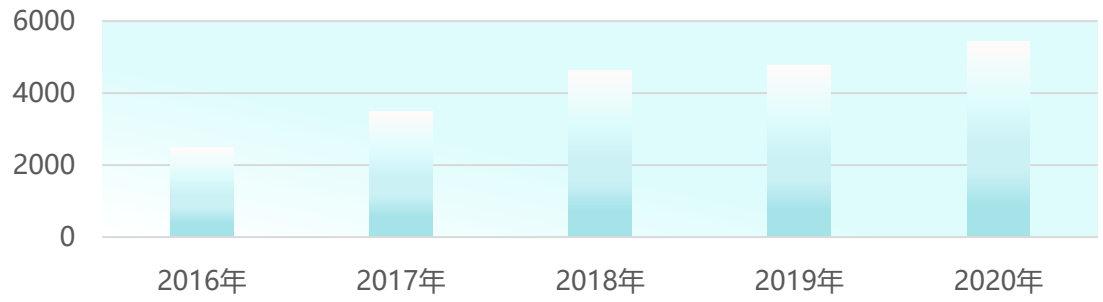


传统存储TESS 软件定义存储SDS 超融合基础架构HCI

来源：IDC

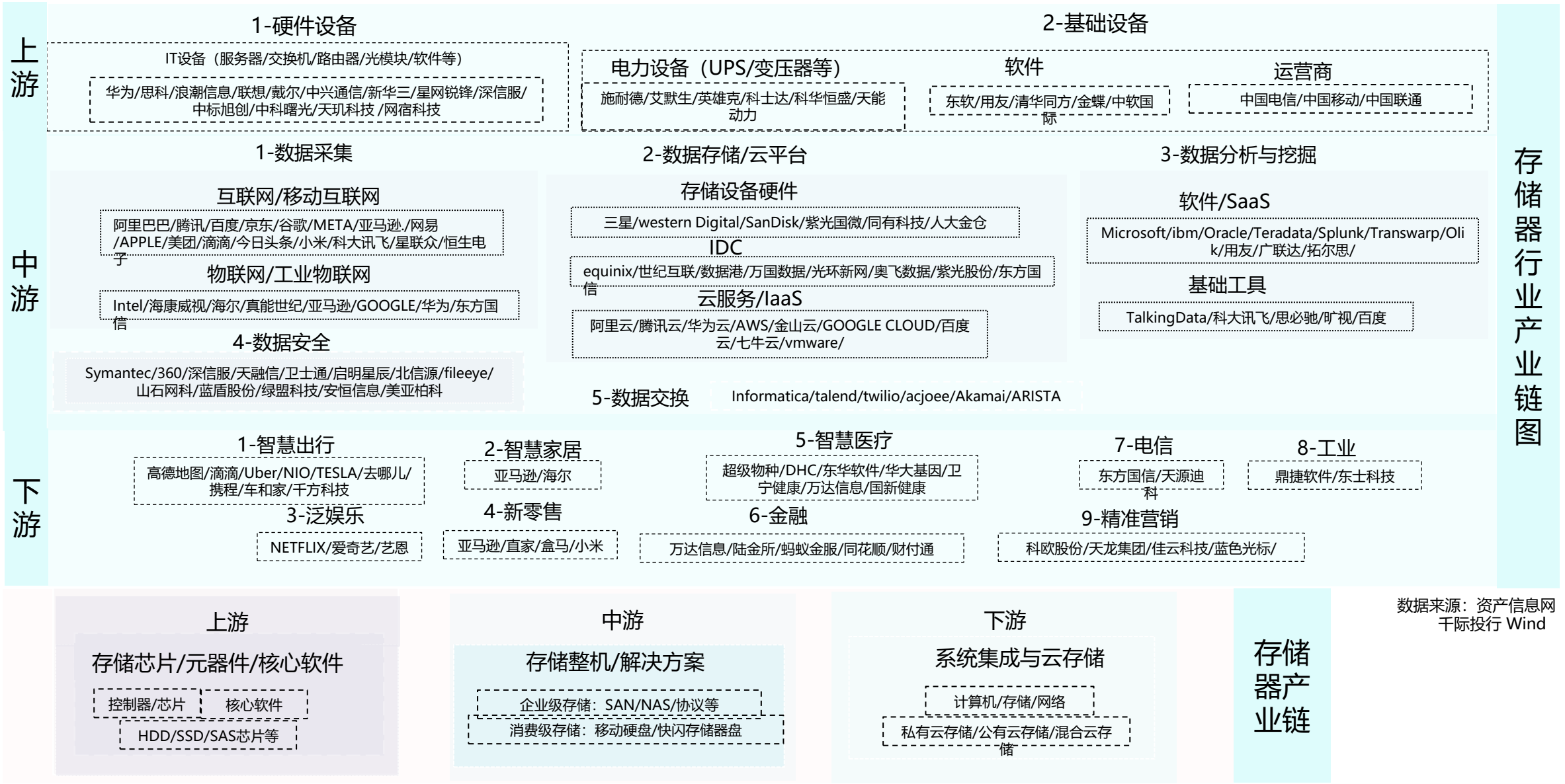
2024年未来年复合增长SDS 12.8%，HCI 13%

中国信息存储市场2016-2020年市场规模(亿元)



数据来源：IDC前瞻产业研究院

1、现代企业级存储综述-存储器产业链（2022）



数据来源：资产信息网 千际投行 Wind

1、现代企业级存储综述-数据单位与关联性

序号	存储单位	中文简称	英文名称	英文简称	换算（byte=1）	十进制换算	数据说明
1	位	比特	bit	b	0.125		
2	字节	字节	byte	B	1		
3	千字节	行字节	kilobyte	KB	2^10	10的3次方（kilobyte）	
4	兆字节	兆	megabyte	MB	2^20	10的6次方(megabyte)	
5	千兆字节	十亿/千兆	gigabyte	GB	2^30	10的9次方(gigabyte)	一个手机的存储如512G
6	太字节	万亿	terabyte	TB	2^40	10的12次方(terabyte)	一块1T数据的硬盘
7	拍字节	千万亿	petabyte	PB	2^50	10的15次方(petabyte)	一个云资源池的数据级是PB级
8	艾字节	百亿亿	exabyte	EB	2^60	10的18次方(exabyte)	数据中心占全球总数据约20%（其余数据在终端或边缘，大部分在终端）(全球500万个机架)
9	泽字节	十万亿亿	zettabyte	ZB	2^70	10的21次方(zettabyte)	全球数据量总计44ZB（gartner）
10	尧字节	一亿亿亿	yottabyte	YB	2^80	10的24次方(yottabyte)	预计2030年每年增1YB（4万亿台高端手机256G存储能力）
11	珀字节	千亿亿亿	brontobyte	BB	2^90		
12	诺字节	一百万亿亿亿	nonabyte	NB	2^100		
13	刀字节	十亿亿亿亿	doggabyte	DB	2^110		
14	馈字节	万亿亿亿亿	corydonbyte	CB	2^120		
15	约字节	千万亿亿亿亿	xerobyte	XB	2^130		

大数据范畴

特殊领域使用：如天文学，宇宙等衡量，基本上待使用。
问 题：真有一天能日常能使用到吗？会有那么一天？

编码	字节数（英文）	字节数（中文）
GB2312	1	2
GBK	1	2
GB18030	1	2
ISO-8859-1	1	1
UTF-8	1	3
UTF-16	4	4
UTF-16BE	2	2
UTF-16LE	2	2

网络带宽的计算单位是：bps， 比特位每秒，也就是表示一秒钟传输多少位（bit），1Kb=1000bps，1Mb=1000Kb

2025年， 50亿人使用计算机上网.....

虽然单位够大，但数据量越来越大的情况，也可能科学界会创造新的容量单位，以大单位，小数字来表示数据量

1、现代企业级存储综述-存储器历史演进里程碑

公元前202年前	公元前202-8	1725年	密度界限	1928年	1932年	1942年	1946年	1947年	1947年	1951年
龟甲、骨、石、竹、木、帛、绳	纸	打孔卡&纸		磁带	磁鼓内存	数字计算机	静电记忆管（数字存储器）	延迟线存储器	磁芯存储器	磁带（用于计算机存储）

1956年	1962年	1965年	1966年	1971年	1978年	1980年	1987年	1994年	1995年
磁盘驱动器	移动硬盘	光盘	DRAM	软盘	RAID	闪存	数字音频磁带	CF存储卡、FC光纤通道	闪存驱动器

1997年	1999年	2000年	2006年	2006之后
MMC存储卡、固态硬盘	SD存储卡	U盘	云存储	分布式存储、DNA存储、纳米存储、量子存储、全息存储、PCM(变相存储)

来源：https://www.chinastor.com,

存储技术持续演进

存储是记录与分享的过程
存储的技术是在交流电普及与创新基础上最快发展
存储技术发展最快是20世纪
存储技术是一种整体技术方案，整个组件的突破
存储的发展前期主力是美国（以IBM为代表）
现代海量数据也在带动存储的改变，比如下一代存储
曾经发明过的技术现代都可以用，现代方案更多，但效能性与经济性是使用何种存储的重要衡量标准

1、现代企业级存储综述-存储技术指标

存储主要技术指标	存储容量	字节编址：字节数表示 字编址：字数*字长	
	存取速度	存取时间Ta	CPU发出指标到完成读写操作
		存取周期Tm	主存完成一次完整的读写操作
		主存带宽Bm	每秒进出的最大值 Bm=主存等效工作效率*主存位宽/8
	可靠性	规定时间内，无故障读写的概率 ，MTBF，企业级最少5个9	
	功 耗	耗电能力	
	扩展性	文件大小与数量而引起的可支持的容量变化	
	空间效率	有效存储/裸容量的比例	

- 存储技术：除以上也需要关心存储处理的能力、成本等能力

名词	类型	说 明
IOPS	性能	存储产品的资料中看到关于IOPS的参数,指的是每秒种的I/O次数。
TPC-C	性价比	由服务器和客户端构筑的整体系统的性能，TPCC测试系统每分钟处理的任务数,单位为tpm,(transactions per minute)
JBOD	磁盘	磁盘簇，又称SPAN，Span是在逻辑上把几个物理磁盘一个接一个串联到一起，从而提供一个大的逻辑磁盘。

- 存储名词：除以上之外如RAID、磁盘通道、主机通道、磁盘镜像等

存储技术指标解析	
指标类型	描 述
存储密度	每单位物理容量的比特数，道密度，位密度，面密度。
存取速度	访问数据的延迟与带宽
存储周期	数据可保存与可读取的最长时间
数据成本	每次读取时的成本

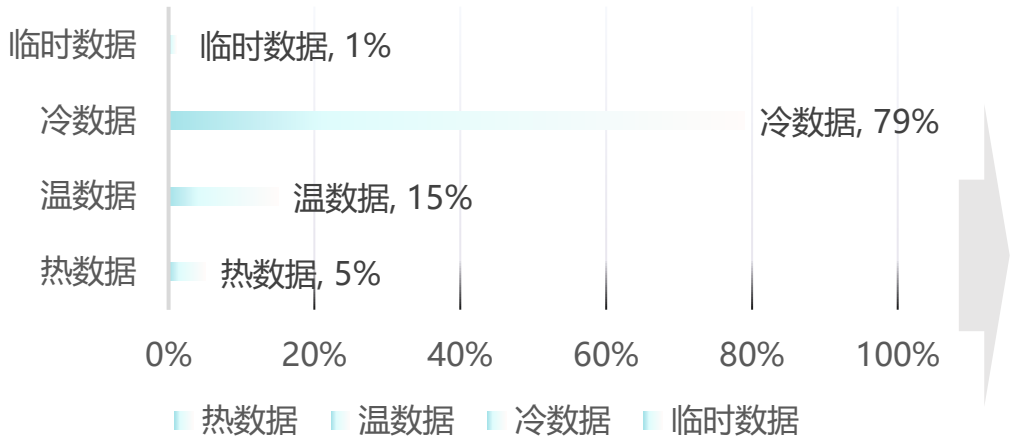
类型	协议	应用
块存储	sata/scsi/iscsi	san,nas,ebs
文件存储	ext3/ext3,xfs,ntfs	pc,serve,nfs
对象存储	http,rest	s3,gcs,rcf

1、现代企业级存储综述-历史数据处理与存储方案

历史数据存放方法

- 内存历史数据库
- 磁盘历史数据库

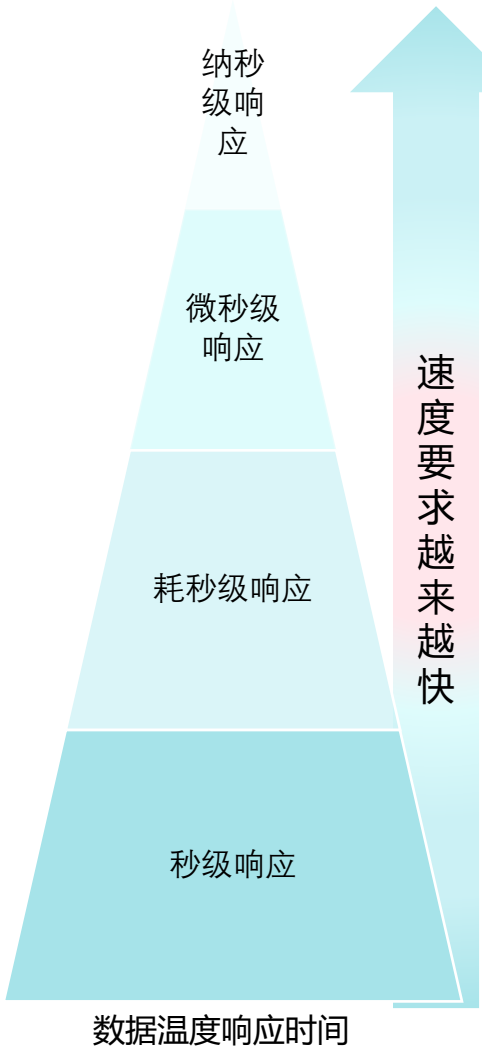
企业数据按访问情况比例



类型一：按热度划分

- 响应的时间和存储可以达到，但依据其它维度做不同的选择，所以这些界限不是特别明显。
- 测试或开发数据在临时数据的范畴

热数据存储	传统存储	SAN、NAS
	软件定义存储	虚拟化存储、Server SAN、超融合架构（HCI）
温数据存储	传统存储（中低端）	
	软件定义存储（中低端）	
临时数据存储	传统存储（中低端）	
	软件定义存储（中低端）	
冷数据存储	磁带、光盘	



1、现代企业级存储综述-外存、内存、Cache、CPU、存储的关系

越小，越少，越快。存储器越贵

越大，越多，越慢；存储器越便宜

L0寄存器

L0 CPU寄存器保存来自高速缓存存储器的字

L1高速缓存（SRAM）

L1高速缓存保存取自L2高速缓存的高速缓存行

L2高速整缓存（SRAM）

L2高速缓存保存取自L3高速缓存的高速缓存行

L3高速整缓存（SRAM）

L3高速缓存保存取自主存的高速缓存行

L4主存（DRAM）

L4主存保存取自本地磁盘的磁盘块

内存 L4 可达60ns

L5本地二级存储（本地磁盘）

L5本地磁盘保存取自远程网络服务器上的磁盘文件

L6远程二级存储（分布式文件系统、Web服务器）

速度减

Cache L1速度延迟可达1ns

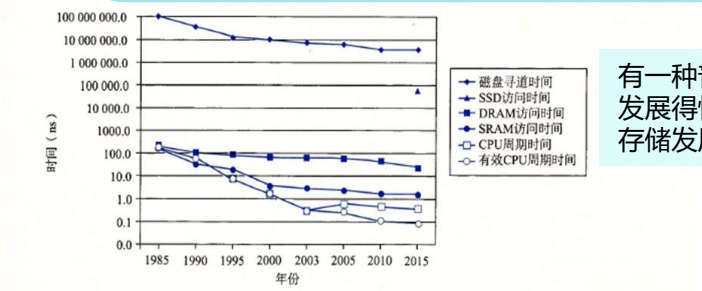
Cache L2速度延迟可达3ns

Cache L2速度延迟可达15ns

SSD外存/辅存 速度可达 25us

NVM外存/辅存 速度可达 50ns

(PCM/STT-MRAM/3D-Xpoint)
非易失性存储器



来源：公共网络

有一种普识（如左图）：存储与CPU发展起来，没有CPU发展得快。
存储发展速度很快，读写速度已在向内存靠拢。

介质

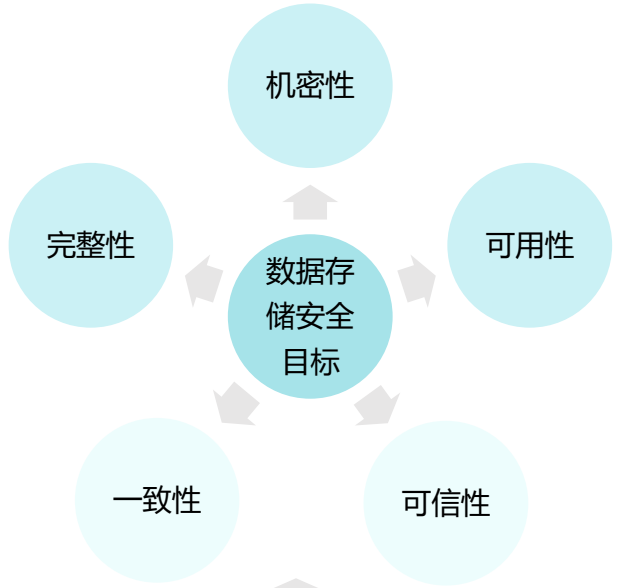
软盘、光盘、DVD、硬盘、闪存、U盘、CF卡、SD卡、MMC卡、SM卡、记忆棒（MemoryStick）、xD卡等

1、现代企业级存储综述-常用存储类型比较

常用存储类型比较

比较项	块存储	文件存储	对象存储
定义	光纤联接服务器提供存储服务	使用文件系统，有目录树结构	将元数据与数据作为一个对象
传输单位	块	文件	对象：元数据与数据
传输协议	FC、iSCSI	CIFS、NFS	基于HTTP/HTTPS的REST、SOAP、API
元数据	固定属性	固定文件属性	自定义元数据
优势	交易数据	简单访问、易管理	内容仓储、文件分享
IO支持	随机读/写	随机读/写	追加写，随机读
访问情况	iSCSI访问、磁盘挂载	NFS、CIFS访问、局域网共享	REST访问、公网传输与共享
设备	cinder、硬盘	ftp、NFS服务器	swift等键值存储
特征	分区，格式化	大文件	高速与共享性
对存储的操作	磁盘读、写	文件级打开、修改、保存、删除	对象上传、下载、查询、删除
提供接口	QEMU Driver，kernel module	Posix	Restful API
最大并发客户端数	数百级	数千级	数千级
扩展性	TB	PB	EB
最大吞吐量	十几GB/S	数百GB/S	数百GB/S
IOPS	百万级	十万级	千级
可靠性	9个9	10个9	11个9
速度	百微秒	毫秒级	数十毫秒
单位容量成本	高	较低	低
分布性	不能异地分布	可分布，性能有瓶颈	分布，高并发能力
文件大小	不限制	适合大文件	不限制
文件级权限管理情况	不支持	支持	支持
典型技术	SAN	HDFS,GFS	SWIFT，,Amazon S3
限制1	难以跨数据中心扩容	元数据与扩展性极限在10亿节点	非高频次操作的数据
限制2	不能共享数据	传输速度低	不兼容多种模式
文件可修改性	即时更改	即时更改	客外对象会被创建
场景	数据库，ERP	数据中心、HPC,企业OA	网络媒体、大数据/IoT、备份/归档

1、现代企业级存储综述-数据存储安全性解决方案



传统安全威协：数据存储安全最大的破坏力是数据库的存储，达90%以上

数据库最大的破坏力是人为操作占70%以上

权限分离、应用与Database权限泳道隔离

采集

存储

分析

交互

应用

销毁

新型数据安全威协

- 数据采集与查询泛滥
- 软硬件漏洞
- 新设施应用场景应用
- 技术管理不到位的泛滥使用

数据存储安全问题解决

事前
事中
事后

事前
事中
事后

1-系统的外围手段方法

- 权限控制
- 数据审计
- 数据加密
- 数据脱敏
- 数据备份
- 数据快照

2-数据存储外围手段系统

- 外围防火墙系统
- 数据库防火墙系统
- 数据库集群
- 数据库主从复制系统
- 分布式数据库系统
- 分布式文件系统
- 分布式存储系统
- 云存储
- 数据备份系统

数据存储安全核心：数据安全性保护、安全损坏后的快速恢复

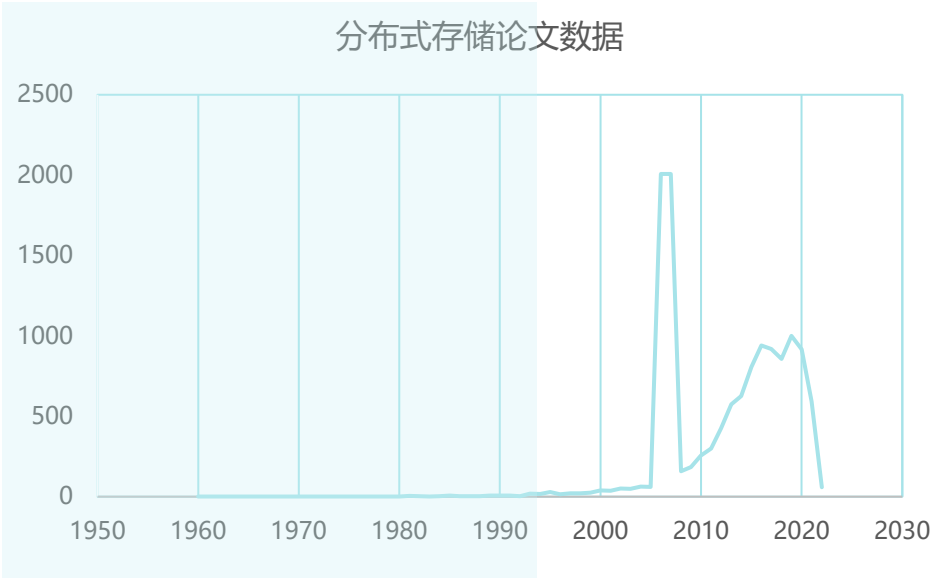
现有趋势解决性方案是云存储方案

数据存储安全问题全周期都存在

目录/CONTENTS

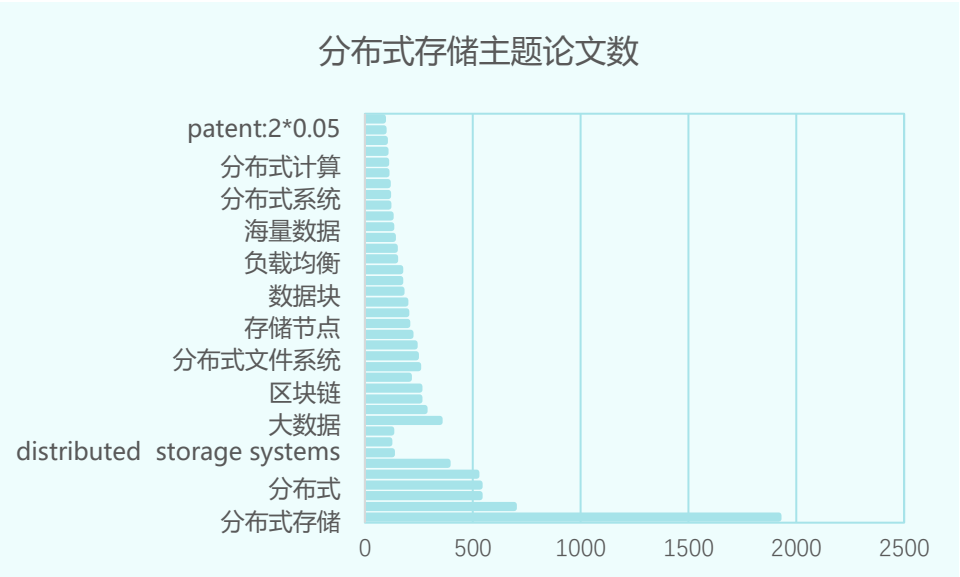
2、现代企分布式存储技术-分布式存储的理论情况

分布式存储的研究是从外国起步、壮大、应用，后期由国内加入研究并正在研究的课题



来自于外文期刊

数据来源：论文知识库



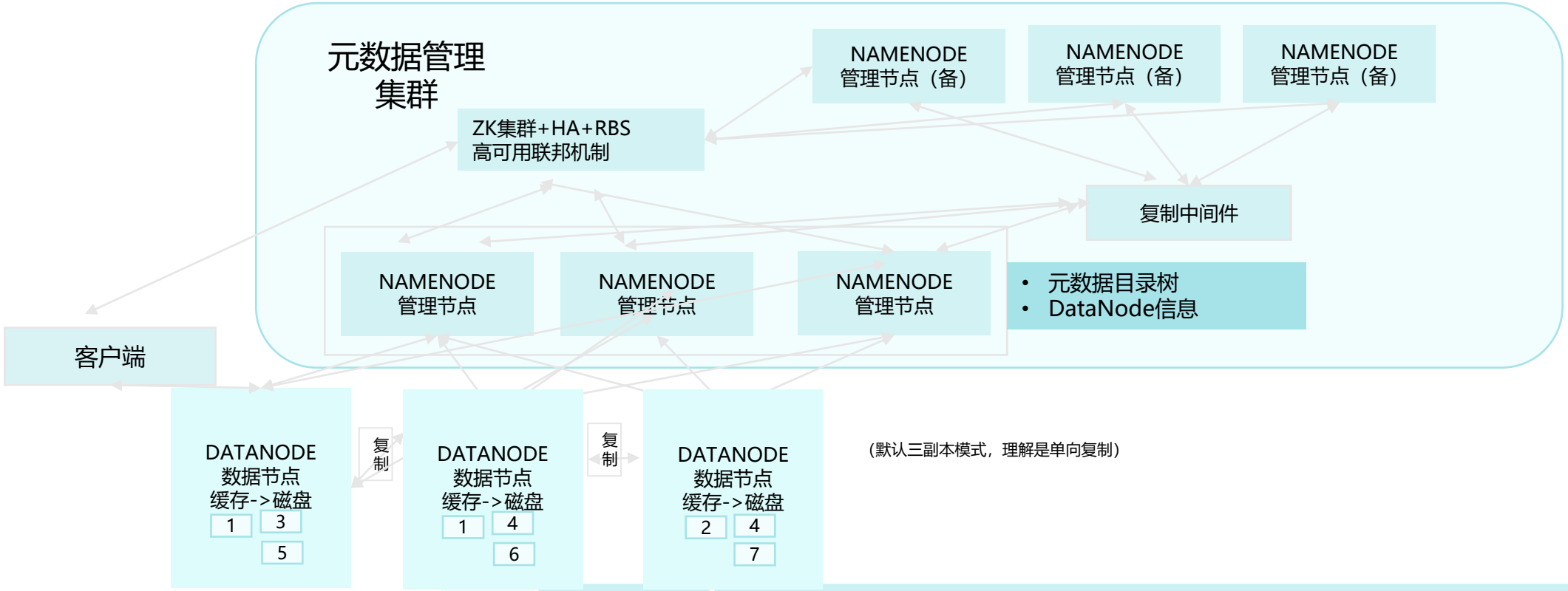
数据来源：论文知识库

从1960年开始有研究论文发表
1994年之前的论文为外国论文
2019年论文数达到1000个，从2020年开始下降

分布式存储是个整体解决方案
实现方案的是组件与集成的问题

2、现代企分布式存储技术-分布式存储的结构

主流一：有统一元数据管理集群的分布式存储HDFS（基于3.0）

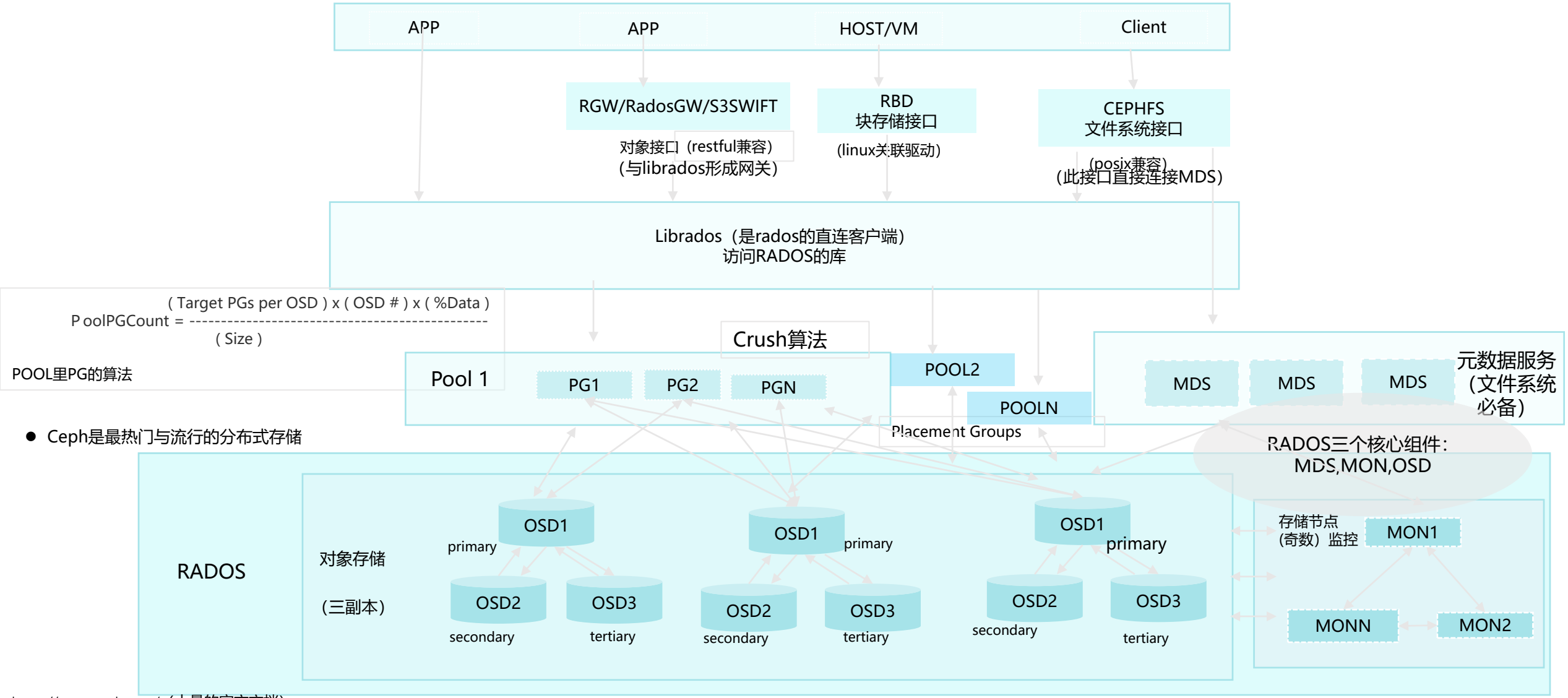


HDF是最重要的大数据存储技术之一也是现代云计算常用的存储之一。

HDFS特性	HDFS的特性标签	分布式、高容错（3副本）、数据海量、高可用、高吞吐、中间件机制、块、文件系统、高可扩展
	HDFS优点	自动化高容错，从中间件到数据节点是副本模式 适合大数据：数据规模PB级、文件规模：百万；节点规模：万节点级 数据一致性 硬件选型廉价（成本选型）
	HDFS缺点	数据低延迟有限；全局锁有限；管理节点内存要求高，小文件存储效率低。架构重、运维复杂
	HDFS适合场景	百T级以上业务，高并发 （Hadoop）

2、现代企分布式存储-分布式存储的结构

主流二：有统一元数据管理集群的分布式存储Ceph

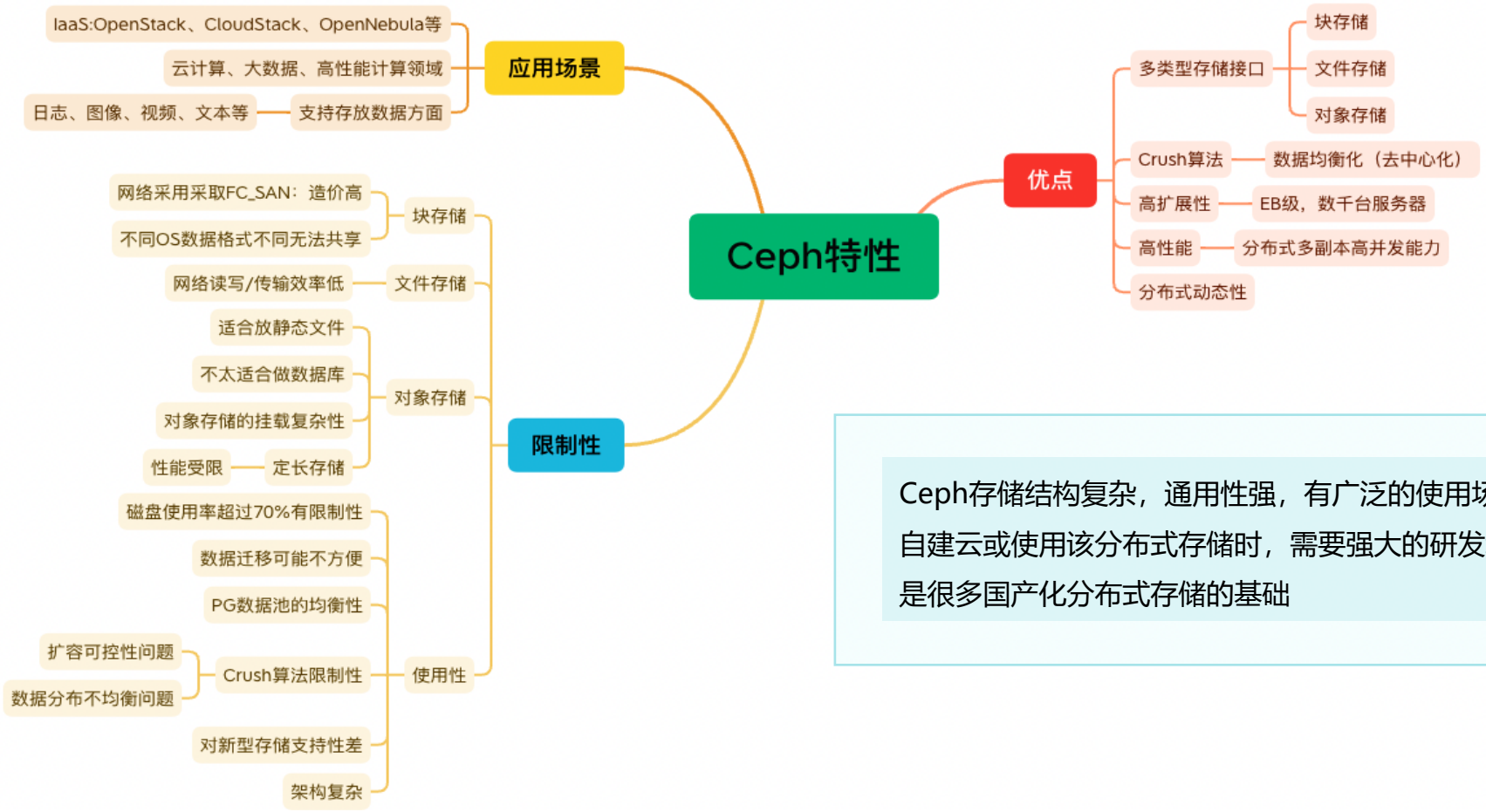


- Ceph是最热门与流行的分布式存储

https://www.ceph.com/（大量的官方文档）

2、现代企分布式存储-分布式存储的结构

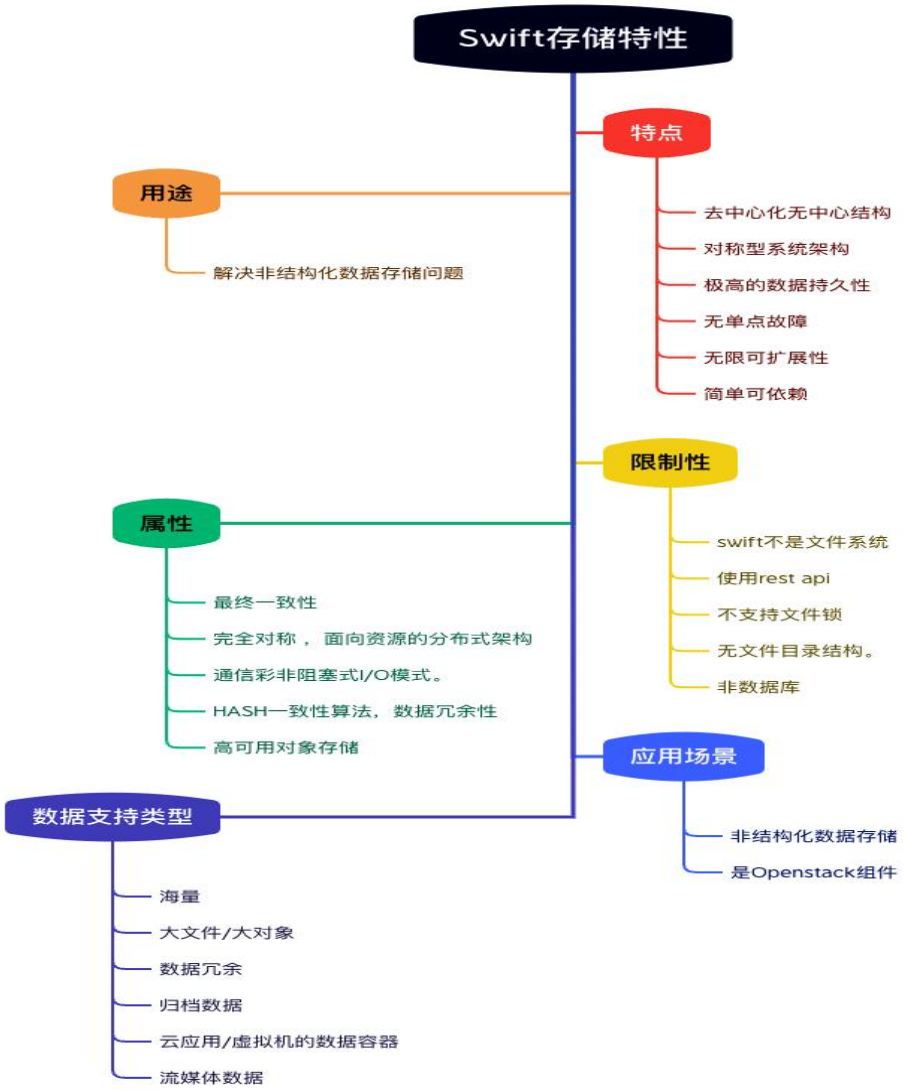
主流二：有统一元数据管理集群的分布式存储Ceph

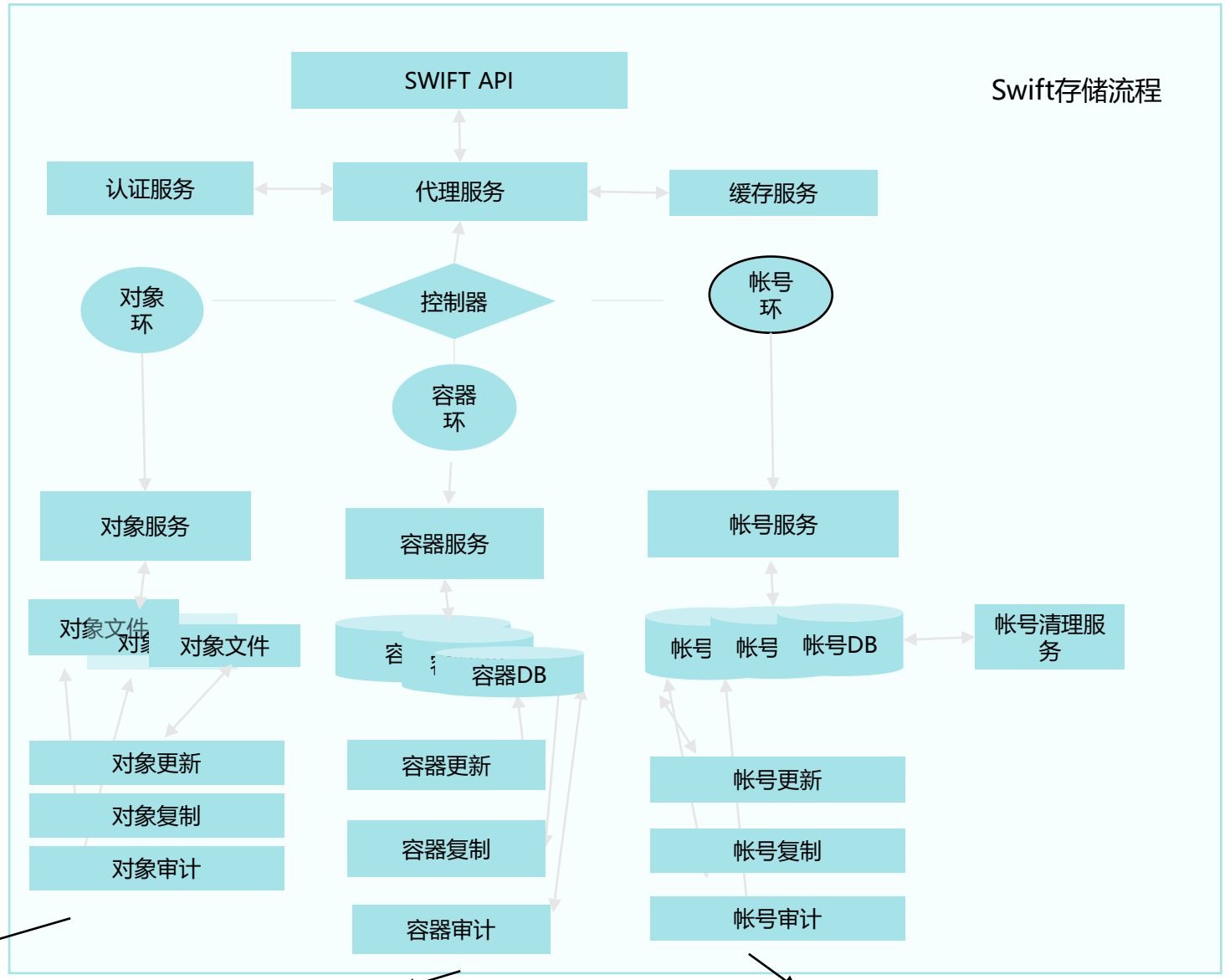


Ceph存储结构复杂，通用性强，有广泛的使用场景
自建云或使用该分布式存储时，需要强大的研发级优化能力
是很多国产化分布式存储的基础

2、现代企分布式存储技术-分布式存储的结构

主流三： SWIFT存储（对象存储）





2、现代企分布式存储技术-分布式存储的结构

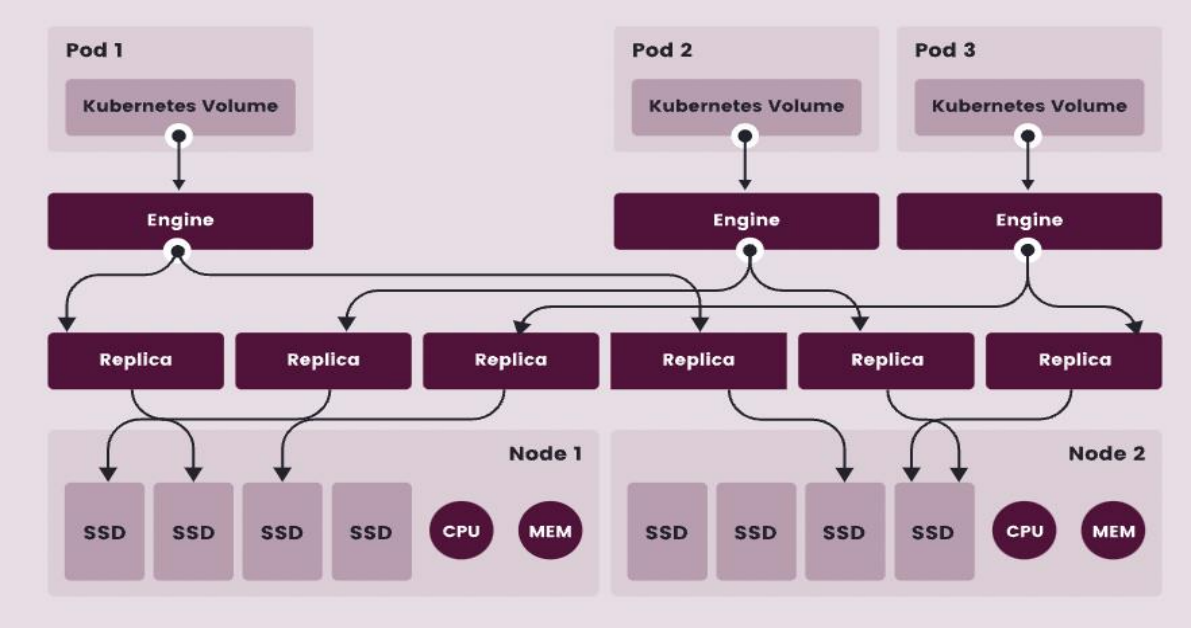
云原生分布式存储： Longhorn（分布式块存储）

Longhorn 是用于 Kubernetes 的轻量级、可靠且功能强大的云原生分布式块存储系统，

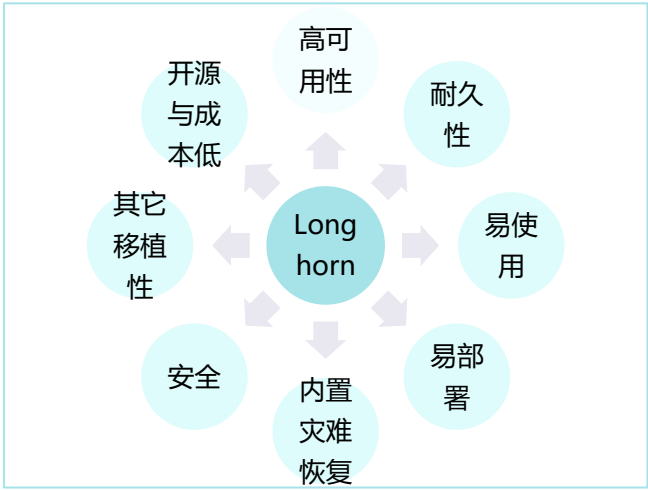
功能特性

无单点故障
GUI 仪表板提升体验
使用Longhorn 卷作为kubernetes集中分布式有状态应用程序的持久存储
跨多个节点和数据中心复制块存储用以提高可用性
将备份数据存储在NFS及AWS S3等外部存储上
创建跨区灾难恢复卷，可以快速恢复主K8S
定期卷快照并备份到NFS或与S3兼容存储上
从备份恢复卷
Longhorn在线升级

Longhorn存储流程



来自longhorm官网



应用优势

与云原生应用程序一起运行良好的分布式存储系统（无需依赖外部提供商）
与 Kubernetes 紧密耦合的存储解决方案
高度可用且持久的存储
没有专用硬件且不在群集外部的存储系统
易于安装和管理的存储系统

Longhorn 是 Kubernetes 持久存储的完美解决方案

2、现代企分布式存储技术-分布式存储的关键技术与类型

分布式存储关键技术		
	网络数据中心	中心是交换机：服务器只负责处理和存储数据，扩展性好。 中心是服务器：无需路由器与交换机。链路冗余高。 交换机与服务器混合：网络结构灵活。
	数据容错	基于纠删码：数据块分割、容错数据修复、优化网络编码 基于复制：数据复制：如何管理更多副本技术；数据组织结构：组织结构P2P、基于元数据
	节　能	硬件节能 ： 计算机部件、数据中心 软件节能：节点管理、数据管理（静态放置、动态放置、缓存预取）
	元数据	元数据管理技术、元数据分配技术
	弹性扩展	数据动态迁移与切换技术，负载均衡技术、节点失效转移与恢复技术
	层级优化	缓存，磁盘、节点、温热数据的预处理技术

分布式存储：一致性解决方案非常重要；是基于网络的分布式，网络抖动的应急与预处理，如最终一致性（常用方法时间戳），强一致性的选择是需要根据业务及成本首先需要考虑的。最终一致，一致性最快多久可以达到一致性效果？机房内边际效用：耗秒级（可以更小）。跨机房：秒级

分布式存储要解决的重要问题：存储的副本一致性问题：解决读检测，不一致时触发自动修复报错，修复不了即报错

分布式存储的副本一致：偏于最终一致性

2、现代企分布式存储技术-主流分布式存储平台比较

现阶段常用分布式存储

序号	产品名称	运营团体	架构组件	分布式架构特点	系 统	一致性	适应场景
1	GFS	google	MASTER、CHUNCKSERVER、CLINENTS	全局统一命名空间机制（类中间件）	文件系统	弱一致性（最终一致）	大型的、分布式的、对大量数据进行访问的应用
2	TFS	淘宝	NameServer，DATA SERVER	命名服务协调	文件系统	强一致 性（W=N, R=1）	海量、非结构化的大数据
3	Ceph	Linux基金会	rados、librados、osdc	去中心化	块存储、文件存储、对象存储	强一致性	云平台、私有云、容器、公有用整合、海量文件
4	HDFS	Hadoop	HDFS Client， NameNode, DataNode、Secondary NameNode	全局主控节点	文件系统	弱一致性（最终一致）	大数据场景（副本延迟）
5	Swift	Openstack	Proxy Server、Storage Server、Consistency Server	去中心化	对象存储	弱一致性（最终一致）	网盘（不支持实时读写编辑、用于上传下载）
6	GlusterFS	Z RESEARCH	gluster、glusterd、glusterfs、glusterfsd	模块化堆栈式	文件系统	弱一致性（最终一致）	大数据应用和视频存储
7	LUSTRE	lustre基金会	MGS、Lne、MDS、mdt、mgt、client、oss	集群和并行架构	文件系统	弱一致性（最终一致）	超算（不适合小文件）。石油、天然气、制造、富媒体、金融等
8	MooseFS	自由软件	Master Server、metalogger Server、Chunk Servers、client	分层的目录树结构（类UNIX）	文件系统	弱一致性（最终一致）	大规模高并发数据存储小文件（<1M)、大文件.(master server有性能瓶颈)
9	MinIO	MINO基金会	MINO NODE	去中心化的无共享架构	对象存储	弱一致性（最终一致性）read-after-write	云原生应用、物联网、私有云，存储海量的图片，视频，文档
10	SeaweedFS	seaweed	NameNode、DataNode	元数据、数据节点分离	文件系统	弱一致性（最终一致）	存储海量小文件
11	Longhorn	SUSE LINUX	数据平面(data plane)、控制平面(control plane)	基于微服务的设计	块存储系统	强一致性	企业级云原生容器分布式存储、轻量级、微服务

分布式存储通用特性

分布式存储系统是基于网络的，多采用弱数据一致性。
分布式系统是横向扩展高并发的存储。
分布式存储多采用元数据与数据分离结构。

分布式存储最终还是在云上，用于云计算，云原生，物联网等应用场景。
多采用文件系统，并不是兼容多种存储类型。
分布式存储涉及分布式系统及分布式数据库

2、现代企分布式存储技术-国内几个分布式存储比较

国内主流混合云技术架构对比

比较项		华为	H3C	深信服	SmartX
分布式存储	产品	Fusion Storage（块、对象、文件）	h3c ONEstor（块、对象、文件 ）	aSAN（文件）	ZBS（块）
	技术来源	始于Ceph，后自主研发为主	始于Ceph，后自主研发为主	始于Glustre FS,后自主研发为主	自主研发
交付方式		软硬件一体化	软硬件一体化	一体机/软件	一体机/软件
集群规模		3-4096节点	4096节点	255节点时性能明显下降	单个集群规模为255节点
分布式存储成熟度		高	较高	较高	高
兼容虚拟化平台		vsphere/hyper-v/kvm/K8S等	vsphere/hyper-v/kvm/cas/xen等	vsphere/k8s/kvm等	vsphere/x8s/kvm等
hypervisor	产品	fusionsphere	H3C CAS	aSV	ELF
	技术来源	基于KVM	基于KVM	基于KVM	基于KVM
数据保护		多副本/N+2-N+4纠删码	2-6副本/N+1-N+4纠删码	CDP技术/多副本/虚拟机备份/应用数据备份/网络行为管理	多副本/异地容灾备份/快照/等
数据自愈		自行并行重构（4T/小时）	并行重构(1T<30MIN)		多节点并发数据恢复
服务兼容		只兼容华为服务器	只兼容华三服务器	主流都兼容	主流都兼容

主流分布式存储技术

来源：广发证券发展研究中心

分布式存储补充说明

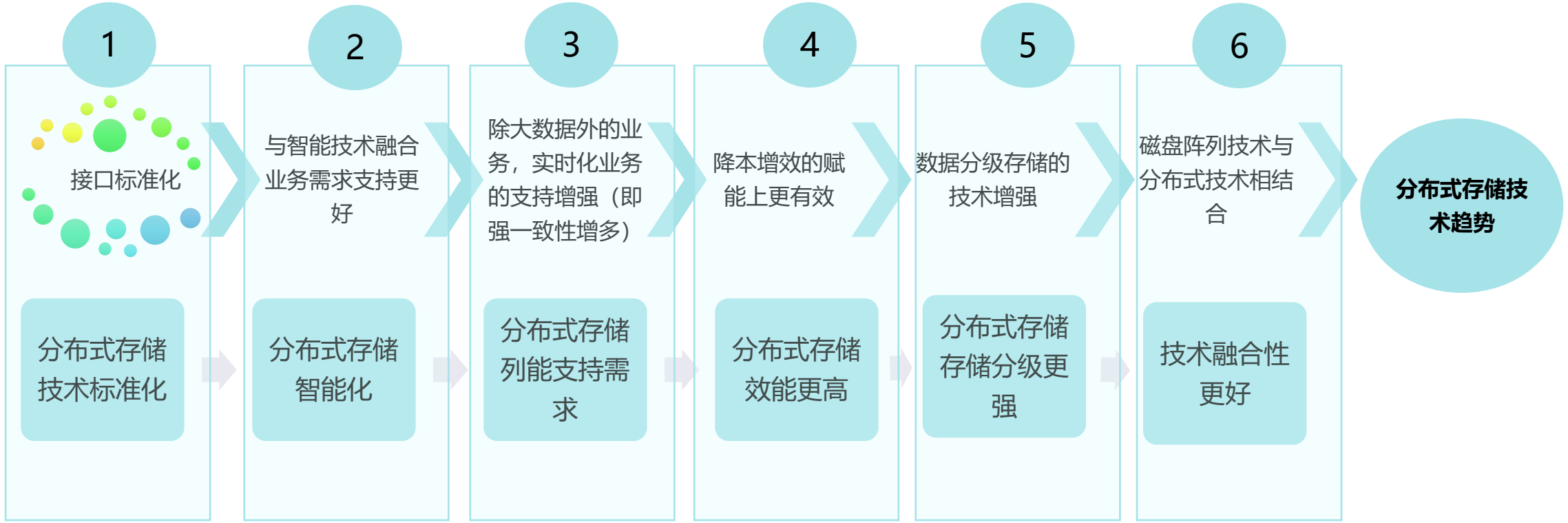
国产化的分布式存储越来越多，并且兼容性增强。

研究的维度可作为选型与测试的参考。

分布式存储相对其它而言维护与架构比较复杂。

云环境下对存储稳定性、性能要求更高。

2、现代企分布式存储技术-分布式存储技术趋势



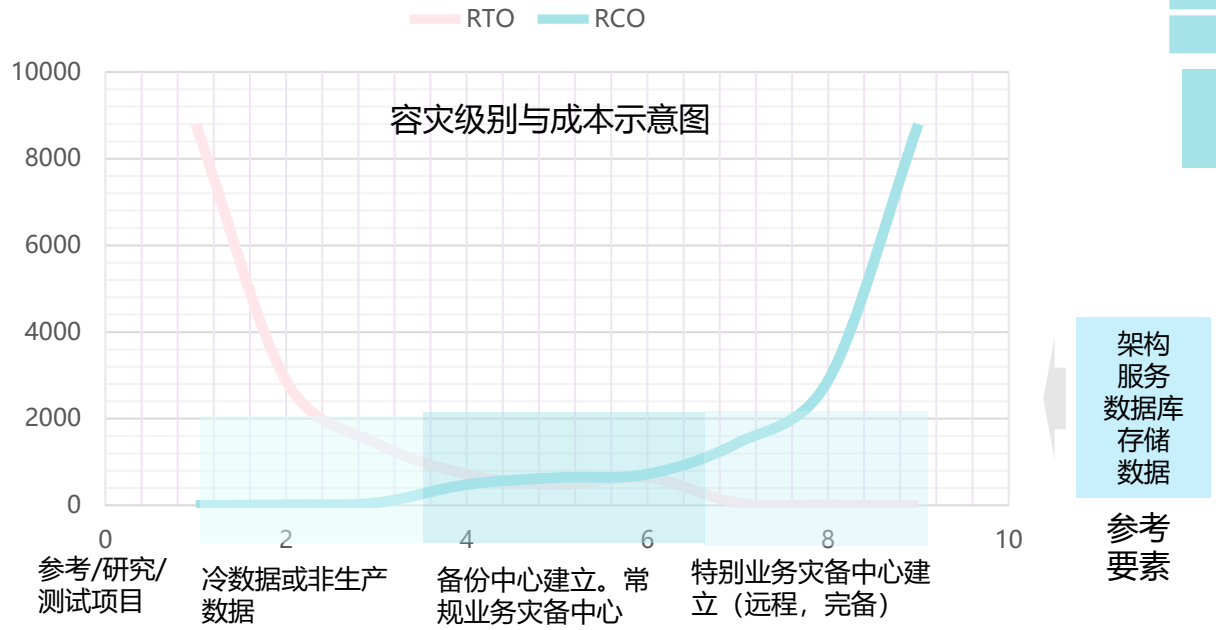
目录/CONTENTS

信息系统风险与威胁无时无刻不在：自然灾害，设备故障、误操作、病毒感染、黑客攻击等。

存储容灾目标：为了服务于业务与服务连续性。

3、数据存储容灾技术-数据存储容灾系统简介-容灾关联指标关系

国际标准	ISO 22301:2012业务连续性管理
国家标准	GB/T30146-2013 公共安全业务连续性管理体系要求
国家标准	GB/T20988-2007 信息安全技术信息系统灾难恢复规范
国际标准	SHARE78灾难恢复标准



- 系统无绝对的可靠性，不同级别的系统，会明确指标数据，并训练可靠性预防与恢复的方案与技术。

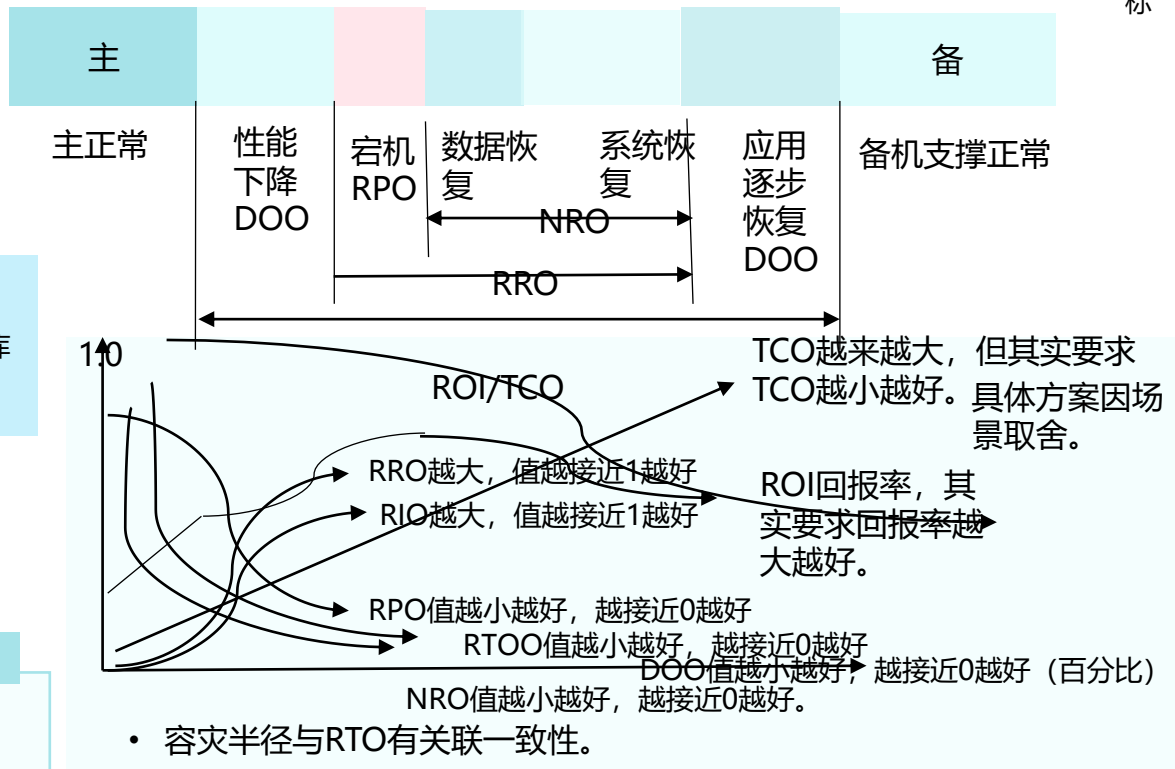
《银行业信息系统灾难恢复管理规范》（2008）

一类信息系统：RTO<6小时，RPO<15分钟
二类信息系统：RTO<24小时，RPO<120分钟
三类信息系统：RTO<7天

容灾与业务关联性的指标		
RTO	恢复时间目标	衡量业务恢复正常所需时间
RPO	恢复点目标	最大数据丢失量（以时间来度量）
RRO	恢复可靠性目标	最大数据恢复/切换成功率
NRO	网络恢复目标	网络切换到备机的服务时间
RIO	恢复完整性目标	最大状态恢复率（百分比）
DOO	降级服务目标	最大降级服务率(百分比)
ROI	投资回报率	投资获得回报价值
TCO	总成本	总体成本

业务连续性最常用指标要求

服务SLA的重要指标



- 容灾半径与RTO有关联一致性。

3、数据存储容灾技术-数据存储容灾

数据存储容灾的目标

- 业务连续性
- 数据可靠性

数据存储环境

- 普通IDC
- 云计算IDC
- 虚拟化

数据存储配套技术

存储硬件系统
- 磁盘柜
- 磁盘阵列
- NAS设备
- 存储服务器
- 磁带库

存储网络
- 存储卡
- 存储网线
- 存储卡系统

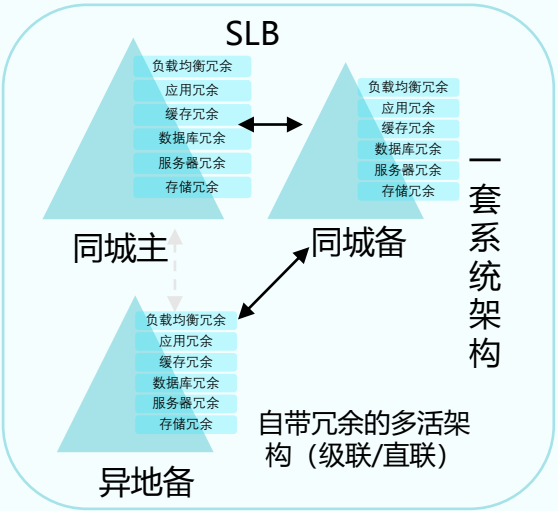
存储系统
- 数据管理
- 数据镜像
- 数据复制
- 路径冗余
- 数据快照

存储辅助系统
- 双机高可用
- 存储虚拟化
- 数据备份容灾
- 数据归档
- 数据迁移

存储应用解决方案
- 数据扩容
- 数据集中存储
- 集群高可用
- 双存储
- 数据备份
- 数据容灾
- 存储虚拟化
- 安全
- 数据归档
- 数据迁移
- 数据复制

数据存储容灾落地

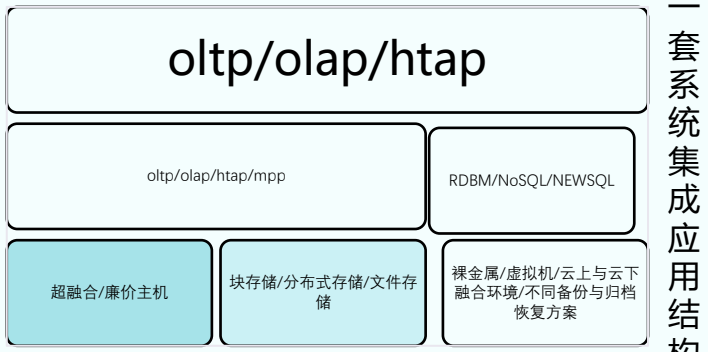
容灾型系统架构



主要使用数据库/数据复制技术

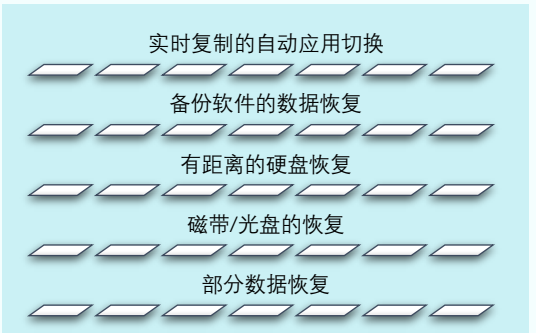
部分为去中心化架构

不同软硬件配置的集成型应用系统



按功能解耦解决系统容灾

有梯度的数据恢复技术



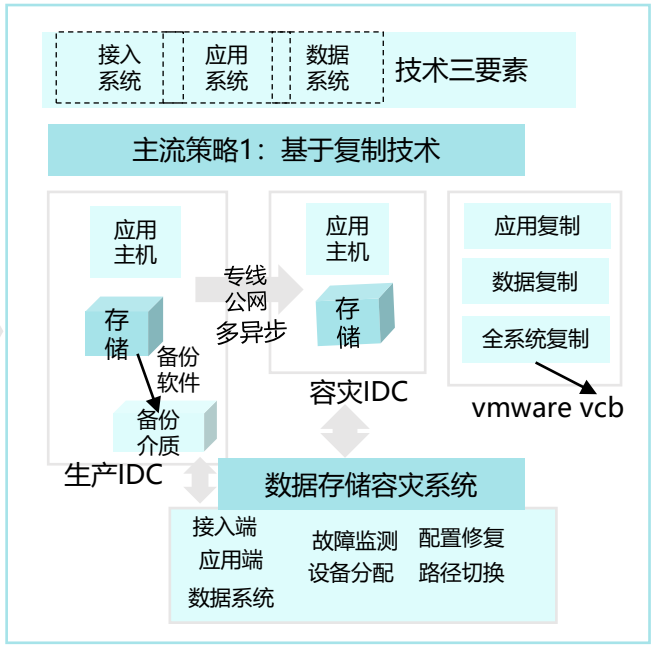
按应用要紧程度配置恢复方法

在有基础工具的基础上：数据存储容灾的实质落地是依靠：系统架构，系统集成与数据恢复技术来完成。

3、数据存储容灾技术-主流数据存储容灾系统（2022）

系统架构

数据存储容灾系统与技术



存储容灾软件工具

容灾类型	软件名称	技术	备　注
(远程)数据镜像工具	IBM PPRC	数据同/异步	数据存储级镜像
	IBM XRC	数据同/异步	
	EMC SRDF	数据同/异步	
	HDS TureCopy	数据同/异步	
CDP工具	EMC RepliStor	准CDP	SNIA标准真CDP必须三符合标准 1.可以捕获任意的数据变化； 2.至少可以备份到另外一个地方； 3.可以恢复到任意时间点。
	IBM Tivoli CDP		
	AppAssure Replay		
	EMC RecoverPoint		
应用复制工具	Oracle DataGuard	数据同步	利用数据库（应用）技术
	IBM DB2 HADR	数据同步	
	DSG（国产）	数据同步	
	DBSync（国产）	数据同步	
全系统复制工具（云计算环境）	Vmotion	封闭/虚拟化	虚拟化高可用技术/虚拟机时迁移
	Xen LIVE Migration	基于共享存储	
	Nomad	高可用管理集群	
	Vmware VCB/VDR	备份/校验	

■ 容灾驱动力为CDP（Continuous Data Protection）连续性数据保护

数据存储容灾系统关联性

灾备（备份/恢复/验证）软件
灾备介质
异构存储灾备
共享式/集中式灾备平台
灾备池/灾备湖
灾备中心/多云灾备中心

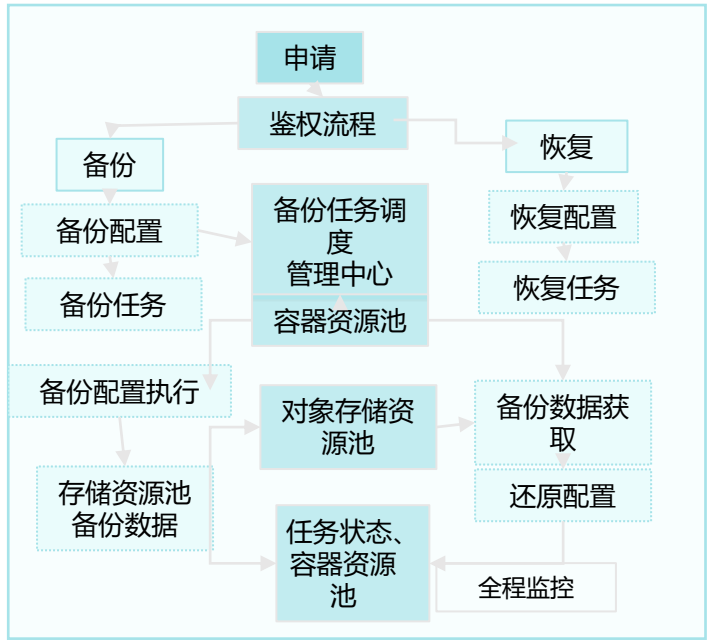
灾难恢复即服务（DraaS）

Gartner 2019 年灾难恢复即服务魔力象限

供应商	产品特征
iland	Secure Cloud Console
Sungard AS	Sungard AS Cloud Recovery
Infrascale	Infrascale Backup&Disaster Recovery(IBDR)
IBM	IBM DRaaS
Intervision	与Carbonite、Zerto和VMware合作。
Expedient	与Zerto、Cohesity和VMware合作
TierPoint	与VMware、MicrosoftZerto、Nutanix和Dell合作
Recovery Point	Gartner DRaaS MQ的领导者

3、数据存储容灾技术-高效容灾备份系统与技术

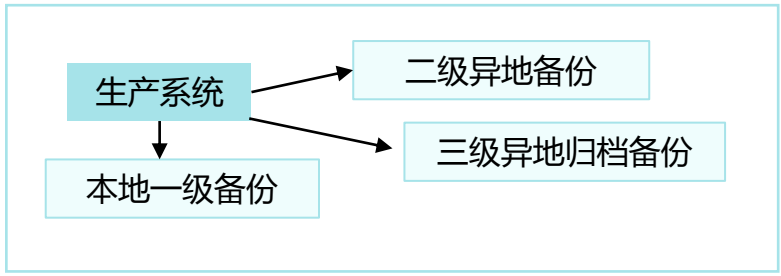
云计算备份即服务流程



容器间的资源池：任务间互相隔离
资源控制颗粒度细
资源动态调整
备份能力可伸缩
特性：云计算特性+数据加密+数据解压缩技术
容器化技术的能力
异构化数据

主流云计算公司存储即服务情况

云公司	备份简介	备份产品服务名	备份对象	优　势	场景
AWS	集中管理和自动执行各种 AWS 服务的备份工作	AWS BACKUP	所有类型数据	利用99.999999999% 的数据持久性保护备份。分钟级扩展、高效支出的数据保护、高效数据传输	AWS全场景
阿里云	阿里云统一灾备平台，是一种简单易用、敏捷高效、安全可靠的公共云数据管理服务	AWS HBR	支持ECS （文件，MySQL, Oracle, SQL Server，SAP HANA），NAS， OSS，Tablestore 等阿里云上数据源备份	数据源多、经济（重删、网络流量小）、操作简单、备份容灾归档迁移一体化	阿里云多场景
IBM		IBM CLOUD BACKUP	所有类型数据	基于WEB的 UI管理、允许裸机复原、细粒度恢复、deltapro去重、智能压缩	200多种操作系统与应用,不限数据中心类别
Microsoft Azure	帮助防御勒索软件的集中式备份服务与解决方案	Azure备份	备份 Azure 虚拟机、本地服务器、SQL Server 和 Azure 虚拟机上的 SAP HANA、Azure 文件存储和 Azure Database for PostgreSQL。	集中管理、保证应用程序一致、多工作负载、本地冗余LRS、异地冗余 GRS、区域冗余ZRS存储备份。	
华为云	为云内的云服务器、云硬盘、文件服务	华为云 CBR	云服务器备份存储库、云硬盘备份存储库、SFS Turbo备份存储库、混合云备份存储库		云服务器整机、磁盘部分数据、文件系统数据保护、云上备份云下业务数据

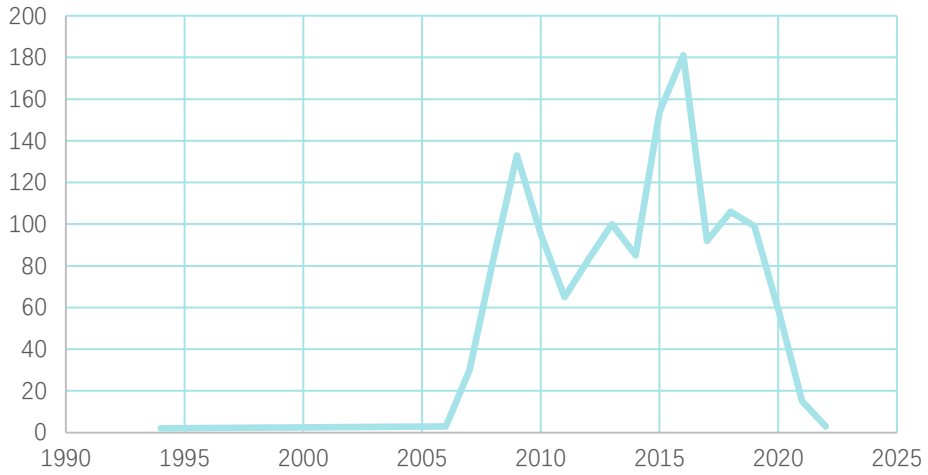


云公司存储即服务特性	与云计算服务高效结合，提供全程程序与数据保护，并一致性。
集中化	服务灵活，并且支持多样性，备份能力与公司支持的服务能力一致性。
一体化	属于基于备份、恢复、冗余、容灾技术集中一体化技术的集合体与融合体。
整体目标类似	子系统，实现的计算方式与软件、策略有差异。

目录/CONTENTS

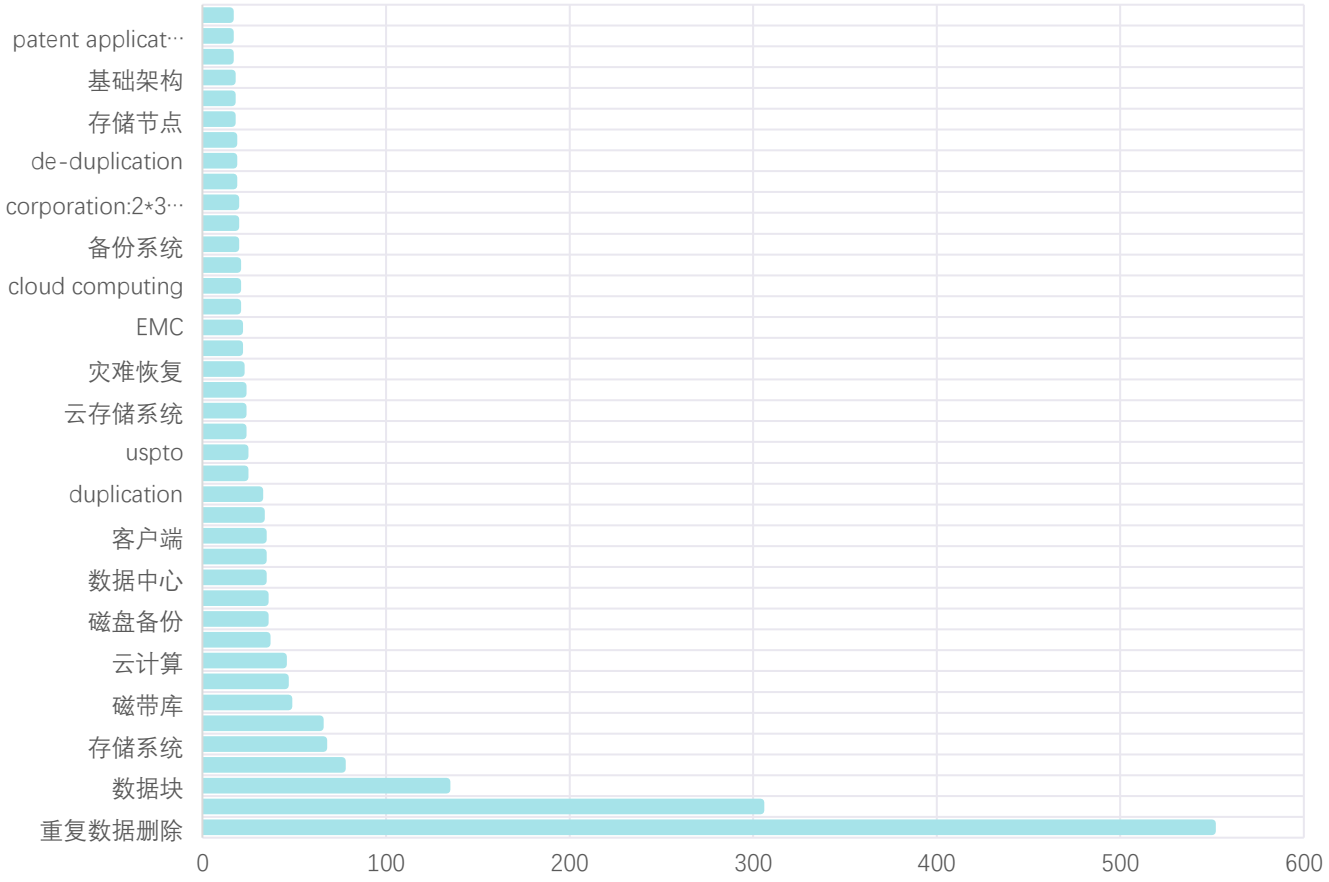
4、数据存储系统的删冗技术-数据存储的删冗技术现状

重复数据删冗技术论文发表情况



数据来源：论文知识库

与数据删冗关联的技术论文发表类型涵盖情况

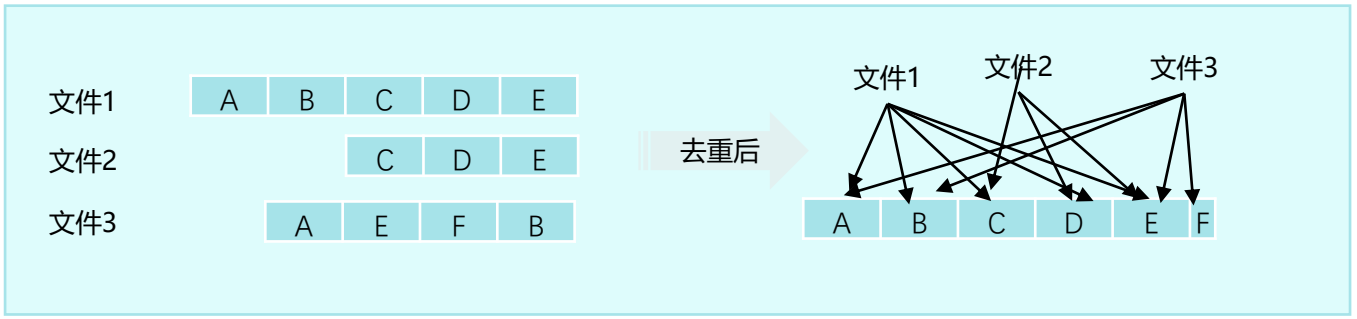


数据来源：论文知识库

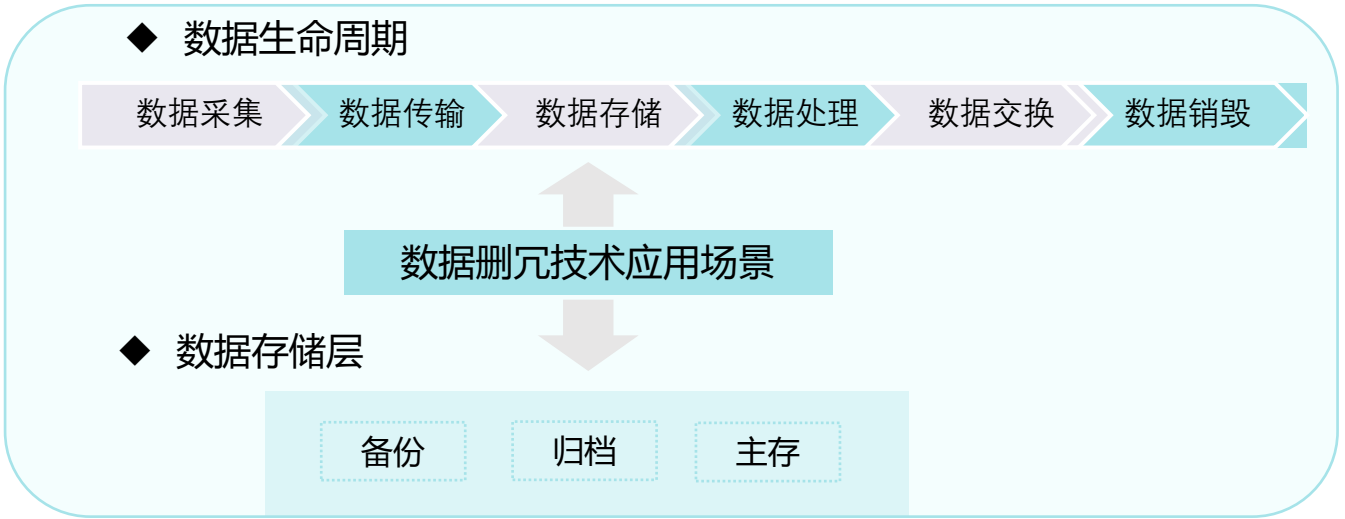
数据删冗技术是持续改进现已成熟的技术类型
数据删冗包含首先直接目标是删除冗余数据，但同时需要保护现有生产系统性能与数据完整性
数据删冗包含技术，系统、产品及故障解决，包含了环境（如数据中心、云计算）等所有的环境与设施

4、数据存储系统的删冗技术-数据存储的删冗技术应用

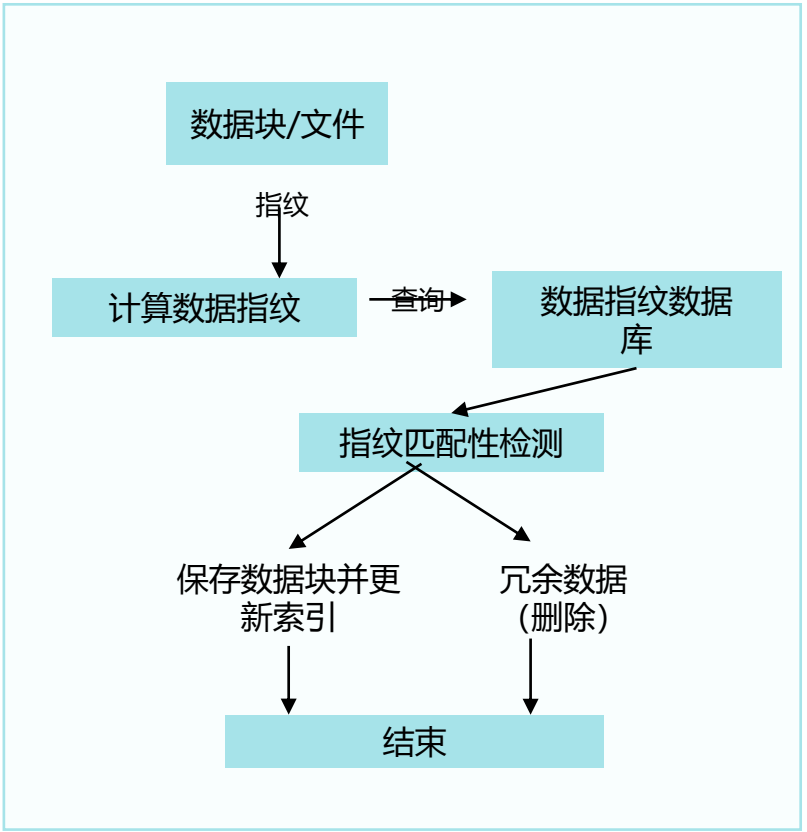
数据删冗原理：只保存唯一一份备份的数据段



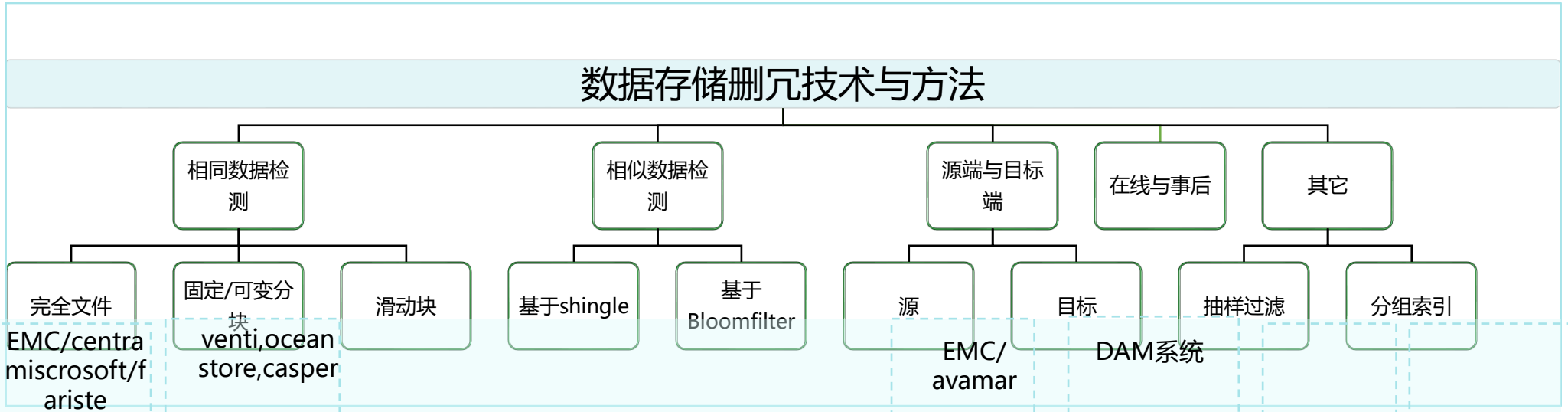
数据删冗：数据存储与数据传输



数据删冗原理流程



4、数据存储系统的删冗技术-数据存储的删冗技术应用



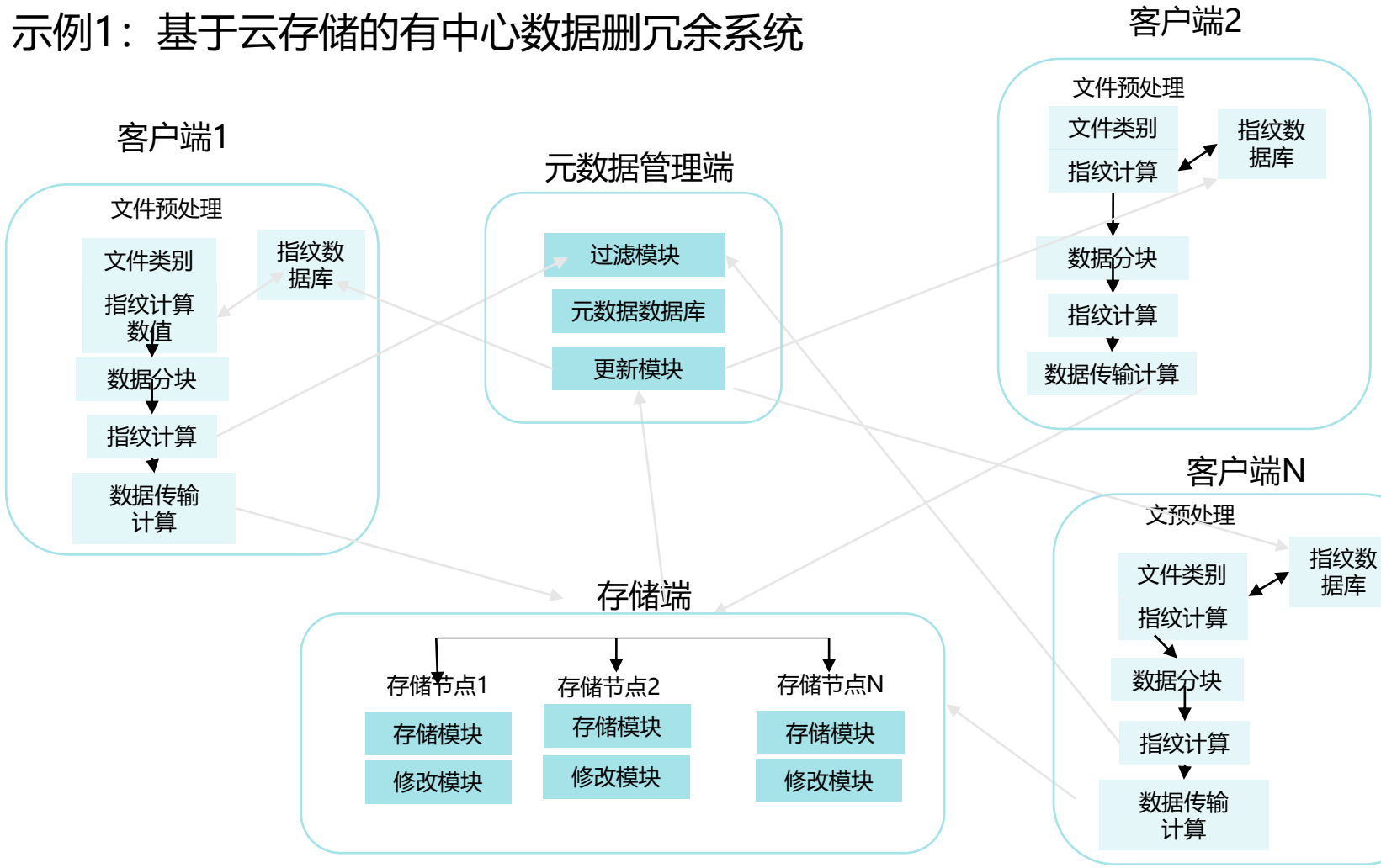
技术应用系统	EMC/centramicrosoft/fariste	venti,oceanstore,casper		EMC/avamar	DAM系统	
优点	快速、效率高	快速、多元化		快速，开销小	对源端透明，不占源端资源	减少内存数据	减少内存数据
缺点	颗粒粗，应用有限	对更新敏感		会占源资源，影响性能	会占网络带宽	无法应用到主存	无法使用块存储

删冗关键技术	基于IO优化的删冗技术	DDFS系统/bloomfilter D2D/TTTD Debar/改进DDFS	缓解IIO操作
	删冗可靠性技术	副本可靠性 R-admad/ecc nec/hydrastor/纠删码	数据可靠性

删冗技术主流使用的技术
加速数据分块技术 加速索引查询技术 减缓磁盘I/O技术

4、数据存储容灾系统的删冗技术-高效删冗的系统设计方法

示例1：基于云存储的有中心数据删冗余系统



删冗系统的组件说明

客户端	文件数据块分块、计算、设置指纹传输
元数据服务	对比指文数值，确定数据块是否上传，并反馈信息
存储端	接受客户端传输的数据块，并把信息传输给元数据服务，更新元数据库

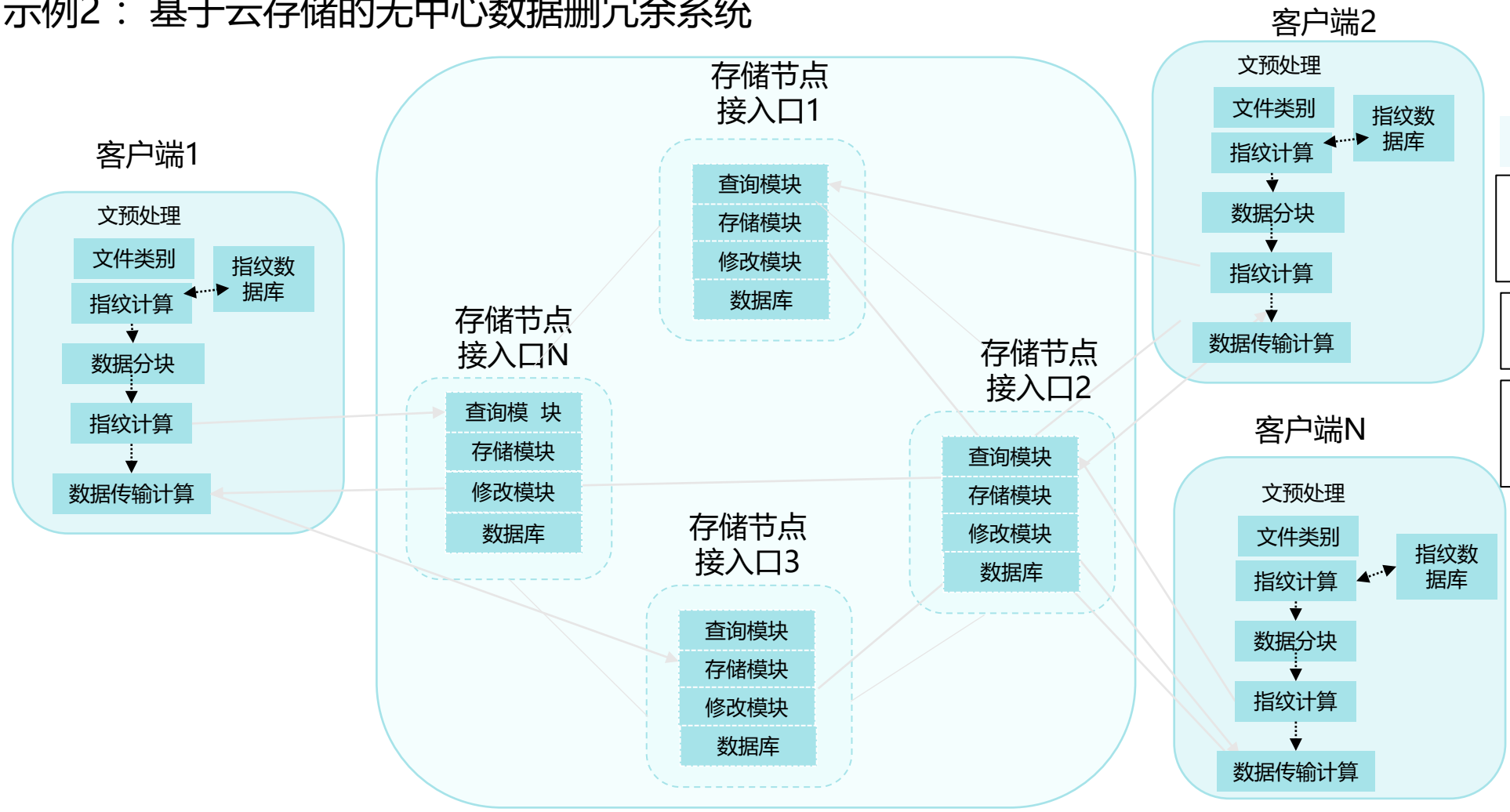
高效删冗系统三标准：删冗率、扩展性、IO吞吐率

高性能删冗系统基于：磁盘I/O，可扩展性、容错性、负载均衡

该删冗余系统是分布式架构

4、数据存储容灾系统的删冗技术-高效删冗的系统设计方法

示例2： 基于云存储的无中心数据删冗余系统



无中心删冗系统

无中心的架构相当于环形架构（各节点的地位相同。数据一般通过HASH散列来存放）

客户端：文件数据块分块、计算、传输给存储指纹值到节点。

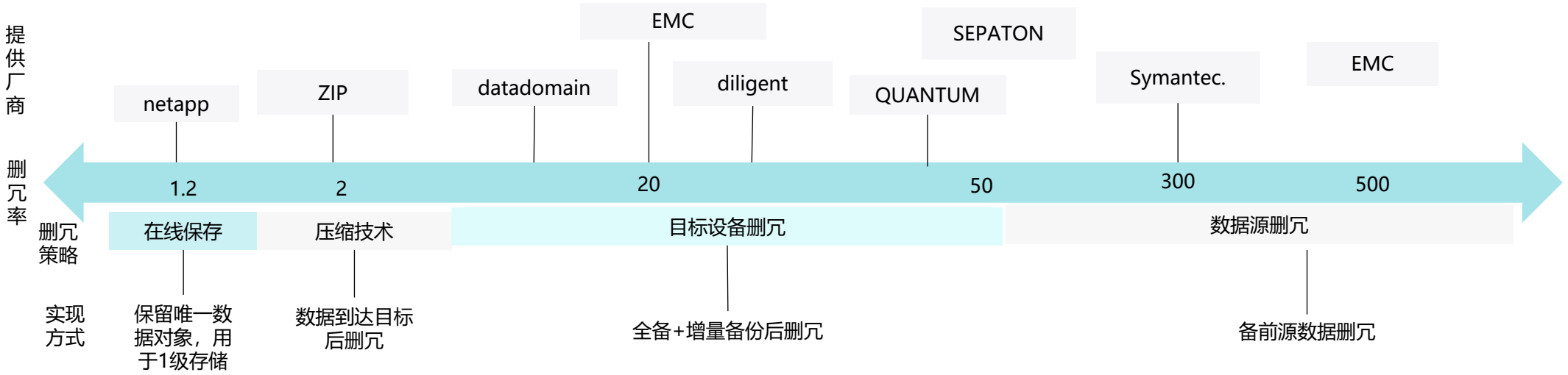
存储端：任意一个节点接受查询，并依据字典表，并存储数据块，修元数据

高效删冗系统三标准：删冗率、扩展性、IO吞吐率

高性能删冗系统基于：磁盘I/O，可扩展性、容错性、负载均衡

该删冗余系统是无中心分布式架构

4、数据存储容灾系统的删冗技术-主流删冗方案对比



单一存储与云存储的数据场景区别

项目	单一存储	云存储
存储设备	设备可靠性高，价格高，数量较少	设备可靠性低，价格底，数量超大
数据量	本地存储，数据量TB级	分布式存储，数据量可达PB级
备份时间	备份时间窗口大，空闲时间多	备份时间窗口小，用户要求响应度高
数据周期	历史版本较少，重最新数据	历史版本多且都需要保存
故障恢复	可接受较长时间恢复，分钟可接受	不接受恢复时间，要求秒级

数据删冗策略选型

删冗策略	业务类型	场景
边备份边删冗	网络传输	网络直播、视频监控
先备份再删冗	数据计算	科学计算
先删冗再备份/先备份再删冗	数据存储	云存储、WEB服务器

删冗效率因素

删冗率效率：数据块变长切块技术可提高删冗率
切的块越小删冗率越高

压缩率效率：数据块越小删冗率越高

云存储

4、数据存储容灾系统的删冗技术-数据删冗余技术的问题及预防

数据删冗本质是一个解决数据共享的问题，用于数据容量的效能管控

数据删冗余与**数据意外损坏**冲突

数据意外损坏的数据恢复

通常的解决办法：数据恢复技术

恢复方式

备份数据的恢复
快照恢复
复制数据的恢复
临时文件恢复
高级恢复

意外损坏低：但是解决办法依旧是适当的数据冗余还原法

- 删除技术与传统技术交叉结合保全数据

数据损坏的其它原因

介质老化 | 误操作 | 程序BUG | 感染病毒 | 天灾

数据删冗时注意事项

本身是受损的数据	原数据是坏的，冗余数据是好的，把坏的数据当好的。
应用与磁盘数据不一致	内存数据是好的，磁盘数据是坏的，冗余数据也是坏的

源数据校验技术

删冗产品校验能力 | 删冗方案与规划中的校验 | 原应用系统的本身数据校验

不适合数据删冗的场景

医疗影像 | 视频流（重复低） | 地球物理（重复低） | 高价值类（用户，金融）

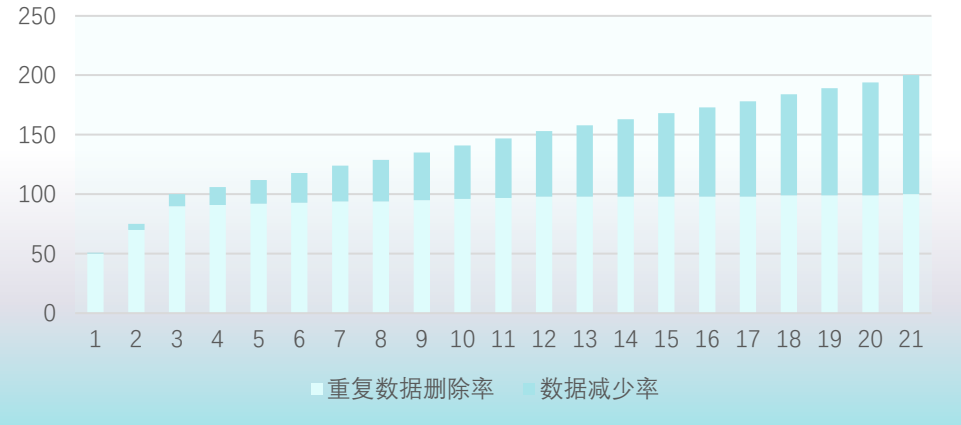
疑 问

有必要的数据删冗，数据冗技术也是必须掌握的，同样我们也需要掌握数据如何冗余，冗余度是多少，数据删冗技术的可靠性是多少？及工具功能的选择能力。

4、数据存储容灾系统的删冗技术-数据删冗余技术的问题及预防

数据删冗技术的其它因素

数据删冗与收益递减率并非正比关系



收益递减情况：重复数据删除越多，数据减少的收益就越少

重复数据删除与数据压缩

重复数据删除与数据压缩都是减少存储空间，过程与本质不同。

主存与辅存数据存储删冗方案应该不同

性能与恢复要求不同，实施方案不同。

数据删冗率原因多

数据类型　数据更新频率　数据保留期限　备份策略

分类与分方案可提高数据删冗效率

目录/CONTENTS

5、数据存储纠删码技术-数据存储纠删码技术简介

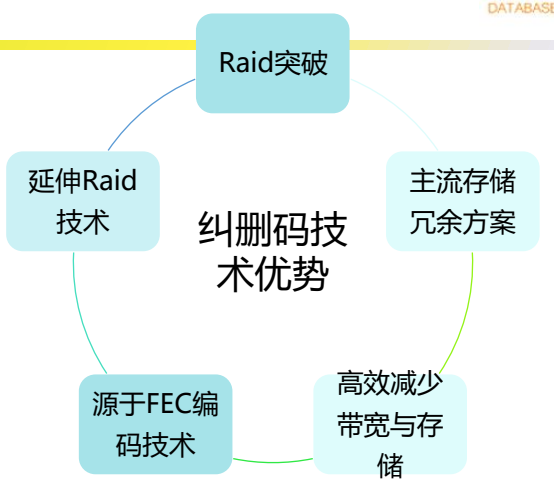
- 纠删除码技术从2001年开始
- 2010年到2020年间有最多研究与应用
- 随着存储技术成熟，2020年后研究创新性输出减少

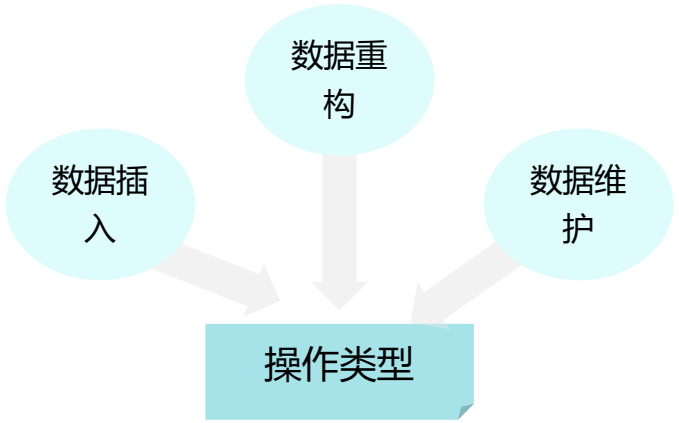
- 纠删除码本身技术与原理研究量最大
- 纠删除码技术多应用在存储系统，数据块等领域

5、数据存储纠删码技术-数据存储纠删码技术简介

分布式数据存储冗余性挑战

类　型	挑　战
数据可读性	保证可靠性前提下，简化编码结构，降低数据解码杂度
数据可读性	如何保证用户在多模式下的访问性能
数据维护通信量	如何减少参与修复的节点数同时降低每 个节点的上传数据量
数据维护通信量	如何保证数据可靠前提下，降低数据维护通信量
数据分配复杂度	如何简化数据可靠度与节点 可靠度之间关系，有效分析数据存储分配量
服务节点选择	如何不做迁移且满足可用性，尽可能多关闭节点
服务节点选择	如何解决访问概率与数据失效概率
负载均衡	如何解决各节点性能与数据均衡





纠删码技术

纠删码技术是一种数据恢复技术，是数据容错主要方案之一，它通过在原始数据中加入新的校验数据，使得各个部分的数据产生关联性。在一定范围内的数据出错情况下，通过纠删码技术都可以进行恢复。纠删码方法中的 RS 码(Reed-Solomon Code)是最广泛使用的一种编码方式。

纠删码术语

名词	英文	说　明
纠删码	erasure coding	前向错误纠错技术（FEC），根据纠删码算法与原始数据，算出冗余数据存储，保证数据可恢复性。
编码	enconde	计算出纠删码数据的过程
解码	decode	通过纠删码机制与纠删除码数据计算并恢复原始数据的过程
修复	repair	数据重建，从若干磁盘中恢复出若干磁盘的数据过程
MDS性质		保证N=K+M个磁盘中什么问题K个磁盘可以恢复出K个数据盘。是纠删码重要性质
系统码	systematic codes	编码后只包含校验数据，不包含原始数据；（信息位与校验位分开）
非系统码	non-systematic codes	编码后包含原始数据与校验数据（信息位与校验位交叉）
编码矩阵	Generator Matrix,GM	编码矩阵就是单位矩阵和范德蒙德矩阵的组合
数据块大小		按一定比例的原始数据或校验数据组成，总数据块 = 原始数据块 + 校验块
条带	stripe	是由若干个相同大小的数据块构成的序列，分为数据块和校验块
水平码		校验数据存放于单独的校验磁盘的编码方法，每个条带都是水平方式存储于n 个磁盘中，相同条带的数据块位置相同
垂直码		校验数据分布于所有的磁盘中，没有单独的校验盘，每个条带倾斜地将每个数据块分布于磁盘上不同位置上
原始数据	Original Data	原真实数据
容错率	Fault Tolerance Rate	m(纠删码块）/（K（原始数据块）+m(纠删码块））
冗余度		（K（原始数据块）+m(纠删码块））/k(原始数据块)
更新	update	原始数据修改，校验码跟着计算的过程

5、数据存储纠删码技术-数据存储纠删码技术的应用

主流云存储厂家使用的纠删除码策略

主流云存储厂商EC编码方式		
产品	使用EC方式RS（K,M）	冗余度(K+M/K)
Google GFS	RS(6,3)	1.5
Facebook HDFS	RS(10,4)	1.4
Microsoft Azure	LRC(12,2,2)	1.33
EMC ECS	RS(12,4),RS(10,2)	1.33,1.2
阿里云 盘古	RS(8,3)	1.375
Ceph	RS(10,4)	1.4

- 国内主流云存储（如华为云，腾讯云、青云）等也使用了EC纠删码技术
- 云存储中使用纠删码也主要是为了解决AZ之间的网络传输性能

纠删码正在研究的方向

容错转换机制	数据热时副本，数据冷时RAID5
	数据热时副本，数据冷时纠删码
系统支持多种纠删码并互相转换	权衡存储冗余度与读写性能

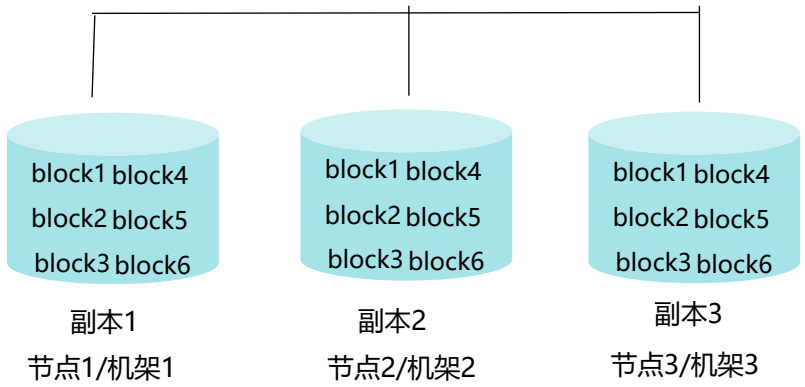
- 上述机制与策略还在研究中，未在生产应用。

5、数据存储纠删码技术-数据存储纠删码技术的应用-Hadoop HDFS

基于Hadoop 3.3.3

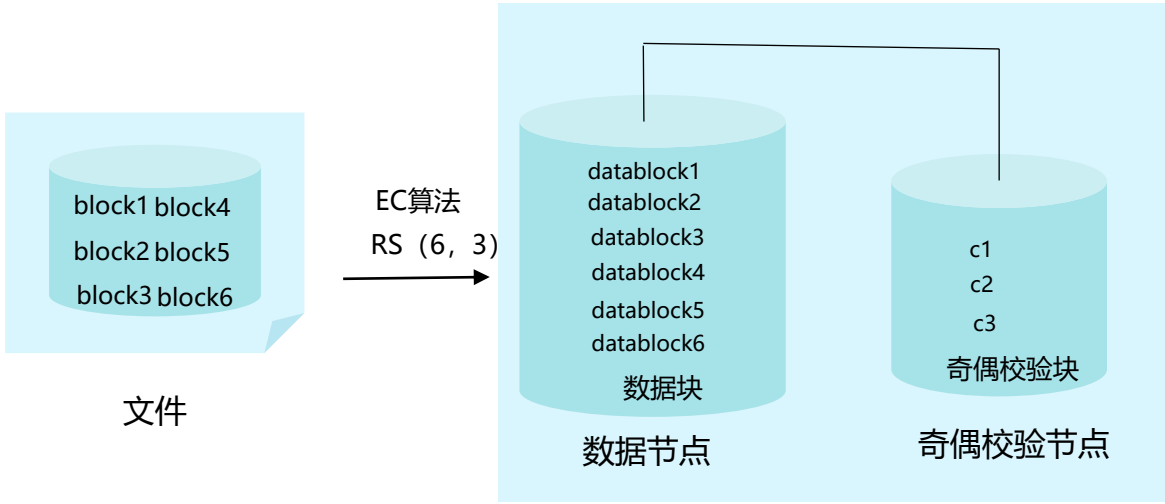
HDFS集成EC是为了提高存储效率

HDFS三副本机制



HDFS 默认大小在Hadoop2.x/3.x版本中是128M，1.x版本中是64M
3复本的复制因子为3，三个复制数据块需放三个不同机架
数据额外开销200%并因同步多占网络与IO开销
三副本是对等全量复制

HDFS EC机制



优 点

- 原占有6*3的数据块，经EC RS(6,3) 奇偶校验后只需要6+3个数据块空间
- 支持在线EC
- 自动将小文件发送到DATANODE中

HDFS EC局限性 某些HDFS操作不支持EC

即hflush，hsync，concat，setReplication，truncate和upress，不支持EC

查询不支持办法：客户端可以使用 StreamCapabilities;_API 来查询 OutputStream 是否支持 hflush（） 和 hsync（）

支持性

- 布局：支持带有条带化的EC。未来支持连续性EC
- 策略支持：RS-3-2-1024k、RS-6-3-1024k、RS-10-4-1024k、RS-LEGACY-6-3-1024k、XOR-2-1-1024k
- 混合支持：复制也支持，EC也支持，但有EC情况下，复制因子只能为1（即不能交错使用）
- 允许XML文件自定义EC策略
- 支持ISA-L 代表英特尔智能存储加速库（但需要OS级先开启）

5、数据存储纠删码技术-数据存储纠删码技术的应用-Ceph纠删码

Ceph支持的EC代码库

Jerasure erasure code（默认）
ISA erasure code
Locally repairable erasure code
SHEC erasure code
CLAY code

Jerasure erasure code提供一般的RS码和CRS码两种编码方式

编码库与三副本性能比较

比较项	三副本	RS（10，4）	LRC（10，6，5）	SHEC（10，6，5）
容量开销	3X	1.4	1.8X	1.6X
恢复开销	1X	10X	5X	5X
可靠性	高	中	中	中下

Ceph纠删码

Ceph pool默认是复制配置

最简单的EC配置是RAD5,至少3个主机

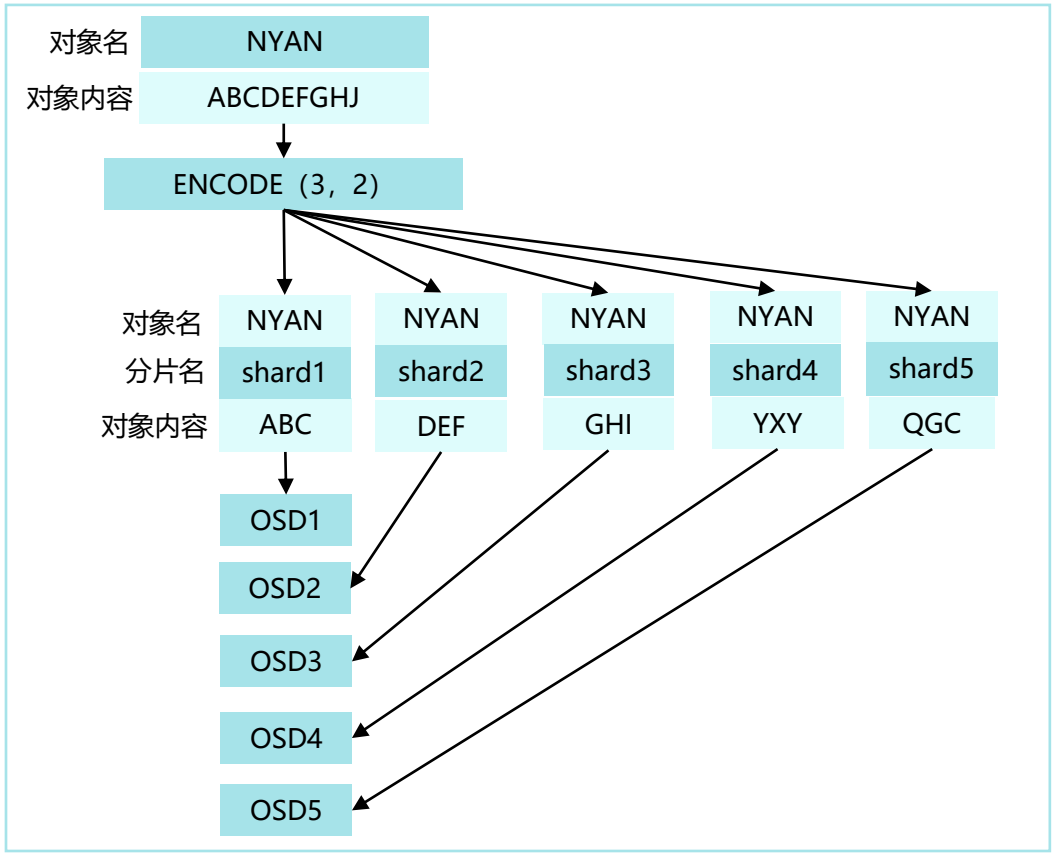
默认是副本复制模式，可以更改配置为EC

不是所有的应用都支持纠删码池，RBD 只支持副本池而 radosgw 则可以支持纠删码池

Ceph 从 Firefly 版本开始支持纠删码，但是不推荐在生产环境使用纠删码池

如果此时有数据丢失，Ceph 会自动从存放校验码的 OSD 中读取数据进行解码

Ceph纠删码逻辑流程图（通用）



流程说明

ENCODE(K,M），K=3，M=2

一般纠删码的分片最多冗余是K个，如果故障发生多于K时，将会真正丢失数据。（K是对象块数，M是冗余度），ENCODE（3，2）的情况下就是最多坏3个OSD，超过会丢失数据。

数据将在主 OSD 进行编码然后分发到相应的 OSDS上去

计算合适的数据块并进行编码

对每个数据块进行编码并写入 OSD

5、数据存储纠删码技术-数据存储的冗余技术对比

比较项	纠删码技术（N+M）	RAID技术	三副本技术
定义	纠删码将数据存储为数量众多的条带；每个条带包含数据块和校验块，并被放置于多个机架的多个物理节点上。	RAID通过条带化实现EC，它将逻辑上顺序的数据（例如文件）划分为较小的单位（例如位，字节或块），并将连续的单位存储在不同的磁盘上	三副本机制来保证数据的可靠性，每一个数据块被复制为3个副本，然后按照一定的分布式存储算法将这些副本保存在集群中的不同节点上。
本质	纠删码一种编码容错技术	RAID 是一种虚拟化技术（多磁盘管理）	数据同步技术
使用原因	降低数据冗余， 提高跨机房数据传输的网络使用率 降低数据冗余成本	提升整个磁盘效能 提升磁盘总性能、总容量 提升磁盘数据吞吐率、提升数据传输效率 通过数据校验提供容错功能	保持地理位置接近用户，从而减少延迟； 提高系统的可用性和鲁棒性， 通过扩展性来提供读查询，从而增加读取吞吐量
机 制	机制与 RAID 5/6 类似， 多位校验算法，节约磁盘空间；	磁盘数据条带化，并行读取磁盘数据，镜像或存储奇偶校验实现数据冗余	三副本数据节点存放位置 、数据一致性 、数据复制 数据恢复节点使用
关键技术	分片、编码、解码	镜像、条带、数据校验	协调、复制、元数据数据管理
特征	常见算法：Reed-Solomon（RS）； 参数RS(M,N)，M数据块，N校验块，最多容忍N块数据丢失 数据指纹	硬件故障隔离，避免了网络修复可能导致的稳定性问题； 可自动避让业务，保证业务无感知； 本地修复时数据延迟小，只使用本地RAID带宽，不消耗网络带宽； 抵御故障能力强，每个节点都能抵御一个或多个硬盘故障；而三副本最多抵御两个连续的硬盘故障。	存储系统自动确保3个数据副本分布在不同服务器的不同物理磁盘上 存储系统确保3个数据副本之间的数据一致
类型	EC,RS，LDPC、MDS	RAID0 、 RAID1 、 RAID3 、 RAID5 、 RAID6 和 RAID10。	/
磁盘利用率	大于60%	50% ~ 90%	30%
方式	数据分片、编码，数据传输、解码	物理磁盘合并成一个更大的虚拟设备	工具复制
可靠性	允许配置校验数的节点个数失效	只容忍磁盘故障，不能容忍节点故障。 一个RAID（RAID 5)组只容忍1个磁盘失效。	允许两个副本失效
数据重建	通过解码完成	例如RAID 5（校验码与数据放每个磁盘 ）， 数据均衡分布每个盘，通过校验算法重放数据。	直接从其它副本COPY
SSD磁盘寿命	总写次数少，SSD消耗少	SSD消耗最少，可延长SSD寿命	与次数多，SSD消耗多

5、数据存储纠删码技术-数据存储的冗余技术对比

比较项	纠删码技术（N+M）	RAID技术	三副本技术
优 势	低开销，高容错、高可靠	效率高，（个别磁盘坏，用户无感），大容量，高性能、可管理	写入效率高，无多余计算
缺 点	由于 E C纠删码存在比较严重的写放大问题，小块数据的写性能严重不足；随机写，特别是改写和重构(Rebuild)时产生的 I/O 惩罚较大	无法重构，无法代替备份；冗余数固定、不灵活	存储效率低，成本非常高、稳定性、木桶效应（IO不均衡）；冗余度高；
场 景	云存储、比如磁盘阵列系统、数据网格、分布式存储应用程序、对象存储或归档存储	企业服务器的标配	在虚拟化、私有云、数据库等块存储场景 主要应用于分布式存储：高性能计算、大数据视频云应用场、大数据分析应用场景
应用领域	EC主要运用于存储阵列、数字编码领域、P2P例如磁盘阵列存储（RAID 5、RAID 6），云存储（RS），大文件。（grid存储、peer-to-peer存储、云存储）	DAS，NAS，SAN	块存储，小文件；云存储。
计算开销	高	比纠删码小但比副本大	几乎无
网络销耗	较 高	无	较低
恢复效率	较 低	最低（EC比RAID5数据恢复效率高很多）	较高
应用限制	通常仅适用于视频等P2P场景、备份、容灾等对ＩＯ性能要求不高的业务场景	无	提供接口，支持裸设备及额外附加软件
扩展能力		仅适用于TB级存储	跨设备负载轻，支持PB级存储
故障恢复速度	恢复速度比RAID，比副本慢	磁盘更换，重建恢复周期长，恢复过程影响性能	故障可失效立刻转移切换，无需要等待
应用厂家	google,facebook,miscroft,emc,阿里云、华为，腾讯	如 EMC 、 IBM 、 HP 、 SUN 、 NetApp 、 NEC 、 HDS 、 H3C 、 Infortrend	EMC、IBM、HP、华为、XSKY、新华三、浪潮

RAID研究前沿：异构RAID技术（大规模使用是AFA技术）（All-Flash Array）
RAID 5/6是最简单的纠删码
其它冗余技术：镜像技术与快照技术

目录/CONTENTS

6、数据存储应用-超融合解决方案-综述

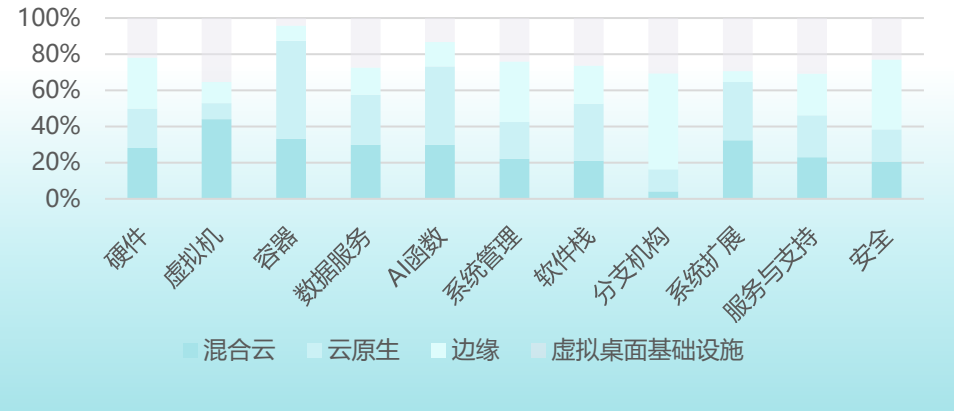
超融合基础设施 (HCI) 软件通过在服务器硬件上运行的单个实例提供虚拟化计算、存储和网络（Gartner)

超融合关键技术能力要求

硬 件	虚拟机	分支
硬件配置	支持hypervisor平台	远程办公、分支机构、边缘
第三方硬件能力优化	支持热迁移、快照及HA,DR高级特性	满足性价比，可用性，管理要求
硬件平台认证	多Hypervisors混合支持	
支持最新配件	混合应用场景支持	软件栈集成
网络支持无中断扩展		OS
硬件故障处理能力	容器	ERP，DB、BI
	支持Docker	NOSQL
数据服务	支持k8s	VDI
存储功能与备份	支持容器持久化	PaaS
容灾与高可用	支持云原生应用	
压缩与重删优化		服务与支持
带宽、延迟与IOPS优化	系统管理	打包模式、软件模式
性能与容量存储分层	监控、管理、故障诊断	监控、解决问题工具、处理流程
	部署、配置服务	边缘计算
AI 函数	API管理	
算法、机器学习（自动化）		
故障检测与纠正	系统扩展	安全
性能优化与通知	系统规模	角色、权限管理
	集群互联协议	
	计算与存储扩展	

数据存储超融合是一种系统组件融合化的产品
产品以超融合一体机、超融合服务器等
数据存储超融合形成一个超融合基础设施
基于超融合的研究热度持续中。
超融合关键词：超融合架构;软件定义存储;策略驱动;全闪存 数据中心; 超融合; 存储架构; 数据存储;虚拟存储;超融合基础架构； 超融合服务器

超融合关键技术能力在不同场景的权重

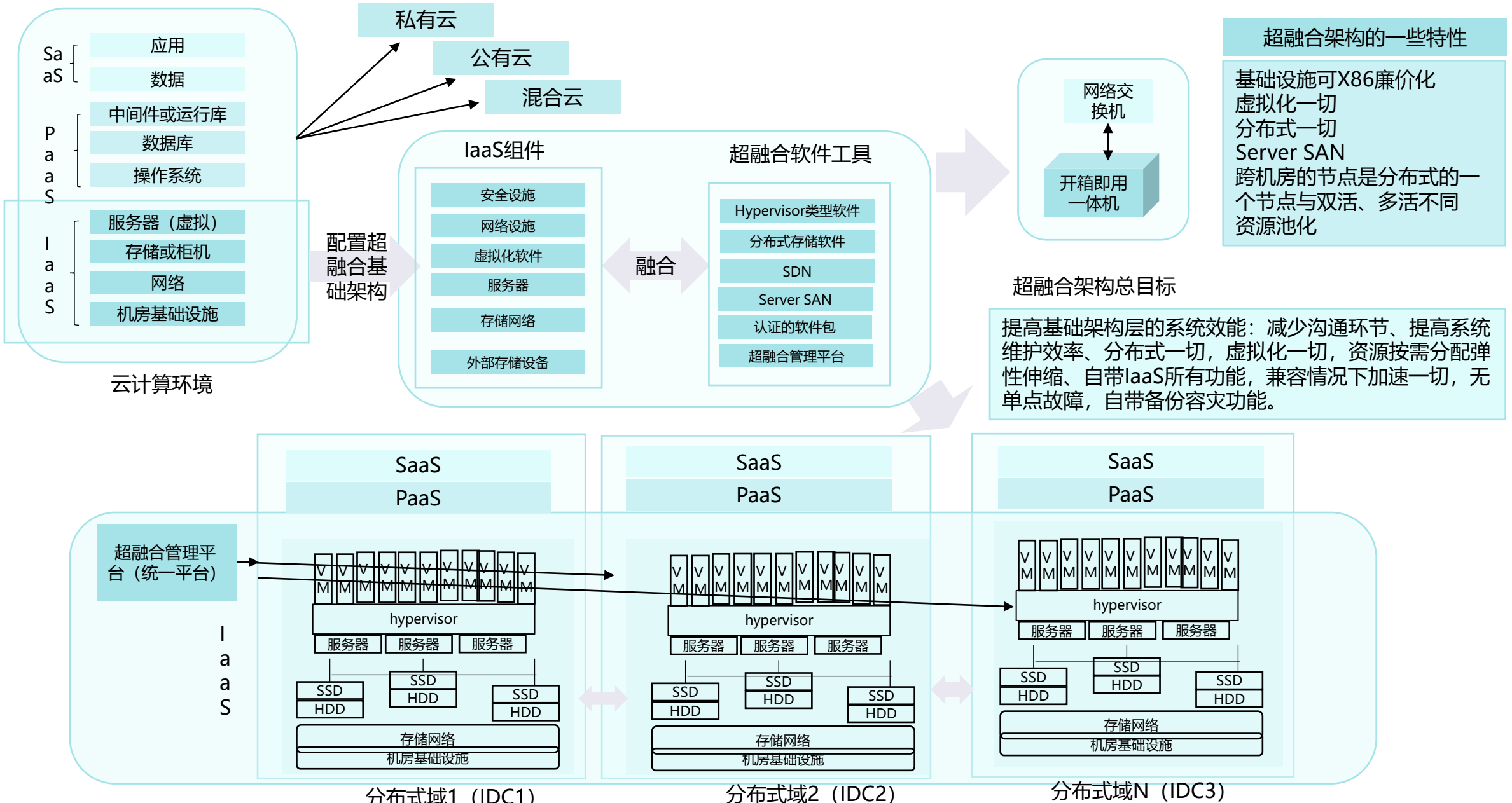


数据来源：GARTNER ,2021.11

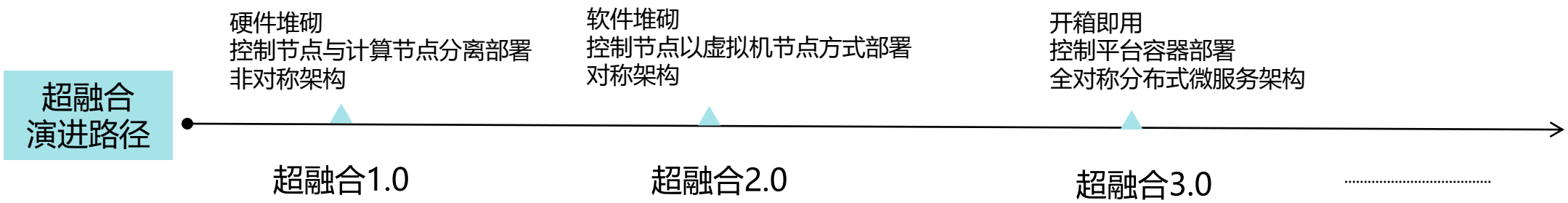
关键能力	混合云 hybrid cloud	云原生 cloud-native	边缘 edge	虚拟桌面基础设施VDI
硬件	9%	7%	9%	7%
虚拟机	15%	3%	4%	12%
容器	8%	13%	2%	1%
数据服务	12%	11%	6%	11%
AI函数	9%	13%	4%	4%
系统管理	12%	11%	18%	13%
软件栈	8%	12%	8%	10%
分支机构	2%	6%	26%	15%
系统扩展	11%	11%	2%	10%
服务与支持	6%	6%	6%	8%
安全	8%	7%	15%	9%
总计	100%	100%	100%	100%

数据来源：GARTNER ,2021.11

6、数据存储应用-超融合解决方案-超融合基础架构



6、数据存储应用-超融合解决方案-演进路径及优势与局限性



超融合优势（对比传统基础设施、普通融合、集成、云计算虚拟化）

比较优势项	描 述
按需要采购	初始投资小，可先在一个节点部署后期扩展
快速交付	几十分钟交付部署
管理极简	统一界面可视化管理
弹性扩展	无单点故障线性扩展
支持简单	单一厂商减少多层沟通
超强稳定	存储深度融合优化
兼容性强	软硬件出箱已兼容
维护极简	减少中间环节的沟通
组件可靠	认证的软件硬件
架构优越	IO本地化，提高访问速度（部分超融合产品）；
自带容灾体系	如容灾、恢复、快照功能齐全
数据容量自动均衡	数据变更后，容量快速恢复分布均衡
节约成本	资源高复用，系统融合度高
异构节点支持	部分超融合产品支持异构节点

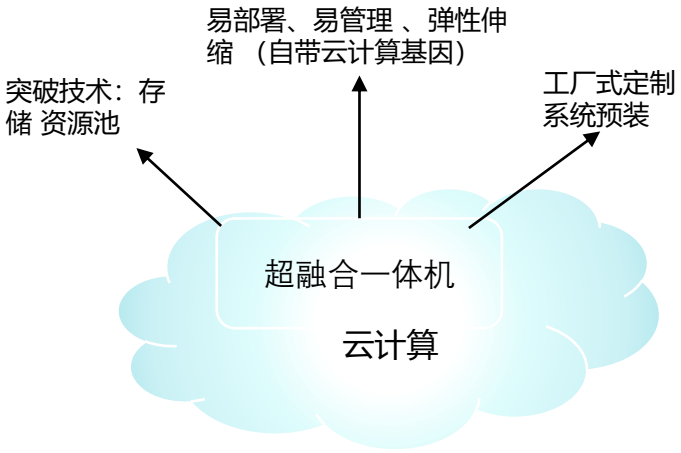
现代的超融合技术已与过往不同，可使用廉价X86机型也可支持裸金属机型。
超融合由IaaS、分布式存储、虚拟化技术发展的一种创新，是IaaS向前推进的重要技术路线。
超融合的本质在虚拟化基础上降低技术门槛，降低使用复杂度，并让用户用得起。

超融合局限性

局限性	描 述
与原有系统硬件融合	选择硬件要求高，与现有架构融合难度加大
与原有系统管理融合	难与现有资源统一调度与管理，规模有限
聚合性能并不简单	超融合架构的分布式支持业务性能复杂化
超融合架构有场景要求	更适合中小企业的私有云
超融合有特性要求	不适合计算密集型，容量密集型
超融合有应用要求	已在副本容灾机制，部署本身副本应用浪费（HADOOP等）
没有改变技术本质	传统采购模式，弹性不足，仅提供IaaS,整体云计算环境技术门槛高

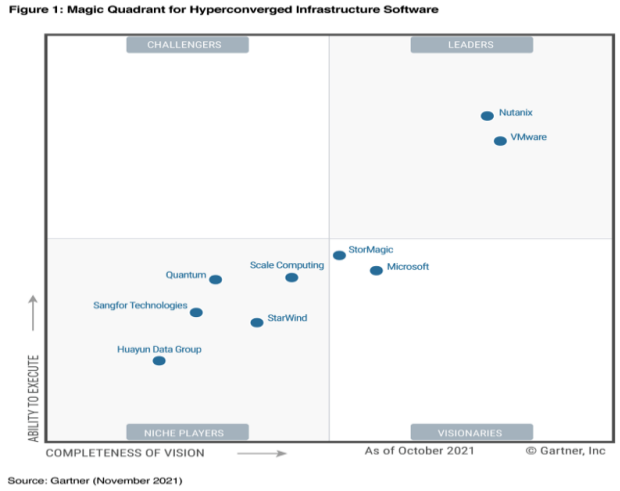
超融合并非全能，但优越性明显。
在局限性要求下，充分发挥超融合的场景，可以让企业即省钱，又可以省心，提高了系统的效能。

6、数据存储应用-超融合解决方案产品-超融合一体机



- 超融合一体机：以节点为单元横向扩展模式。云计算扩展要求灵活。
- 超融合一体机是云计算整个方案的一部分。不能完全满足所有云计算业务需求。

2021年超融合魔力象限（2021.11）



主流超融合产品

厂家	Nutanix	Vmware	EMC	StorMagic	Dell	华为	H3C	sangfor（深信服）	浪潮
总部	美国	美国	美国	中国	美国	中国	中国	中国	中国
产品名称	Nutanix	Vmware HCI	VxRail	SvSAN	Dell EMC VxRail	Fusion Cube	UIS-Cell	aServer2000	InCloud Rail IR5280M6
产品类型	一体机	软件	一体机	软件	一体机	一体机	一体机	一体机	一体机
管理平台	Prism	vCenter	vCenter	SvSAN	vCenter	Fusion Cube Center	CAS	OpenStack Horizon	InCloud Rail
软件定义存储	NDFS	VSAN	VSAN	服务器SAN	VSAN	Fusion Storage	基于Ceph	基于Glusster FS	incloud Storage

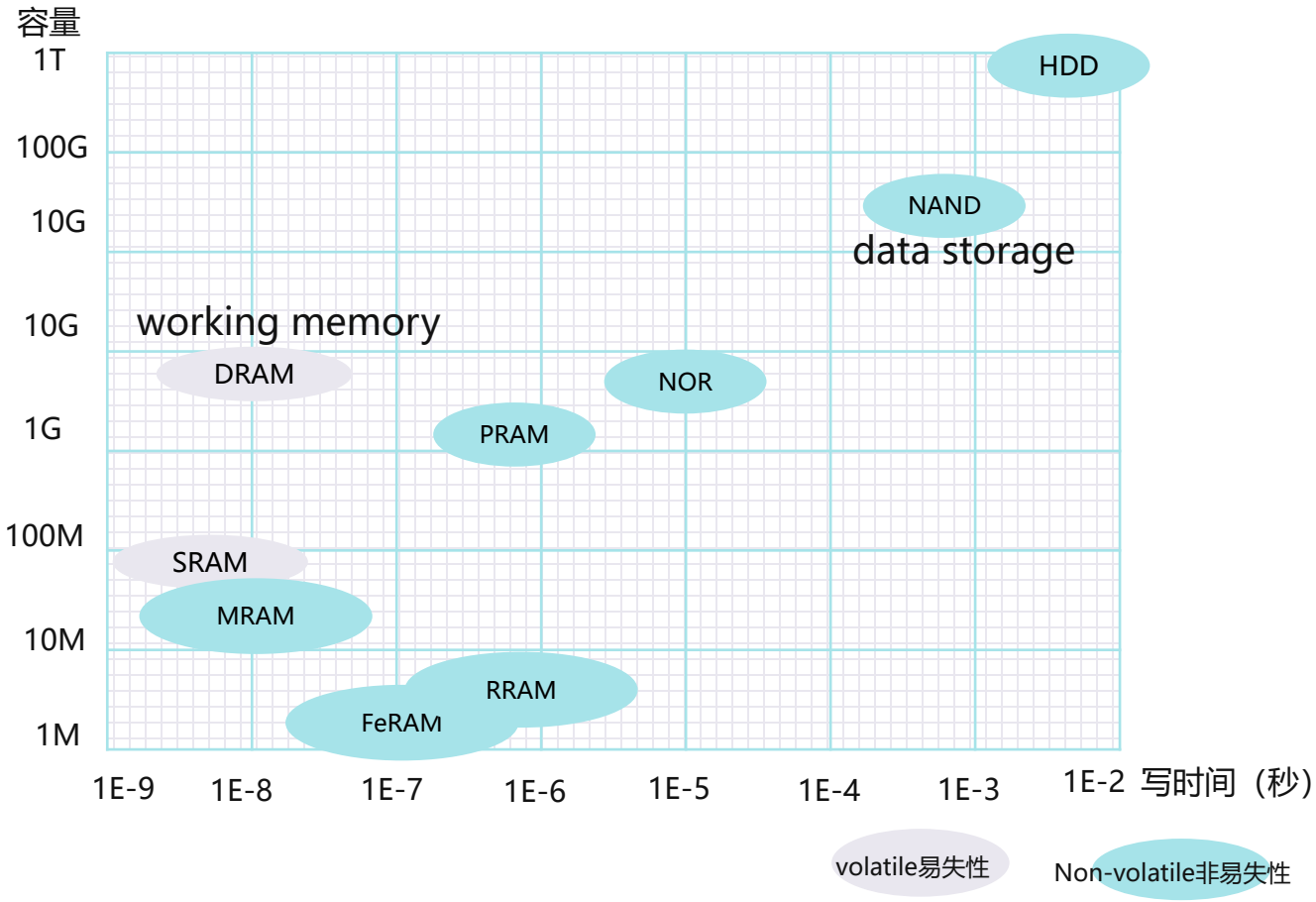
目录/CONTENTS

7、数据存储的未来-下一代存储的迷思

下一代存储技术：存储器、存储服务器/存储方案、存储服务

下一代存储器（侯选）		
碳、磁、铁电、阻变、相变等（半导体）	CBRAM	导电式随机存储器
	NRAM	纳米随机存储器
	CeRAM	电阻式存储器
	STTRAM	自旋扭矩转换随机存储器
	RRAM/ReRAM	阻变存储器
	PCM/PRAM	相变存储器
	3D Xpoint	3D磁存储器
	FeFET	铁电栅场效应晶体管
	SRAM	静态随机存储器
生物存储器	蛋白存储器	
量子存储器	基于镱的金属等	
全息存储器	基于银盐等	

分类	MRAM	SRAM	DRAM	FLASH	FeRAM
读速度	快	最快	中	快	快
写速度	快	最快	中	低	中
阵列效率	中/高	高	高	中/低	中
可升级能力	好	好	有限	有限	有限
单元密度	中/高	低	高	中/低	中
非易失性	是	否	否	是	是
耐用性	无限	无限	无限	有限	有限
单元泄漏	低	低/高	高	低	低
低电压	是	是	有限	有限	有限
复杂度	中	低	中	中	中



上游存储器决定中下游的方案

总体存储器需要很长时间才能代替

新型存储器发展对未来更有意义，也可能是颠覆式的

7、数据存储的未来-DNA存储的发展现状

需求侧现状	产生数据越来越快，数据越存越久
问　题	现使用的存储技术针对需求现状越来越有局限性（速度，容量、保存周期、成本，大规模并行复制与处理等）
解决问题方案之一	DNA存储技术，使用生物学解决信息学问题，是跨界关联性高效方案

DNA 信息存储通过编解码、合成、编辑和测序等过程，实现数字信息写入、存储与读出

维 度	现有传统存储（磁性、光学、固态）	DNA存储
密 度	传统硬盘存储每立方厘米为10的13位（bits），内存为10的16位（bits）	DNA存储是10的19位 （1克DNA能装2.2亿部高清电影）
寿 命	10多年	至少上百年甚至千年

古生物科学研究表明， DNA 保存的基因数据在没有特别人工干预的情况下能保存万年之久

时　间	DNA存储进展
1964	DNA信息存储概念首次提出
1964-2012	不断有实验数据输出：数字，文字 ，诗歌，图片，歌曲等写入DNA并数据恢复
2012	数字化转换编码（霍夫曼编码等）、重叠法、等不同格式存入DNA并恢复，确定可行性
2016	基于DNA的存储系统体系架构（A DNA-based archival storage system）
2017	证实DNA可存放视频短片
2019	存储 1000MB的数据到DNA并实现了提取
2019	dot(dna-of-things)存储架构构产生
2019	DNA喷泉码压缩算法
2020	基因组录音机
2020	澳利用保存运动图象
2020 年 11 月	第一个 DNA 数据存储联盟

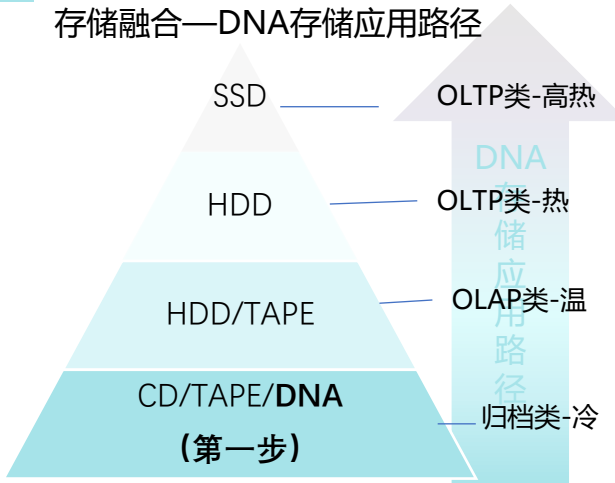
DNA存储的数据操作流程

DNA存储编码技术：
哈夫曼编码，喷泉码、LZMA、
纠错码：汉明纠错码、RS码纠错、LDP码纠错

我国对DNA存储的研究处于起步阶段，于2018年开始研究扶持（合成生物学）。

参与DNA研究的企业：华为，华大基因
参与DNA研究的学校：东南大学，华中科技大学，天津大学，国防科技大学，及军事院校。

存储融合—DNA存储应用路径



现阶段DNA存储的研究工作

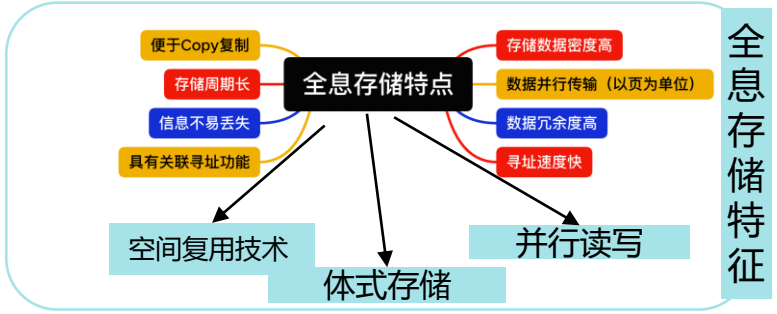
- DNA存储的研究已上升到国家战略，是现阶段研究的热点

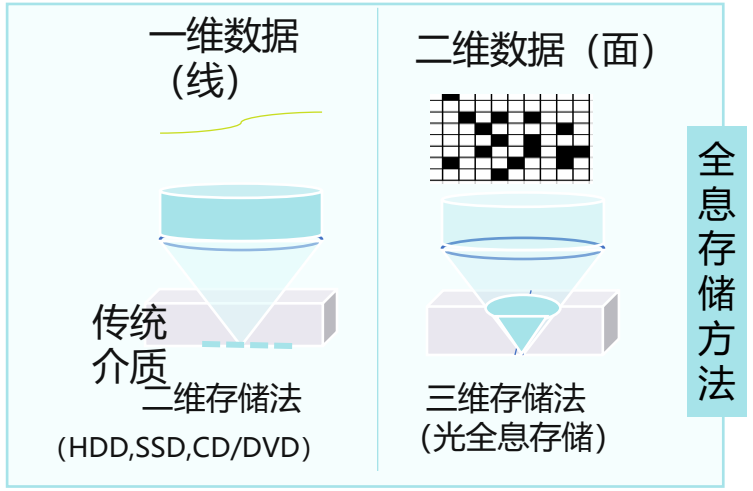
优点	缺点	产品形态	未来落地
存储密度大，稳定性好、能耗低、存储时间长（寿命）、易备份（PCR技术）、抗电磁干扰，维护低成本	成本高，存取耗时长，技术难点多。信息检索与操作有限制类：归档数据	DNA硬盘 DNA光盘 DNA磁带	技术预备性，可实现规模性，鲁棒性，可编程碱基配对支持分子计算与数据库操作的可能性。因此DNA是最有效的分子存储材料 。但落地至少还得5年以上，DNA存储潜力巨大

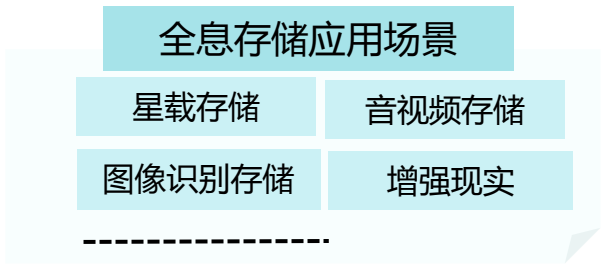
编码方式	充分利用DNA存储能力的高效编码方式
纠错机制	读写存高保真的纠错机制
生物机制	DNA本身的批量、准确合成技术确保DNA对数据的保存。
随机存取	存取的能用性

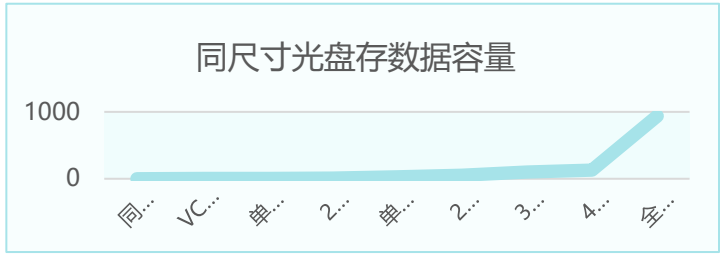
7、数据存储的未来-全息存储发展现状

全息存储产品：处于研发阶段，有企业正在参与技术研发（主力是高校）

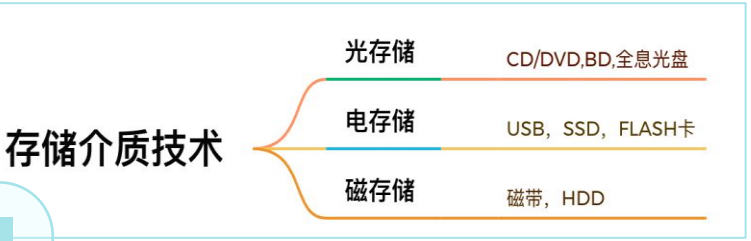






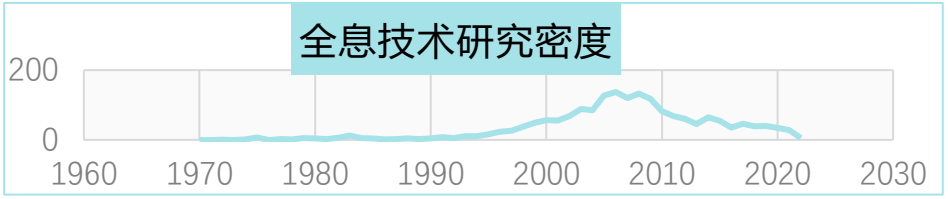


- 同尺寸光盘全息光盘是普通DVD的200倍



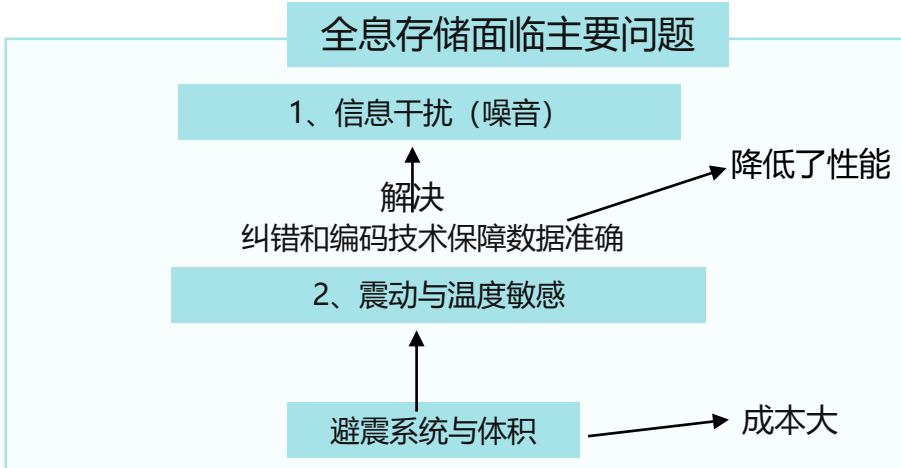
日本Optware公司 为代表的同轴全息数据存储系统和以 InPhase 公司为代表的离轴全息数据存储系统并 行的市场化探索格局。（两个2010年左右破产）（同轴更有利）

Facebook 于 2014 开始就着手建造总容 量可以存储 1000 PB 的蓝光光盘数据库（能耗比磁盘存储低80%，保存时间30-50年，是磁存储5-8倍）



体全息存储试验样机演示的最大存储密度大约为２．４Ｔｂ／ｉｎ２（１mm 厚存储材料），该值比理论极 限值４０Ｔｂ／ｉｎ２小一个数量级

全息存储的一种：体全息存储（相位全息存储，偏振全息存储。是未来的一种方向可能。



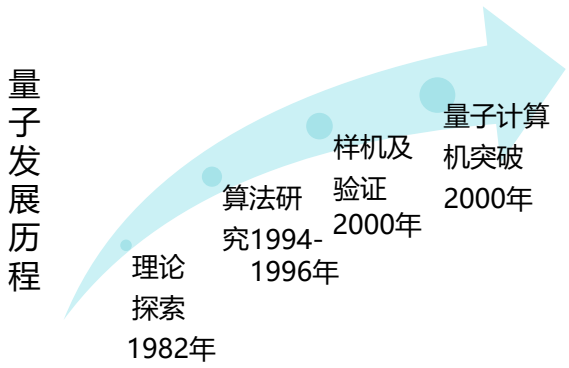
全息存储具有未来可能性

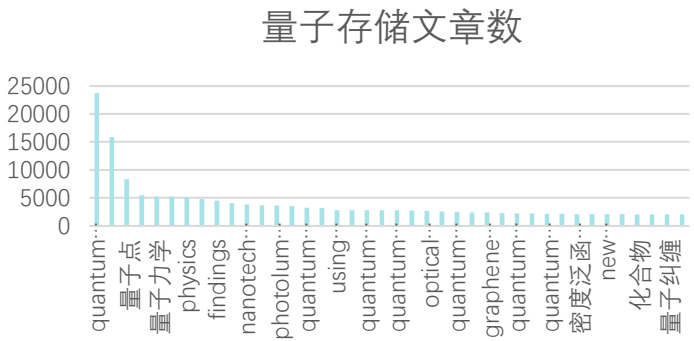
如全息存储商业化，将会代替BD类存储

同轴全息光存储的基础理论与关键技术研究（2019国家重点技术）

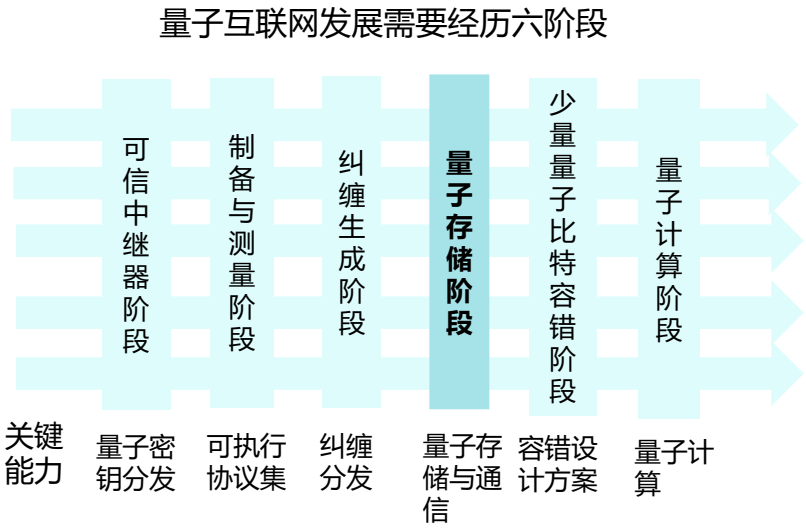
在光存储中，全息存储还在研发，是下一代存储的可能技术

7、数据存储的未来-量子/量子存储

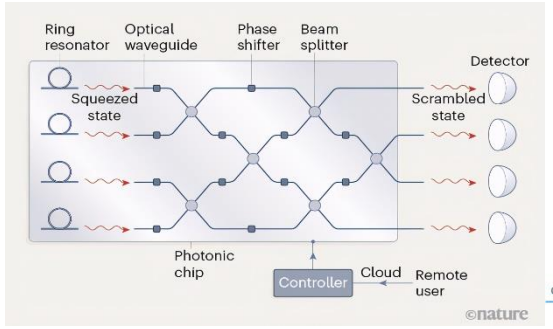




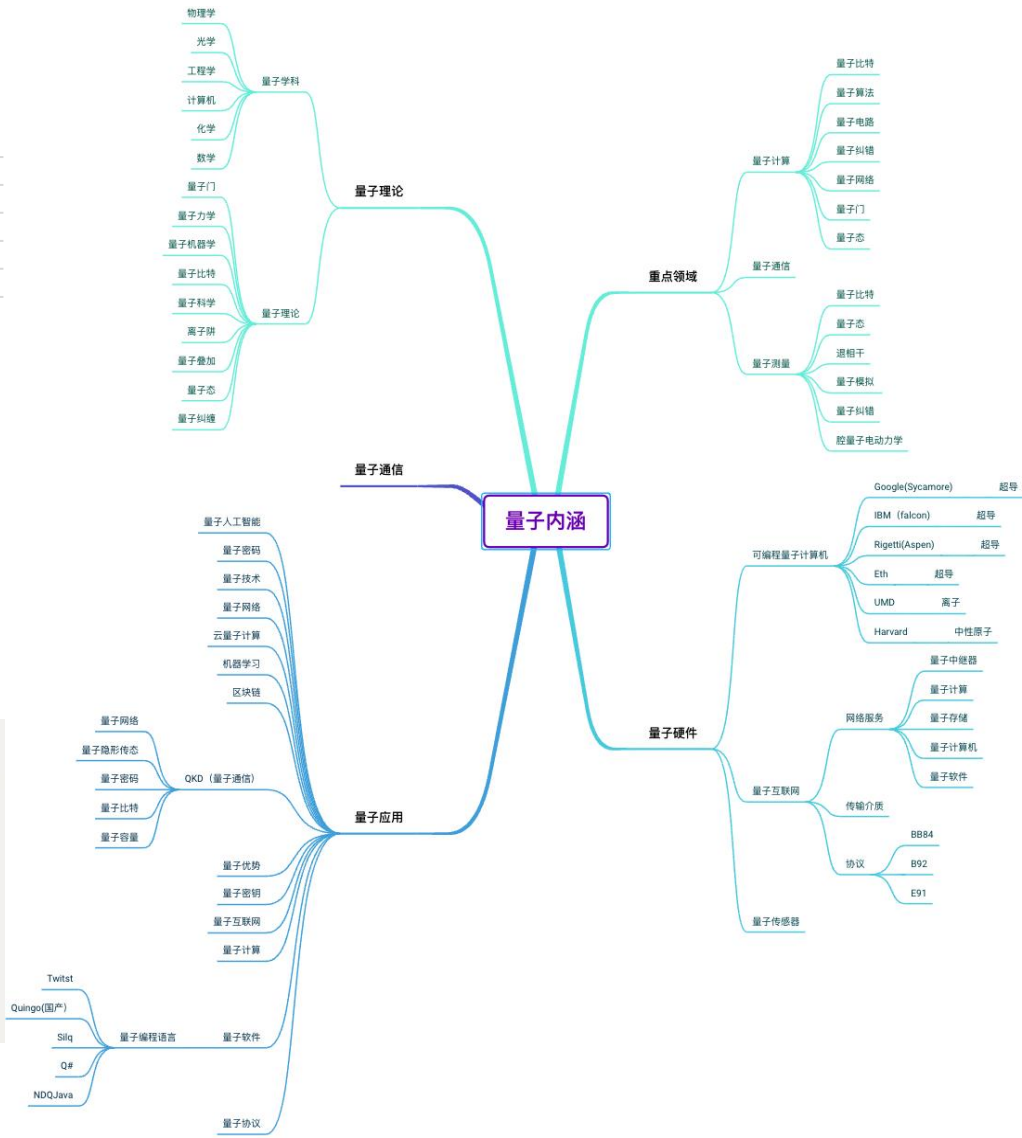
来源于论文知识库



重点关注：量子计算与量子通讯领域



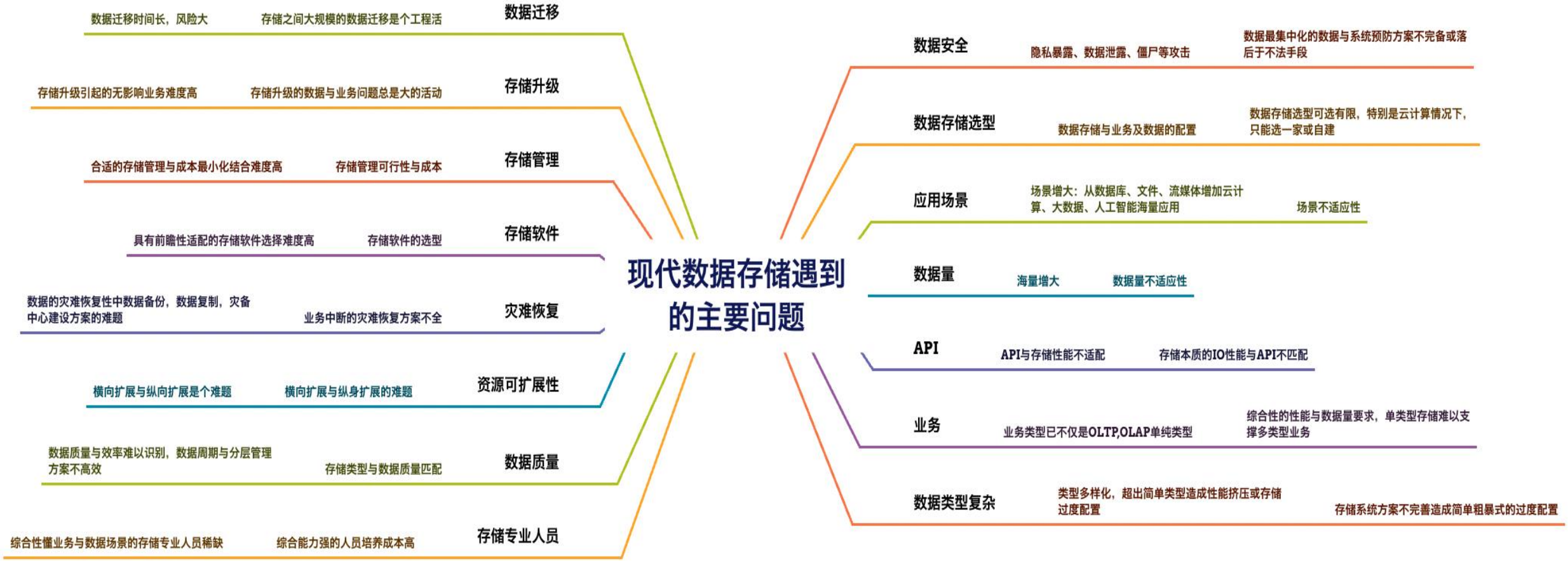
光子芯片上实现的量子算法 Arrazola《nature》



我们着重关注：量子计算领域，在云计算与人工智能等发展成熟的情况 下，量子如何颠覆性的替代我们正在使用的各个组件与元素，实现效率极大化，是我们时刻关注的。

7、数据存储的未来-现代数据存储遇到的主要问题

现代数据存储遇到的主要问题：存储措施面对突如其来的大规模数据具有短期不适应性



存储的技术影响到整个业务（大规模）：需要整体统筹与解决

即使是云存储，需要考虑提供云存储的厂家实力或者及业务适配性

数据使用方便性与数据存储方案有关


openGauss
OceanBase
ArkDB
RASESQL
HotDB
StellarDB
QianBase xTP
云树Shard
GoldenDB
DolphinDB
MatrixDB
DynamoDB
SinoDB
FastData
Galaxybase
KunDB
SQL Server
vertica
D B 2
GBase
Oracle
达梦数据库
Spacture
开务数据库
UbiSQL
神舟通用
TDSQL
TiDB
Tapdata
StarRocks
KingbaseES
2010
2014
2018

THANKS