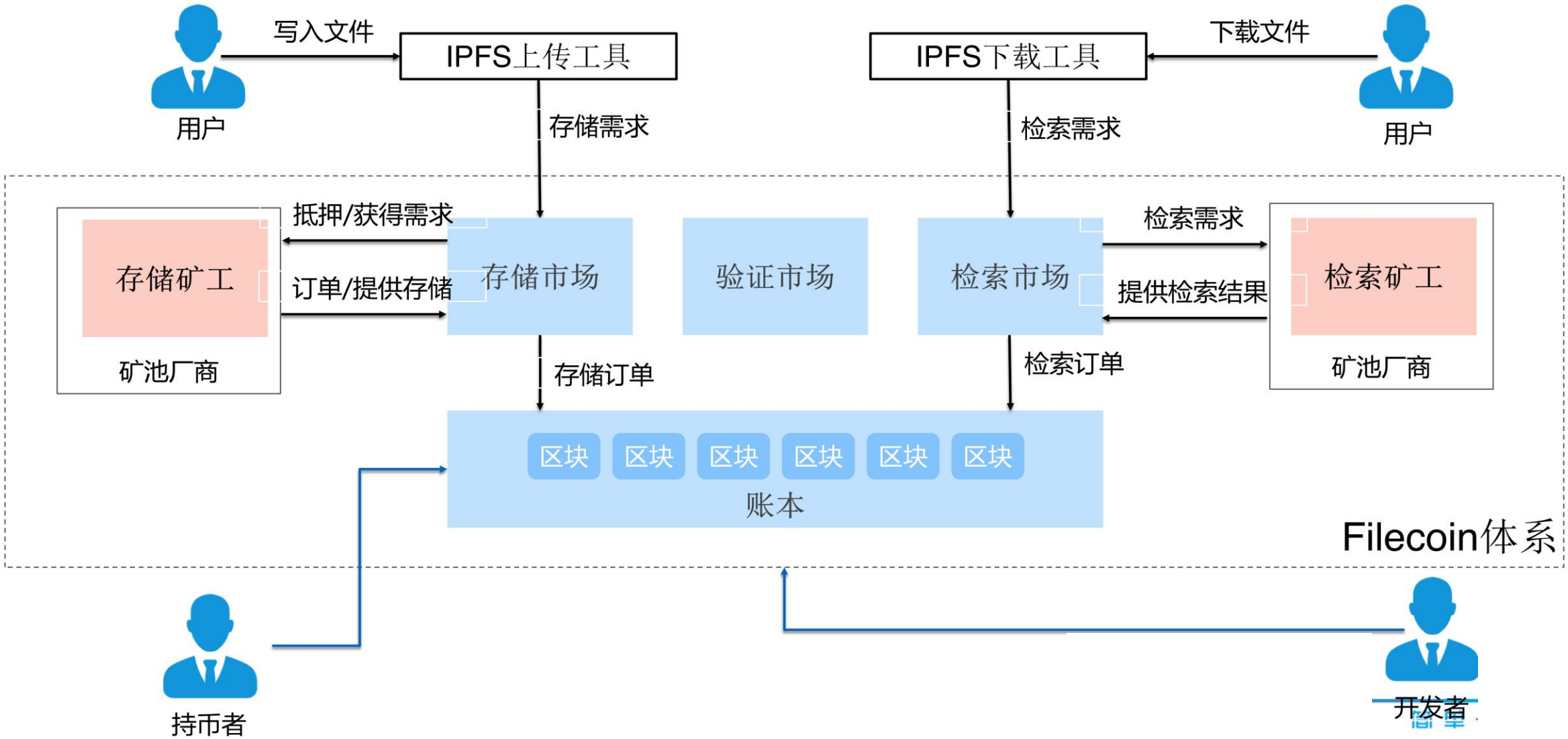
CCIT ECT一站式解决方案

- 一. Filecoin体系结构
- 二. IPFS存储需求以及部署
- 三. IPFS存储应用特点
- 四. 算力软件
- 五. 网络带宽
- 六. 运维
- 七. 收益
- 八. CCIT提供一站式服务(数据中心、系统软件、方案设计、咨询顾问、维护.)

一.Filecoin体系结构

能力&目标



Filecoin是什么

定位&角色

Filecoin是一个开源的云存储市场，协议和加密货币

- 目标是要创建一个功能强大的分布式动态云存储服务平台
- 直接竞争对手是亚马逊云、阿里云、腾讯云、七牛云、金山云等云存储服务商

角色	收益/用途	能力/要求	备注
用户	上传、保存、下载文件	基于IPFS协议	IPFS是基于HTTP的BT协议
矿工	区块打包奖励，存储收益，检索收益	提供相应资源池	前期主要收益在区块打包奖励，后续体系运转后在于存储和检索
持币者	币的增值	通过Filecoin官方购买FIL币	FIL币发放受控，矿工被分配代币
开发者	平台运营收益	研发能力	Filecoin平台开发/运营者

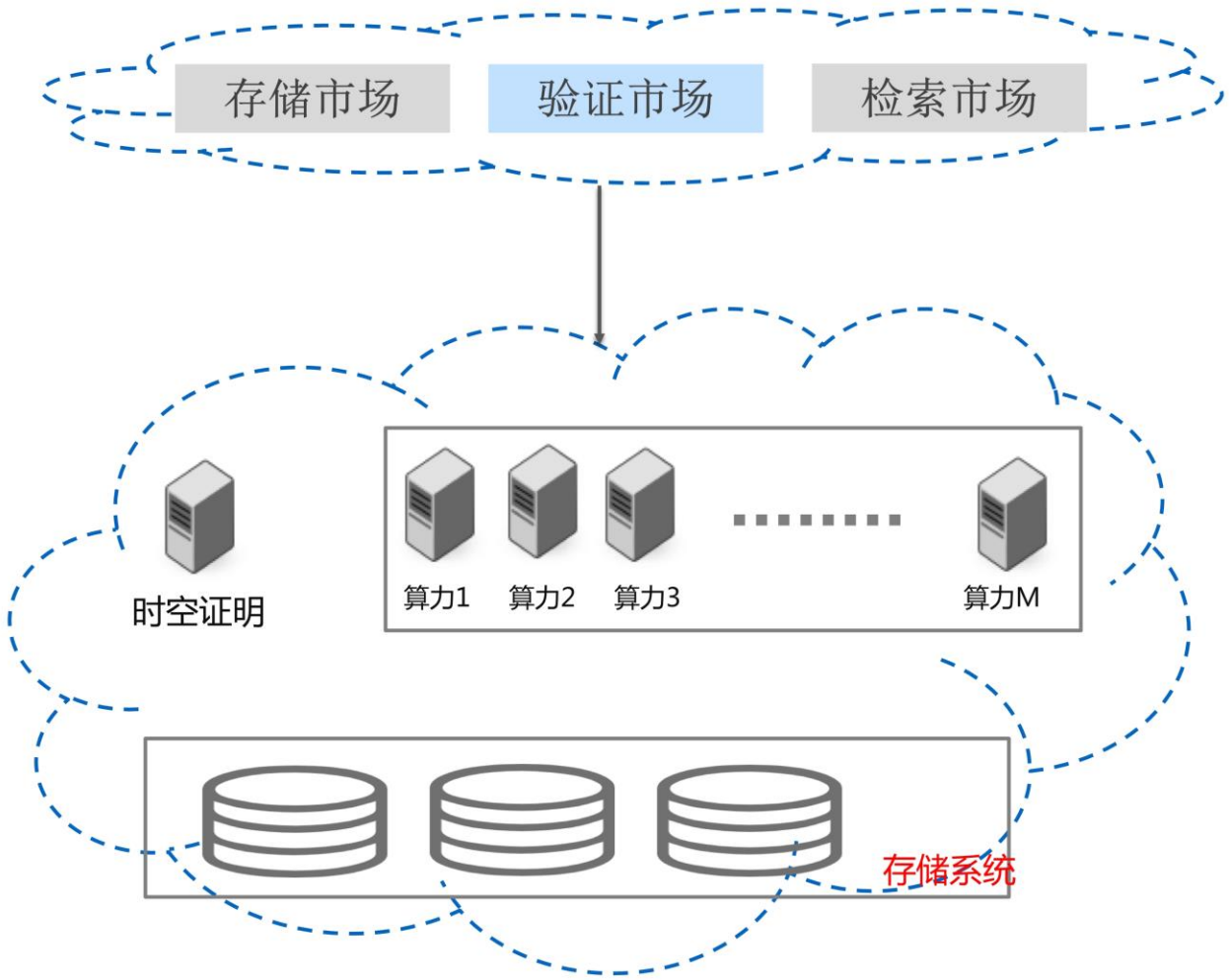
矿池是什么

矿池等厂商的定位

类别	来源	作用/目的	备注
服务器	自己采购，客户采购	提供矿池物理资源能力	
Filecoin软件	Filecoin官方，开源软件	实现整个体系的软件系统	位于github
矿池	运维&运营	接入到Filecoin体系的资源方（BT的超级种子）	资源单元，用户可指定文件冗余参数，将数据写入多个矿池

典型矿池技术方案

背景&需求



Filecoin市场

- 验证市场：验证矿工存储使用情况（发起惩罚）
- 存储市场：发起/管理待封装的存储与文件
- 检索市场：文件下载市场

矿池厂商/资源

- 时空证明：每天48次全量证明，检索并校验存储资源
- 算力：区块打包，将数据编码打包，通过创建格式化的区块，获取系统代币(区块）奖励。
- 存储系统：持久化”算力” 节点计算的区块相关数据，并为时空证明节点提供检索支持

存储系统—额外说明

- 一般用分布式文件系统实现，如七牛kodo、ceph、moosefs等；
- 未标准化：每个矿池厂家的实现均有差异
- 数据丢失问题：矿池丢数据后做相关惩罚/丢算力
- 存储规模问题：存储技术受限，很难突破10PB

增强—专业/技术语言

时空证明&区块

时空证明-window post

- 证明存储正常，保证不被惩罚
- 每天48次，向验证市场汇报全量的证明材料
- 单次证明：算力节点完成区块封装后通知证明节点，证明节点进行验证
- 单次任务任务：
  - ✓ 文件检查：遍历，确认文件在不在（提前做）
  - ✓ 读64字节的文件：
  - ✓ 获取偏移量：读9MB文件8个，第一次；9MB文件8个，第二次
  - ✓ 10*8次读取32G大文件中的32Byte校验数据
- 全量证明任务
  - ✓ 遍历所有扇区，一个分区2349扇区
  - ✓ 在一个扇区内，执行单次证明任务的数据读取与证明过程
  - ✓ 一次全量认为要求30分钟内完成
  - ✓ 执行证明任务的同时，算力节点持续写入封装区块

封装&算力&出块

- 完成区块封装，封装完成后将数据写入存储系统（归档），并向时空证明节点汇报（winning post 66*8）
- 一个任务产生的文件
  - ✓ 一个32GB大文件：含打散后的原始数据，校验数据
  - ✓ 1个64字节文件 + 8个9兆文件：元数据、校验数据等
- 算力规模：一般一个矿池，数百乃至上千台算力节点
  - ✓ 存储集群写入带宽/吞吐 要求很高
  - ✓ 持续大压力数据写入
- 出块/区块：根据算力规模，由系统分配（抢）出块权，从而获取filecoin（收益）
- 数据结构与相关概念
  - ✓ 存储碎片：是客户所存储数据的分配单位，数据是可以任意划分为许多片，并且每片都可以有不同集合的存储矿工来存储。
  - ✓ 扇区：是存款矿工向网络提供的一些磁盘空间。矿工将客户数据的碎片存储到扇区，并通过他们的服务来赚取令牌。
  - ✓ 密封：存储矿工为未来的证明准备碎片；算力节点的任务；
  - ✓ 证明：存储矿工证明他们正在存储所承诺的碎片（数据）

二.IPFS的存储需求

应用场景特点以及与传统存储应用的差异

大吞吐，封装/出块效率的保证：IPFS算力集群庞大，worker节点需要及时写入大文件，常规要求100Gbps+吞吐，以保证封装数据的**写入效率**

随机读，证明效率保证：每日48次（半小时）一次的全量证明，要求平台良好的随机读取效率

单写全读模型：无热点数据，存储分级、缓存等读写优化技术失效

超大容量：IPFS要求数据持续可读（证明），不删除，周数据数PB级，年度数百PB级至EB级别

数据安全&低TCO：年百PB数据量长久可访问保存，需要高出盘率，以降低存储成本

大规模集群的高效运维：数百PB至EB级存储系统的高效运维

云分布式海量存储系统

无需关注底层技术，安心聚焦IPFS存储应用

百万级

长期信赖的海量客户

9年+

完全自主研发的核心技术

10000亿+

总文件数

高可靠

达到 11 个 9 的可靠性

易扩展

平滑扩展至 EB 级别或更多

低成本

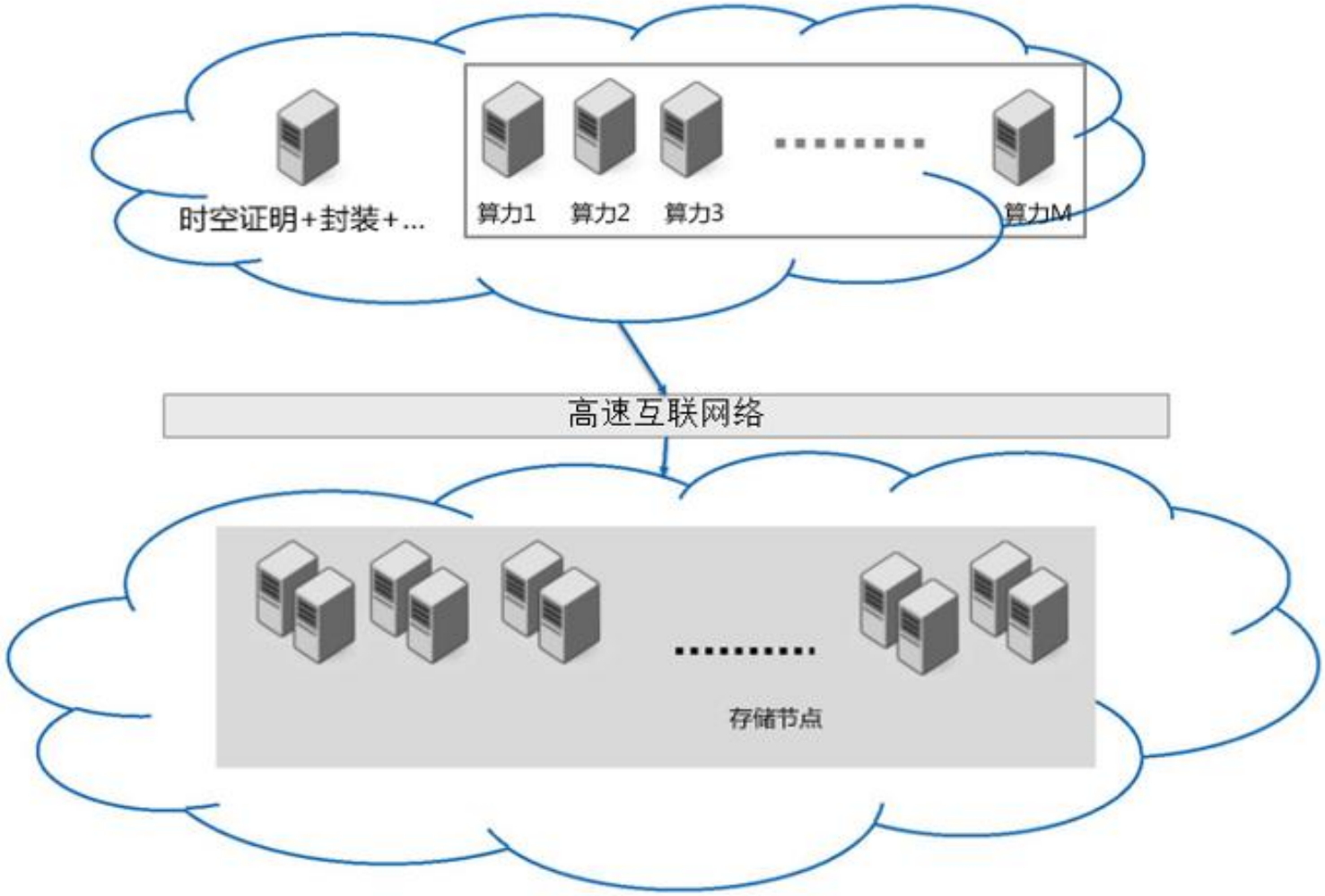
冗余度低至 1.14

数据智能化

弹性智能数据修复，全局恢复耗时显著降低

可随需求扩展的对象存储方案

纯软件方案 / CCIT一体机方案



算力集群

- 时空证明节点：每日多次遍历存储并随机读
- 封装节点：挖矿产生文件，并写入存储集群
- ...

存储集群

- 存储节点*N：单节点磁盘配置 16TB SATA*36；
- 可随需扩容，建议不少于 36 台；
- 支持小集群模型，弹性适应大/小 集群规模；

弹性集群规模

灵活应对存储小集群、大集群场景

存储一体机

一体化管理运维
开箱即用
快速上线扩容

小集群场景

- 最少4台，4+2:1冗余：支持任意故障1台服务器，或任意两块磁盘，不影响数据写入与时空证明
- 集群规模可按需配置：根据出盘率、故障容忍性与存储容量/性能需求，设计集群规模与EC模型，如6+2，8+2，15+3等。

大集群场景

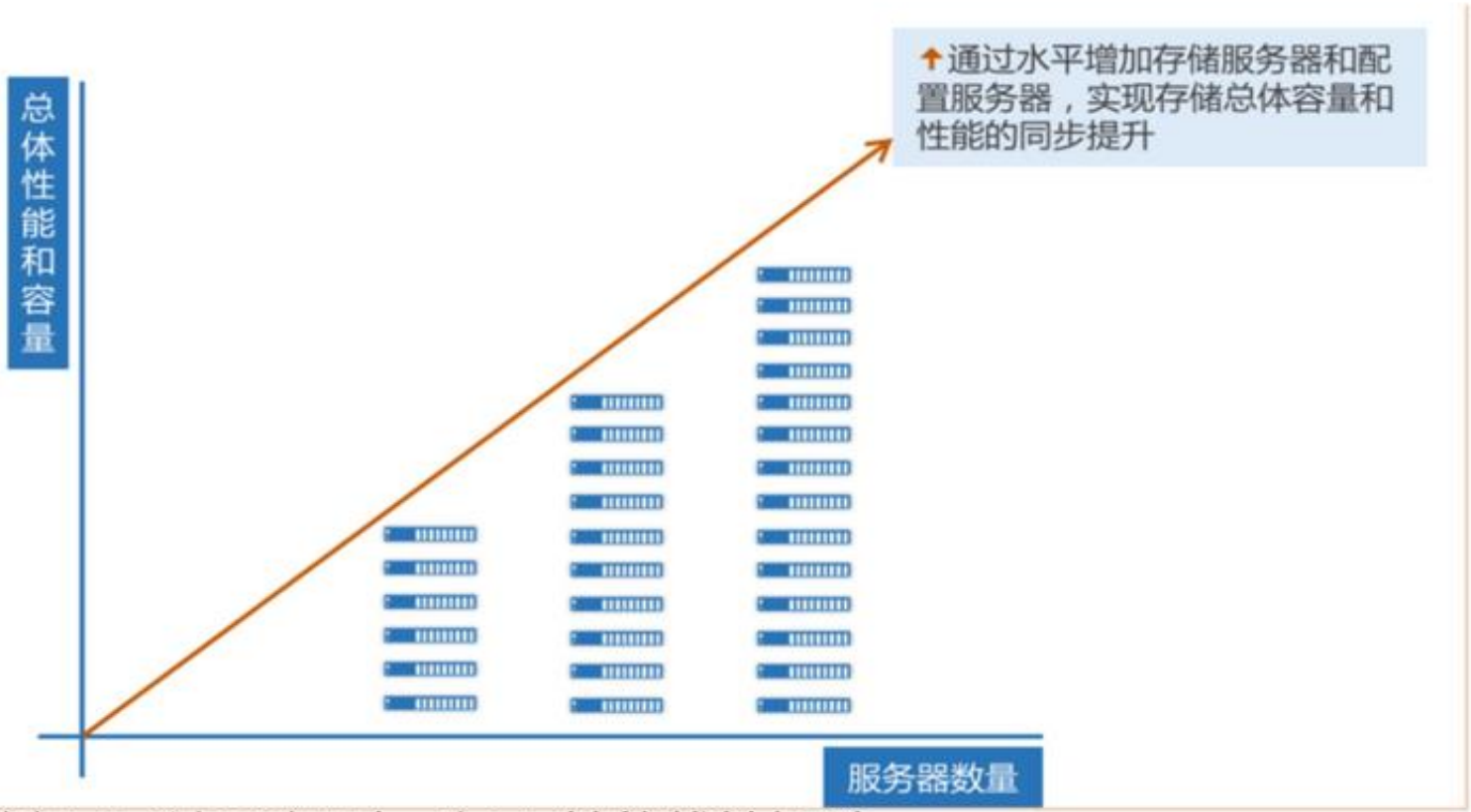
- 推荐36台+：单节点磁盘配置 16TB SATA*36；EC模型为29+3，存储一体机数量可按需增加；
- 1PB+日增算力：更多存储节点提供更高性能

扩容场景

- 最小扩容单位：单存储节点
- 建议扩容单位：根据存储需求与EC规则，扩容对应的安全数量
- 扩容可更改EC；满足提升出盘率、容错性的新要求；应对算力爆发式增长等场景

存储体量、吞吐能力 线性提升

容量、写性能、读性能线性提升



- 单集群容量可达EB级别，容量/性能线性提升
- 可创建多个命名空间，用于不同矿池
- 对象存储模型，从key直接到value，无目录树查找损耗，满足千亿文件数低延迟访问

低成本，高可靠

保障数据可靠性、可用性

- **利旧建议**： 基于纠删码（28+4）技术， 五副本的容错能力，出盘率87.5%
- **一体机**：纠删码推荐29+3， 四副本的容错能力，出盘率90.625% ，成本接近单副本
- 弹性EC：支持小集群模型，通过调整纠删码模型，满足小集群到大集群的演化
- **高利用率**：购买 100PB 空间，实际可用高达 87.5PB/90PB（传统双副本只有 50PB）
- **故障域隔离**：小集群基于服务器；超大规模集群基于机柜或交换机，提升可靠性

专为 IPFS 优化

从Lotus到云存储，全面优化提升性能

- 存储写入，提供SDK，一个函数调用完成从文件复制到http上传的改变
- Winning PoST，专门针对 IPFS 数据读做性能优化，66*9次随机读合并为一个请求
- Window PoST，20秒内读完一个Partition（2349个32G的sector，18.8s 读完）

方案收益

顺应数字化时代的到来，拥抱云计算技术

- **高性能**：专为IPFS优化，Window Post，**20秒内**读完一个Partition(2349个32G的sector)
- **高扩展**： 按需**线性扩展**，单集群可达数百PB至EB级
- **高稳定**：按配置可容忍多台服务器同时挂掉，但**不影响读写**功能与性能
- **低成本**：独创纠删码技术，利用1.10冗余度，4副本的容错能力；且出盘率可达90.625%

三.IPFS存储应用特点

应用特点

- 节点间相互独立
- 专有共识机制，特有算法
- 定期证明机制
- ……

IT需求思考

- 计算和存储过程， 应用复杂
- 过程时间限定，对于IT产品性能、可靠性要求较高

1 数据接收

2 数据处理

3 区块同步

4 周期性证明

Filecoin IT 构建基本要素

去中心化存储架构、应用过程复杂，IT构建面临多方面挑战

服务器、存储、网络

30%

可靠

数据安全因素要求IT设备（存储）稳定至关重要，停机等故障将面临质押罚没、无法赢得区块奖励等.

30%

高效

算力、数据通路的高效性影响矿场是否能将本地存储转换成有效存储算力（矿场收益）

25%

技能

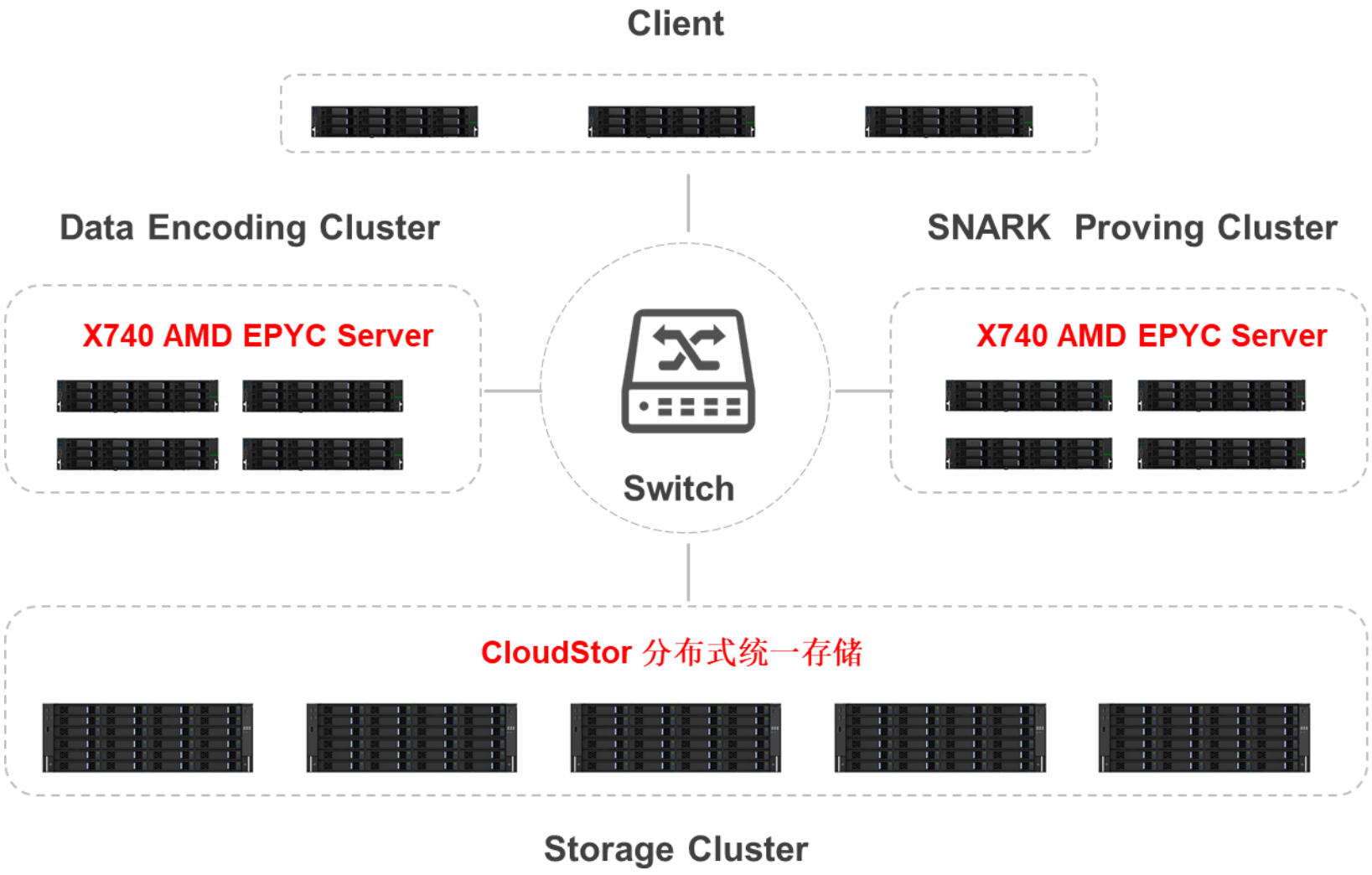
要求有较高的基础架构设计、架构优化、技术规划以及强有力运维能力。

15%

成本

控制IT采购（主要是存储）、运维成本，确保能够跑赢数字货币，保护投资人的利益。

构建高效区块链存储的IT方案—曙光服务器



构建要素

- 高效、可靠IT产品
- 合理IT架构（计算/存储、集群构建）
- 精细化计算和数据流处理管理、优化

IT产品线全家福—曙光服务器

机架式服务器（AMD EPYC）

A620-G30
旗舰双路

A620-G40
支持ROMA/MILAN

A320-G30
2U单路

人工智能&去中心化存储 服务器（AMD EPYC+GPU）

X745/X785
3双宽与单宽GPU服务器

离线训练
X740-A30

刀片、融合服务器(AMD EPYC)

TC4600E（支持液冷）
刀片服务器

CX55-G40刀片节点

TC4600T
高密度服务器

TC4600E-LP G4
液冷刀片服务器

存储产品线

DS 系列多控统一存储

CloudStor分布式统一存储

DBStor 备份一体机

CloudStor分布式统一存储

基于可控 CloudStor
构建区块节点的最佳存储产品

CloudStor

分布式统一存储

1个集群

容量

EB级，通过验证的国内最大原型系统

2种架构

全对称

中小规模提升性价比

非对称

大规模提升扩展性

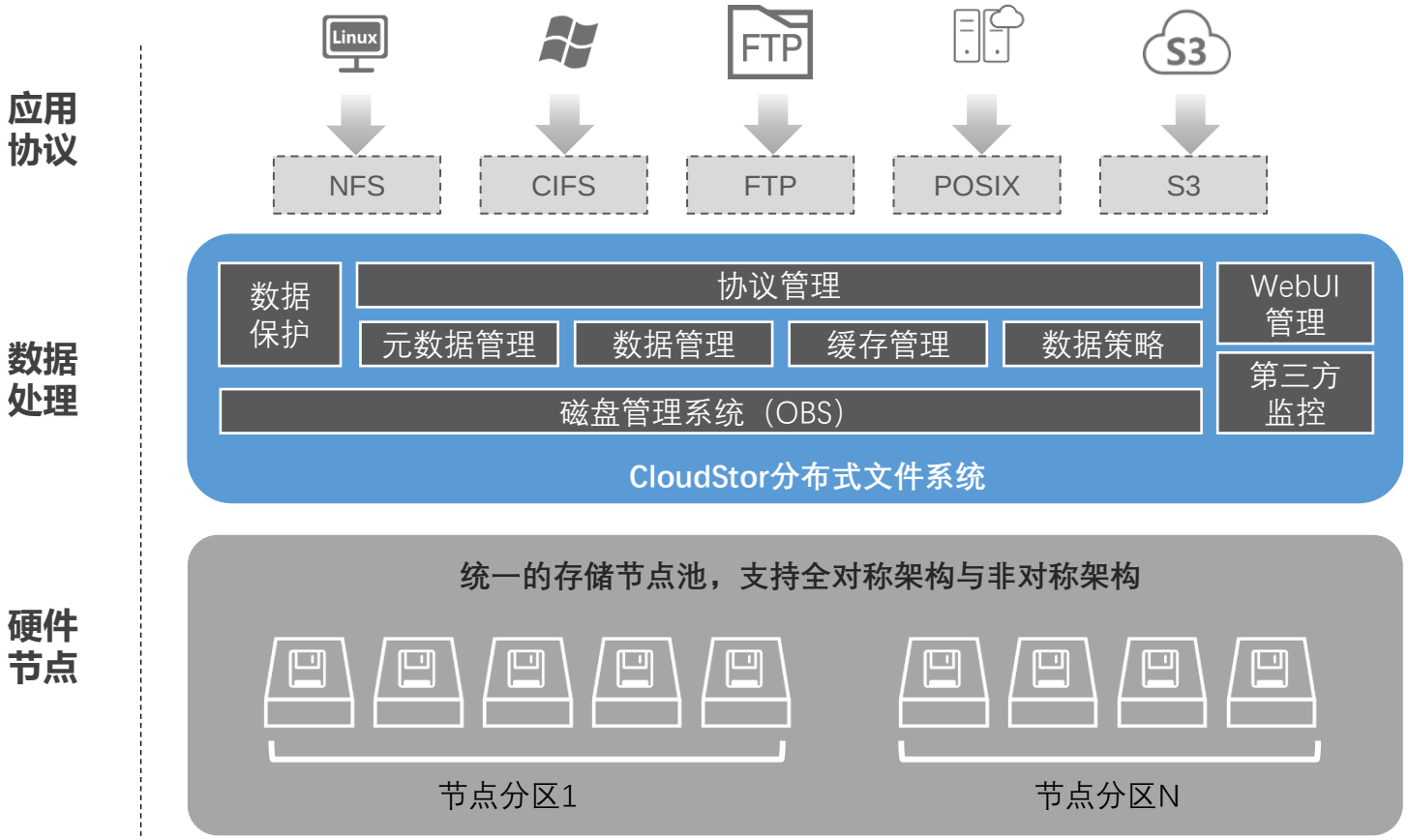
3类硬件

性能型　均衡型　容量型

4种协议

文件　块　对象　HDFS

CloudStor 产品架构



1款软件 2种架构

- CloudStor默认为全对称部署，支持非对称部署
- 非对称部署中，存在单独的元数据节点，要求3个及以上
- 建议按照如下规模选择部署方式

  3≤存储节点数≤64，全对称部署

  64 < 存储节点数 < 100，实际情况讨论

多维度冗余设计，稳定可靠

部件

电源、风扇、硬盘等

最多允许任意4硬盘故障

网络

网卡、交换机等

冗余拓扑

节点

最多允许任意4节点故障

支持机柜级冗余

数据

副本/纠删码冗余机制

元数据副本冗余

磁盘分组/节点分区

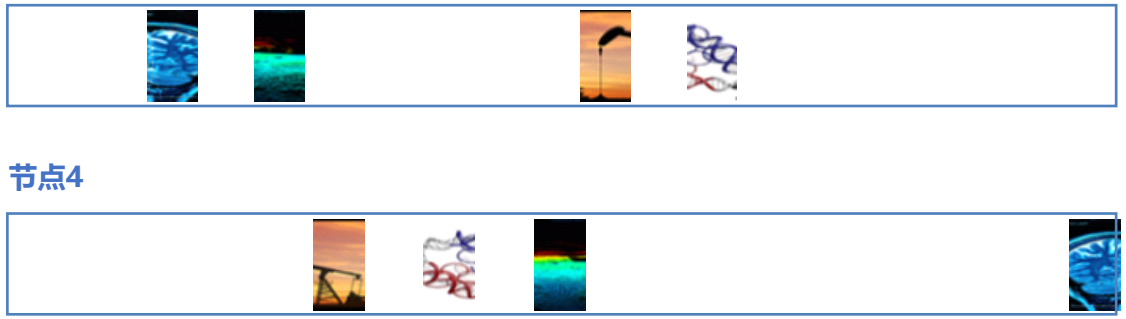
数据冗余 多副本 vs. 纠删码

文件切成对象，分布到不同节点的不同磁盘上

节点1

节点2

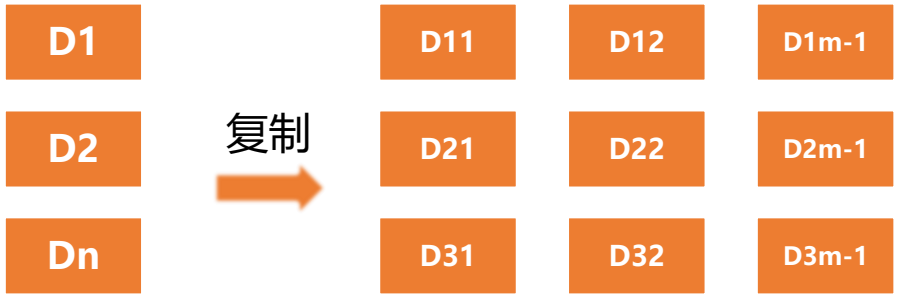
节点3



节点4

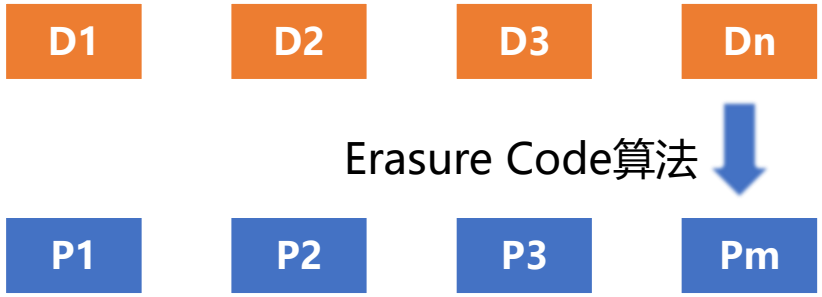
多副本

空间利用率=1/M



纠删码

空间利用率=N/(N+M)



纠删码冗余 N+M:b

N

数据对象个数

切片分段后的原始数据，N≤16

M

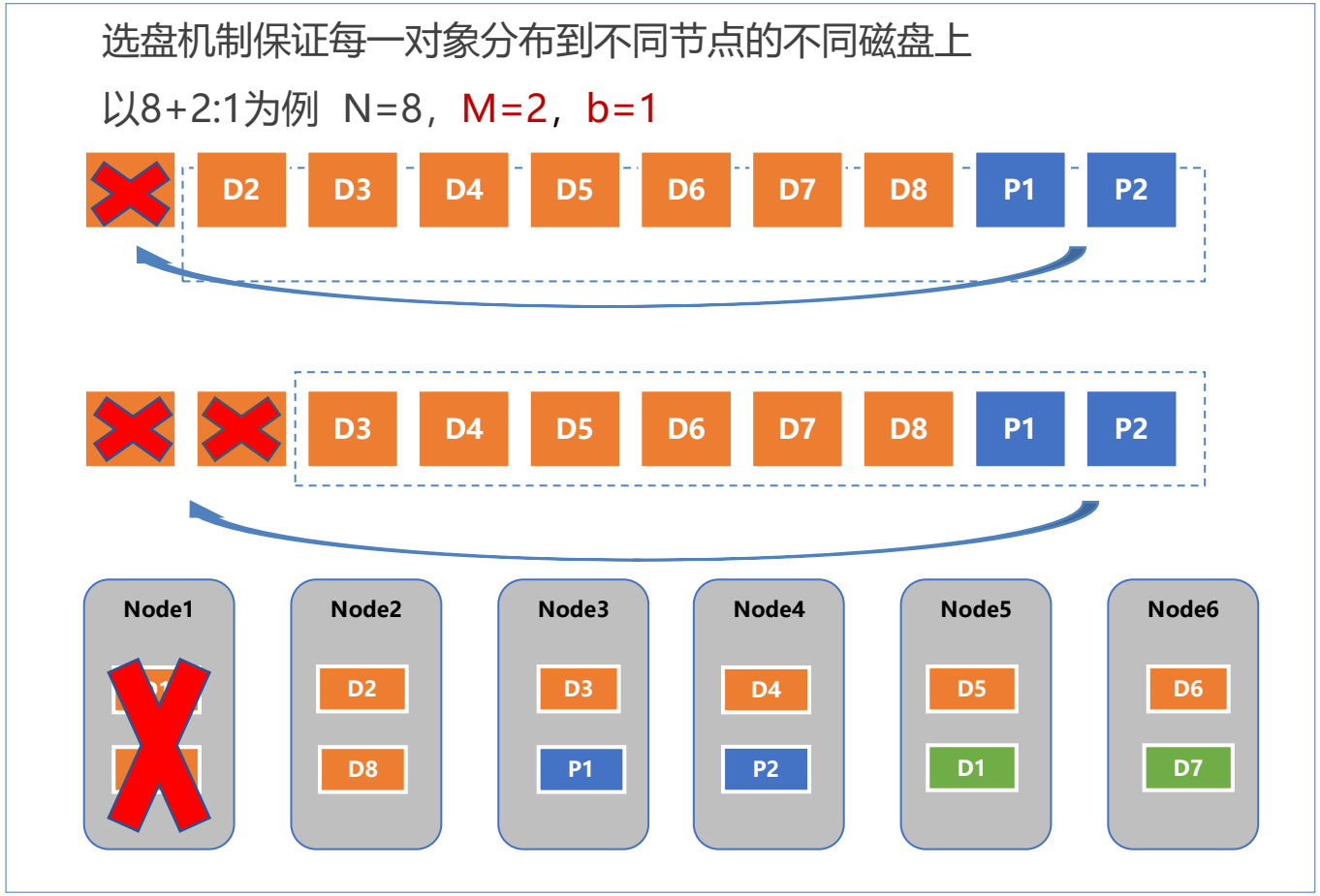
校验对象个数，冗余数据

允许故障的磁盘数，M≤4

b

决定数据对象在节点如何分布

允许故障的节点数，b≤4，且b≤M



元数据采用多副本方式存储，若数据采用N+M:b模式，则元数据默认采用（M+1）副本；且至少为3副本

常见的数据冗余模式

冗余方式	存储节点数量	空间利用率
双副本	≥3	50%
三副本	≥5	33.3%
4+2:1	≥3*	66.7%
6+2:1	≥5	75%
4+2:2	≥8	66.7%
8+2:1	≥6	80%
8+2:2	≥12	80%
10+2:1	≥7	83.3%
10+2:2	≥14	83.3%
16+1:1	≥18	94.1%
16+4:2	≥12	80%
16+4:4	≥24	80%

多副本

M副本，M=2~4

空间利用率为1/M

任意允许故障的磁盘数或节点数为（M-1）

EC纠删码

N+M:b，空间利用率为N/（N+M）

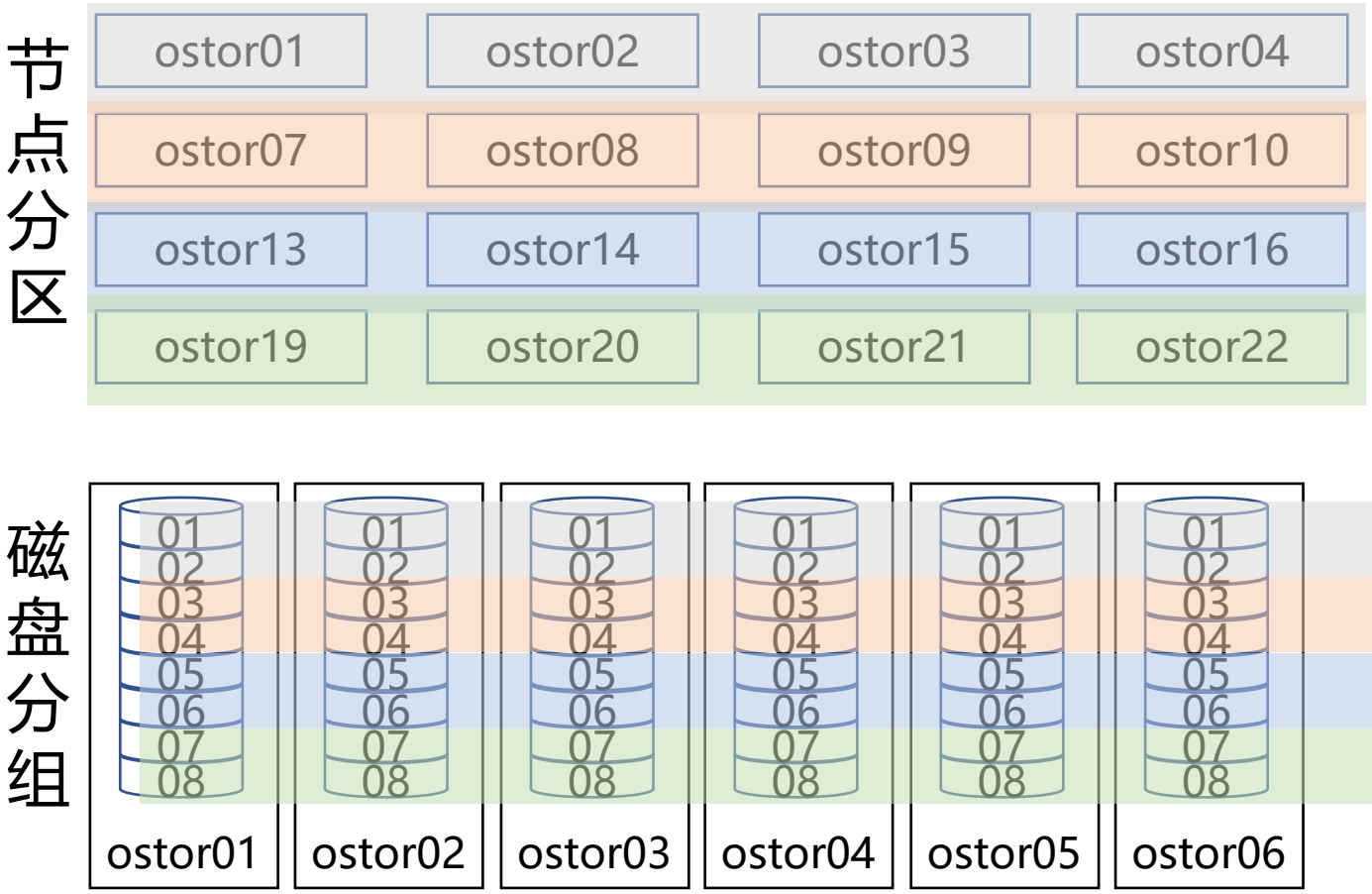
允许任意同时故障的磁盘数为M，M≤4

或允许任意同时故障的节点数为b，b≤4

推荐的最小节点数为[b(N+M)/M]+b

*注：4+2:1时，推荐的最小节点数为4；3个节点配置4+2:1，一个节点故障后，无法保证数据及时修复，请务必谨慎配置！！

双重机制，进一步缩小故障域



节点分区

节点组成不同的分区（节点池）

物理节点相互隔离，故障互不影响

每一节点池配置N+M:b，满足b的要求

缩小节点的故障域

磁盘分组

同一节点池中的磁盘池化（分组）

磁盘相互隔离，故障互不影响

每一分组的磁盘遵从N+M:b，满足M的要求

缩小磁盘的故障域

文件快照，数据安全的再保障

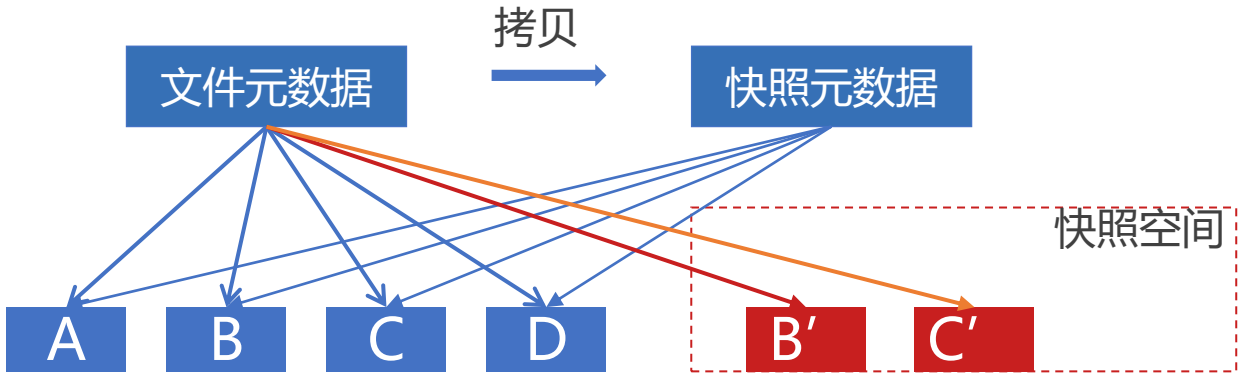
快照

支持目录/文件级快照

一次性或策略性快照

Web界面可实现快照回滚

单一集群文件快照数量上限为20000个



实现原理

ROW，首次写重定向

①元数据记录文件布局（Layout），主要是数据块的存储位置

②某一时刻触发快照，IO挂起，元数据拷贝一份，记录原始Layout；拷贝完成后立即解除暂挂

③之后某时刻，文件某一数据段发生更改，按照规则，将变更内容写到另外位置，同时元数据更改，重定向到新位置

配额管理，存储精细化设置

多粒度多层次

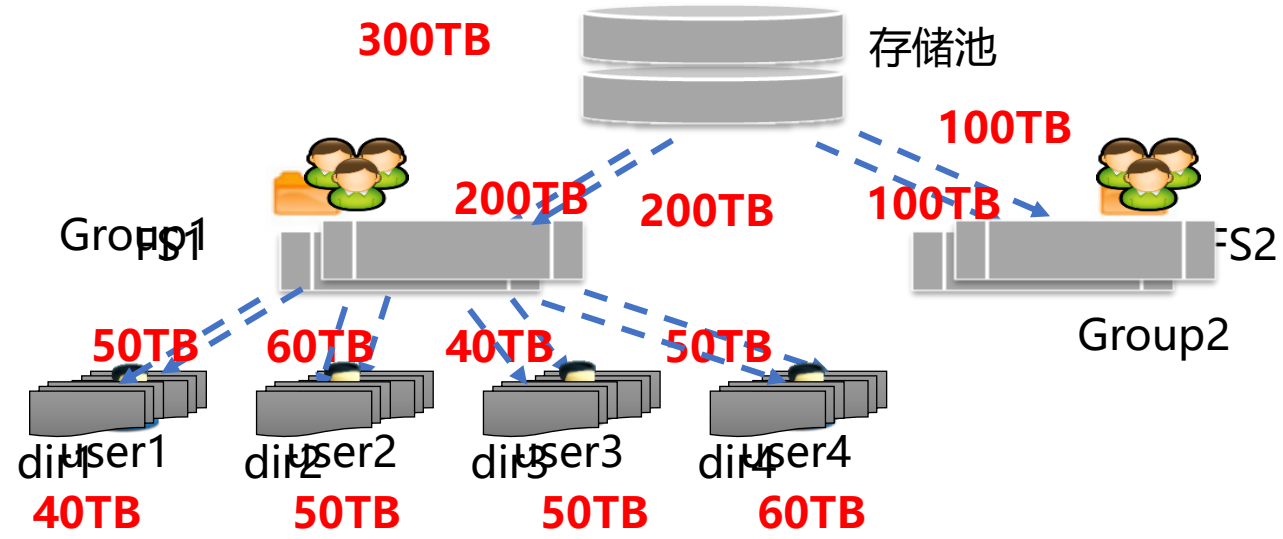
- 基于文件系统
- 基于目录
- 基于用户组
- 基于用户

多种配额类型

- 统计配额：仅监控，不限制
- 限制配额
  - 文件数/容量
  - 软阈值/硬阈值

写入精准控制

- 软阈值：仅告警
- 硬阈值：无法写入



3款硬件平台，按需选择

OStor-K30-212

- 2U12盘位
- Hygon 7100系列CPU，单路
- 元数据盘：≥1块 SATA SSD
- 数据盘：支持3.5”及2.5”规格HDD；支持SATA SSD
- 缓存盘：SATA SSD
- 小规模场景

OStor-K30-424

- 4U24盘位
- Hygon 7100系列CPU，单路
- 元数据盘：≥1块 SATA SSD
- 数据盘：支持3.5”及2.5”规格HDD；支持SATA SSD
- 缓存盘：SATA SSD
- 适中规模场景

OStor-K30-436

- 4U36盘位
- Hygon 7100系列CPU，单路
- 元数据盘：≥1块 SATA SSD
- 数据盘：支持3.5”及2.5”规格HDD；支持SATA SSD
- 缓存盘：SATA SSD
- 大规模场景

CloudStor配置

推荐配置：OStor-K30-436（98001291）

单节点规格	
处理器	1*Hygon 7151
内存	64G
数据盘	35*14T 7.2k SATA HDD
元数据盘	1*1.92TB SATA SSD
系统盘	2*600G 10K SAS
网络	2*10G 以太网
电源	高效能冗余电源
上架	托轨

高性价比配置：OStor-K30L（Filecoin特供节点）

单节点规格	
处理器	1*Hygon 3185
内存	64G
数据盘	36*14T 7.2k SATA HDD
元数据盘	1*1.92TB SATA SSD
系统盘	1*600G 10K SAS
网络	2*10G 以太网
电源	高效能冗余电源
上架	托轨

- 明确要求裸容量还是有效容量，若要求有效容量，优选高的空间利用率EC算法
- 节点数量大于6，存储利用率大于80%；节点数量越多，利用率越高，最高不超过94%

CloudStor在IPFS存储应用中的优势

全自主研发

硬件-软件 完全自主研发，可快速响应业务和客户的需求。

区块链存储定制化

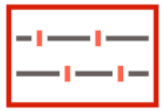
针对区块链存储从硬件-存储系统层面进行大量定制、优化的工作，提高相关流程性能。

TOP客户应用案例

TOP 10 用户已有数百PB规模性部署案例，稳定性 、性能业内同类最优。

业务无感知扩容

节点分钟级在线扩容，智能化根据业务负载进行数据迁移，业务无感知。



多集群管理

一套管理页面管理多套存储集群，极大降低运维压力。

性能

独有高性能内核态驱动和客户端，IO路径短、延迟低，相比NFS、CIFS性能提高20%

CloudStor针对性优化

业务层优化

- 优化目录共享机制
- 优化读写模式

文件IO层优化

- 优化IO读写策略
- 调整数据重建策略
- 裁剪应用无效IO

介质层优化

- 选择适配硬盘类型
- 优化硬盘工作机制

硬件配置优化

大文件顺序写，完全随机读IO特性 → 优化内存配置

提升持续数据IO的可靠性 → 优化数据和业务网配置

基于数据粒度、IO特性 → 优化SSD：SATA容量配比

性能参数

业务阶段	存储响应时间	业务响应时间
Winning-POST	0.02s	1）只有66*32=2112个IO 2）参考业务程序能持续保持100并发，那么需要（2112）/100=2.12轮，存储每轮的响应时延10ms以内，那么WindowPost一共耗时2.12*10ms=0.02s 3）实际WindowPost耗时，取决于并发数是多少，以及是否能维持住大并发 【人人矿场实际业务，当前WinningPost在秒级】
Windows-POST	4.228s	1）读32GB的sector文件，大约2349*10=23490个IO，每个IO 16KB 2）读9.2MB的sector的索引文件，大约2349*8*1=18792个IO，每个IO 4KB 3）参考业务程序能持续保持100并发，那么需要（23490+18792）/100=422.8轮，存储每轮的响应时延10ms以内，那么WindowPost一共耗时422.8*10ms=4.228s 4）实际WindowPost耗时，取决于并发数是多少，以及是否能维持住大并发 【人人矿场实际业务，当前WindowPost在1分钟左右】

推荐配置

节点数	得盘率	可故障节点数	有效容量	总价格（万）	每TB价格
6	0.72（80%*90%）	1 (8+2:1)	2116 TB （14T*35*6*0.72）		
12 （推荐）	0.72（80%*90%）	2 (8+2:2)	4232 TB （14T*35*12*0.72）		
14 (推荐)	0.75（83.3%*90%）	2 (10+2:2)	5145 TB （14T*35*14*0.75）		

四.算力软件

初成和天茹/1475，黑奔，原力驱，点对点，星际联盟，石榴矿机，新钛云，时空云，萤火虫等业内算力公司有深入的合作，可以根据客户需求，来合理化推荐算力软件.

排名	矿工	标签	有效算力 / 占比	24h出块奖励	24h挖矿效率	24h算力增量	地区
1	f01248	智合云(ZH)	74.24 PiB / 3.63%	9,342.79 FIL	0.12 FIL/TiB	54.91 TiB	美国
2	f02770	时空云&灵动	72.63 PiB / 3.55%	9,763.60 FIL	0.13 FIL/TiB	-64.00 GiB	N/A
3	f09652	RRmine.com	51.71 PiB / 2.53%	7,233.84 FIL	0.14 FIL/TiB	194.47 TiB	新加坡
4	f025002	ipfs.so(合盈)	41.73 PiB / 2.04%	5,678.91 FIL	0.13 FIL/TiB	-96.00 GiB	N/A
5	f023530	星际矿池-天秤座	35.07 PiB / 1.71%	5,183.30 FIL	0.14 FIL/TiB	86.00 TiB	中国
6	f02438	原力区	34.18 PiB / 1.67%	3,875.15 FIL	0.11 FIL/TiB	269.53 TiB	中国
7	f09037	星际大陆	31.23 PiB / 1.53%	4,443.76 FIL	0.14 FIL/TiB	84.72 TiB	新加坡
8	f01782	hellofil.com	30.52 PiB / 1.49%	4,538.27 FIL	0.15 FIL/TiB	-32.00 GiB	中国
9	f01235	星际大陆	28.89 PiB / 1.41%	3,095.90 FIL	0.10 FIL/TiB	95.22 TiB	新加坡
10	f02303	鑫兜科技	24.89 PiB / 1.22%	3,057.60 FIL	0.12 FIL/TiB	0.00 B	德国
11	f049911	麦客存储为您服务	23.16 PiB / 1.13%	3,151.87 FIL	0.13 FIL/TiB	19.94 TiB	N/A
12	f030347	--	20.53 PiB / 1.00%	2,581.72 FIL	0.12 FIL/TiB	315.25 TiB	中国
13	f035364	HashCow	20.44 PiB / 1.00%	2,791.23 FIL	0.13 FIL/TiB	309.06 TiB	N/A
14	f024563	蝶链科技	20.29 PiB / 0.99%	2,449.47 FIL	0.12 FIL/TiB	-32.00 GiB	N/A
15	f02775	时空云&灵动	20.23 PiB / 0.99%	3,000.62 FIL	0.14 FIL/TiB	32.00 GiB	N/A
16	f020330	ipfs.so(合盈)	19.73 PiB / 0.96%	2,525.92 FIL	0.13 FIL/TiB	0.00 B	中国
17	f01012	蛮牛科技	18.56 PiB / 0.91%	1,747.08 FIL	0.09 FIL/TiB	76.31 TiB	中国香港