概念定义

内容描述文件：对数据资产包含的文件列表、文件格式等的解释文件。数据资产内容包括内容描述文件和内容文件。

数据资源：内容描述文件、文件内容、文件分片等大文件数据的统称

分片数据：将内容文件基于门限方案等方式进行分片切割后的数据为内容文件的分片数据。一定数量的分片数据应能够恢复出内容文件。

分块数据：ipfs存储数据的最小单元，当数据大于设置的块大小时，将切割成多个数据块存储。

互信节点：存储节点网络中，由同一提供方提供的节点，或在同一数据中心的节点。

邻居节点：存储节点网络中，一组互信节点以外的其他节点为互信节点的邻居节点。

成员节点：存储节点网络中所有节点均为成员节点。

需求点分析

能力支撑

- 大文件存储安全性、高可用性
    - 数据只能追加不能删除
    - 多副本分布式存储、副本数控制
    - 文件存储节点应能够提供文件相关存储证明（PoR\PoSt）
- 大文件存储隐私性、保密性
    - 大文件应加密存储
    - 大文件存储应支持业务隔离性，在数据生成方节点可以存储一份大文件数据的完整内容。其他联盟方节点上将大文件数据分片分布式存储，一个文件数据分成m个分片，分布式存储在不同的节点上，每一个节点上只有该文件的一个或部分分片，t个分片可恢复出整个文件(t<m)。
- 大文件数据独立性、自解释能力
    - 结构化数据可解析
    - 一个资产的所有内容数据可自解释
- 内容数据需要和链上资产关联绑定

总体框架设计

接口服务

- 大文件数据的存储、查询
- 文件分片情况查询
- 文件副本数及分布查询
- 节点列表，节点增删

数据模型定义与解析

一个数据资产内容应包含：内容描述文件（des_file）+内容文件（content_filelist）。基于内容描述文件提供数据资产文件的解释。数据资产内容文件的追加、删除、变更等应构建新的内容描述文件

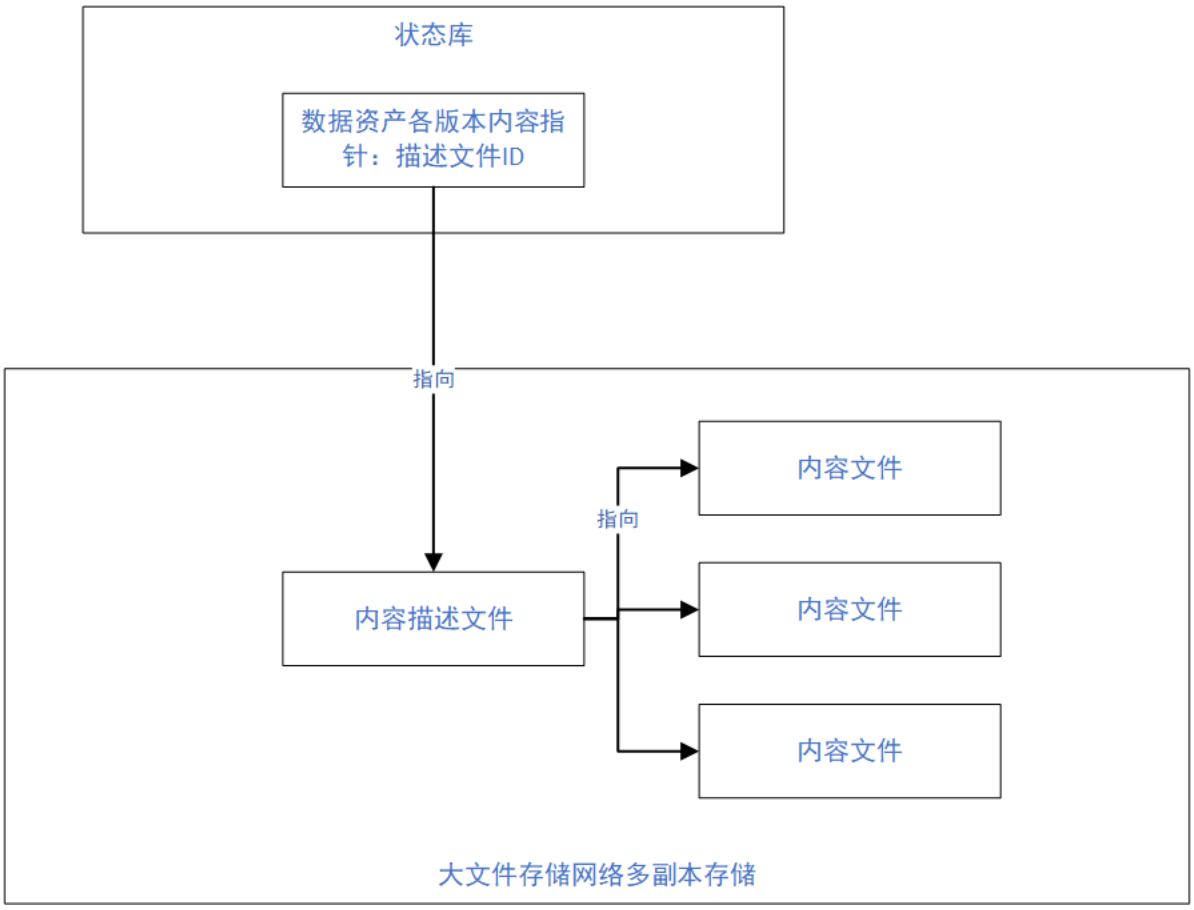
- 内容描述文件要素：
  - 描述文件ID：内容描述文件的标识。内容描述文件也是存储节点中的大文件。
  - 结构化数据列表：数据资产包含的结构化数据列表list{文件名，文件ID，文件格式}；
  - 非结构化数据列表：数据资产包含的非结构化数据列表list{文件名，文件ID，文件格式};
- 内容文件：即数据资产的各项大文件数据，通过数据存储服务及数据分发服务在多个节点全量存储或分片分布式存储。

能力支撑

- **数据资产大文件数据存储**：接收文件存储请求，构建内容描述文件des_file，并将des_file与content_file存储到存储层。
- **数据资产内容追加**：存储追加的大文件内容，并基于旧版内容描述文件des_file($Version_n$)及新追加的大文件ID构建新版内容描述文件des_file($Version_{n+1}$)，返回最新des_fileID。
- **内容文件修改**：大文件内容及对应描述信息的修改，同样基于旧版描述文件构建新版文件，返回最新的描述文件ID，历史版本的des_file及历史大文件并不删除；
- **内容文件删除**：同上，构建新版本描述文件，旧版本数据并不删除

内容锚定与多版本管理

数据资产内容支持多个版本，状态库中各个版本的内容均指向存储节点网络中对应版本的内容数据描述文件ID。



问题分析

- 内容描述文件是否要加密？加密存储后难以解析，且内容文件本身是加密的，内容描述文件中只包含文件ID、格式等，无敏感内容，可不加密。
- 数据资产内容需要支持并发的增删改查，如何保证内容描述性文件的并发安全性？数据资产内容增删改引发的内容描述性文件的变化并不在原文件中更改，均基于某个内容描述文件标识指向的版本创建新的内容描述文件，资产的最新内容已状态库中指向的内容描述文件的标识为准。
- 过于数据资产非结构化数据内容定义和解析：同一类数据资产应使用同样的元数据，故此类数据的定义和解析更适合放在合约或链底层进行。另一方面，内容文件一般需要加密存储，在存储节点层做定义和解析难度需要依赖密钥计算等，耦合性过高。

数据分发模块

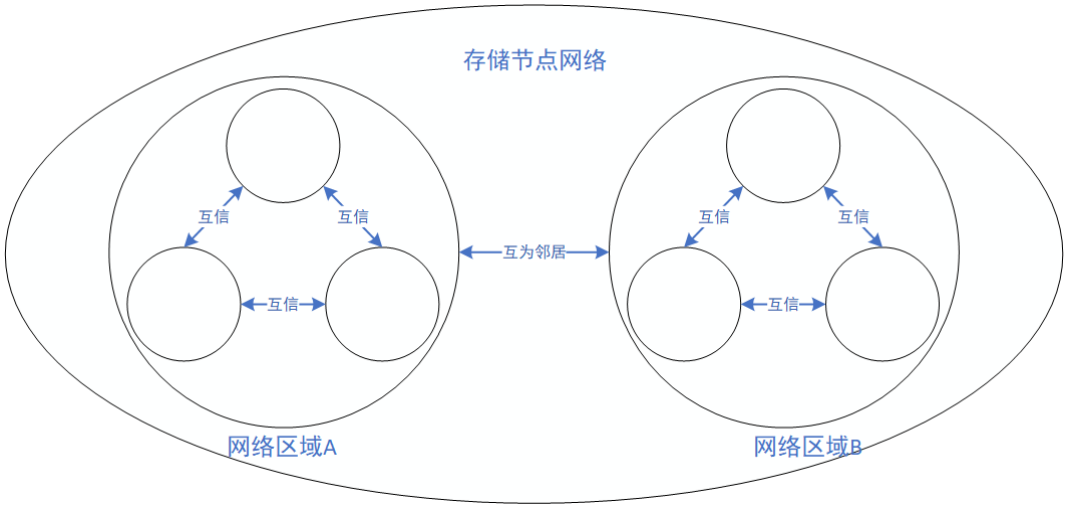
能力支撑

- **大文件副本的分发**：支持数据隔离存储需求，大文件可在一个数据中心的多个节点中多副本存储。
- **大文件分片分布式存储**：将大文件数据分成m个分片，分布式存储在m个节点上，大于等于阈值t(t<m)个分片即可恢复出原大文件。在满足数据隔离性存储的同时，避免某个数据中心异常。

网络拓扑

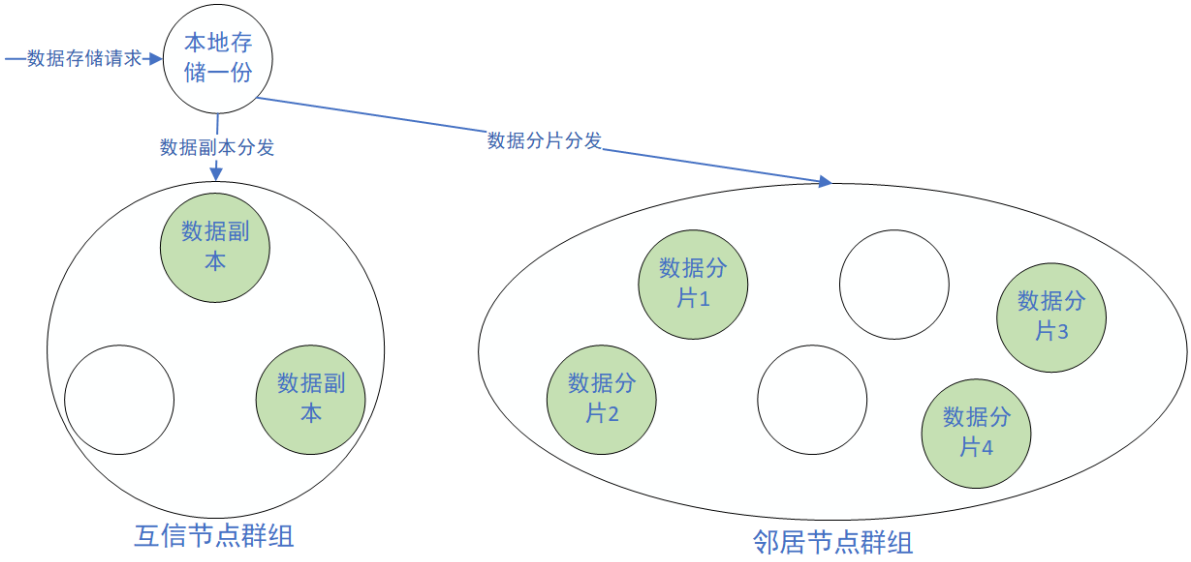
为满足数据生成端用户的数据隔离存储需求，并满足数据的高可用性。将存储节点网络中由同一节点提供方提供的节点或同一数据中心的节点划分为互信节点，其余节点作为互信节点群组的邻居节点。

- 互信节点：由同一干系方提供或维护的存储节点，节点之间互信等级较高，**可直接进行数据副本互备**，另一方面，同一互信节点群组很可能在一个数据中心或网络区域，**群组整体宕机等风险较高**；
- 邻居节点：同一存储节点网络中，某一组互信节点外的其他存储节点为这一组互信节点的邻居节点，**互信节点与邻居节点之间存在一定互信风险。**但互为邻居节点的节点群组在同一网络区域或数据中心的可能较低，**提供数据互备的安全性较高。**



数据分发协议

数据分发协议应保证**分发的均衡性。**为保证**隐私性和数据隔离存储**的需求，需要为副本数据和分片数据分别在不同的节点群组范围内进行数据分发：



- 副本数据分发：在本节点的**互信节点群组**中，选择距数据资源距离最近的**若干**节点进行副本分发。
- 数据分片及分发：由接收客户端数据存储请求的节点负责数据的分片和分发。对一个数据资源的每个分片，在本节点**邻居节点群组**中，分别选择距分片数据距离最近的**一个**节点进行分片数据的分发。
- 数据分发均衡性保证：distance:= neighborID^(CID或pieceID)，在给定节点群组中选出若干距离数据资源距离较近的节点进行副本数据或分片数据的分发。

问题分析

- 内容描述文件及结构化数据文件大小都比较小，如果做分片存储，对存储空间和网络资源等存储代价会明显增大。小文件不分片?
- 区块链建设初期，联盟方较少或资源较少的情况下，全网存储节点数量可能难以满足数据分片存储的需求。可以在所有成员节点中进行数据分片存储，或节点数不足情况下不进行数据分片存储。

索引路由模块

能力支撑

- 有数据资源请求时，基于索引数据快速定位文件资源位置
- 成员节点间基于存储的本地及其他成员节点的数据资源索引进行副本数检测和调度。

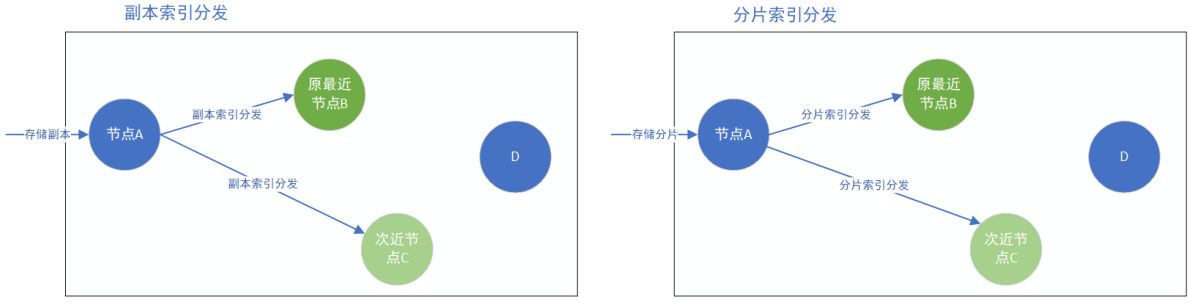
索引存储内容

各节点接收到其他成员节点的provide信息后，**以成员节点为单位分库存储**索引数据：

- 本节点有哪些数据（**是否要考虑分块数据**）：
  - 副本数据：list{CID}
  - 分片数据：List{pieceID：version,CID,all_pieceIDList,分片恢复阈值}。其中version表示此分片所属的分片版本，同一个数据可以有多个版本的分片，不同版本分别对应不同的分片数据内容。
- 部分成员节点有哪些数据：
  - 副本数据：list{CID}
  - 分片数据：list{pieceID：version,CID,all_pieceIDList,分片恢复阈值}。

索引分发协议

- **索引信息分发均衡性保证**：distance:= neighborID^(CID或pieceID)，在给定节点群组中选出**若干**距副本数据或分片数据距离较近的节点进行索引信息的分发。
- **副本索引数据**：节点存储副本数据后，本节点的数据资源索引信息发送到与该节点距离较近的若干节点保存。
- **分片索引数据**：存储分片数据的节点将以上索引信息发送到与分片数据最近的节点。



问题分析

- 当有新的成员节点加入时，节点间的相对距离可能发生变化，从而索引数据的分发路径将发生变化。某节点A的最新最近节点主动从节点的最新次近节点中拉取A的数据资源索引数据。
- 当有成员节点退出时，节点间的相对距离也可能发生变化，但索引为多节点分发，某个节点两个最近节点同时退出的可能较小，由次近节点升级为最近节点即可。
- 选择与节点距离相对较近的若干节点进行索引数据的分发，会导致分发的均衡性较低。对于数据存储业务量大的节点，与其最近成员节点存储索引的频率也更高。若考虑均衡性，基于索引数据进行副本调度时复杂度会加大。

副本控制模块

能力支撑

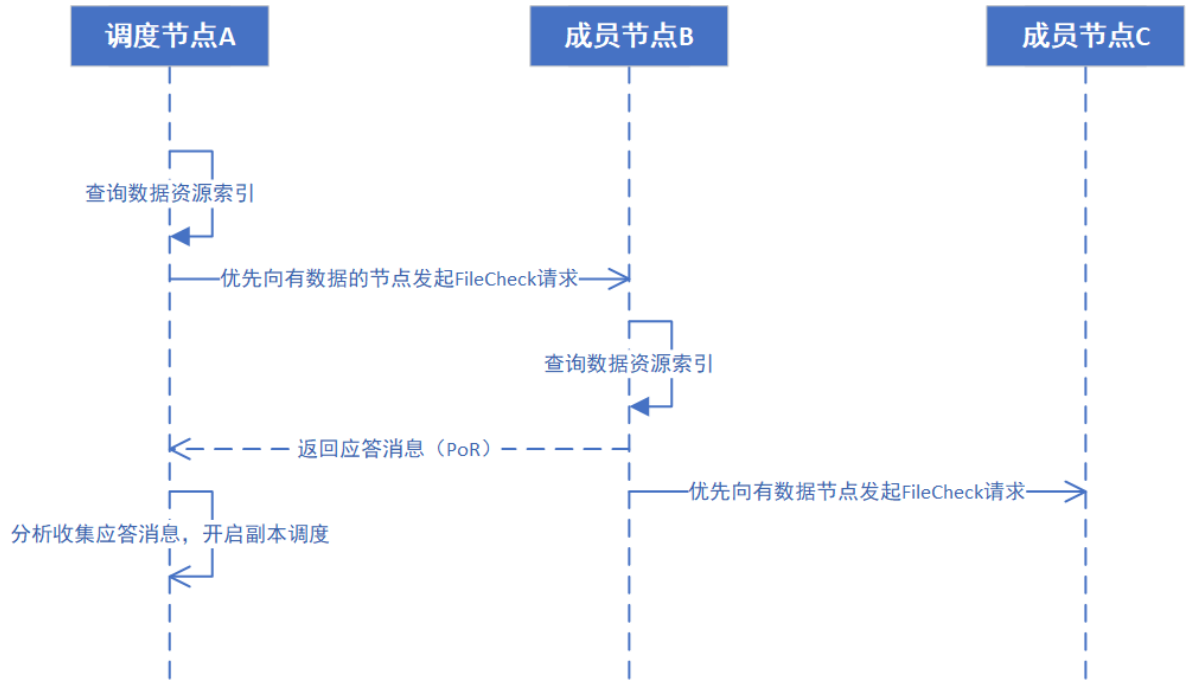
- 检测存储节点网络中的数据资源的副本数量及分片数量，有缺失的数据进行资源的重新调度
- 检测存储节点网络中是否有节点因网络、故障等原因退出网络，对退出节点的数据资源进行重新调度

副本控制协议

- 节点宕机或网络故障后，距此节点距离最近的节点负责故障节点数据资源的副本数检测及再调度。
- 各节点对本节点存储的其他节点数据资源索引数据依次进行数据副本或分片数检测，对于检测发现有数据丢失的情况，将相关数据标识**提交给与检测节点最近的节点**进行副本再调度。

副本数检测及调度过程

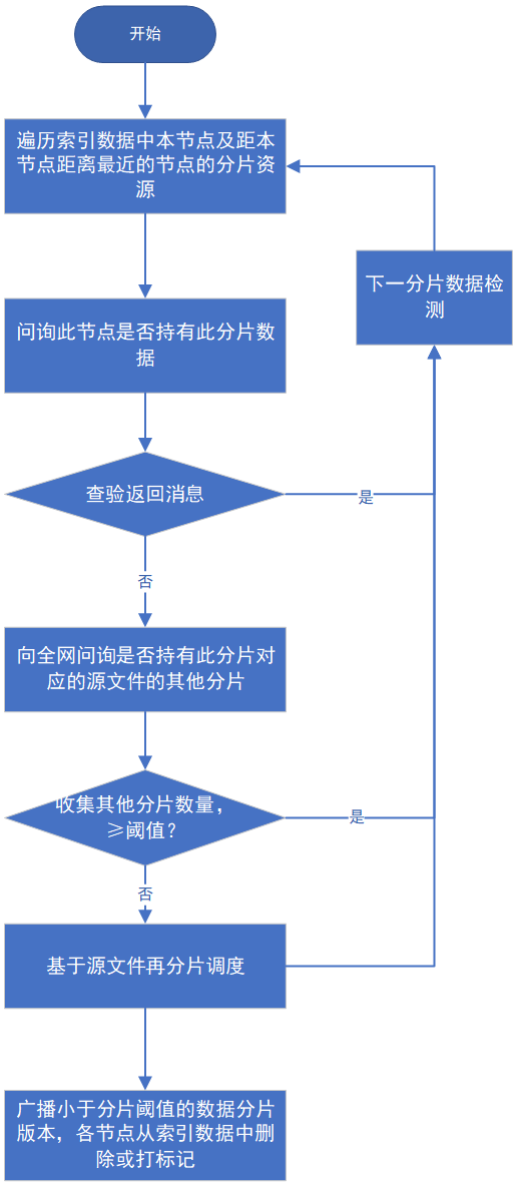
- 调度节点查询本地索引数据中是否有存在该数据副本的节点，优先向以上节点发送副本检测消息；
- 收到消息的成员节点同样先查询本地索引数据确定本地是否有数据，有则向调度节点发送应答消息回复PoR证明；
- 收到消息的成员节点查询本地索引数据中是否有存在该数据副本的其他节点，优先向以上节点进行副本检测消息的递播；
- 调度节点异步收集各节点的应答消息，直至得到足够数量的副本存在性证明，则不进行后续的副本调度。
- 若调度节点收到所有节点的应答消息后无法得到满足阈值数量的副本存在性证明，或在一定时间内无法收集所有消息，则发起副本数据再分配，补齐副本数。



分片数据再调度过程

- 分片数据检测：各节点分片数据的检测和调度，由本节点以及距本节点距离最近的节点共同负责检测和调度。调度节点基于索引数据依次遍历索引数据中的分片数据，向分片资源所在节点请求校验分片数据是否存在。若分片数据不存在或分片资源所在节点失联，则发起分片再调度。分片再调度过程包括可用分片数检测及重新分片分发。
- 可用分片数检测：调度节点向成员节点发送缺失的分片对应的源文件的所有分片列表，并收集成员节点返回的分片ID数量，若剩余分片数量满足文件恢复所需阈值，则取消分片数据的再次调度；

- 重新分片分发：若收集的现存分片数量小于阈值，则由调度节点在本地或成员节点网络中获取完整副本数据，基于完整副本数据重新进行分片分发，并**通知各节点删除或标记恢复失败的分片版本。**



问题分析

- 由距节点距离最近的成员节点负责副本检测，有可能同一个节点分别是多个节点的最近节点，会存在检测及调度任务分配不均的情况。
- 实际场景中，不同的互信节点群组间网络经常不稳定，造成部分节点某个时间段“失联”，会导致数据副本重复冗余存储。
- 失活或异常节点数据资源再调度过程均交付与该节点距离最近的节点进行，依赖关系较重
- 还要考虑节点数不够的情况，数据资源再调度过程中有可能存在节点数不足情况
- 一份文件的现存分片数量无法恢复出原文件后，该版本的分片将成为垃圾数据

数据获取模块

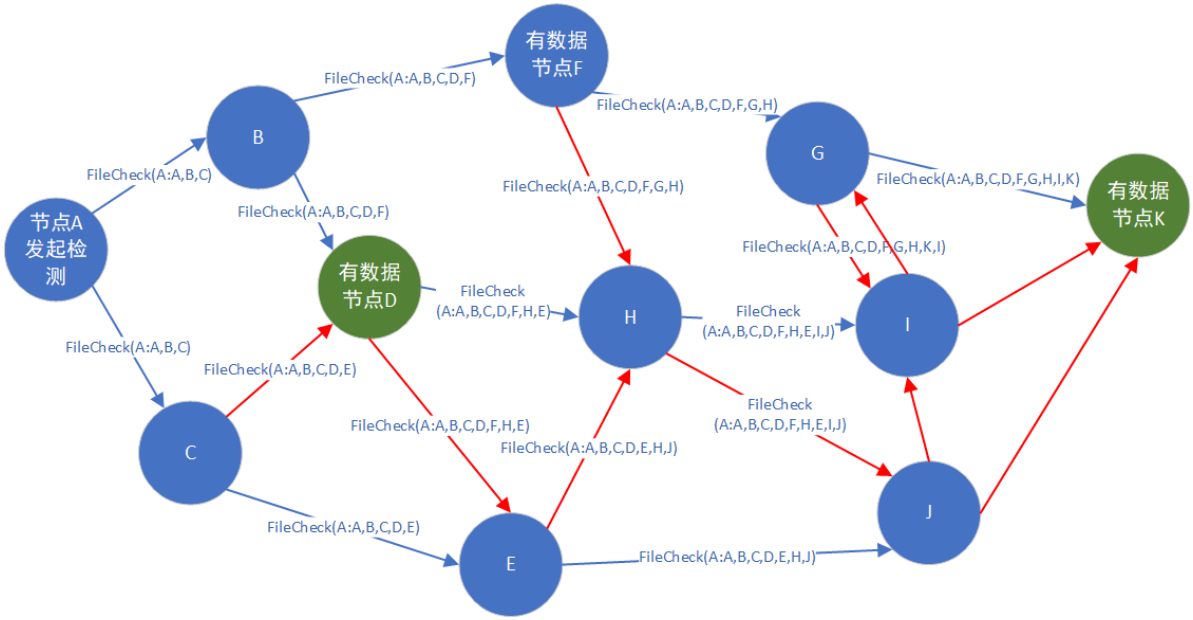
能力支撑

- 接收数据获取请求，若本地有相应数据资源，则返回；
- 若无数据资源，应向其他成员节点请求，收到成员节点有效应答消息后主动拉取数据返回；
- 若向成员节点请求数据资源时，无节点可提供有效资源，则发起请求该数据的分片数据，还原出完整数据后返回。

数据资源的查询终止

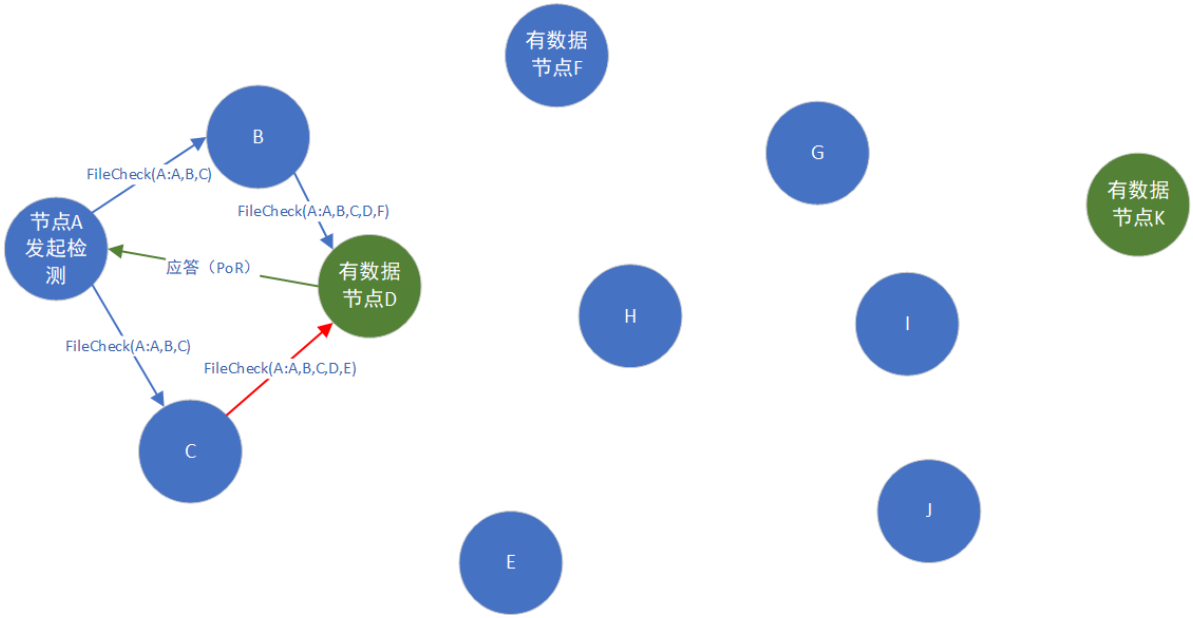
为保证数据资源请求能够有效终止，避免对网络及IO资源的消耗，采用全广播或携带消息关键路径的方式：

- **数据请求消息全广播**：请求节点向全网所有成员节点广播请求某个数据资源，收到请求的成员节点向其应答，此方式**不适用于存储节点数据规模加到的情况**；
- **基于gossip协议+消息关键路径实现查询终止**：各成员节点在发送数据资源请求消息时，应在消息中附加消息传递的关键路径，收到消息的节点不再将消息发送到关键路径中包含的节点，结合重复消息过滤，直至无节点可进行消息传递则终止。如下图所示：蓝色箭头表示消息发送成功，红色箭头表示消息重复被拒收。（A：A，B，C）表示发起副本检测的节点为A，消息关键路径为ABC。



问题分析：

- 当节点规模很大时，数据资源的获取经过的中间路径可能较长。而且当请求节点已收到数据资源后，数据资源请求消息可能还在进行递播。
- 成员节点接收到数据请求后，若查询到本地索引中有节点持有资源，能否直接终止递播？如下图所示，此方式若有数据的D节点为异常节点或恶意节点，不向A进行数据应答，会造成A节点无法获取数据。



数据存储模块

- 与底层存储引擎交互，向上提供数据内容的存储、查询功能；
- 缓存最近的数据内容，用于提供新存储数据内容的即时查询，包括客户端查询请求和数据分发过程中的数据副本拉取请求。
- 服用ipfs数据存储模块，对大容量数据分块存储。

P2P通信模块

p2p通信模块以libp2p为基础，采用安全传输协议，提供节点消息通信功能及连接管理、通信路由管理服务。

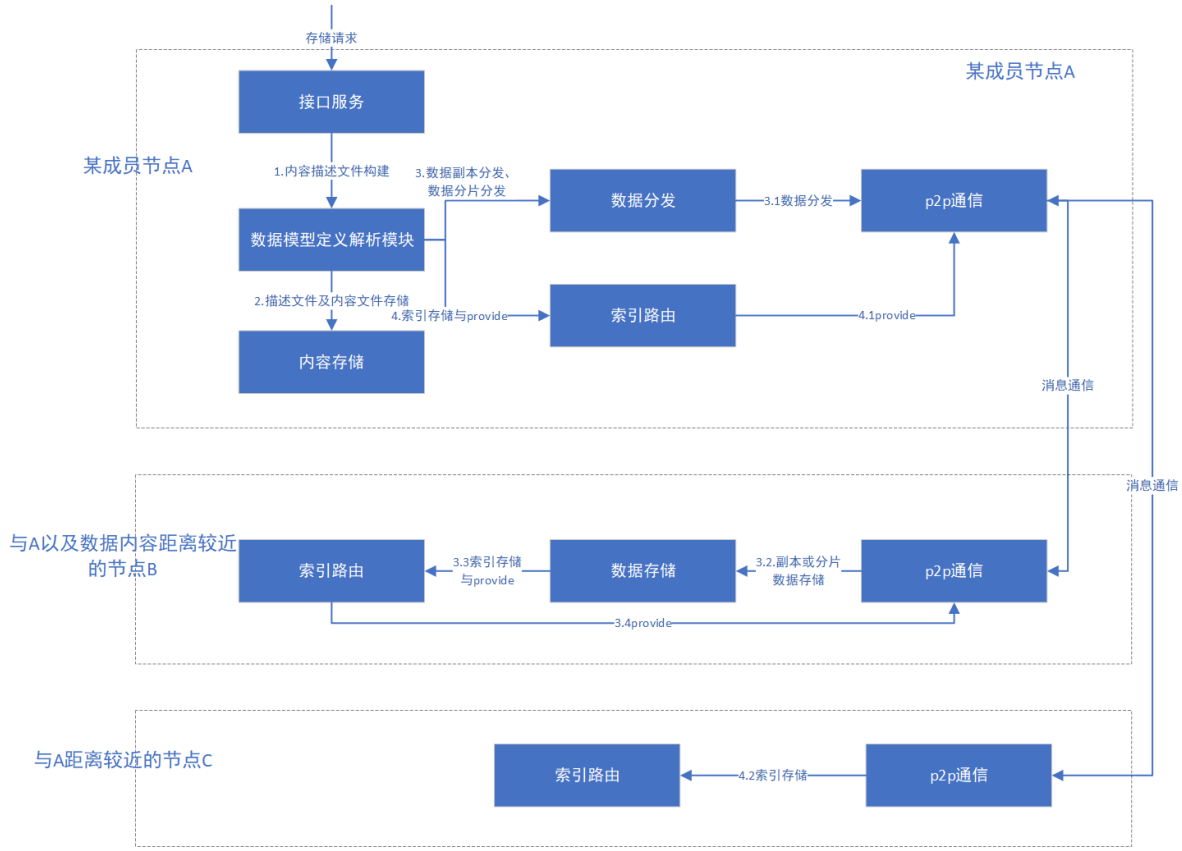
调度总线

调度总线模块负责各个模块间的通信和任务调度。

关键交互模式

数据存储与分发

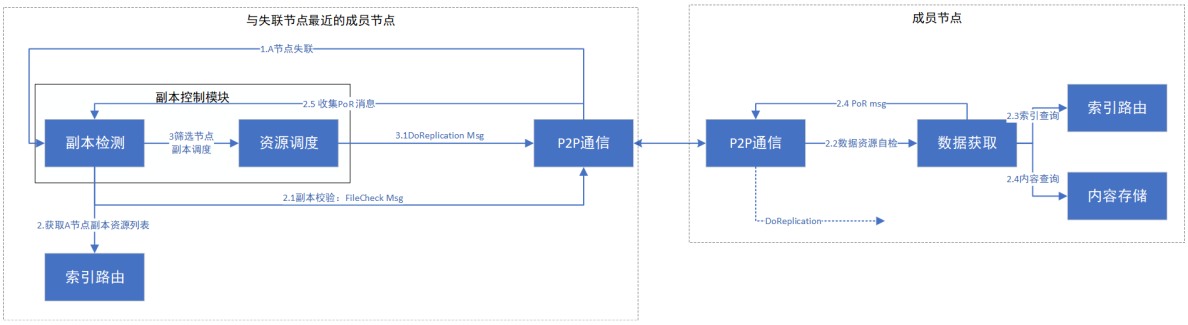
- 节点A接收到数据存储请求后，构建内容描述文件
- 节点A将内容数据及描述文件存储在本地
- 节点A选择与节点A以及存储数据距离较近的n个节点异步进行数据副本、数据分片的分发，接收到副本或分片的n个节点进行数据存储、索引存储和provide
- 节点A选择与A距离较近的n个节点异步进行索引的存储，并选择与A距离较近的n个节点进行索引的provide，接收到provide的节点存储provide内容。



副本控制调度

- 当与节点A距离最近的节点B检测到节点A连接断开，节点B从本地索引路由模块中拉取节点A的副本资源列表，依次向其他成员节点发送各个副本数据的FileCheck Msg，向其他成员节点发起数据资源副本数检测

- 成员节点接收到FileCheck Msg，若本节点有副本数据，则向节点返回文件时空证明PoSt Msg向B节点确认副本数据存在，否则回复副本数据不存在消息FileCheckFailed Msg
- 节点B副本控制模块校验各节点的PoSt，并收集有效副本数量
- 若收集到的副本数少于阈值，则节点B发起副本再次调度，筛选一定数量的无此副本数据的节点，向其发送DoReplication Msg，并在消息中附带有有效副本数据的节点id
- 收到DoReplication Msg的成员节点从有数据节点中同步副本数据本地存储，更新索引数据，并发起新数据的provide，向B节点发送ReplicationACK Msg确认副本数据已存储，并附带PoSt证明。

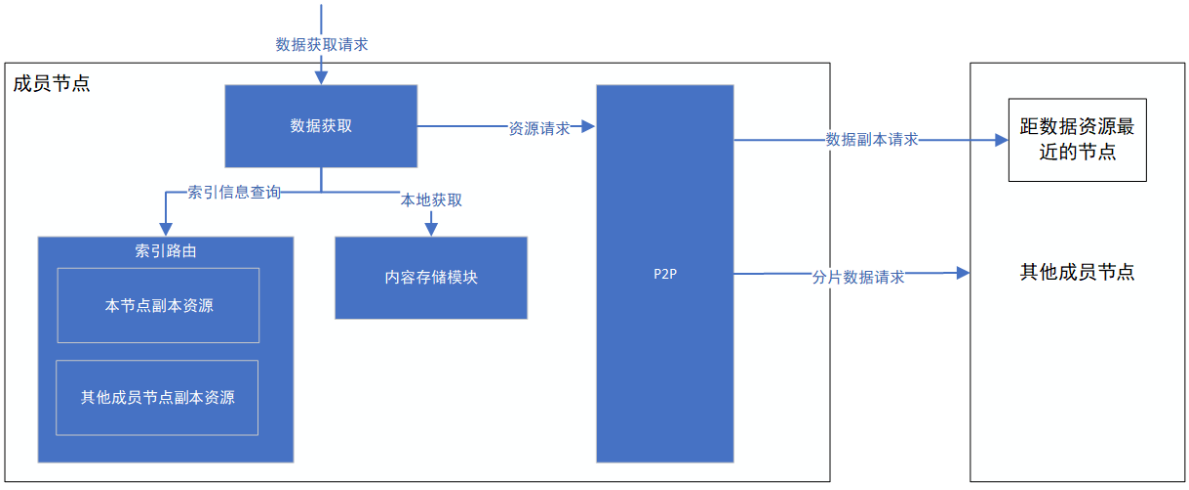


问题分析:

- 当宕机的节点持有数据资源总量较多，副本调度过程中的通信、数据存取等开销会加大，会否影响正常业务。各成员节点的繁忙程度不同，也无法统一设置任务优先级。
- 若副本再调度过程中无足够节点，副本调度模块将存在副本缺失情况的数据标识等信息存储到本地，待有新节点加入时，再调度
- 初期暂时免除PoSt证明及验证过程，由成员节点自行检测副本数据是否存在，后期优化

数据获取

- 节点接收到数据获取请求，先从本节点索引数据中查询本地是否存在；
- 若本节点不存在该数据，则查询本节点索引中保存的其他节点的数据资源列表中是否有该数据，向对应节点拉取
- 若本节点及本节点维护的索引数据中均无该数据，则向距数据资源最近的若干节点发送数据资源请求消息FileWant Msg；
- 接收到FileWant Msg的节点先查询自己及本节点维护的其他节点索引中是否有数据，有则向请求节点返回，否则，将FileWant Msg转发其他成员节点；
- 请求节点接收成员节点返回的应答消息，对于回复拥有该资源的节点，主动向该成员节点拉取数据，并进行索引更新、数据资源的provide；
- 若无节点拥有副本资源，则向成员节点问询是否存在该数据资源的分片信息，基于分片数据恢复出原文件，并对原文件重新进行副本分配。



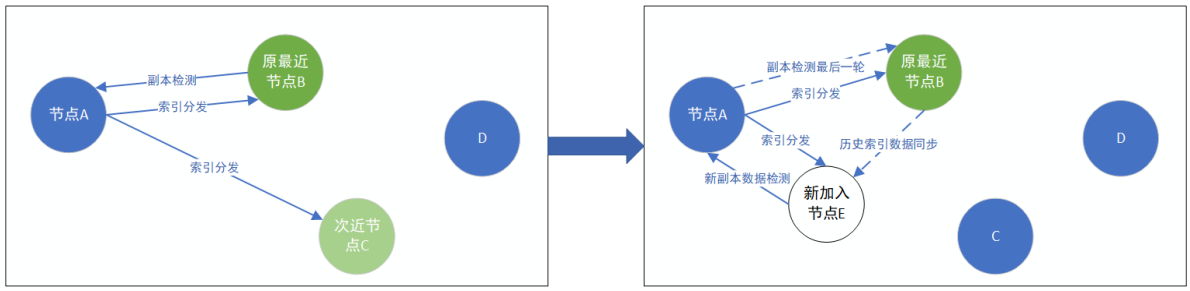
节点的加入和退出

节点的加入和退出会导致节点间相对距离的变化从而导致节点间数据资源、索引数据的分发以及副本检测任务的调度。

新节点加入

当有新节点加入时，节点间相对距离发生变化，主要影响数据资源索引数据的调度以及副本数检测任务的调度：

- 对于副本资源及分片资源索引数据的分发，有新节点加入时，各节点检测新节点是否距本节点距离更近，并将索引数据分发到相对距离更近的若干节点。成员节点间达成共识，原距离较近节点将历史索引数据同步到新的节点。
- 对于副本数检测任务，旧最近节点应基于本地索引数据完成正在进行的一轮检测和调度

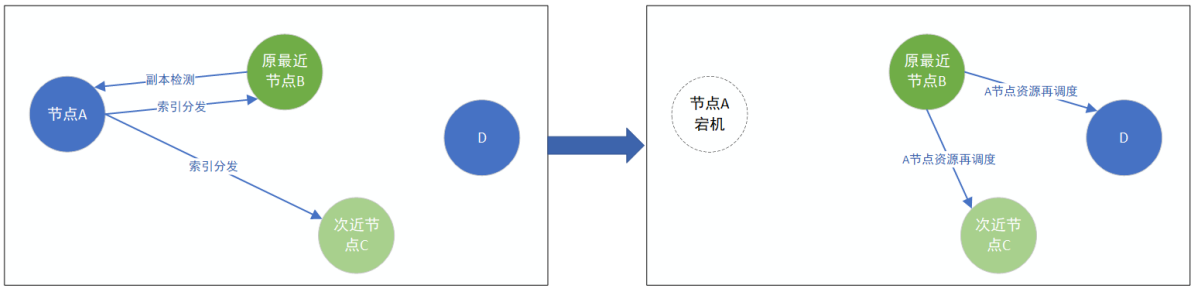


节点退出

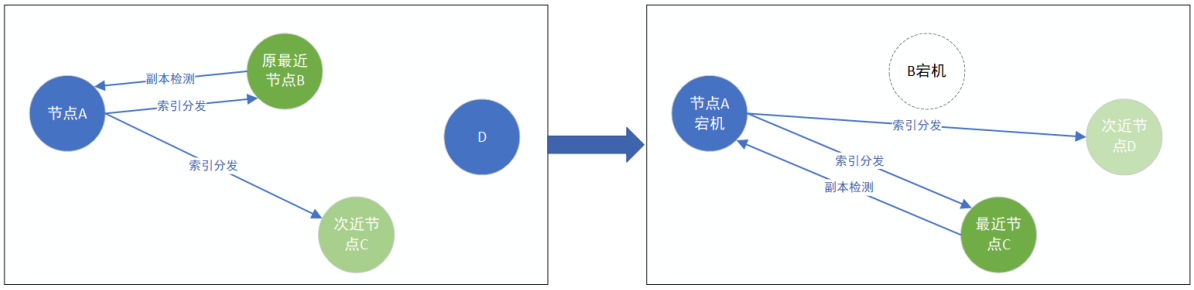
节点宕机等情况退出存储节点网络后，该节点的副本数据、分片数据、索引数据等将丢失。

- 副本或分片数据：与该节点距离最近的节点检测到节点失联后，会基于本地存储的索引数据，对退出节点拥有的数据资源进行检测和再调度。
- 索引数据：由于索引数据分发时采用多节点分发的策略，且索引数据高可用性需求较低，不再进行重复调度。

节点宕机后资源再调度



A节点的最近节点宕机后，索引分发变和副本检测



实施路径

一期：基于ipfs和libp2p重构存储节点，实现数据的多副本存储以及副本数控制

二期：支持数据多副本隔离存储以及数据分片分发与还原

三期：数据资源检测过程支持PoSt证明