

分布式存储系统重复数据删除技术研究

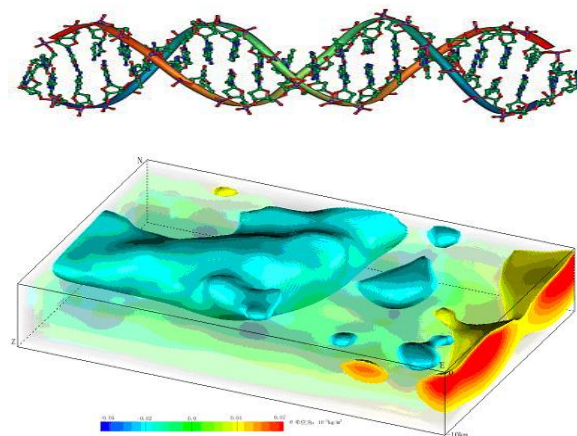
付印金

Email: yinjinfu@gmail.com

数据驱动应用无处不在

■ 科学研究

- ◆ 基因组计划
- ◆ LHC加速器
- ◆ 地球与空间探测



■ 企业应用

- ◆ 邮件、文档
- ◆ 日志
- ◆ 交易记录

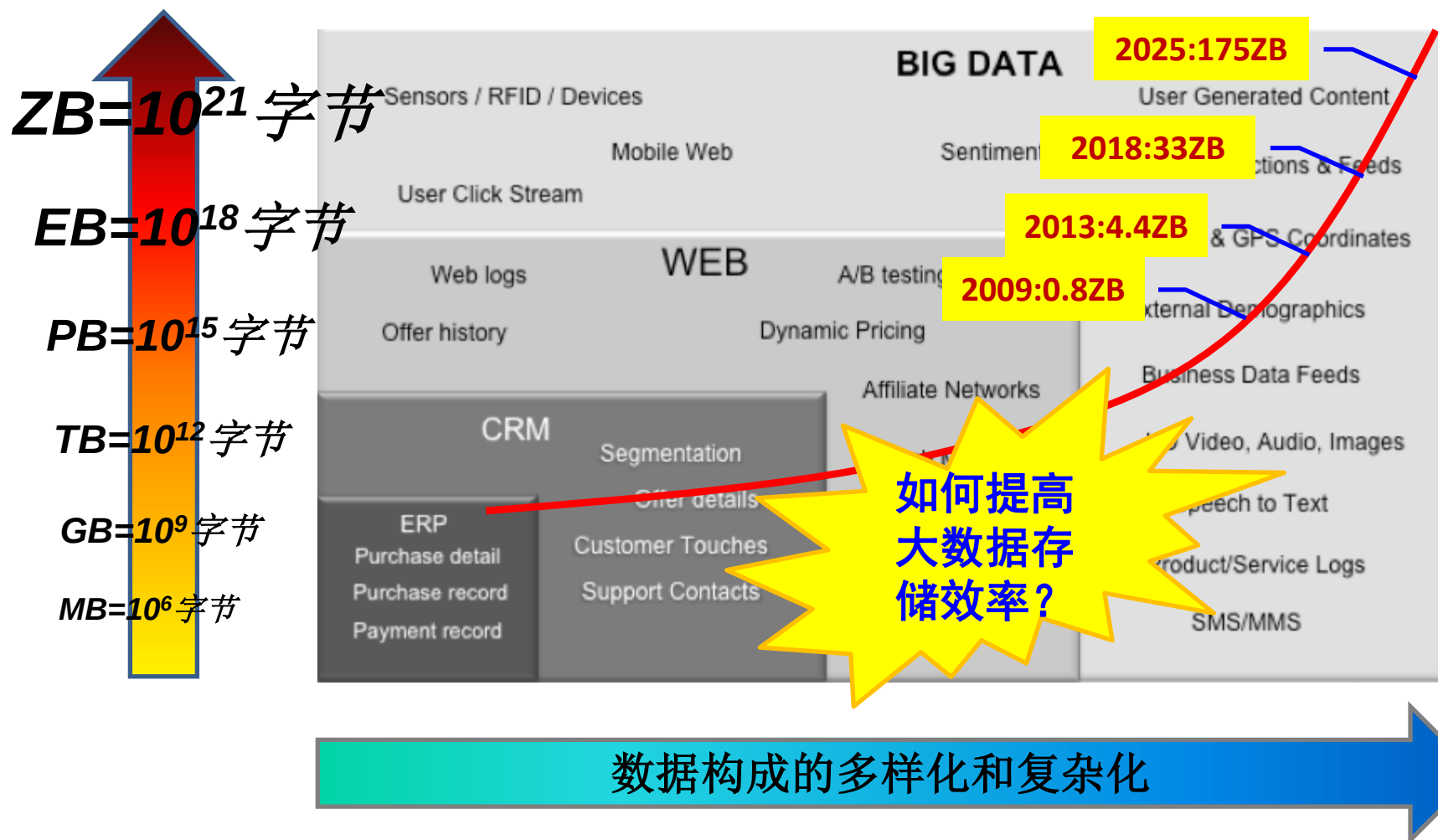


■ Web数据

- ◆ 文本、图像、视频
- ◆ Blog、点击流



大数据管理挑战



Source: IDC Whiter Paper, Data Age 2025.

压缩技术的应用

程序设计 (算法/空间/时间效率)

全文索引 (倒排索引表)

人工智能 (专家系统, 知识树)

编译 (JAVA)

文件系统
(压缩扇区)

密码学 (消除数据的原始特征性)

数据库 (B+树)

归档 (RAR/ZIP)

音频 (MP3)

视频 (MPEG/RM)

存储 (压缩池)

图像 (GIF/JPEG)

通讯 (Modem/网络协议)

电报传真 (CCITT)

压缩技术分类

通用数据压缩
(均为无损压缩)

多媒体数据压缩
(无损和有损压缩)

基于统计模型
的压缩技术

基于字典模型
的压缩技术

图像压缩

音频和视频压
缩MPEG等

Huffman
编码

算术编码

LZ77

LZ78

LZW

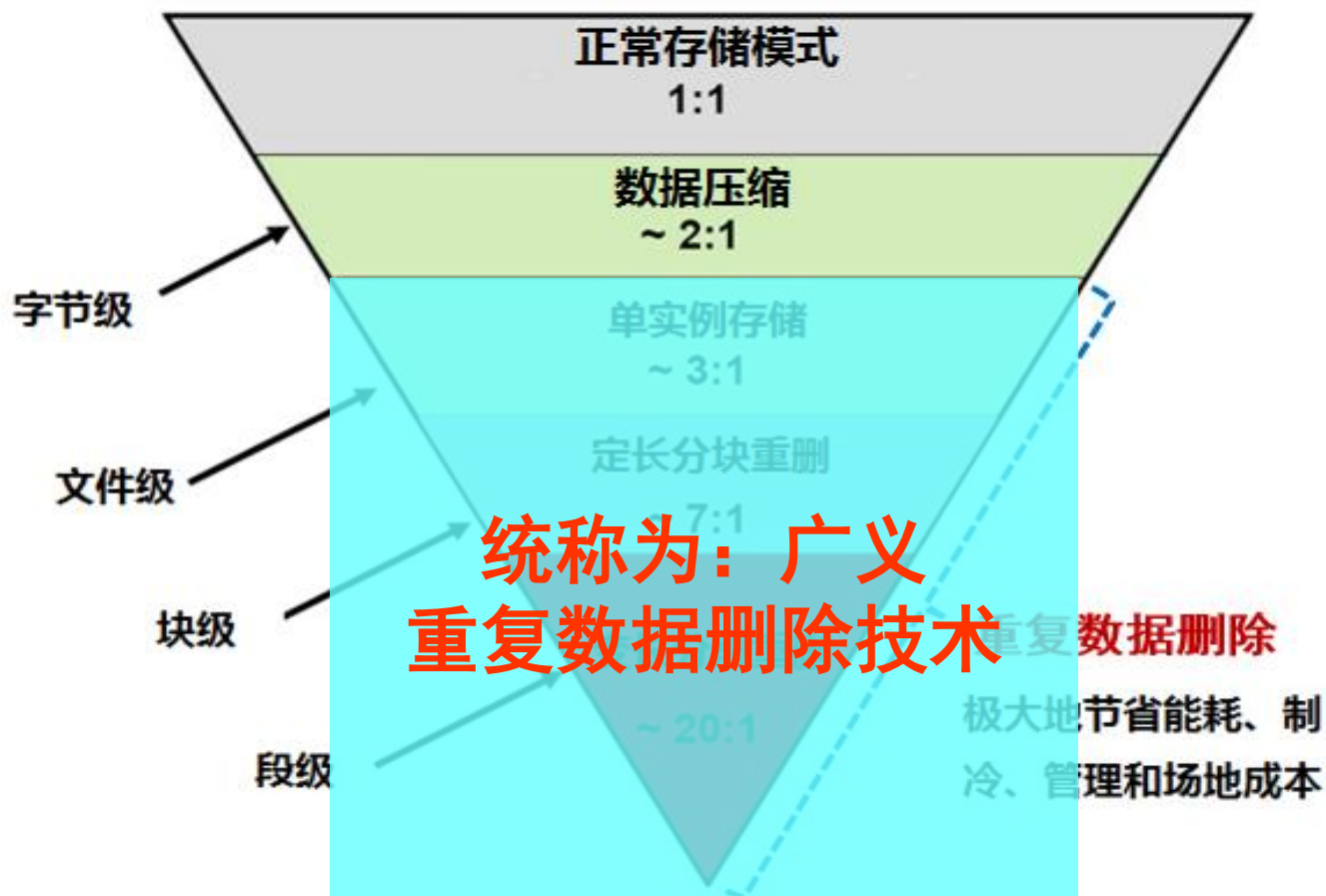
二值图像
CCITT
JBIG等

灰度图像
FELICS
JPEG等

彩色图像
RLE编码
JPEG等

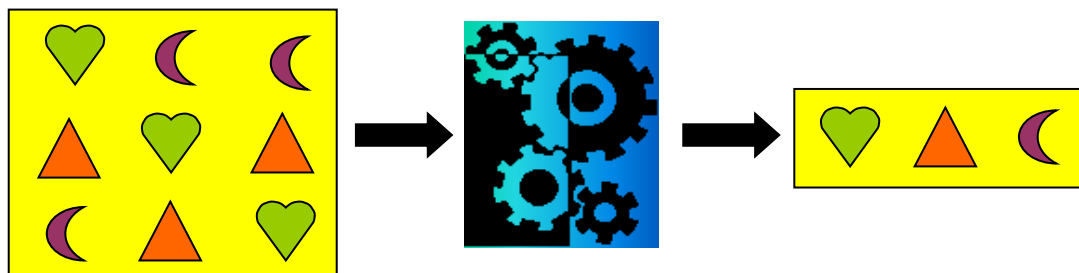
矢量图像
PostScript
WMF
CAD等

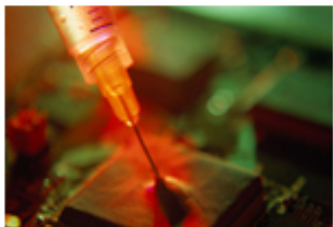
数据缩减技术对比



重复数据删除：一种面向大数据管理的压缩技术

一.重复数据删除技术





■ 外科手术器材制造厂

- 归档文件数据
- 38% 空间节省

■ 全球性的投资管理公司

- 复制 VMware 备份
- 88% 空间节省

■ 全球性的石油公司

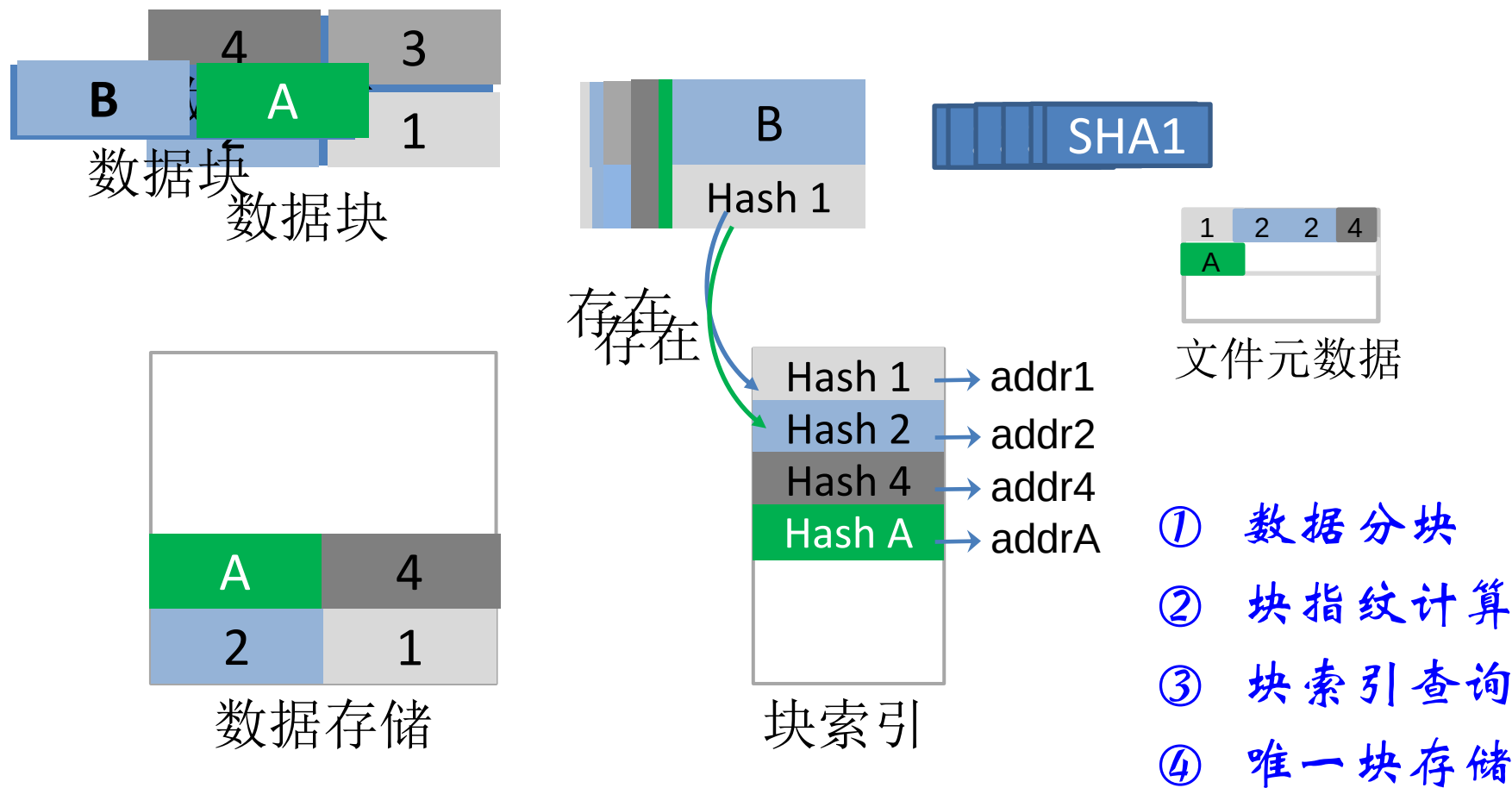
- 用户文件
- 35% 空间节省

■ 试验计量厂商

- 每天的数据库备份
- 98% 空间节省

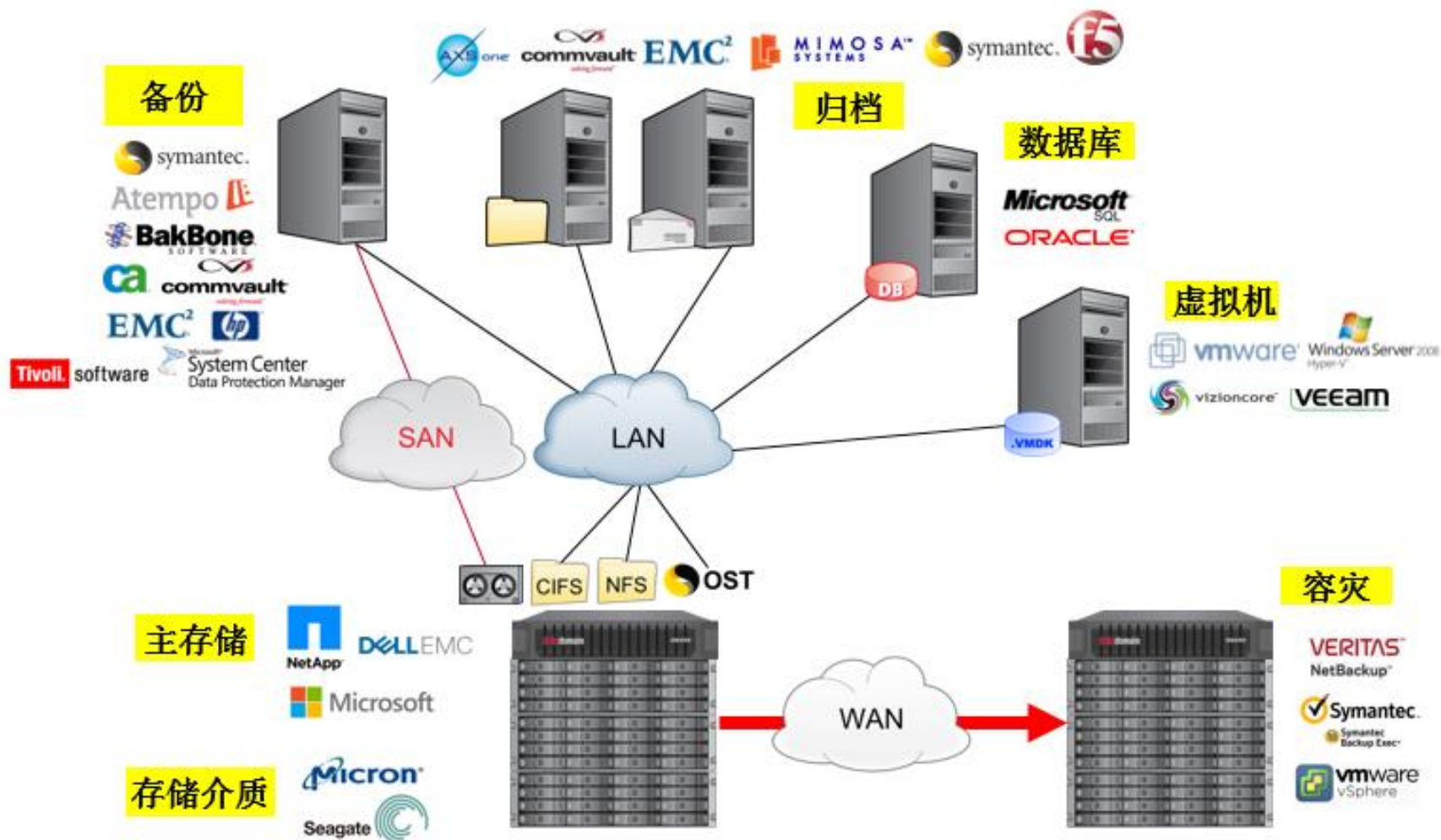
IDC 2010年: 数字世界75%的数据为重复的

重复数据删除处理流程



思想：以时间换空间，以计算换存储

重复数据删除技术应用



重复数据删除技术分类

■ 操作粒度

◆ 文件级重删 vs. 块级重删

■ 操作位置

◆ 源端重删 vs. 目标端重删

■ 操作时机

◆ 在线重删 vs. 后处理重删

■ 操作效果

◆ 精确重删 vs. 近似重删

■ 操作节点数

◆ 集中式重删 vs. 分布式重删

分布式存储系统重复数据删除

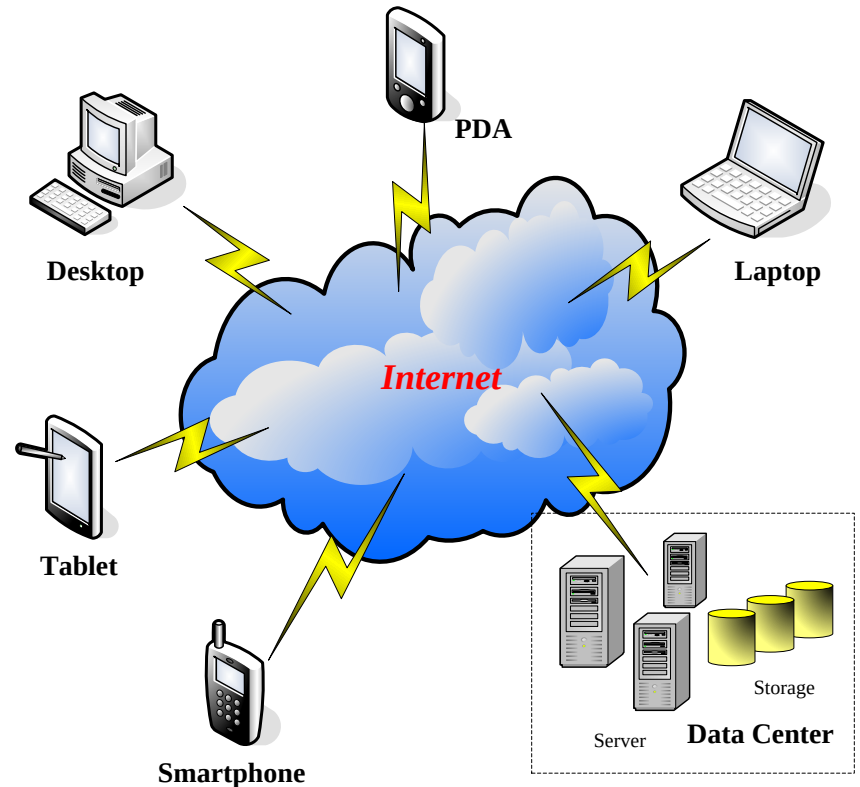
➤ 对海量复杂的大数据集进行可扩展的在线分布式重复数据删除存储处理

- 客户端

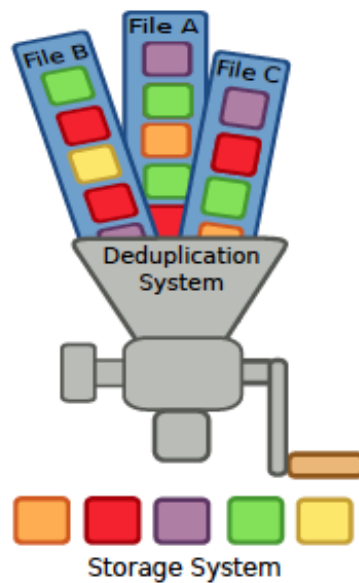
- 源端重复数据删除

- 数据中心

- 分布式重复数据删除



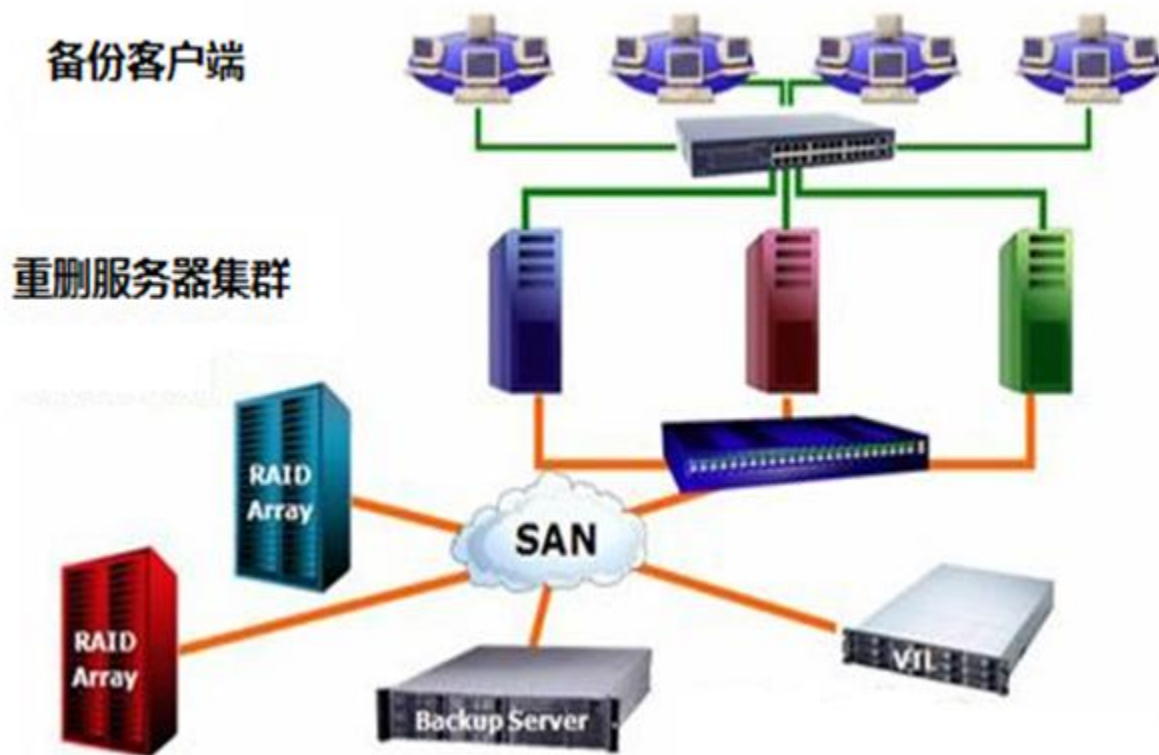
二. 分布式重复数据删除



分布式重复数据删除技术

■ 分布式重删原理

- 数据路由：客户端分配数据到重删服务器节点
- 重删操作：各个节点内独立并行地重删



分布式重删技术挑战

■ 节点“信息孤岛”

- 为节省系统开销，往往只对节点内部的数据进行消重，而不会去执行跨节点的重复数据删除

■ 块索引查询磁盘瓶颈

- 随着存储节点数据容量的扩展，磁盘数据块索引难以全部存放到重删节点有限的内存中，而基于哈希的块索引又不能保留数据局部性

传统分布式重删方法

■ 无状态方法

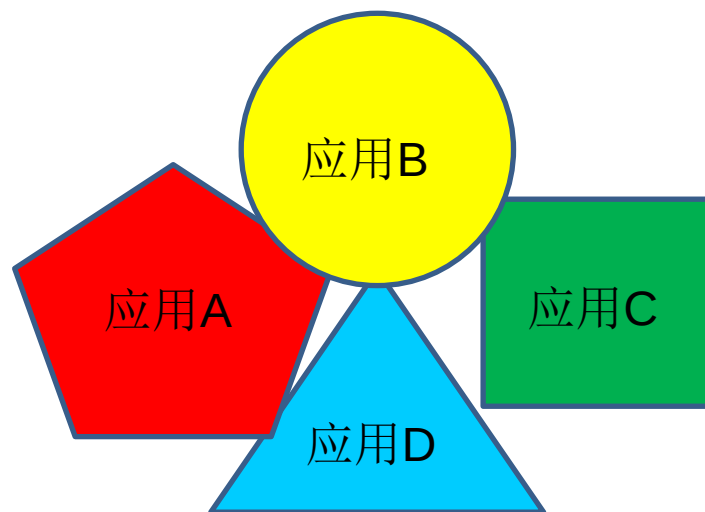
- 基于块指纹按DHT进行粗粒度数据路由
- 简单有效，但数据缩减率低、负载不均衡

■ 有状态方法

- 在数据路由前，需要广播查询粗粒度数据对象与各节点数据的相似度
- 数据缩减率高、负载均衡，但系统通信开销大

■ 基于块指纹的传统数据路由不能很好地平衡容量、性能和扩展性

应用间冗余率可以忽略不计

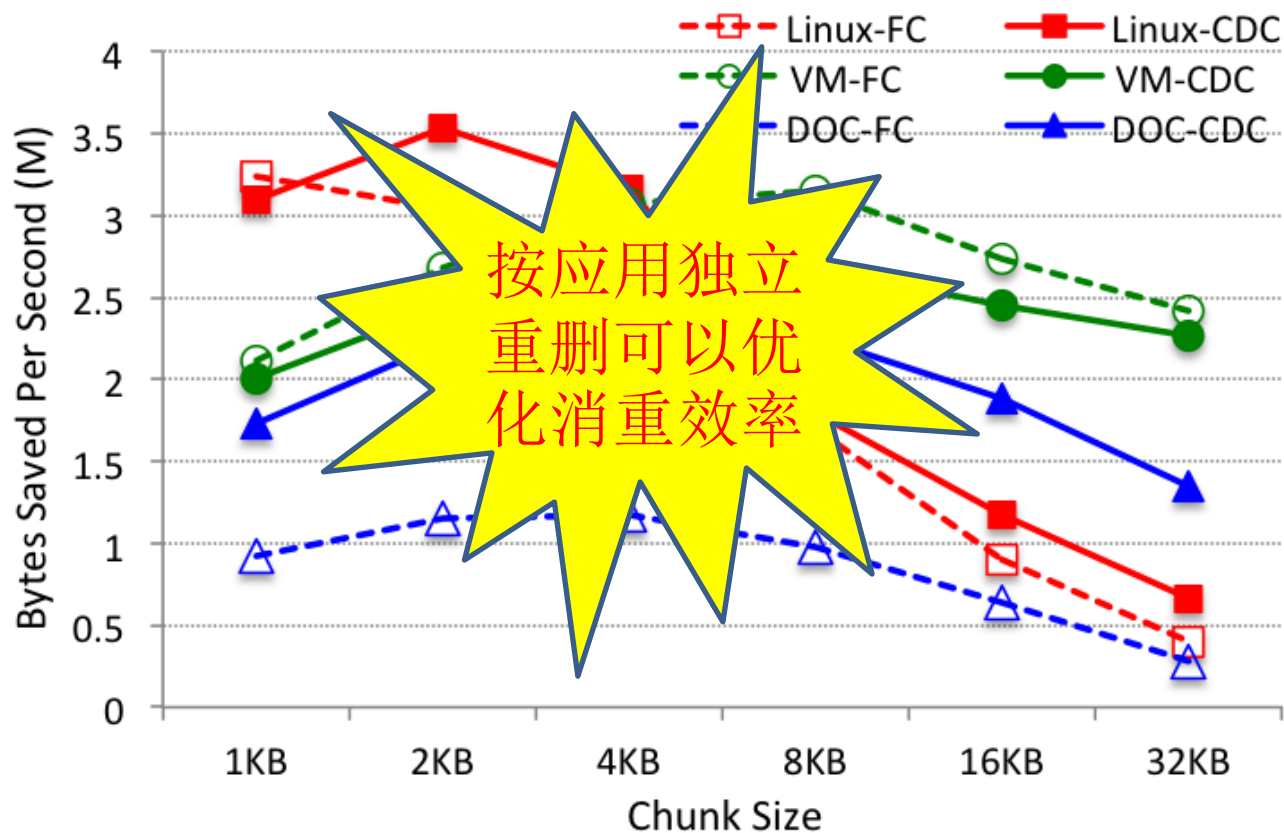


应用数据集	应用内冗余	应用间冗余
研究桌面电脑		1.1%
个人笔记本电脑		0.9%
个人工作站		0.6%
共享家庭服务器		0.7%
总冗余率	7.7%	0.8%

不同应用数据可以独立并行重删

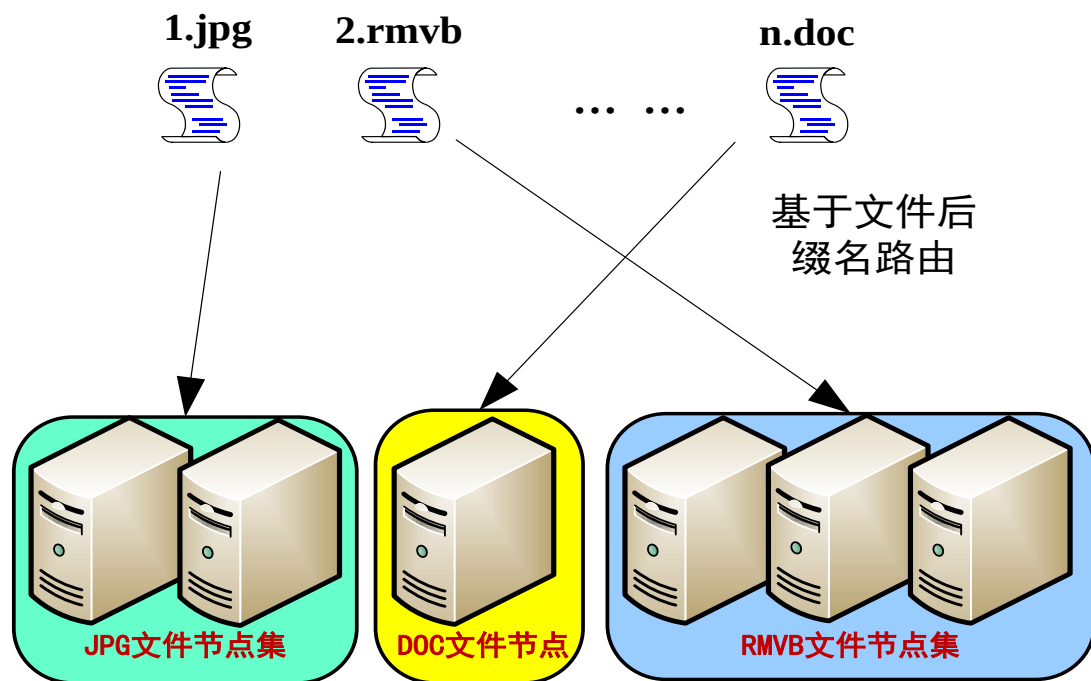
各应用数据消重效率不同

■ 不同应用数据集随着分块方法和块大小变化的消重效率



应用感知路由算法

- 利用“应用间冗余率可以忽略不计”的特性
- 首先对不同类型的应用数据进行隔离分配



手纹技术 (1/2)

➤ 同类应用文件服务器集群内基于超块路由

■ 超块:

- 数据流中一串连续的数据块构成超块

■ 手纹概念

- 取超块块指纹集的 k 个最小者作为特征指纹

■ 手纹 vs. 指纹

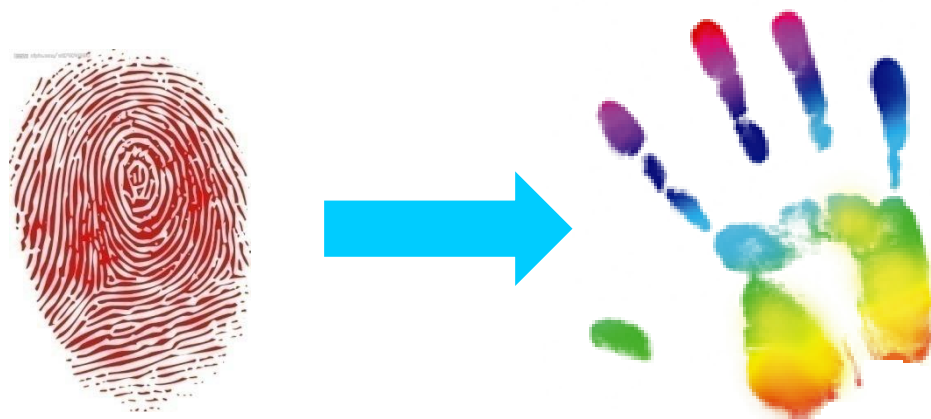
- 更强的相似检测能力

手纹技术 (2/2)

■ Border定理

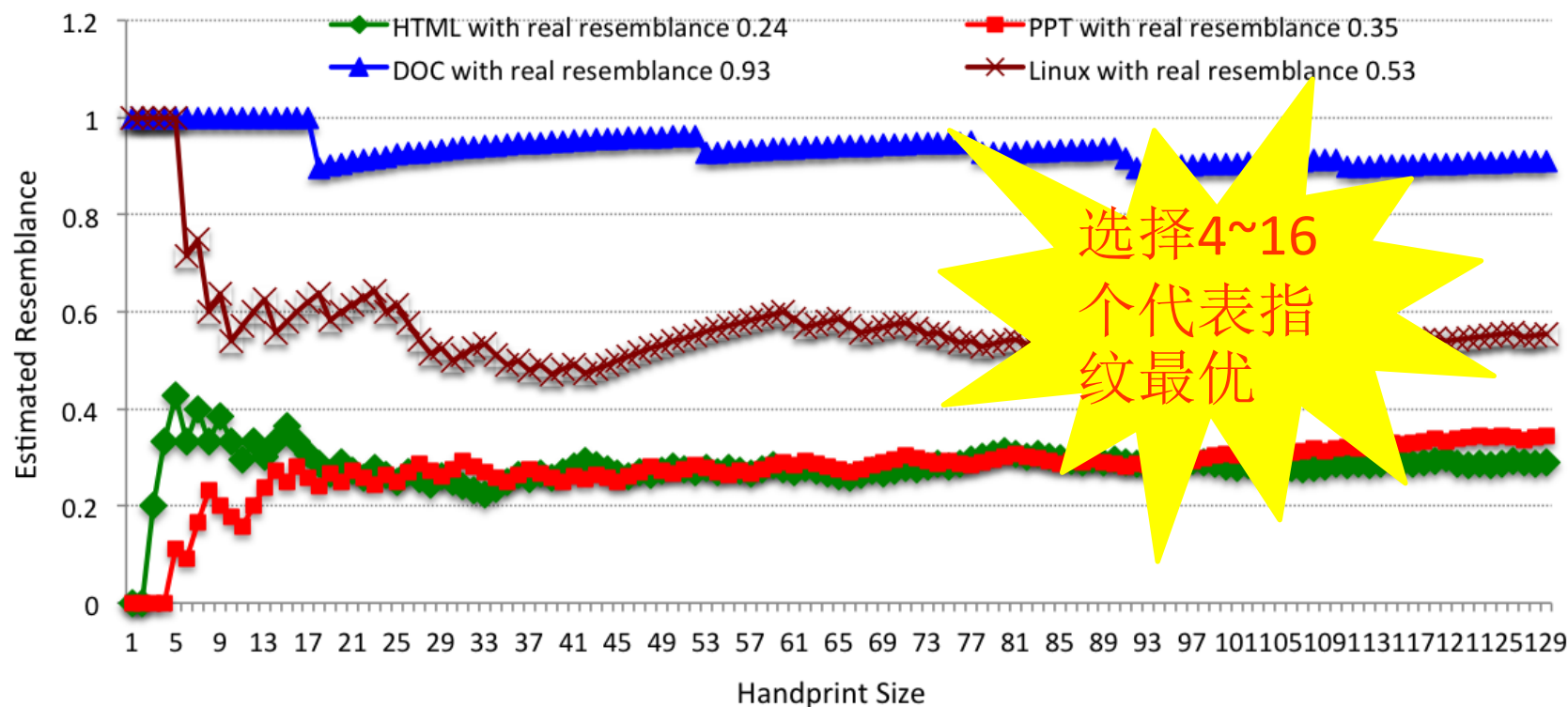
- 如果两个数据块集 S_1 和 S_2 的相似度为 r , 则这两个块集的 k 个最小块指纹中至少有一个为相同的概率是 $1 - (1 - r)^k$

$$\begin{aligned} & \Pr(\min_k \{h(S_1)\} \cap \min_k \{h(S_2)\} \neq \emptyset) \\ &= 1 - \Pr(\min_k \{h(S_1)\} \cap \min_k \{h(S_2)\} = \emptyset) \\ &\approx 1 - (1 - r)^k \end{aligned}$$

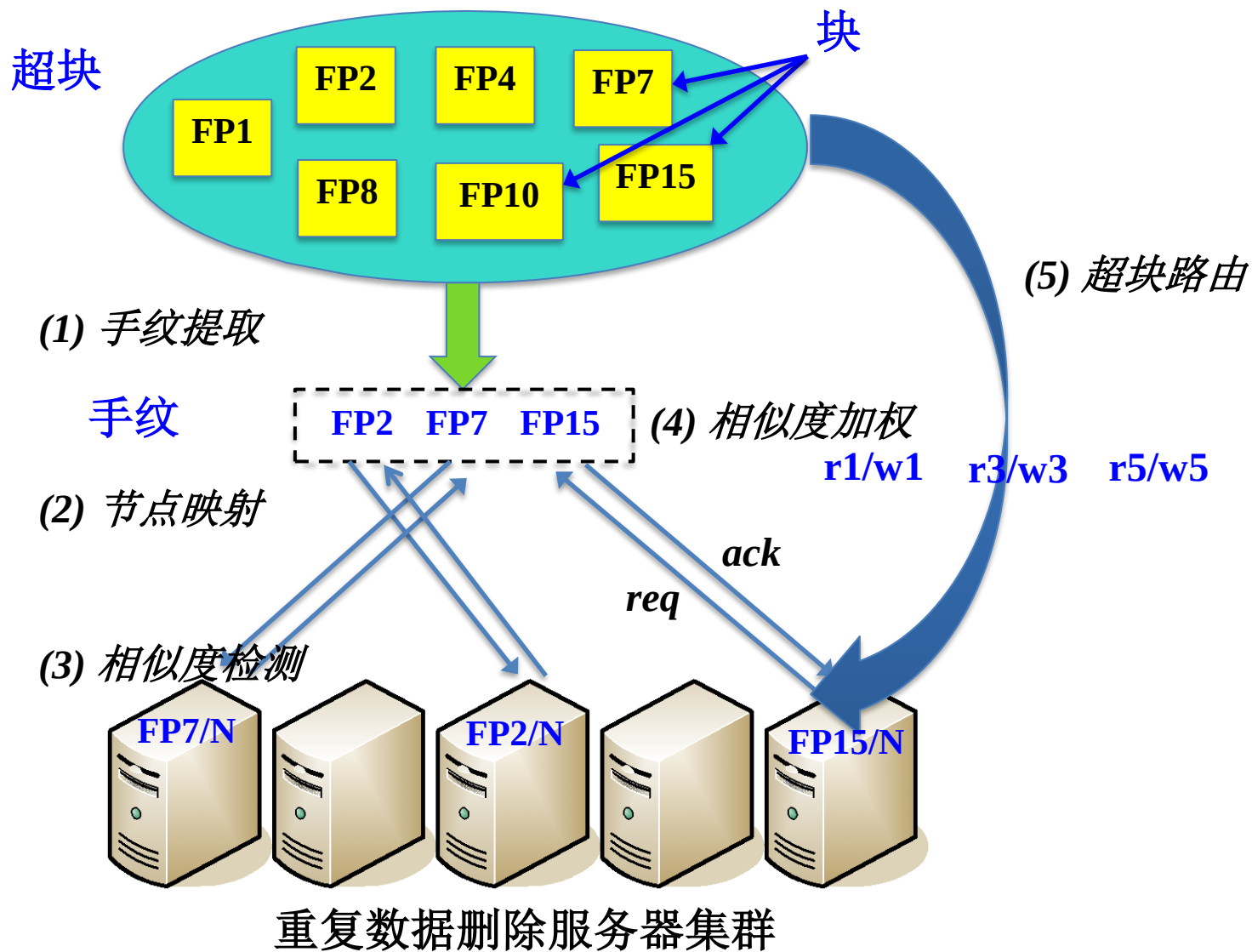


手纹冗余检测能力

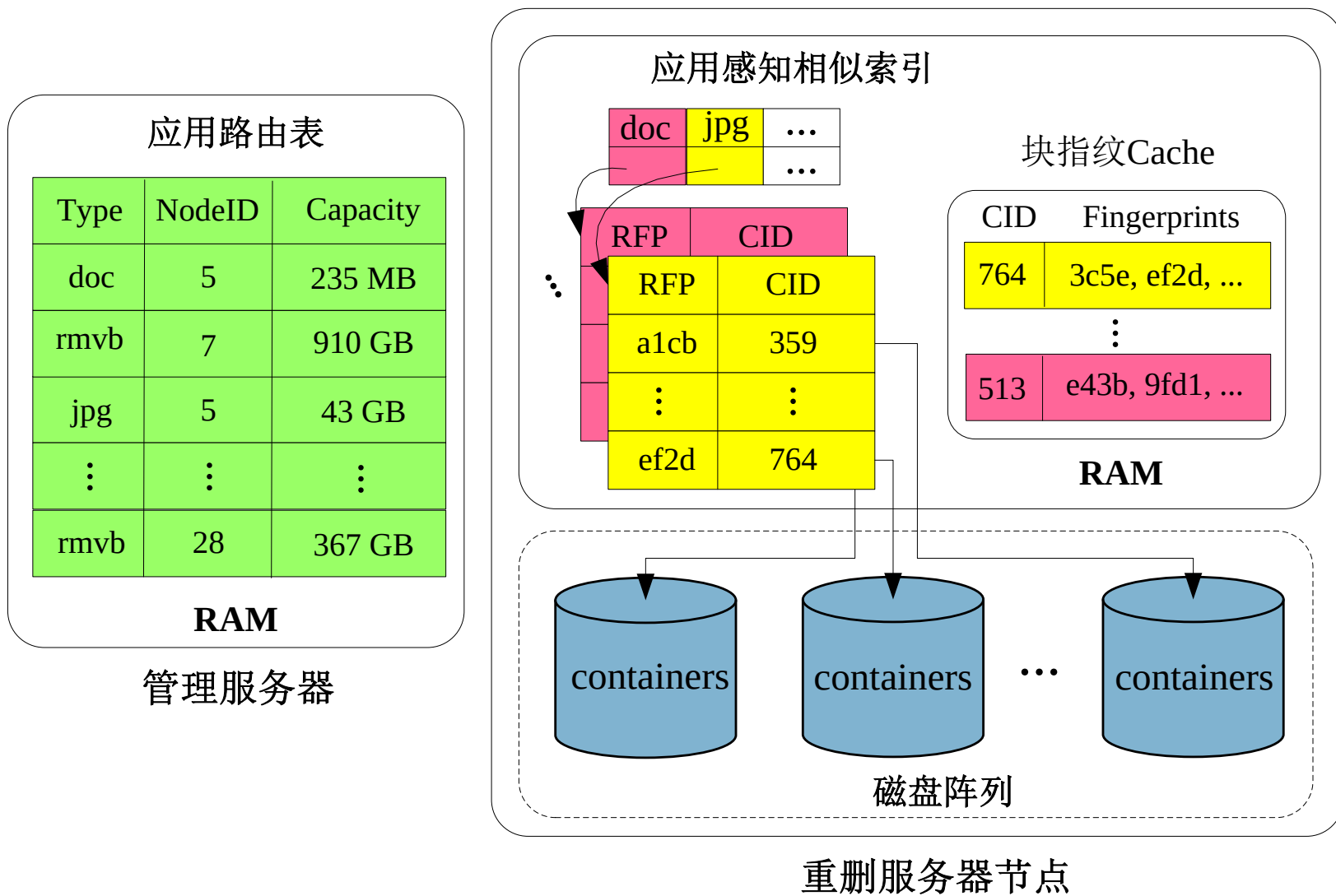
■ 估计值精度 VS. 存储开销



基于指纹的数据路由算法



应用感知数据结构



AppDedupe 分布式重删机制

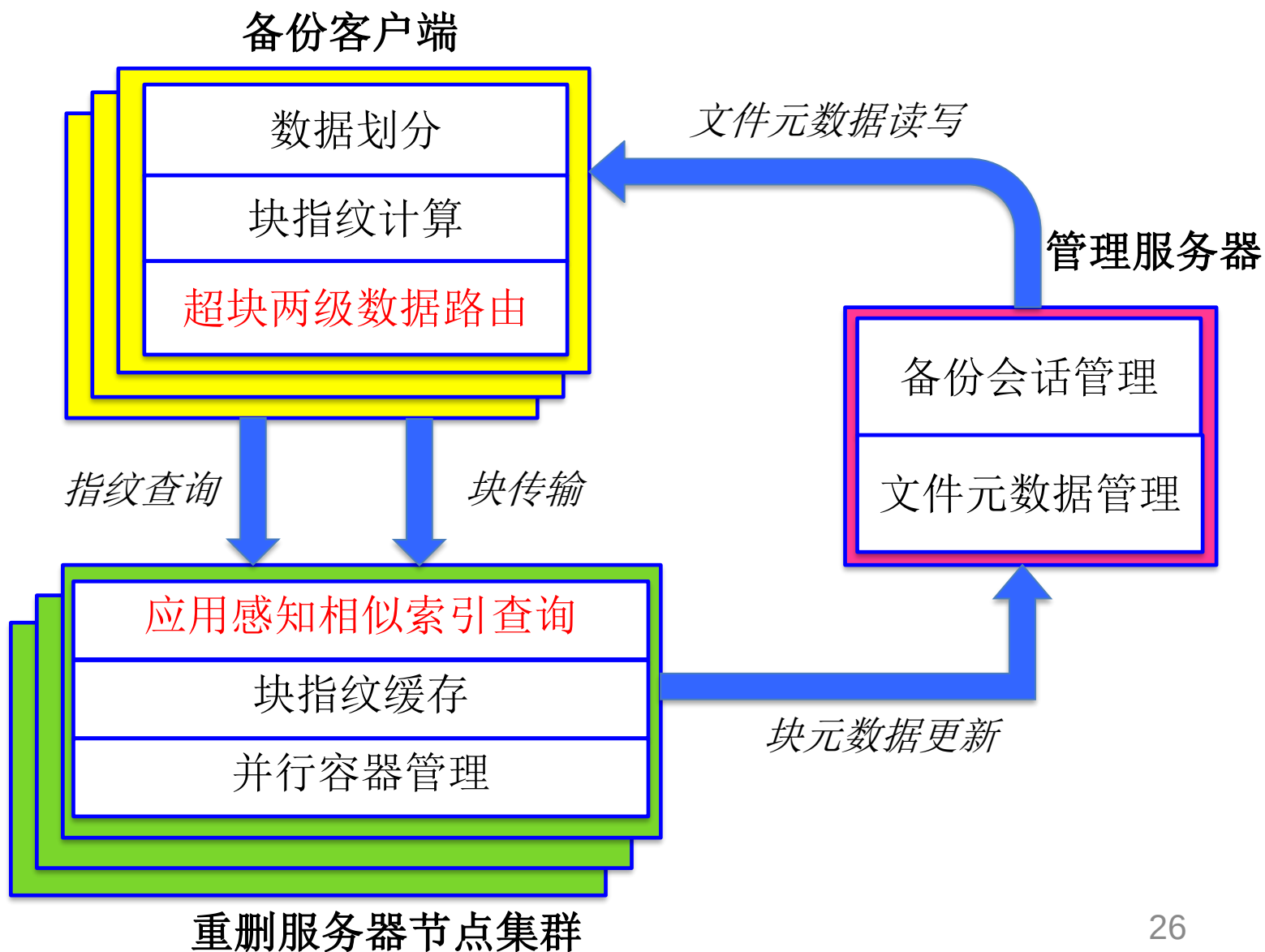
■ 两级数据路由机制来分配数据到重删节点

- 超块粒度分配保持局部性
- 基于应用类型文件路由机制
- 基于手纹的数据路由机制

■ 节点内重复数据删除操作优化

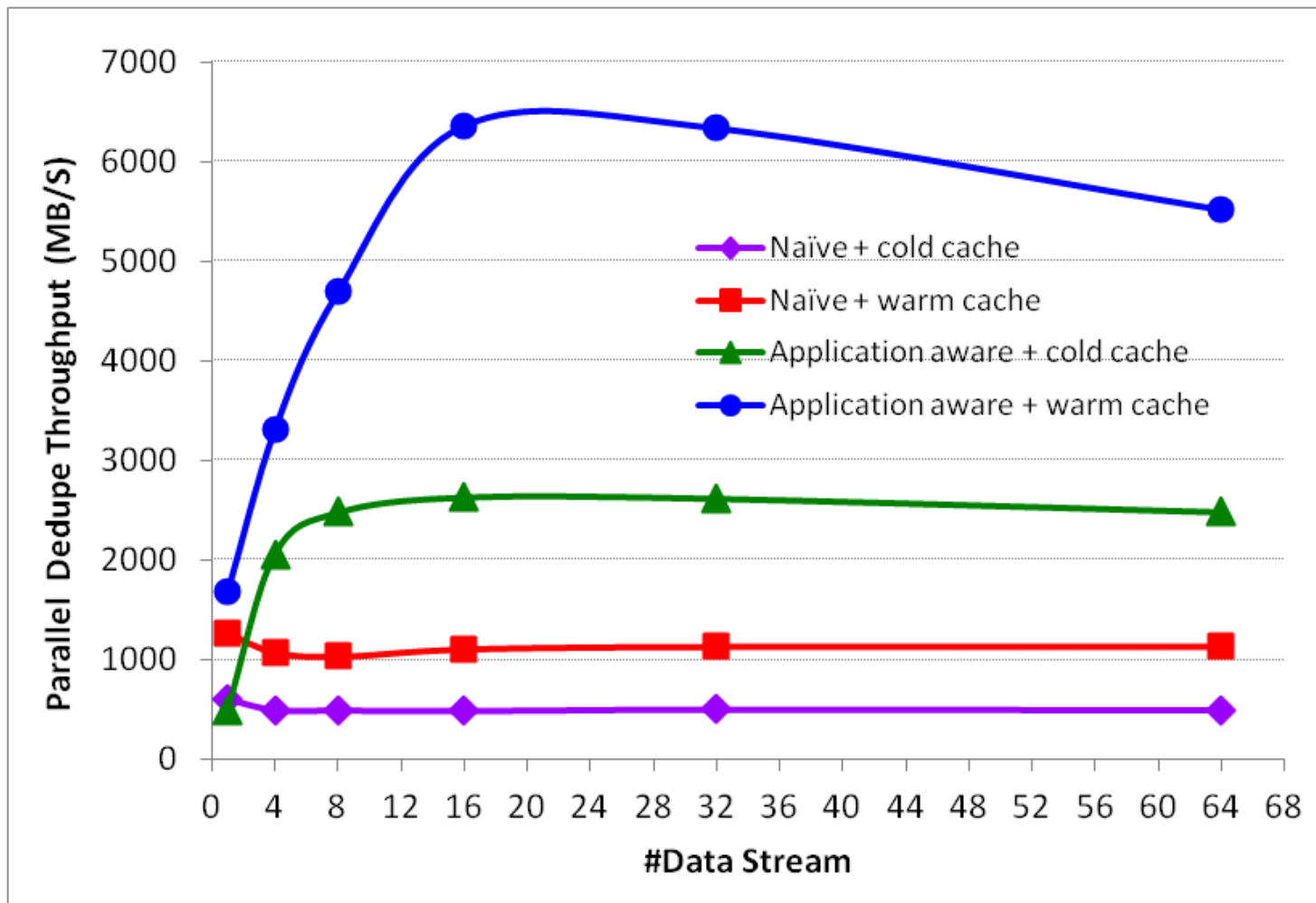
- 块级数据消重
- 相似索引结构
- 基于容器保持局部性

AppDedupe 分布式重删系统设计



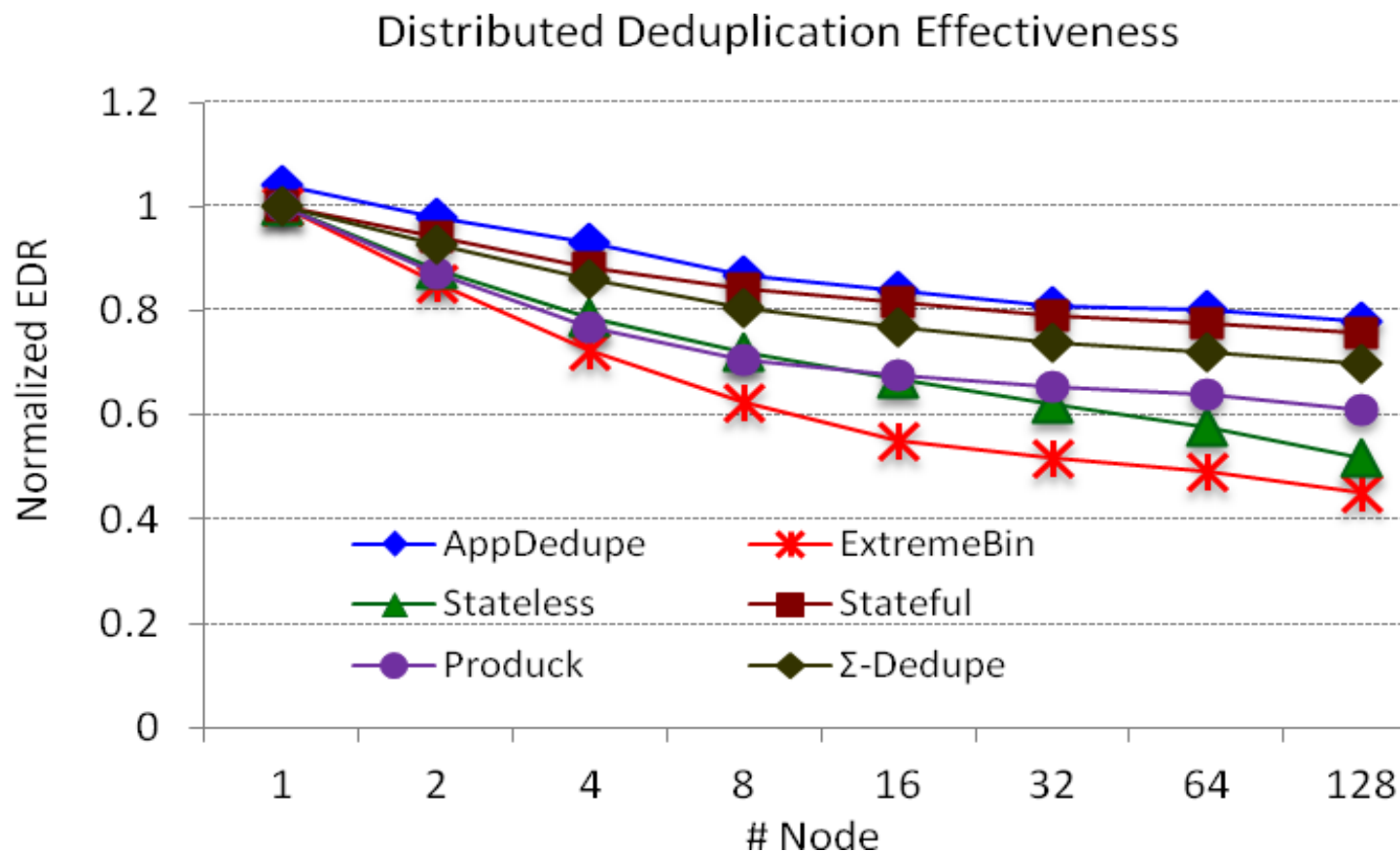
单节点并行重删吞吐量

■ AppDedupe节点内应用感知并行重删性能突出



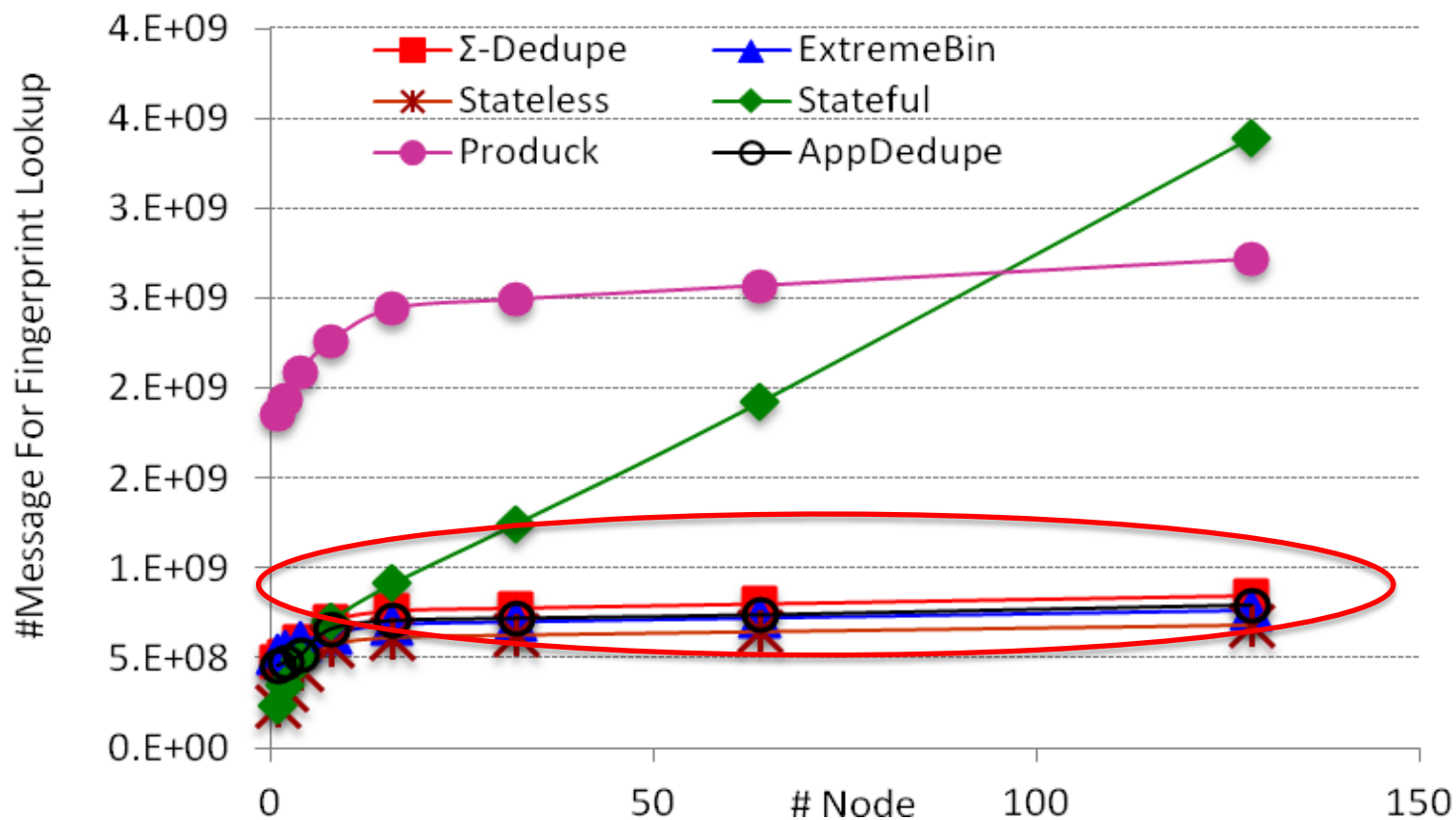
分布全局重删效果

■ AppDedup能获得比传统分布重删方法更高的空间节省

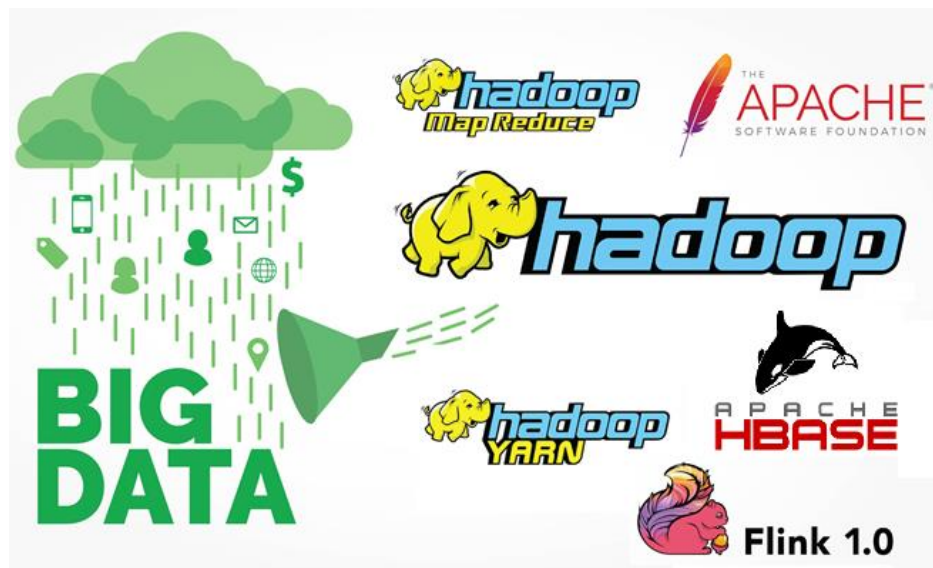


分布重删系统通信开销

■ AppDedupe 能获得近似传统无状态方法的消息开销



三. 可扩展Hadoop集群重删技术



Hadoop平台

■ Hadoop是当前最流行的分布式计算平台之一

■ 主要组件

- ① **HDFS**: 管理海量数据的分布式文件系统
- ② **MapReduce**: 处理海量数据的并行编程模型
- ③ **HBase**: 分布式的NoSQL列数据库

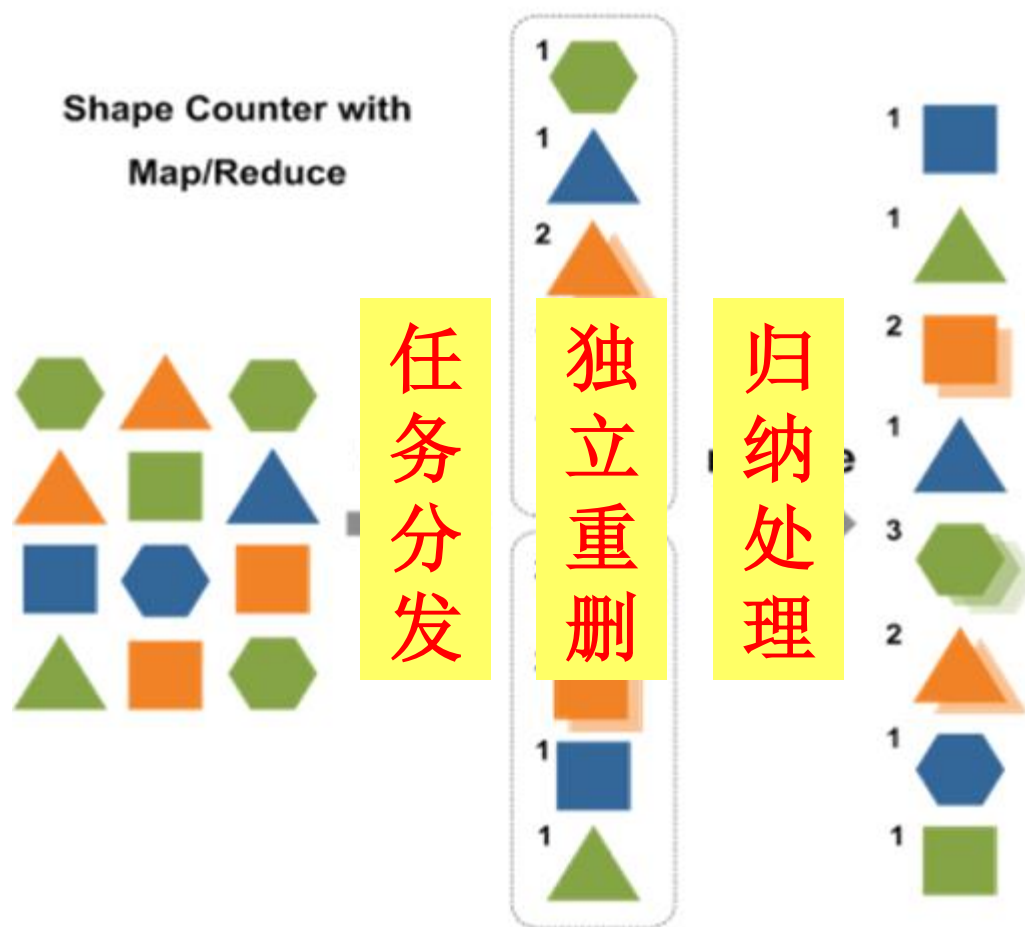
⋮



APACHE
HBASE

Hadoop与分布式重删

■ MapReduce过程能很好匹配分布式重删处理



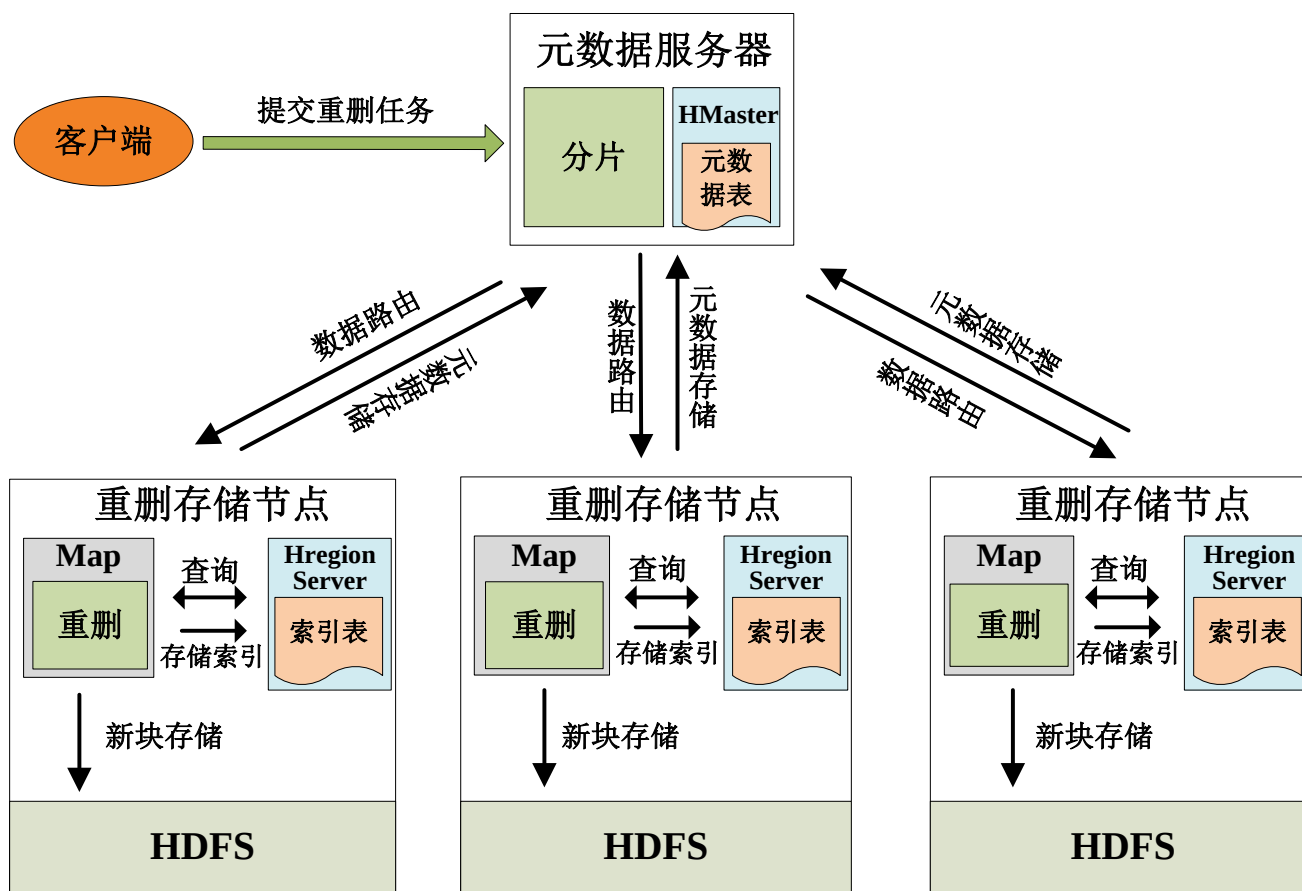
基于Hadoop的分布式重删设计

■ 基本设计思想

- MapReduce负责数据分发、重复检测和删除处理
 - Map函数将数据按MB级粗粒度分发到不同的重删节点
 - 重删节点独立进行4KB数据分块和指纹计算，删除相同指纹值的重复数据块
 - 为避免Sort和Shuffle过程的处理时延，去除Reduce过程
- HDFS负责存储不重复的数据块管理
- 两种可选的数据块指纹存储管理方式：
 - **HBase**: 负责存储全局的块指纹数据库
 - **MySQL**: 负责存储节点局部的块指纹数据库

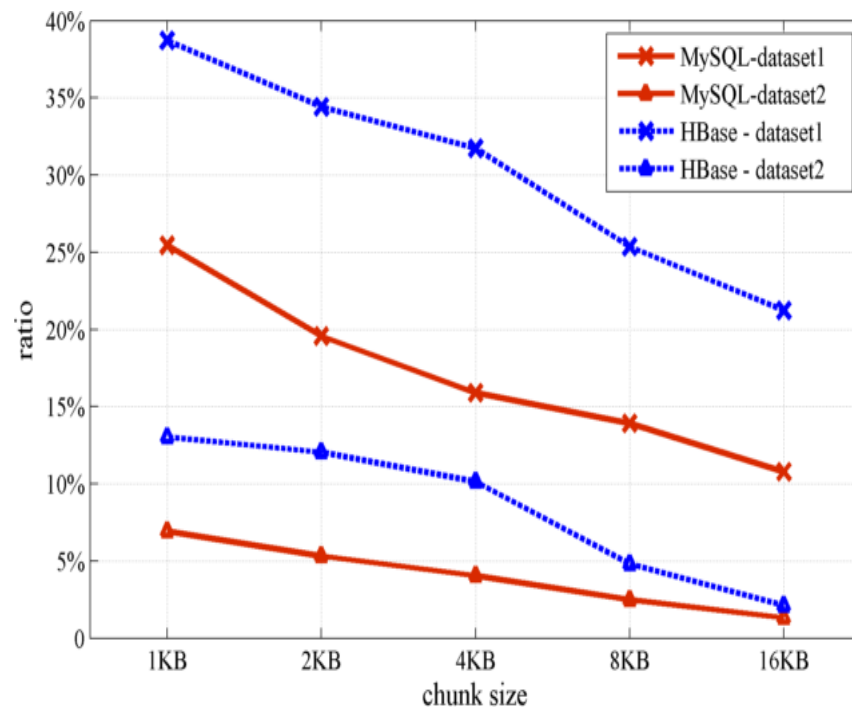
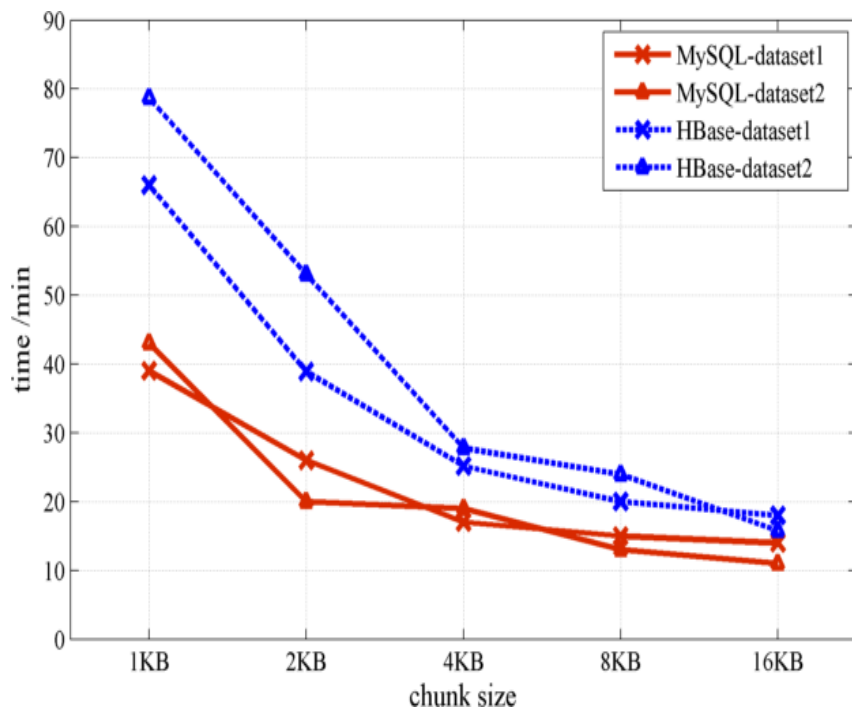
基于Hadoop的集群重删架构

- 客户端、元数据服务器、重删存储节点集群



实验结果分析

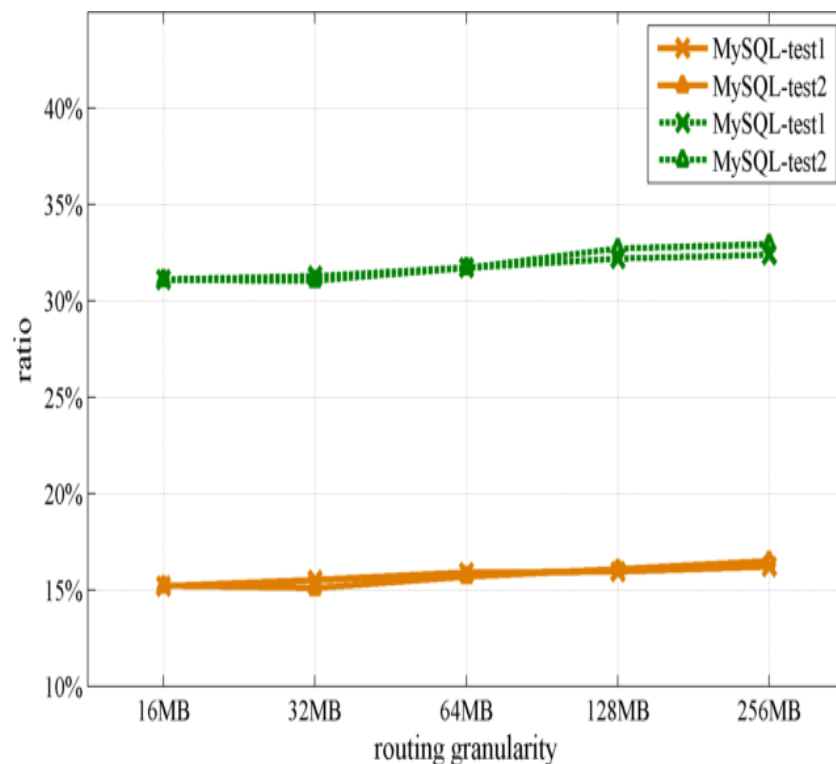
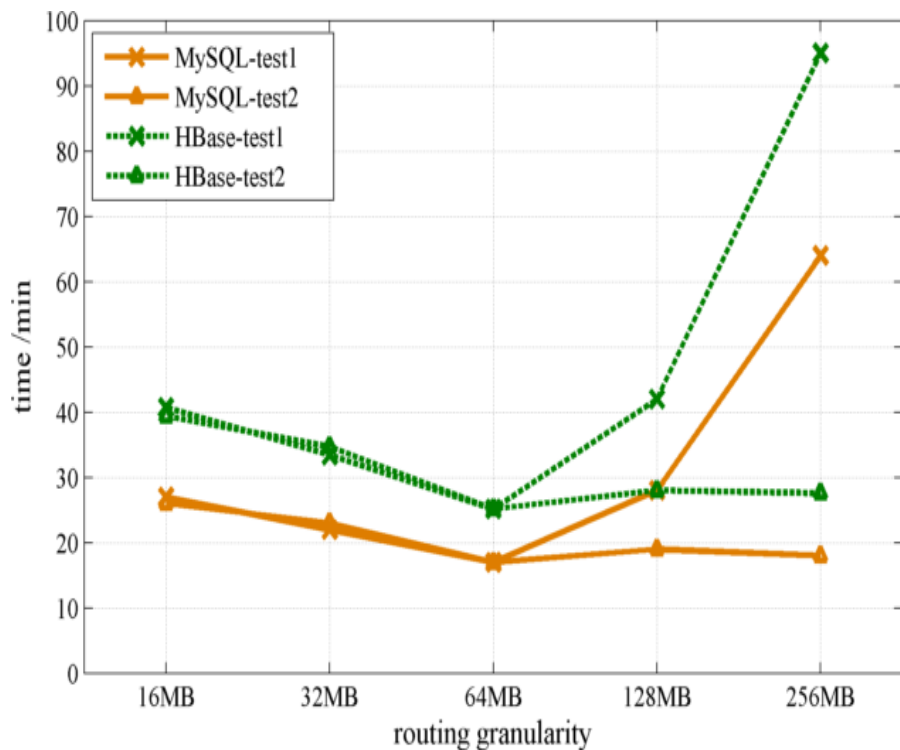
■ 数据分块灵敏度分析



- 数据分块越小，缩减率越高，处理时间越长
- 对比HBase存放全局索引，重删存储节点本地MySQL块索引存储的处理速度快，但缩减率低

实验结果分析

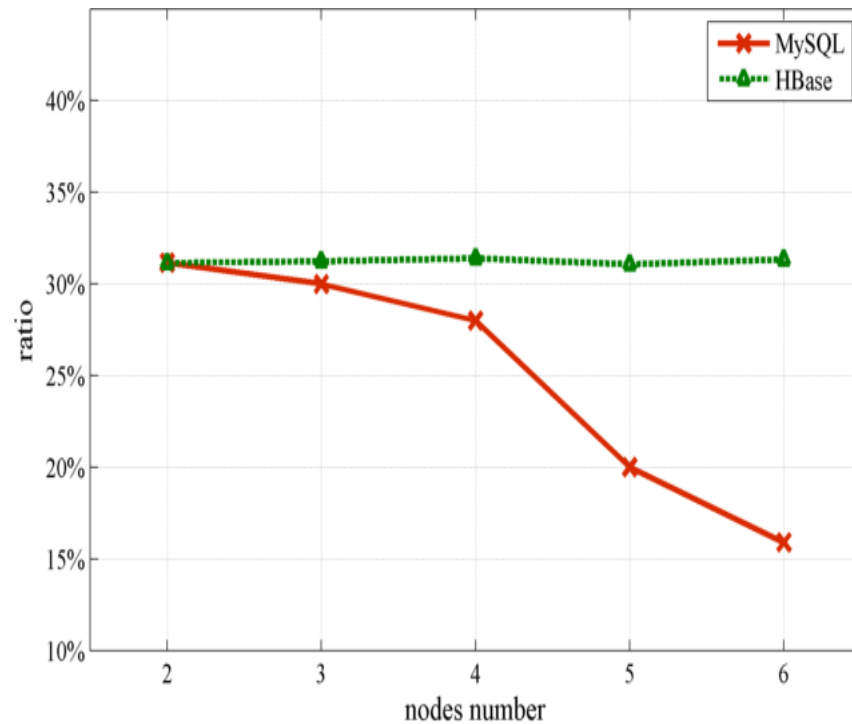
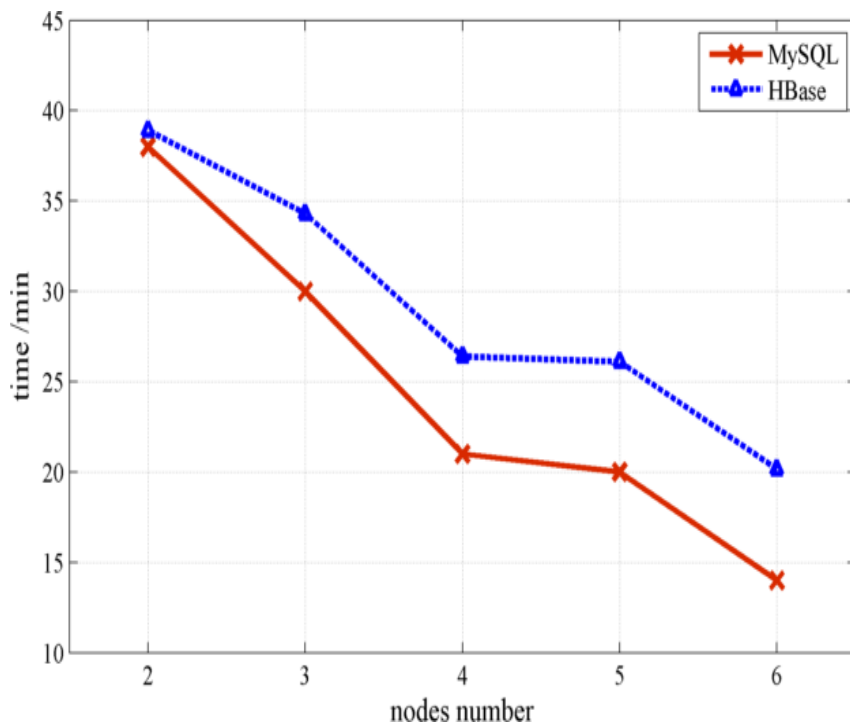
■ 数据路由粒度灵敏度分析



- 路由粒度为 **64MB**（注：HDFS块大小）时处理速度最快
- 数据缩减率基本保持不变

实验结果分析

■ 系统扩展能力分析

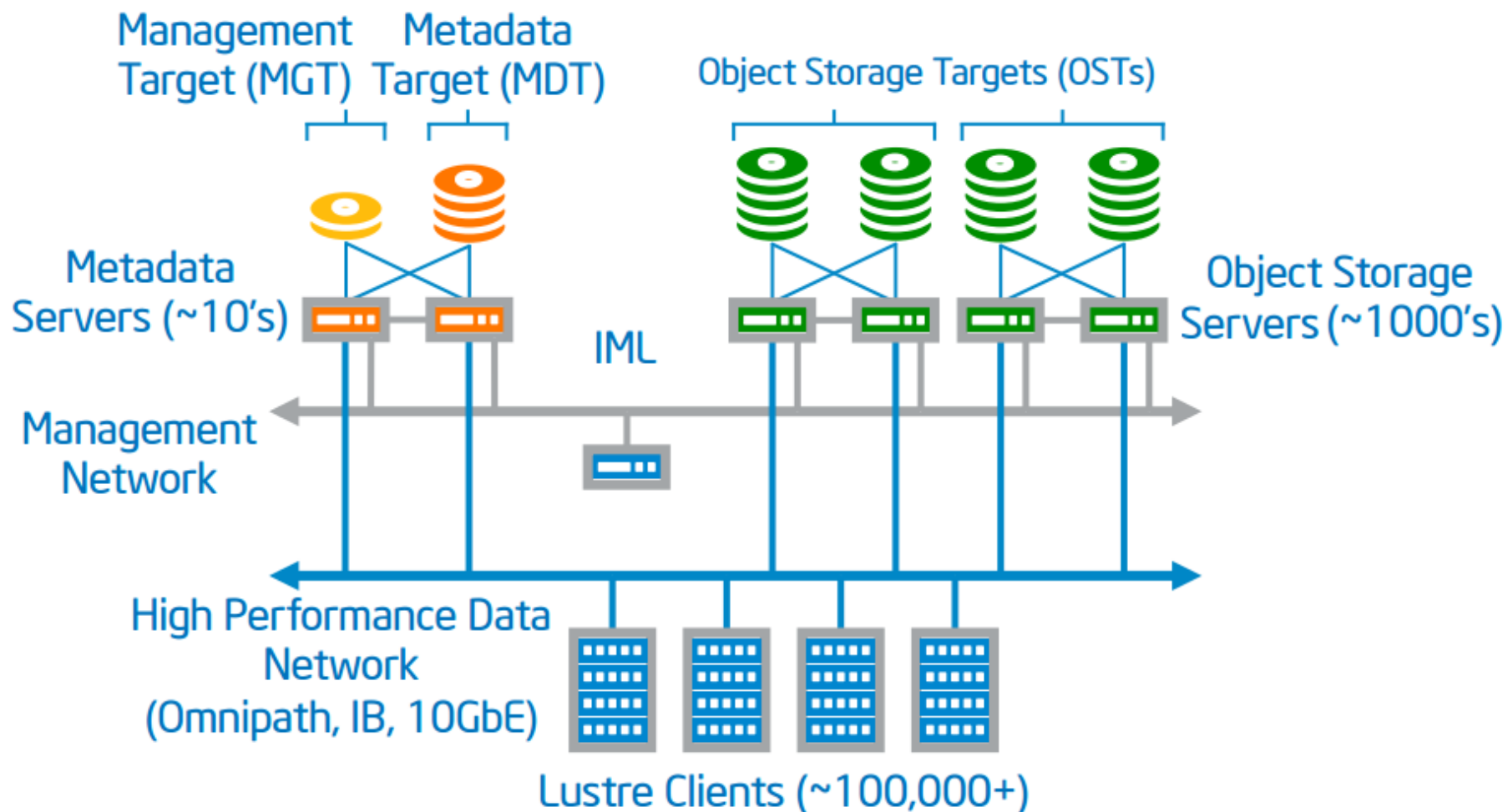


- 在规模较小的情况下，系统性能随节点数增加而近似线性扩展
- 本地MySQL的局部块索引机制性能更好，但随着系统扩展相应的数据缩减率也在递减

四. Lustre文件系统重删功能实现思考



Lustre 文件系统架构



Lustre重删实现方案

- 增加代码库：

- 加密哈希函数：计算块指纹，如Crypto++库
- 数据库引擎：指纹索引和元数据管理，如SQLite DB

- 检测重复数据：

- Clients计算块指纹，通过写请求发送到MDSs处理
- MDSs配置数据库，进行块指纹索引查询
- 可采用Bloom Filter等工具可以加速索引查询

- 重复数据替换：

- 文件级重删：硬链接实现，索引<文件指纹, 文件路径>
- 块级重删：修改Striping EA
 - 块大小=strip size，对象替换，索引<对象指纹, 对象ID>
 - 块大小≠strip size，块替换，索引<块指纹, 块地址信息>，<对象ID, 块指纹>

相关参考文献

1. **Yinjin Fu**, Nong Xiao, Tao Chen, Jian Wang. “Fog-to-MultiCloud Cooperative eHealth Data Management with Application-Aware Secure Deduplication.” *IEEE Transactions on Dependable and Secure Computing*, 2021.
2. **Yinjin Fu**, Nong Xiao, Hong Jiang, Guyu Hu, Weiwei Chen. “Application-Aware Big Data Deduplication in Cloud Environment.” *IEEE Transactions on Cloud Computing(TCC)*, 7(4): 921-934, 2019.
3. **Yinjin Fu**, Hong Jiang, Nong Xiao, Lei Tian, Fang Liu, Lei Xu. “Application-aware Source Deduplication for Cloud Backup Services of Personal Storage.” *IEEE Transactions on Parallel and Distributed Systems (TPDS)*, 25(5): 1155-1165, 2014.
4. Huijun Wu, Chen Wang, **Yinjin Fu**, Sherif Sakr , Kai Lu, Liming Zhu. “A Differentiated Caching Mechanism to Enable Primary Storage Deduplication in Clouds.”*IEEE Transactions on Parallel and Distributed Systems*, 29(6): 1202-1216, June 2018.
5. Bo Mao, Hong Jiang, Suzhen Wu, **Yinjin Fu**, Lei Tian. “Read Performance Optimization for Deduplication-based Storage Systems in the Cloud.” *ACM Transactions on Storage (TOS)*, 10(2), 2014.
6. **Yinjin Fu**, Hong Jiang, Nong Xiao. “A Scalable Inline Cluster Deduplication Framework for Big Data Protection.” *ACM/IFIP/USENIX 13th International Conference on Middleware (Middleware’12)*, Montreal, Quebec, Canada, Dec. 3-7, 2012, pages 354-373.
7. **Yinjin Fu**, Hong Jiang, Nong Xiao, Lei Tian, Fang Liu, “AA-Dedupe: An Application-Aware Source Deduplication Approach for Cloud Backup Services in the Personal Computing Environment.” *The 2011 IEEE International Conference on Cluster Computing (Cluster’11)*, Sep. 26-30, 2011, pages 112-120.
8. Huijun Wu, Chen Wang, **Yinjin Fu**, Sherif Sakr, Liming Zhu, Kai Lu. “HPDedup: A Hybrid Prioritized Data Deduplication Mechanism for Primary Storage in the Cloud .” *The 33rd International Conference on Massive Storage Systems and Technology (MSST 2017)*, May 15-19, 2017.
9. Huijun Wu, Chen Wang, Kai Lu, **Yinjin Fu**, Liming Zhu. “One Size Does Not Fit All: The Case For Chunking Configuration in Backup Deduplication.” *The 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid (CCGrid’18)*. May 1-4, 2018.
10. Qing Liu, **Yinjin Fu**, Guiqiang Ni, Rui Hou. “Hadoop based Scalable Cluster Deduplication for Big Data.” *The IEEE ICDCS SDDCS Workshop*, Nara, Japan, June 2016.
11. 付印金, 肖侖. 《重复数据删除技术》, 清华大学出版社, ISBN:978730256113, 2021.01.

谢谢！
敬请批评指正！