



面向云计算的高性能重复数据删除技术

付印金

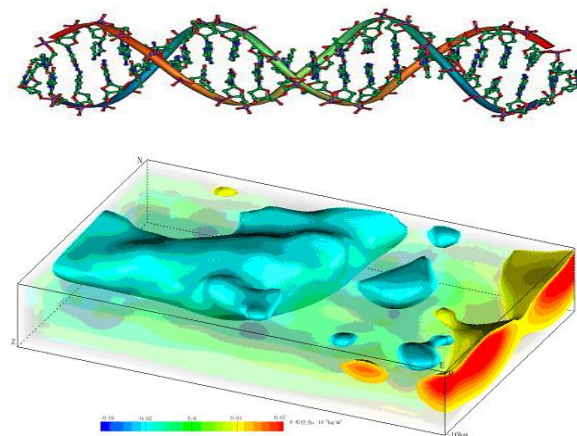
陆军工程大学网络技术重点实验室

Email: fuyinjin@sina.cn

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

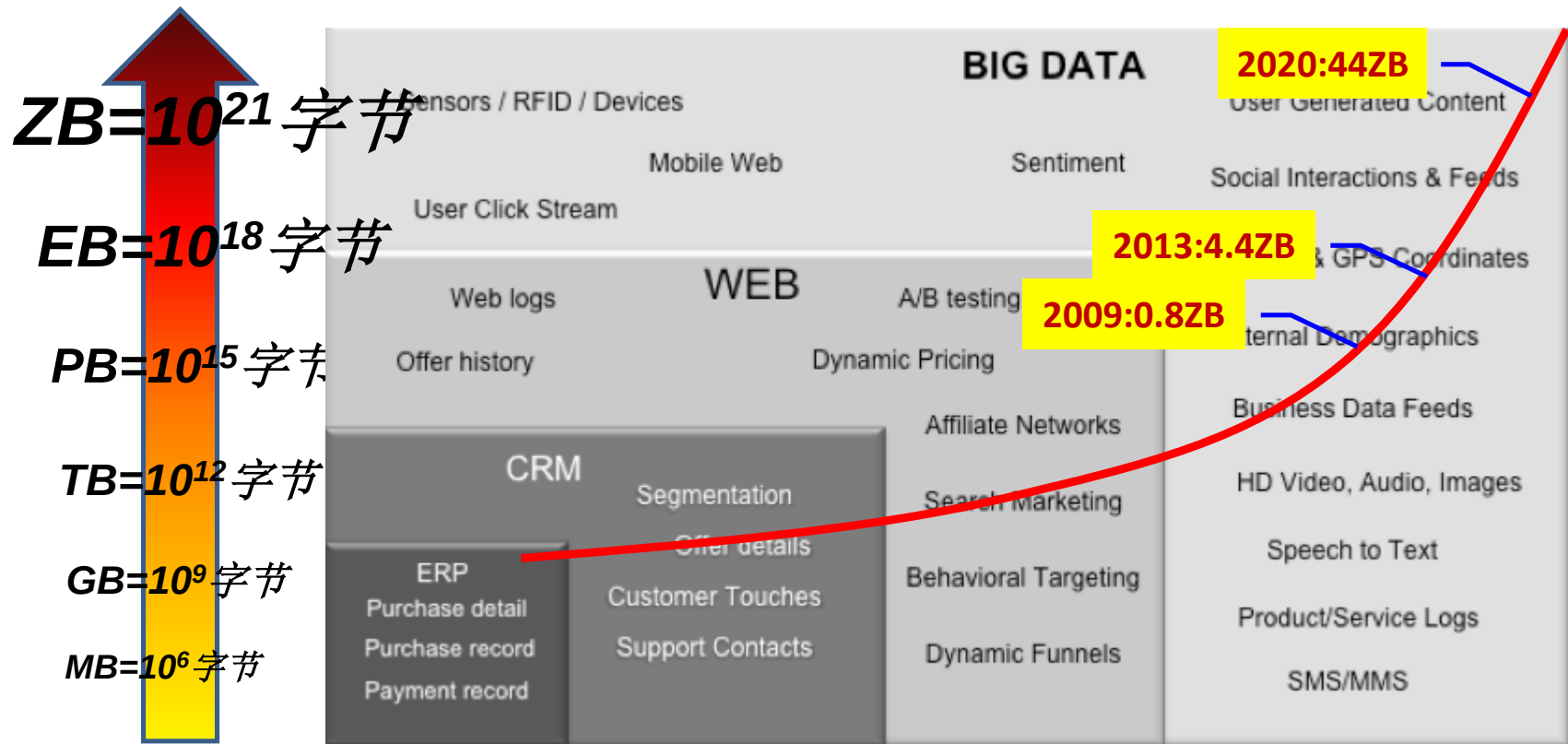


■ Web数据

- ◆ 文本、图像、视频
- ◆ Blog、点击流



大数据管理挑战



数据构成的多样化和复杂化

Source: IDC 2014, the digital universe is doubling in size every two years.

云计算与大数据

■ 云存储是大数据的主要存储管理方式

■ 数据存储集中化

◆ 云接触数据量

■ 20% → 40%

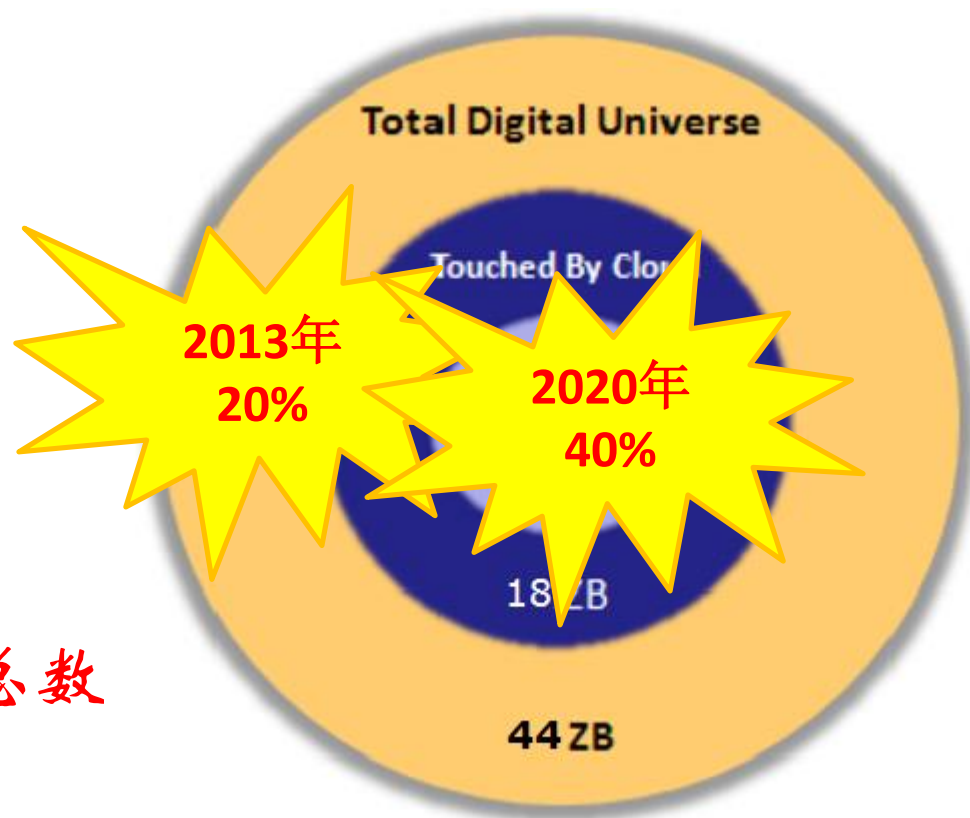
◆ 企业数据产生量

■ 占85%

■ 存储空间资源趋紧

◆ 可用存储容量与总数据量的比例

■ 33% → 15%



Source: IDC 2014

存储空间效率问题

■传统的存储环境

- 企业存储空间的**平均利用率在50%以下**，这个数字还没有除去提高可靠性的存储冗余配置。

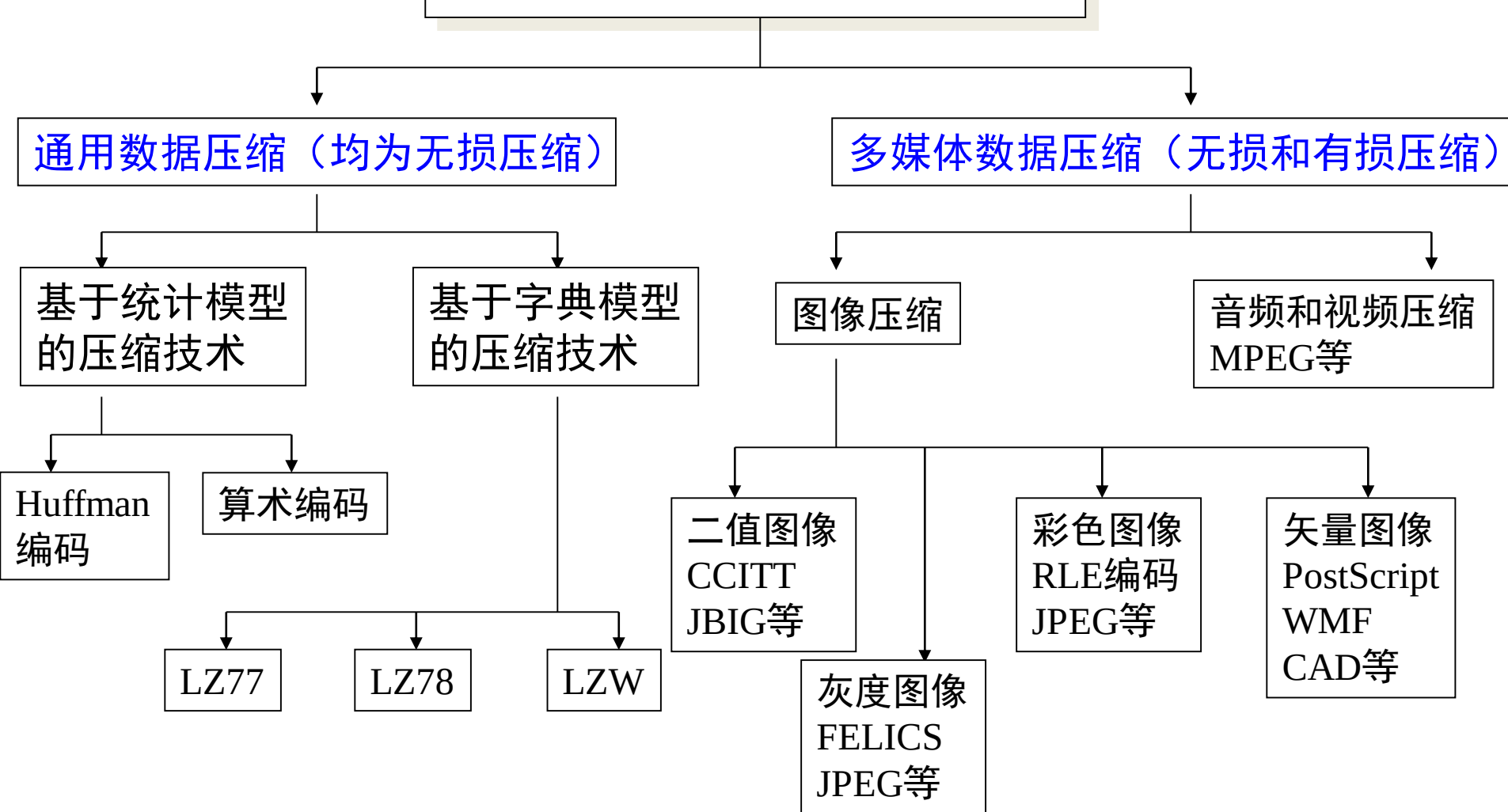
(2009年Symantec的吴锡源，2010年IBM的金微)

■提高存储效率

- 由于**数据缩减技术**的广泛应用，企业存储空间年增长率从2005-2007年间的**60%以上**，到2013-2017年间**降至35%-40%**。

(IDC: Efficiency will hold down storage growth, 2013)

数据压缩技术分类



数据压缩是指在**不丢失有用信息**的前提下，按照一定的算法对数据进行重新组织，减少数据冗余和存储空间

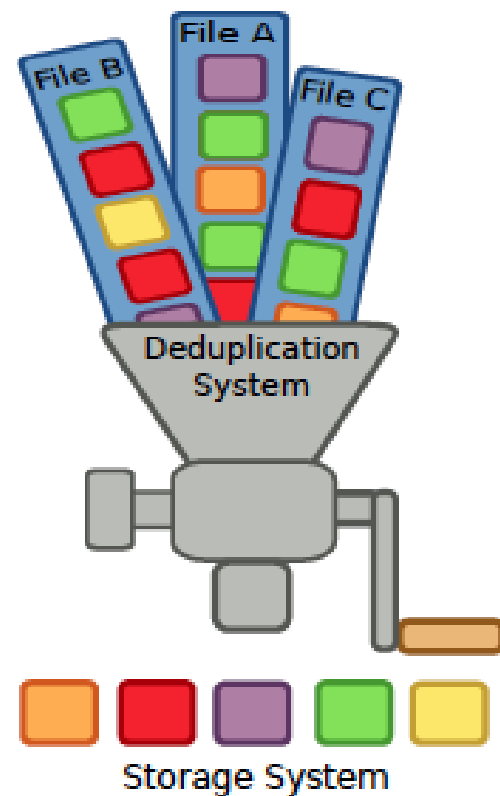
重复数据删除技术

■ 重复数据删除操作包括以下四个步骤：

- 数据分块
- 块指纹计算
- 块索引查询
- 唯一块存储

■ 通过消除数据冗余来优化通信与存储效率

■ 思想：以时间换空间，以计算换存储



用户举例



■ 外科手术器材制造厂

- 归档文件数据
- 38% 空间节省



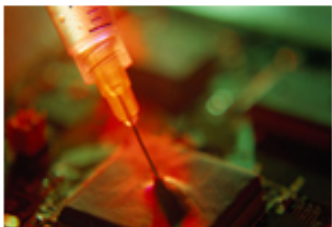
■ 全球性的投资管理公司

- 复制 VMware 镜像
- 88% 空间节省



■ 全球性的石油和天然气公司

- 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

高性能中心在线数据集中有20%-30%的重复数据，在部分应用中可达70%

——SC'12

重复数据删除技术研究

■ 备份/归档系统

- Venti[FAST'02], DDFS[FAST'08], D2D[FAST'09], ChunkStash [ATC'09], IBM ProtecTier[SYSTOR'09], Extreme Binning [MASCOTS '09], HYDRASor[FAST'09], Dedupv1 [MSST'10], DEBAR[IPDPS'10], MAD2[MSST'10], SiLo[ATC'11], EMC GDA [FAST'11], Σ -Dedupe [Middleware'12], Produck[SoCC'12], ALG-Dedupe [TPDS'14], RevDedup[ToS'15], DeFrame[FAST'15], EMC PGC [FAST'17], ...

■ 虚拟机

- CAS[ATC'06], Parallax[EuroSys'08], DEDE[ATC'09], VM Image Analysis [SYSTOR'09, Middleware'11], ...

■ 主存储

- HPCdedup[SC'12], iDedup[FAST'12], POD[IPDPS'14], DEDIS[ToS'16], HPdedup[MSST'17], ...

■ 闪存

- CAFTL[FAST'11], CA-SSD[FAST'11], CacheDedup[FAST'16], ...

■ 内存

- Difference Engine[OSDI'08], XLH[ATC'13], SmartMD[ATC'17], UKSM[FAST'18], ...

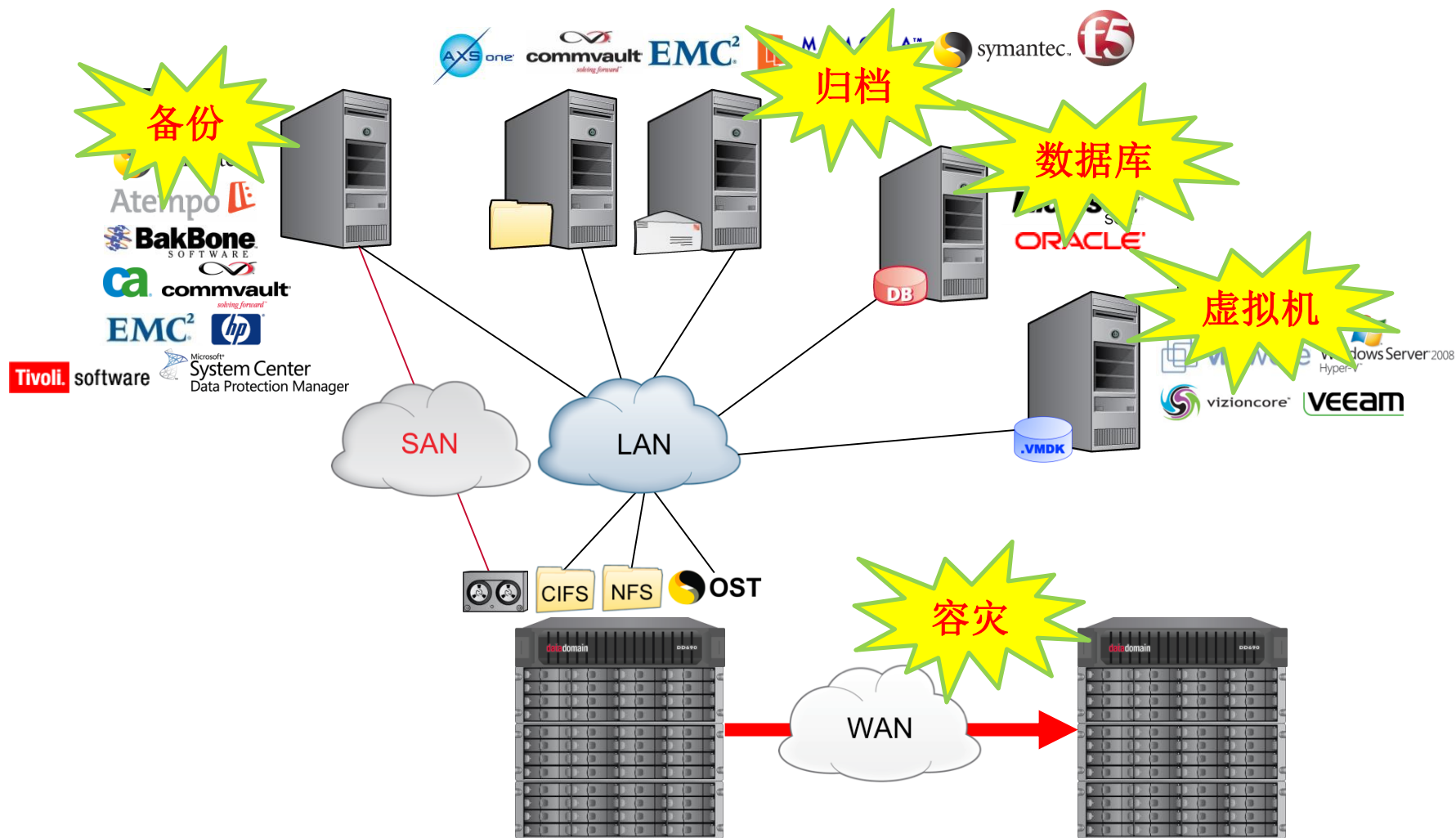
■ 网络优化

- LBFS[SOSP'01], TAPER[FAST'05], SIDC[FAST'12], ...

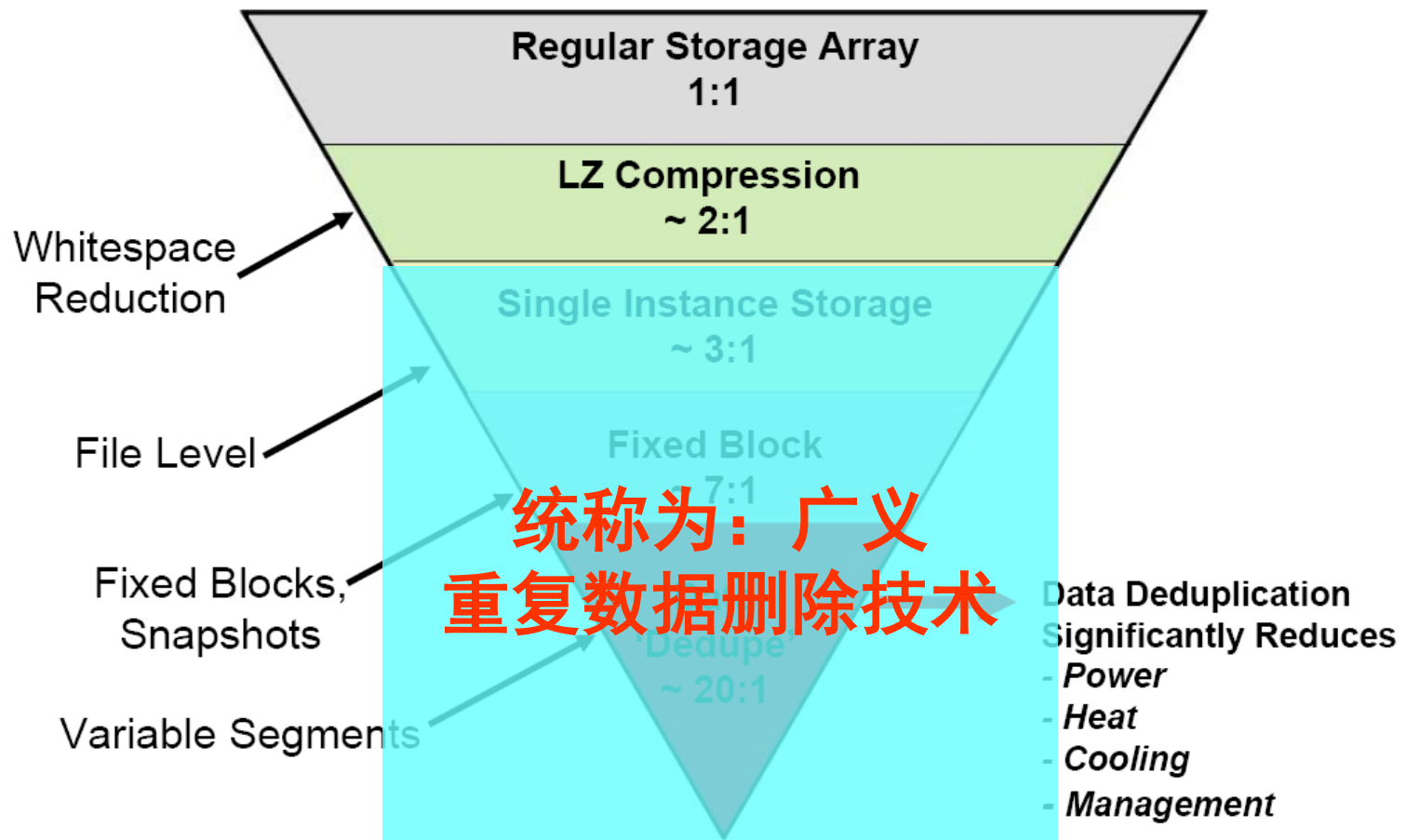
■ 数据库

- dDedup [CMU-TR'16], dbDedup[SIGMOD'17], ...

重复数据删除典型产品应用



数据缩减技术对比



二. 云环境重复数据删除性能优化



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

■ 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

挑战1: 传统单节点重删方法不够智能

■ 开发局部性或相似性来优化重删处理

- 局部性: DDFS[FAST'08], D2D[FAST'09], ChunkStash [ATC'09], HYDRA-Stor[FAST'09], DEBAR[IPDPS'10], MAD2[MSST'10], EMC GDA [FAST'11], RevDedup[ToS'15], DeFrame[FAST'15], CacheDedup[FAST'16], EMC PGC [FAST'17],
- 相似性: IBM ProtecTier[SYSTOR'09], ExtremeBinning [MASCOTS'09], SiLo[ATC'11], Σ -Dedupe[Middleware'12], Produck[SoCC'12], dDedup [CMU-TR'16],

■ 重删处理效率仍然比较低

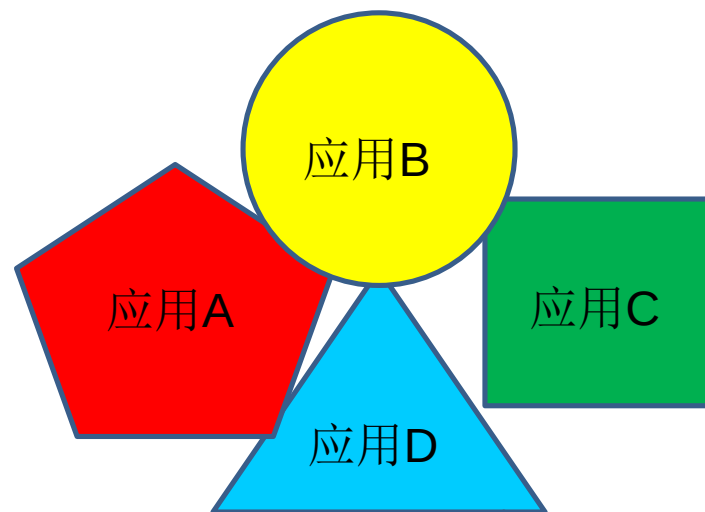
- 无法感知应用层不同数据类型的特征差异
- 无法设计可自适应各种应用数据的智能重删机制

■ 解决办法: 应用感知设计!!!

挑战2: 传统分布重删数据路由不灵活

- **无状态方法** (HYDRA-Stor[FAST'09], DEBAR[IPDPS'10], MAD2[MSST'10])
 - 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** (EMC GDA [FAST'11], Produck[SoCC'12])
 - 在数据路由前, 需要查询数据对象与各节点数据的相似度
 - 数据缩减率高、负载均衡, 但系统开销大
- **基于指纹的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

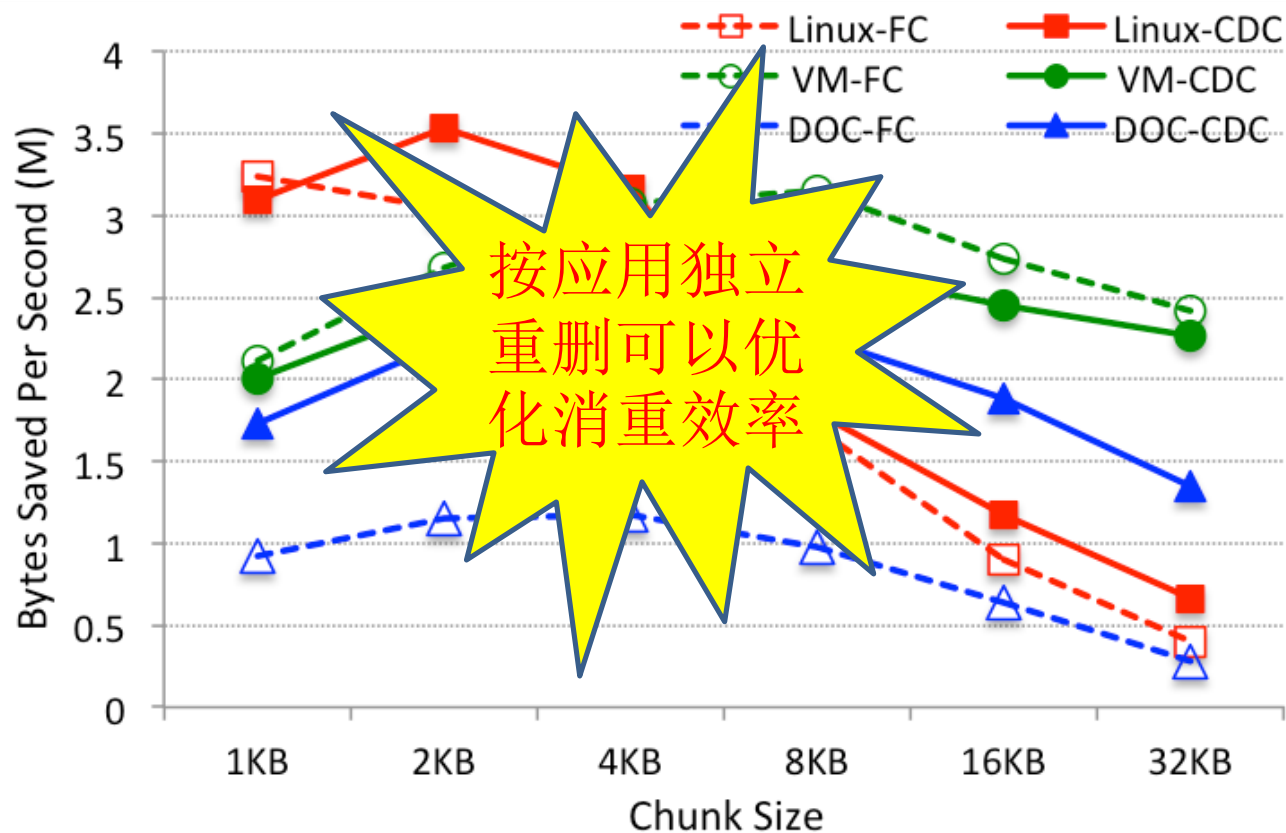


应用数据集	应用内冗余	应用间冗余
研究桌面电脑		1.1%
个人笔记本电脑		0.9%
个人工作站		0.6%
共享家庭服务器		0.7%
总冗余率	27.7%	0.8%

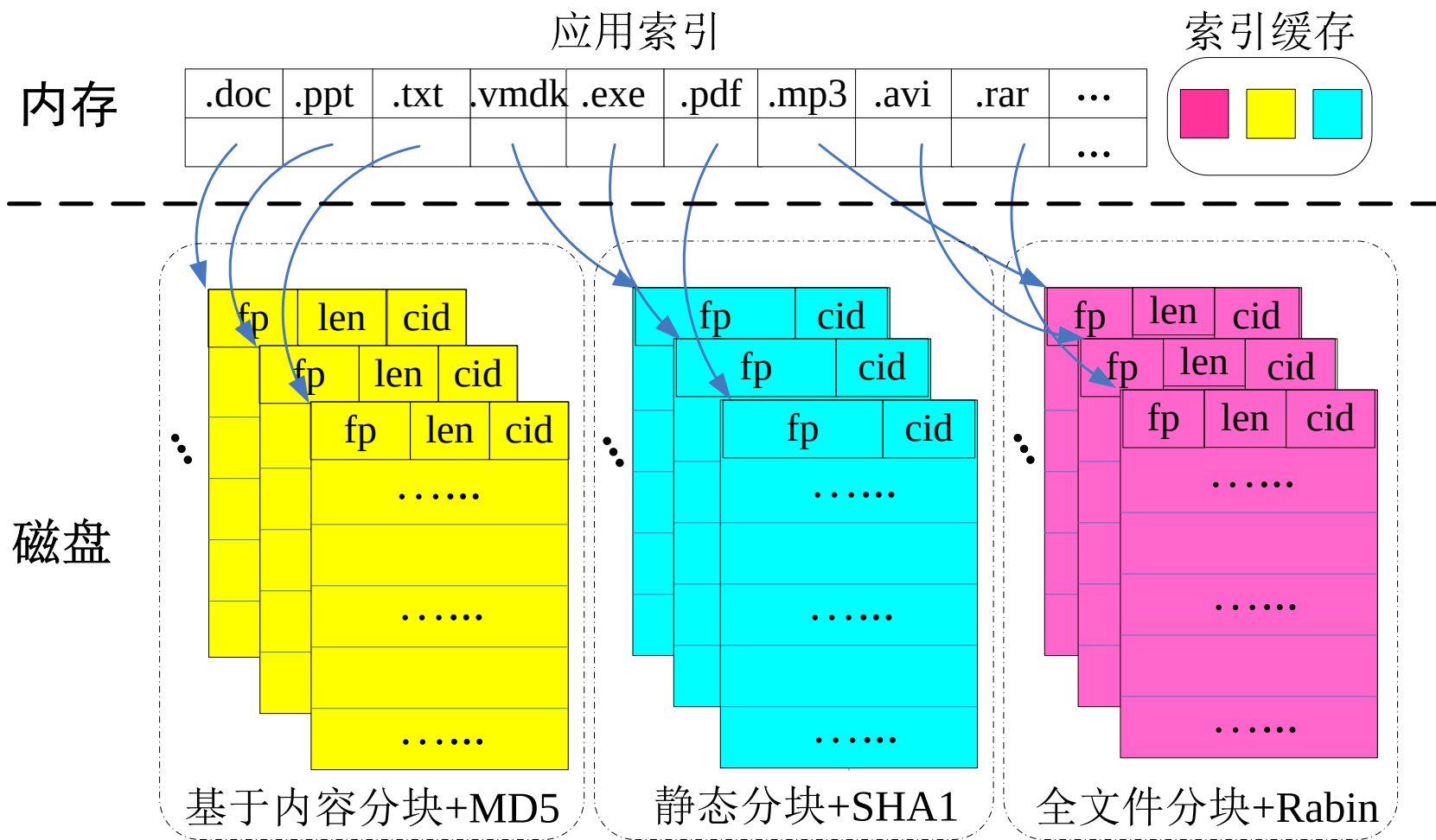
不同应用数据可以独立并行重删

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

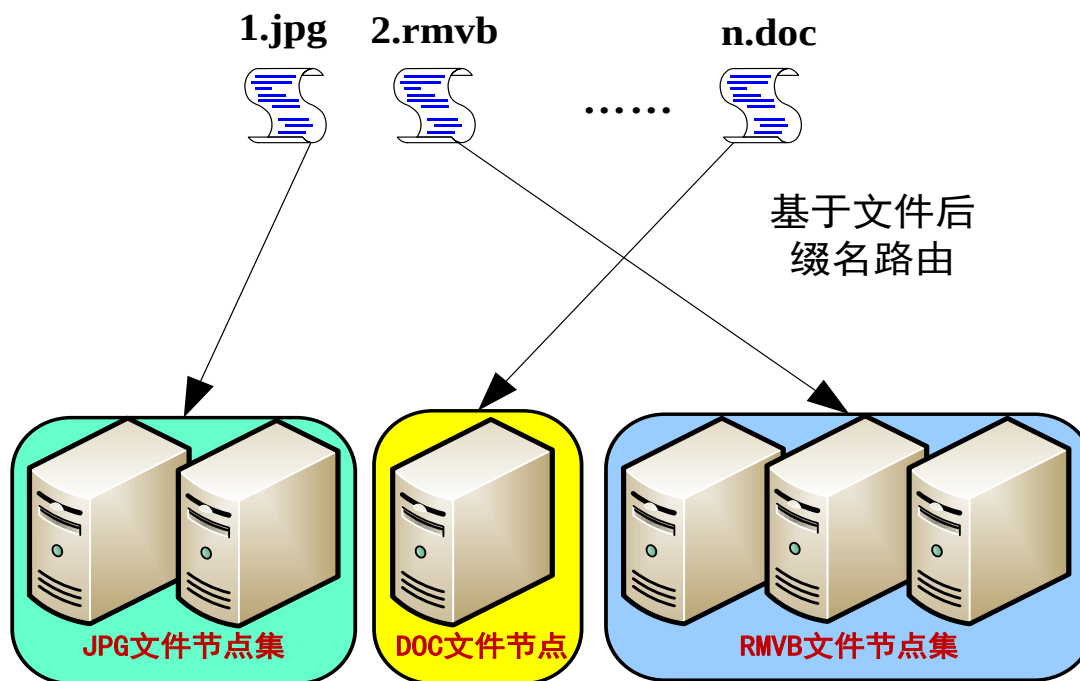


创新1：应用感知索引结构



创新2：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
- 对不同类型的应用数据进行隔离分配

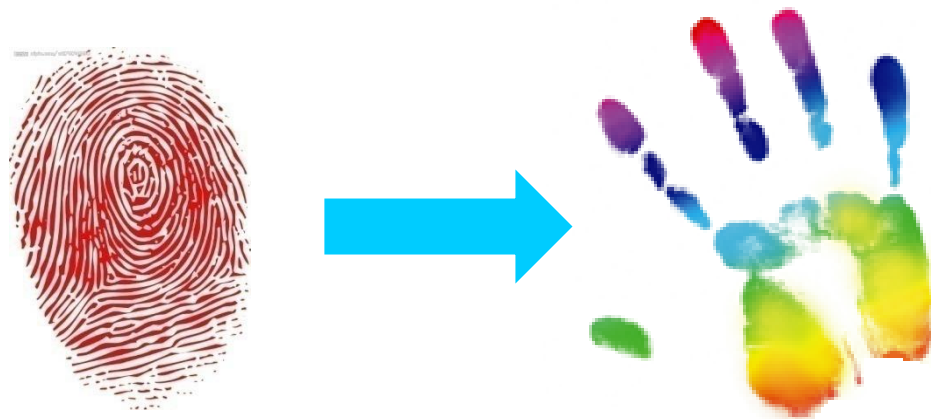


手纹技术 (1/2)

■ 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 $1 - (1 - r)^k$

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



手纹技术 (2/2)

■ 超块：

- 数据流中一串连续的数据块构成超块

■ 手纹概念

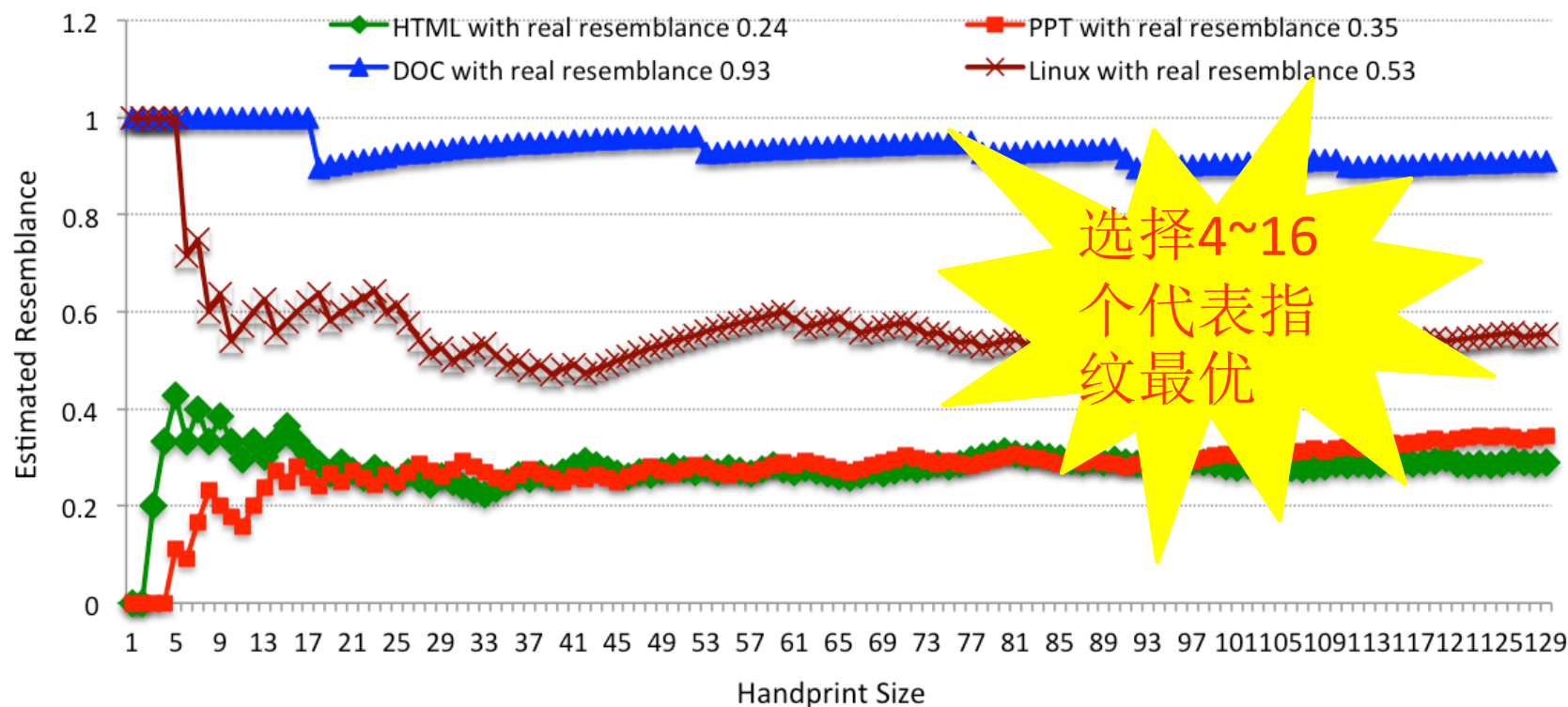
- 超块的块指纹集中的 k 个最小者作为特征指纹

■ 手纹 VS. 指纹

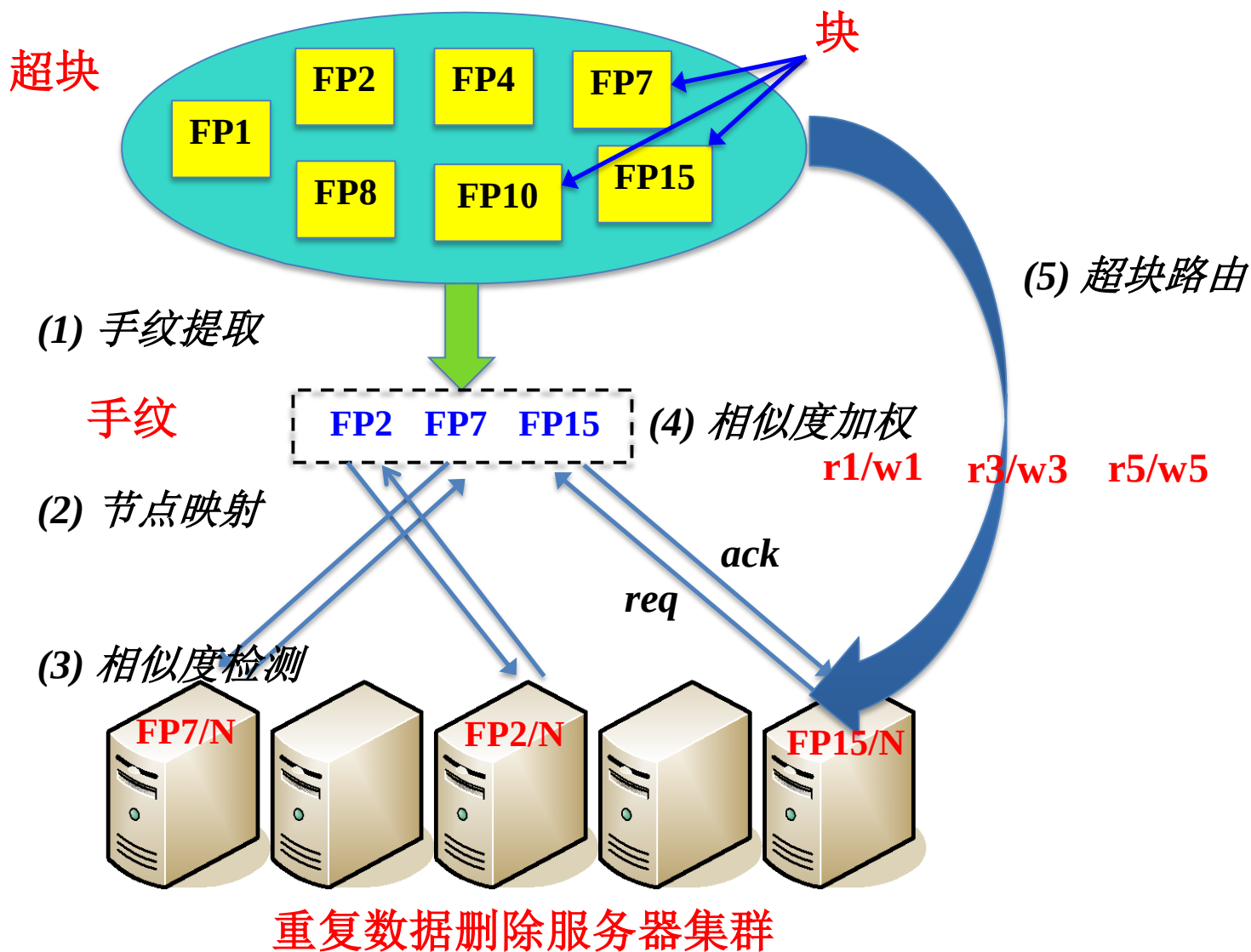
- 更强的相似检测能力

手纹冗余检测能力

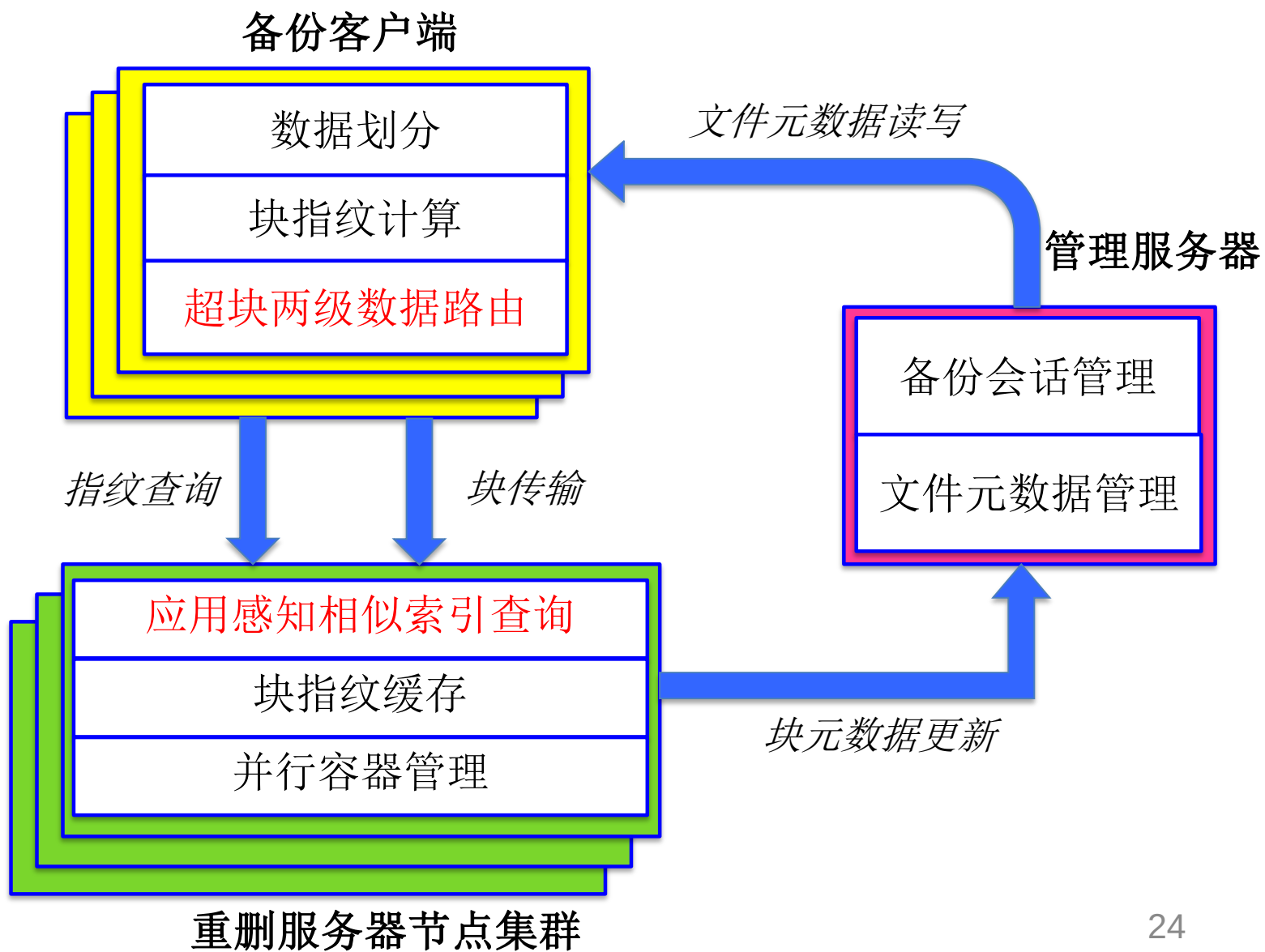
■ 估计值精度 vs. 存储开销



创新3: 基于手纹的数据路由算法

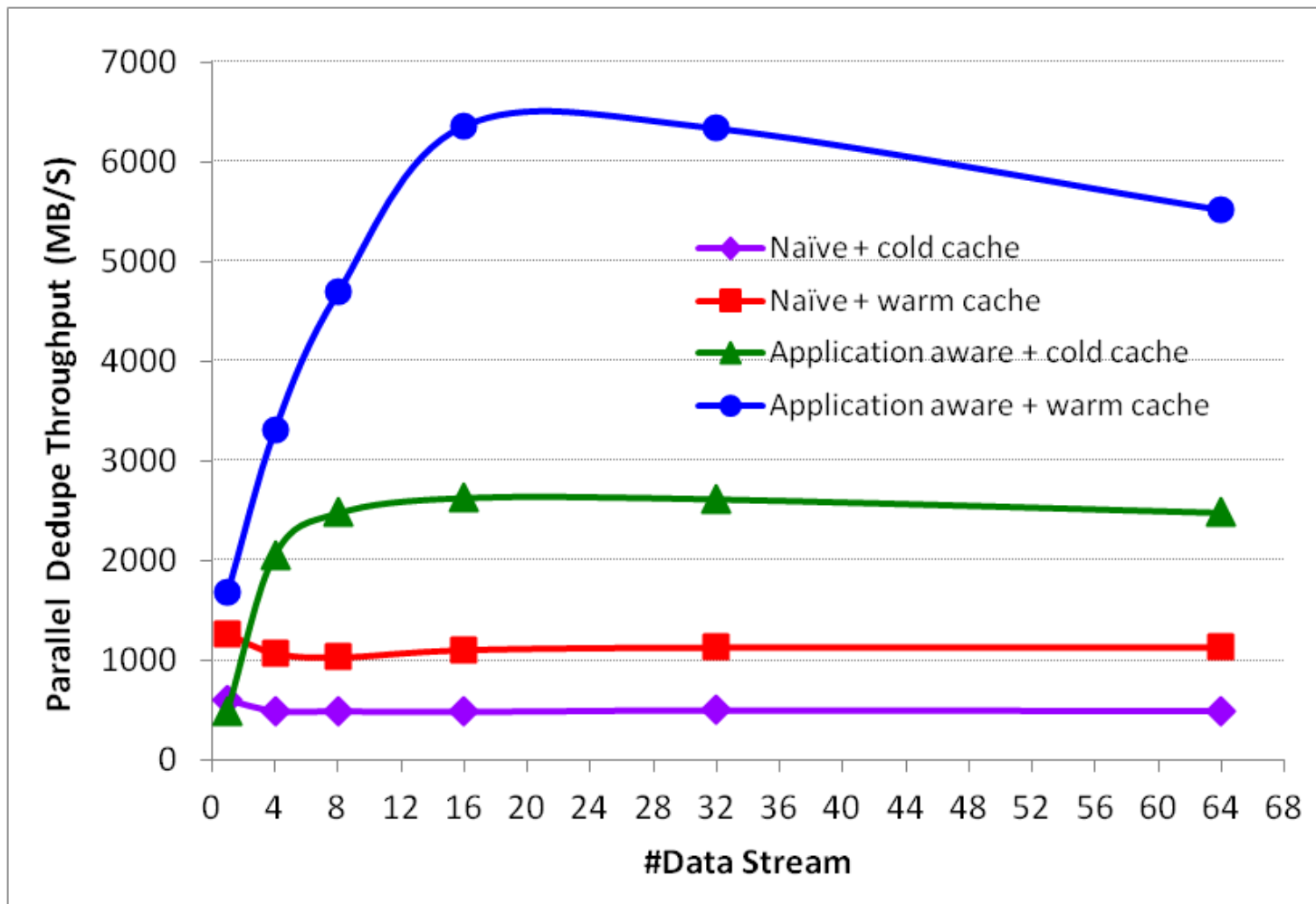


AppDedupe 分布式重删系统设计



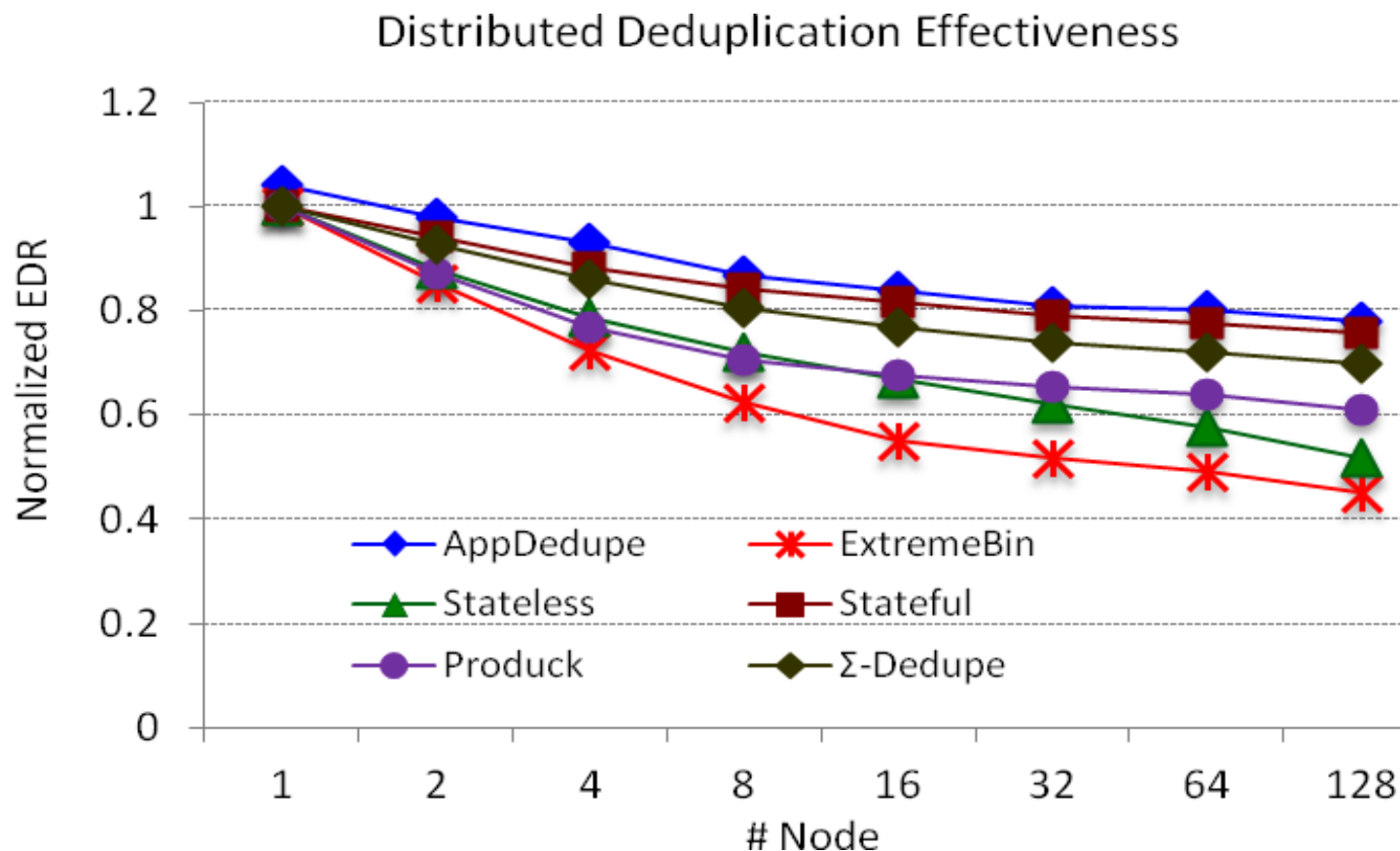
单节点并行重删吞吐量

■ AppDedupe节点内应用感知并行重删性能突出



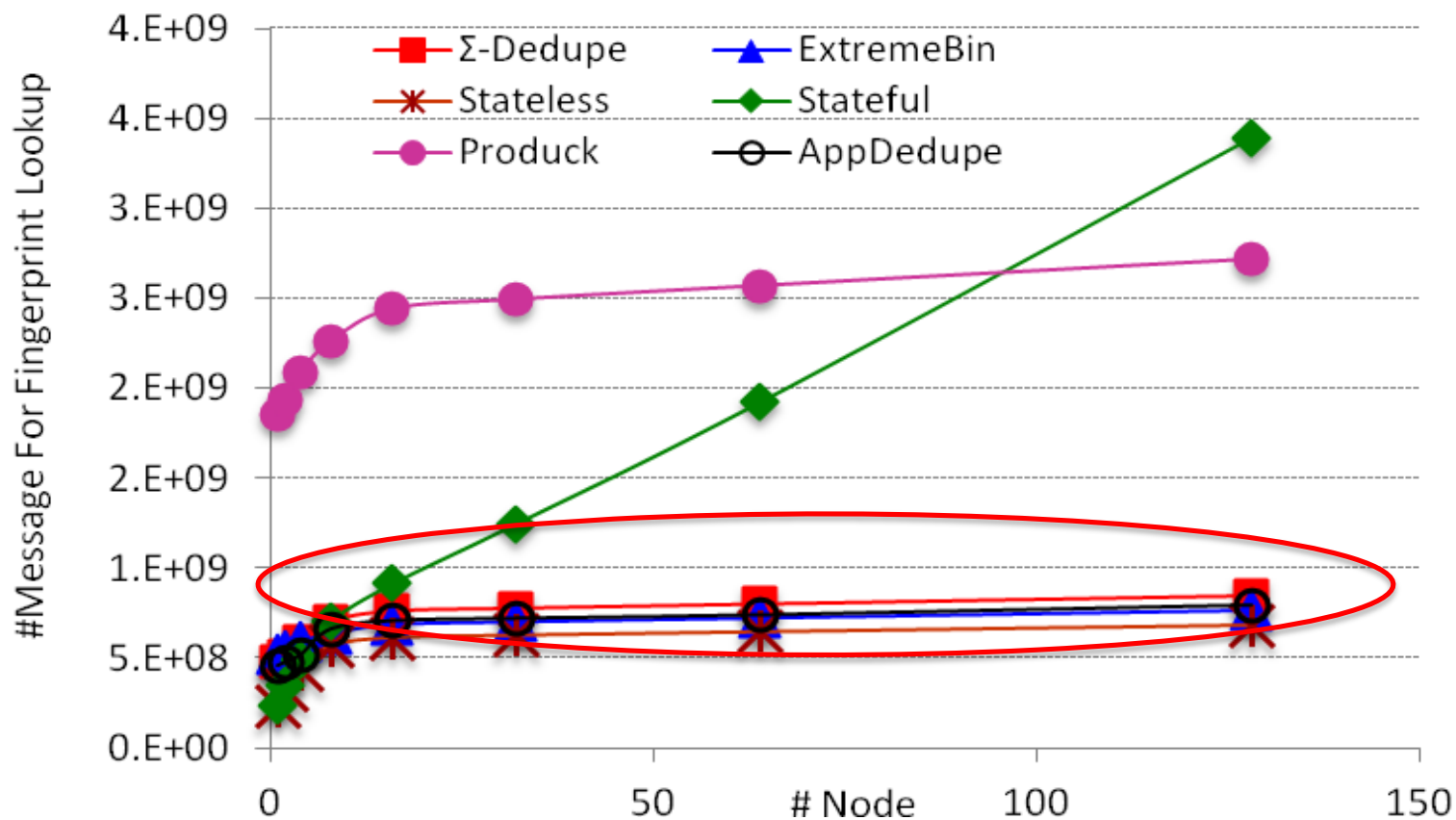
分布全局重删效果

■ AppDedupe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



结论

- 应用感知的重复数据删除设计，可以极大地提升传统单节点重复数据删除机制的处理效率
- 基于应用分类和超块手纹的两级数据路由技术，能极大地提升分布重复数据删除的效率
- 分布式应用感知重删能获得比复杂有状态方法**更高重删率**，且仅有近似简单无状态方法的**低系统开销**。

参考文献

1. **Yinjin Fu**, Nong Xiao, Hong Jiang, Guyu Hu, Weiwei Chen. “Application-Aware Big Data Deduplication in Cloud Environment.” *IEEE Transactions on Cloud Computing* (TCC), PP(99), 2017. (JCR一区顶级期刊)
2. **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. (CCF A类期刊)
3. 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. (CCF A类期刊)
4. 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. (CCF A类期刊)
5. **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.(CCF B类会议)
6. **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. (CCF B类会议)
7. 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. (CCF B类会议)
8. **Yinjin Fu**, Xiaofeng Qiu, Jian Wang. "F2MC: A Data Storage Service with Fog-to-Multi-Cloud Hybrid Computing." *38th IEEE International Performance Computing and Communications Conference (IPCCC’19)*, London, Oct. 29–31, 2019. (CCF C类会议)
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. (CCF C类会议)
10. 付印金, 胡谷雨, 倪桂强, 陈卫卫, 卢继荣. “面向大数据备份的应用感知并行重删存储系统.” 《计算机研究与发展》, 2015, 52(S1): 139-147.

谢谢！

敬请批评指正！