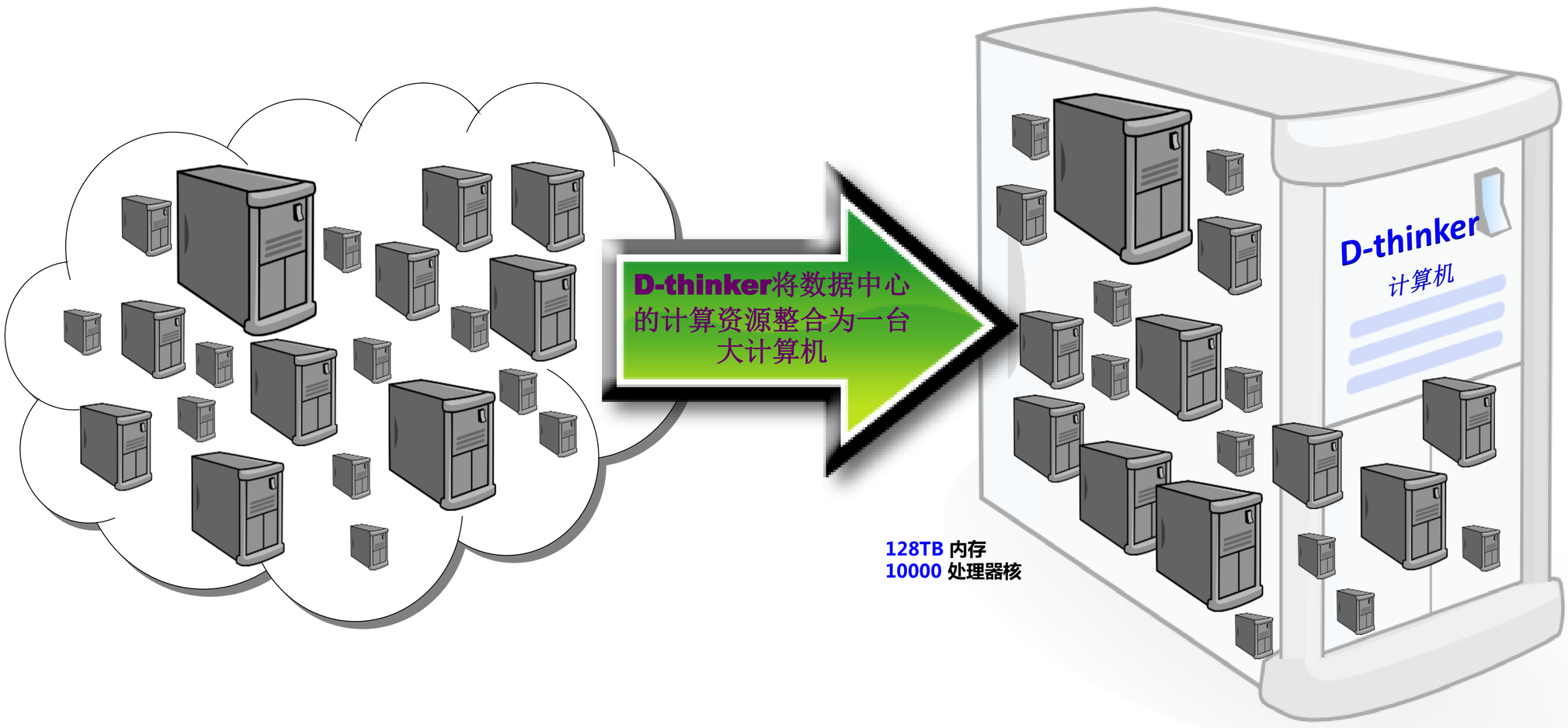
Data Thinker: 高性能大数据计算

HUTUTA TECHNOLOGIES

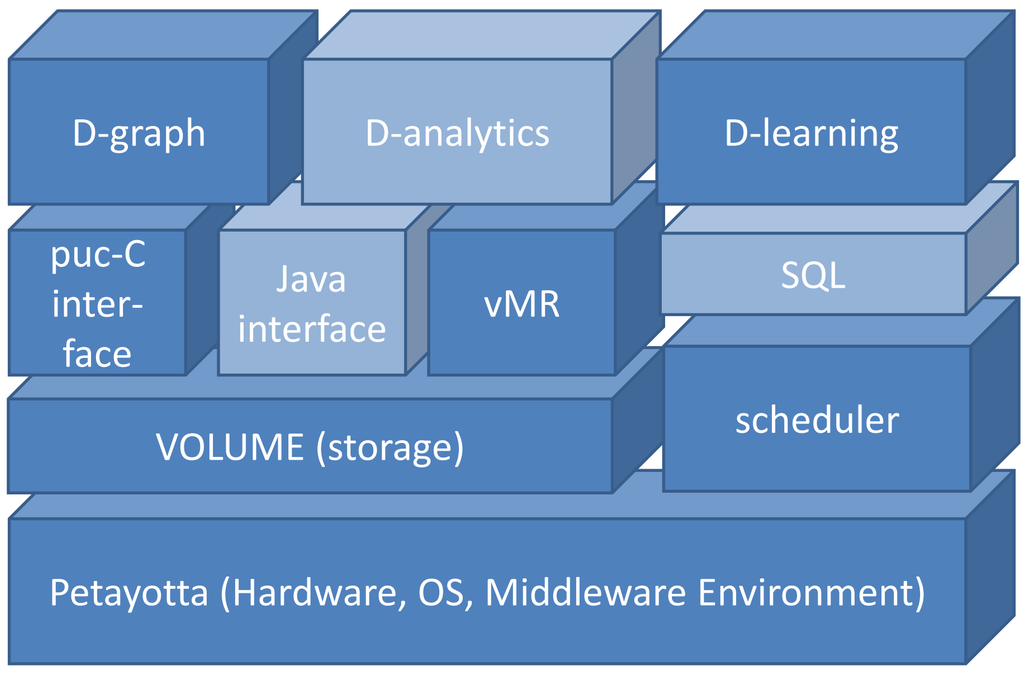
Data Thinker (D-thinker)技术整合众多计算机的CPU、内存及硬盘资源，以经济、高效、可扩展的方式构建高性能、低延时、图灵完备的计算体系，提供对GB-PB量级数据的存储、搜索、挖掘、学习及商业智能处理的能力，其性能较Hadoop高10-100倍、较Spark高2-10倍，已经成功应用于网络日志分析、基因数据处理等领域。该技术也是国内唯一不依赖开源软件、核心技术完全自主开发的大数据技术。



Data Thinker提供了一个基于数据中心的通用、高性能、易用的大规模云计算平台。

特性与优势

- 支持大规模、高吞吐量、低延时、强一致性，适用于各种类型和规模的应用，如大数据分析、科学运算及大规模模拟、事务处理
- 相对其它平台，能更灵活地运算及管理数据
- 能因应不同数据储存规模作出调整，确保网络性能良好
- 提供多种数据检索、图运算、数据挖掘运算程序库或例程
- C、C++、Java、Python等多种语言框架，即将推出支持SQL的高可靠性内存数据库



设计与实现

- 高效、易移植、支持并发编程的指令集D-CISC对D-thinker功能进行抽象
- 统一编址的消息地址空间，支持共享与专有消息区域
- 高效的、支持高度并发存取的共享消息模型，消息可大至TB
- 低开销、高可扩展、易于使用的并发编程支持及并行计算机制
- 基于数据依赖的、易表达、可自动解决依赖关系的任务依赖控制机制

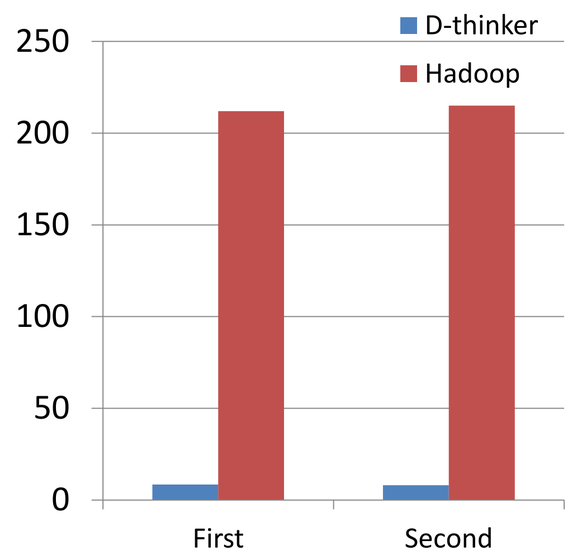
实例与性能比较

K-Means是一个经典的针对多维数据聚类算法，常用于数据挖掘，可以很好的评估一个大数据计算平台对多维数据分析的支持能力。

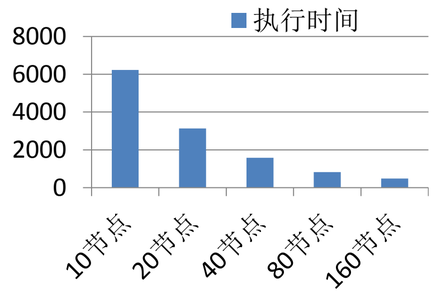
4台服务器进行k-means计算

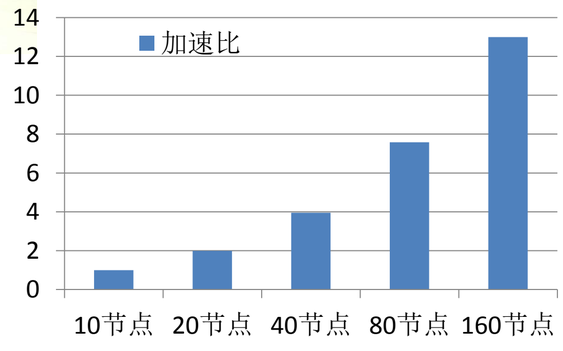
- Hadoop：200+秒
  1. 212秒
  2. 215秒
- D-thinker：8-11秒
  1. 8.39秒
  2. 8.03秒

参照Mahout算法实现对50万个4维点计算20个聚类。

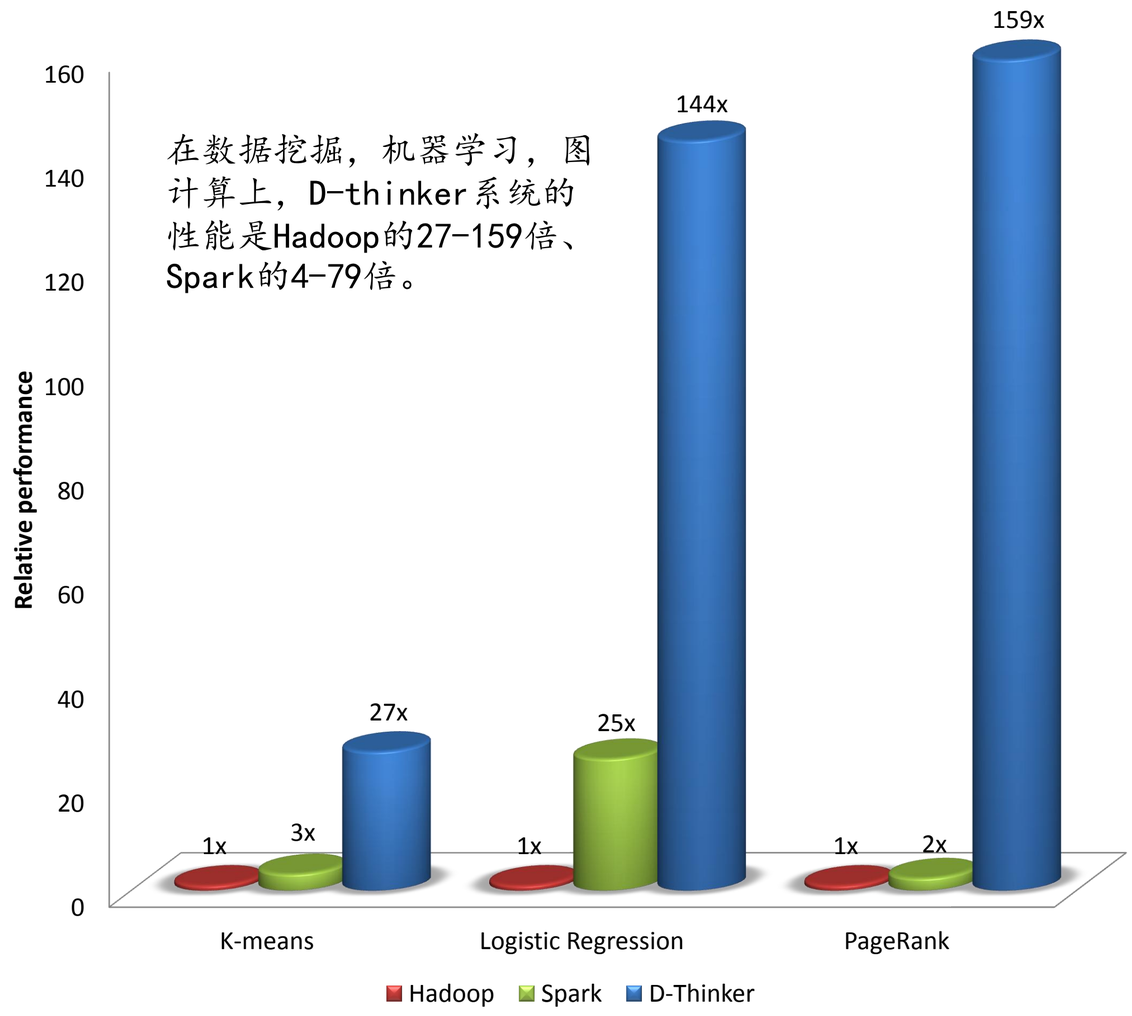


在天河1号上用D-thinker进行大规模k-means计算，节点数从10台服务器依次增长到160台，节点数增加16倍，加速比约为13，接近最优效率。





在数据挖掘，机器学习，图计算上，D-thinker系统的性能是Hadoop的27-159倍、Spark的4-79倍。



D-thinker 网站 http://d-thinker.org D-thinker 论坛 http://tab.d-thinker.org/forumdisplay.php?fid=198

HUTUTA TECHNOLOGIES