高速网络InfiniBand加速大数据应用

刘通

Mellanox亚太市场开发总监

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™

Mellanox公司概况

股票代码: MLNX NASDAQ® TEL-AVIV STOCK EXCHANGE

Mellanox TECHNOLOGIES

- **连接服务器、存储器的高带宽与低延迟网络的领导厂商**
  - FDR 56Gb/s InfiniBand 与万兆/4万兆以太网
  - 降低应用等待数据时间
  - 大幅提升数据中心投资回报率

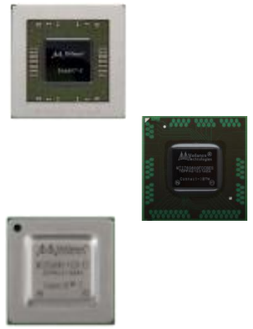
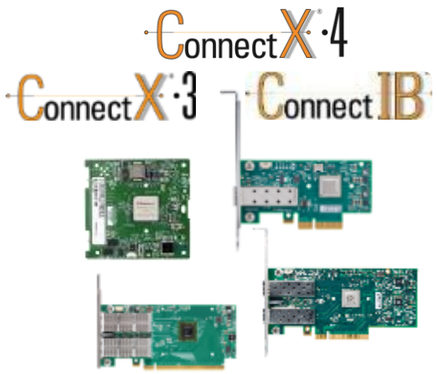
- **公司总部:**
  - 美国加州以及以色列双总部
  - 全球范围内约~1432名员工

- **良好财务状况**
  - 2013年销售近3.9亿美元
  - 现金与投资达3.4亿美元

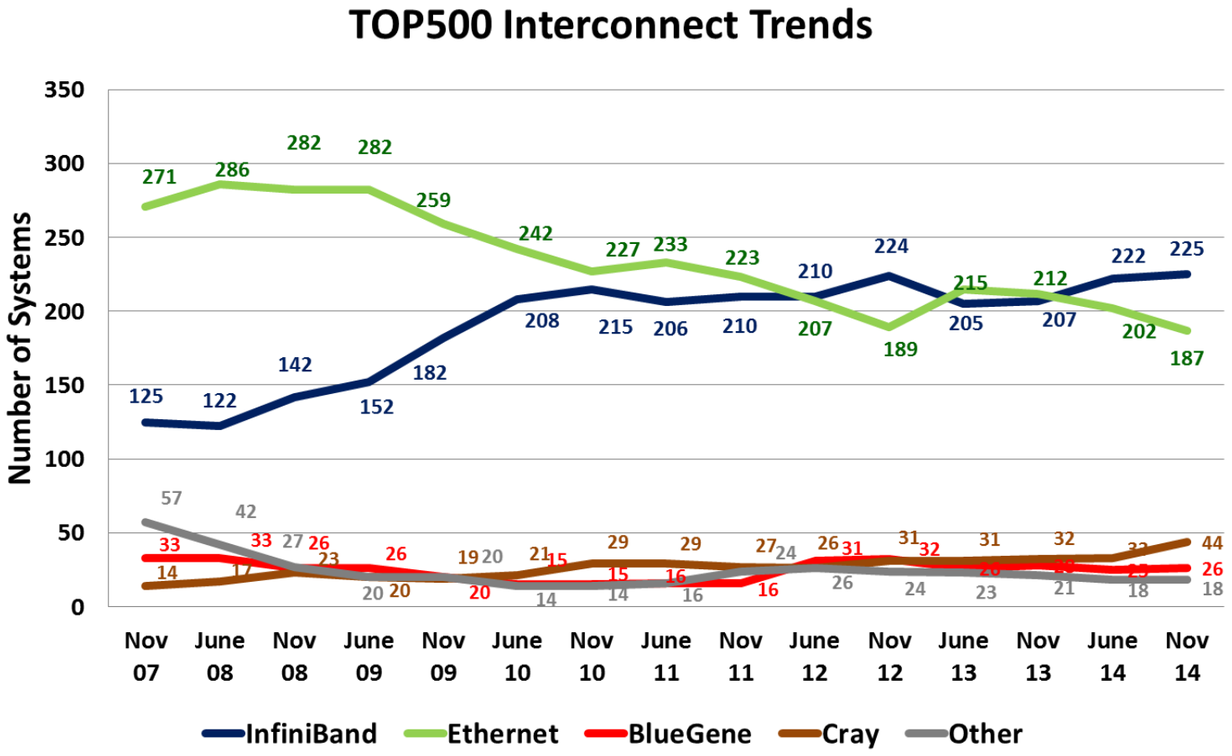
截至2013年9月

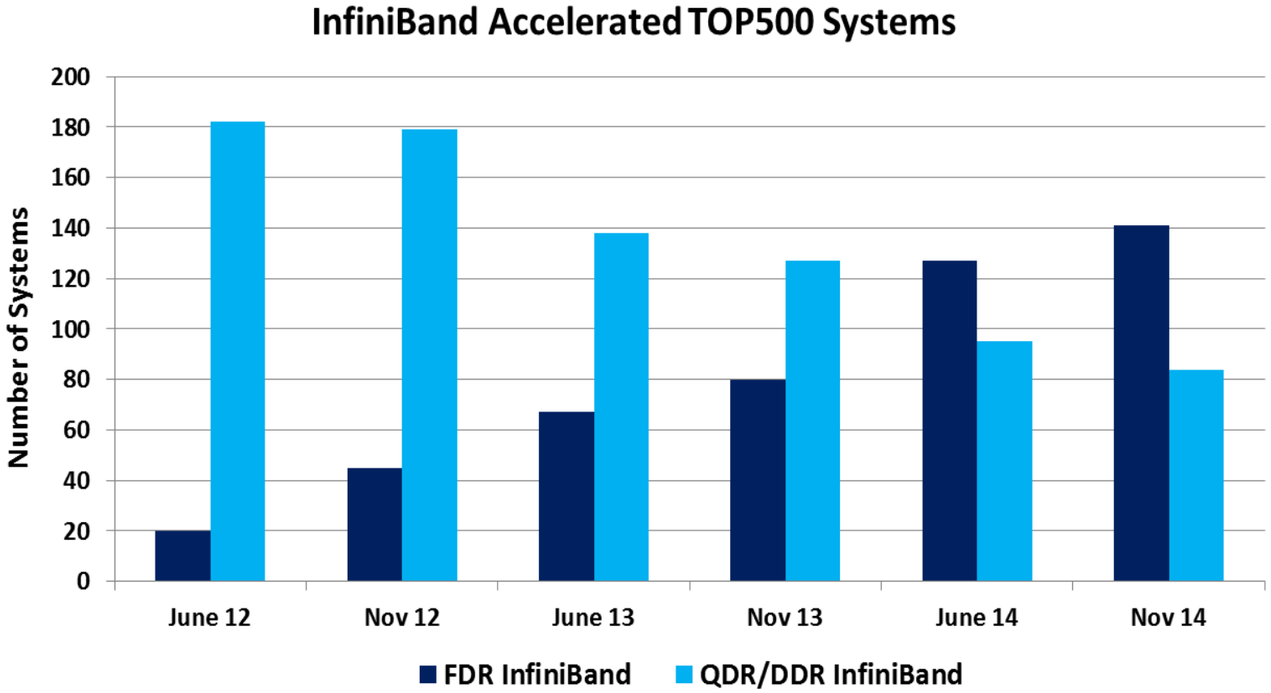
世界领先的端到端网络互连设备提供商

完整的InfiniBand与以太网产品线

芯片	网卡	交换机、网关	Metro / WAN	网线、模块
				

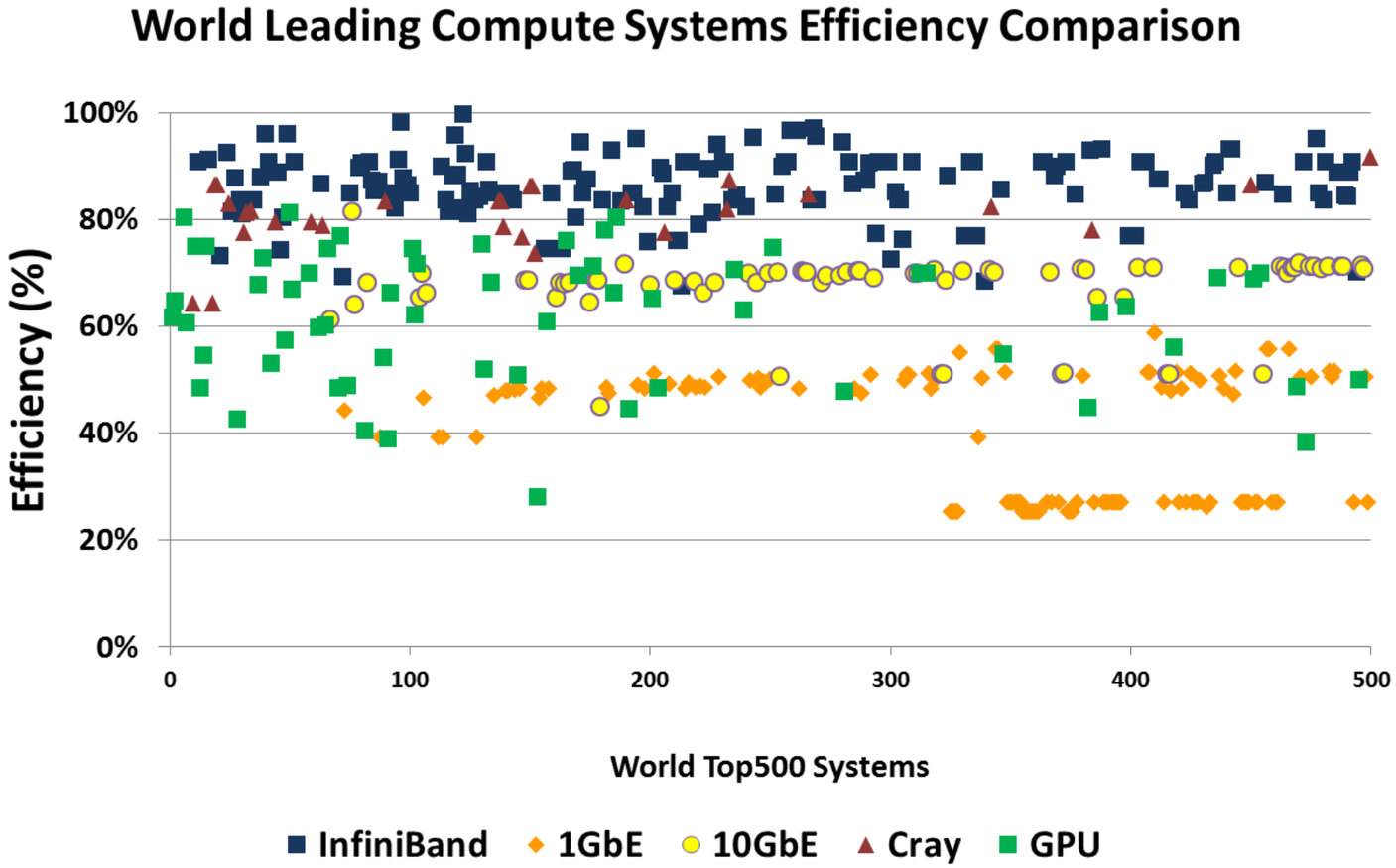
超级计算机TOP500中最高占有率





- InfiniBand是高性能应用的首选网络
- 采用Mellanox FDR InfiniBand 的系统同比增长1.8倍
  - 加速63% 的InfiniBand系统是基于FDR (141 systems out of 225)

InfiniBand提供不可超越的系统效率



平均效率

- **InfiniBand: 87%**
- **Cray: 79%**
- **10GbE: 67%**
- **GigE: 40%**

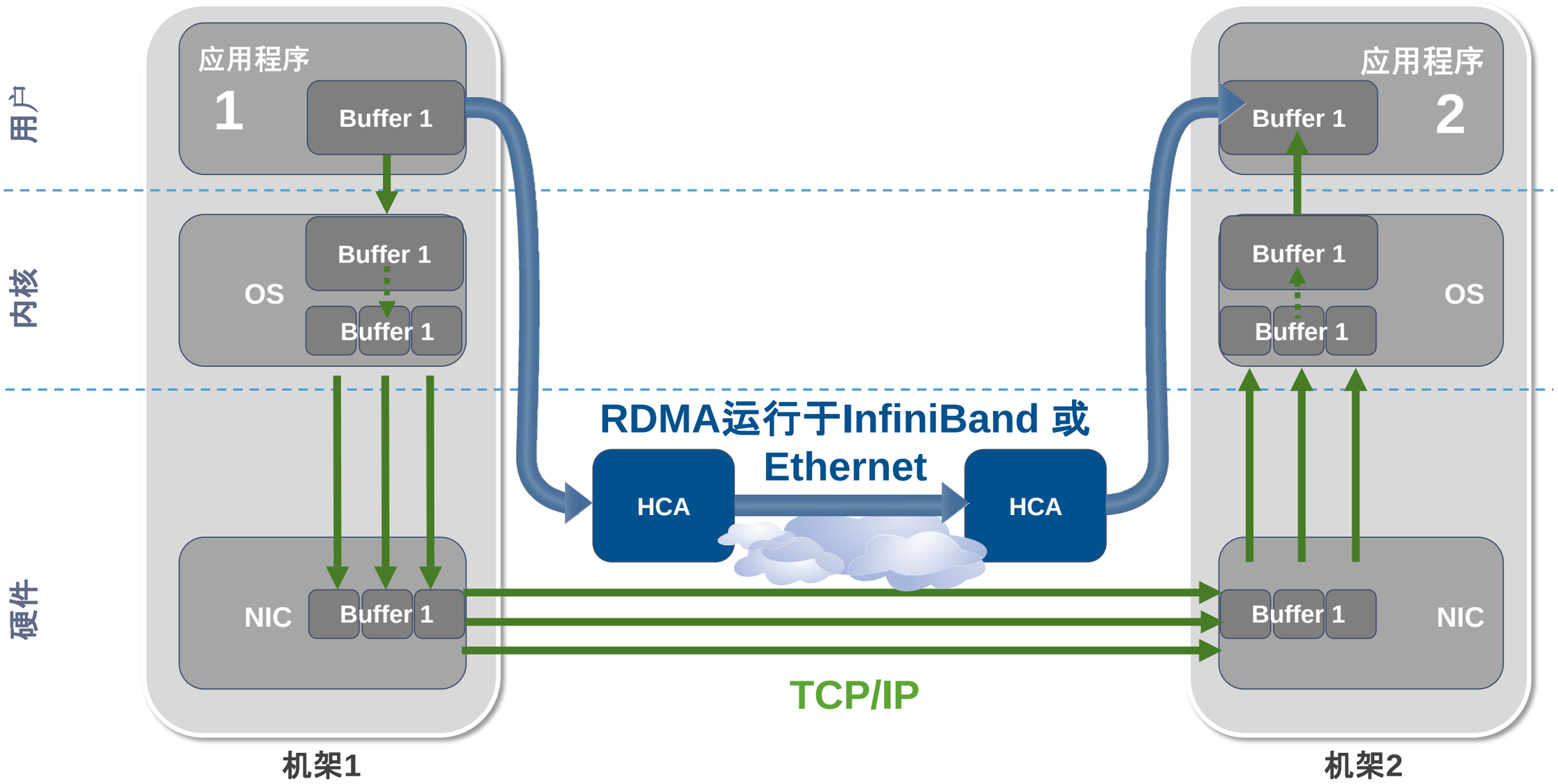
- InfiniBand是实现最高系统效率的关键，平均高于万兆以太网30%
- Mellanox InfiniBand 实现最高效率99.8%

InfiniBand技术优势

InfiniBand 技术的优势和特点

- InfiniBand Trade Association (IBTA) 协会制定规范
  - 开放标准的高带宽、低延迟网络互连技术
- 串行高带宽连接
  - SDR: 10Gb/s HCA连接
  - DDR: 20Gb/s HCA连接
  - QDR: 40Gb/s HCA连接 –现在
  - FDR: 56Gb/s HCA连接 – 2011年底
  - EDR: 100Gb/s HCA连接 – 2014年
- 极低的延迟
  - 低于1 微妙的应用级延迟
- 可靠、无损、自主管理的网络
  - 基于链路层的流控机制
  - 先进的拥塞控制机制可以防止阻塞
- 完全的CPU卸载功能
  - 基于硬件的传输协议
  - 可靠的传输
  - 内核旁路技术
- 远端内存直接访问
  - RDMA-读和RDMA-写
- 服务质量控制（QoS）
  - 在适配器卡级提供多个独立的I/O通道
  - 在链路层提供多条虚拟通道
- 集群可扩展性和灵活性
  - 一个子网可支持48,000个节点，一个网络可支持 2^{128} 个节点
  - 提供多种集群拓扑方式
- 简化集群管理
  - 集中路由管理
  - 支持带内网络诊断和升级

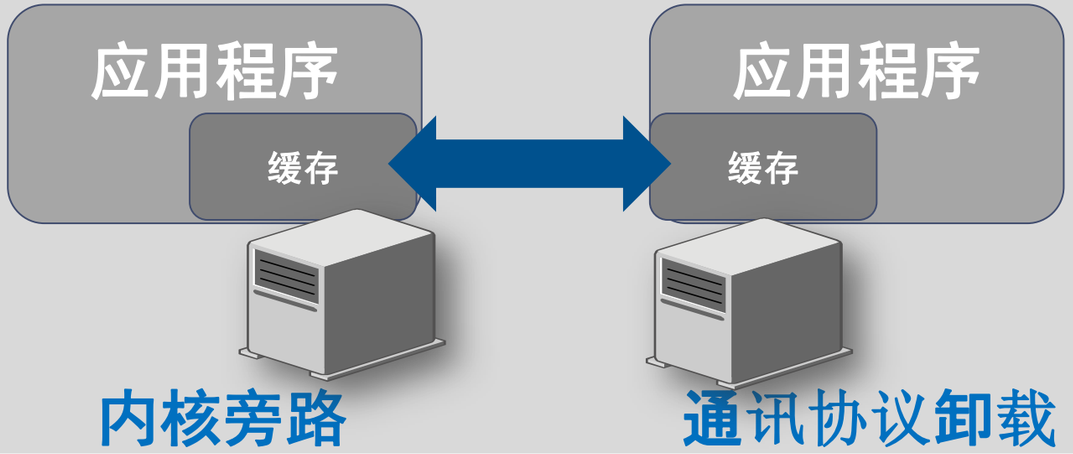
RDMA (远端内存直接访问技术) – 如何工作



Mellanox RDMA远端内存直接访问技术

零拷贝

远程数据传输



低延迟, 高速数据传输

InfiniBand - 56Gb/s

RoCE* – 40Gb/s

* RDMA over Converged Ethernet

加速分布式数据库

迈络思网络加速主流数据库

- **Oracle 数据仓库**
  - 提供4倍闪存
  - 写性能提升20倍
  - 数据吞吐量提高33%
  - 降低能耗10% 到 40%

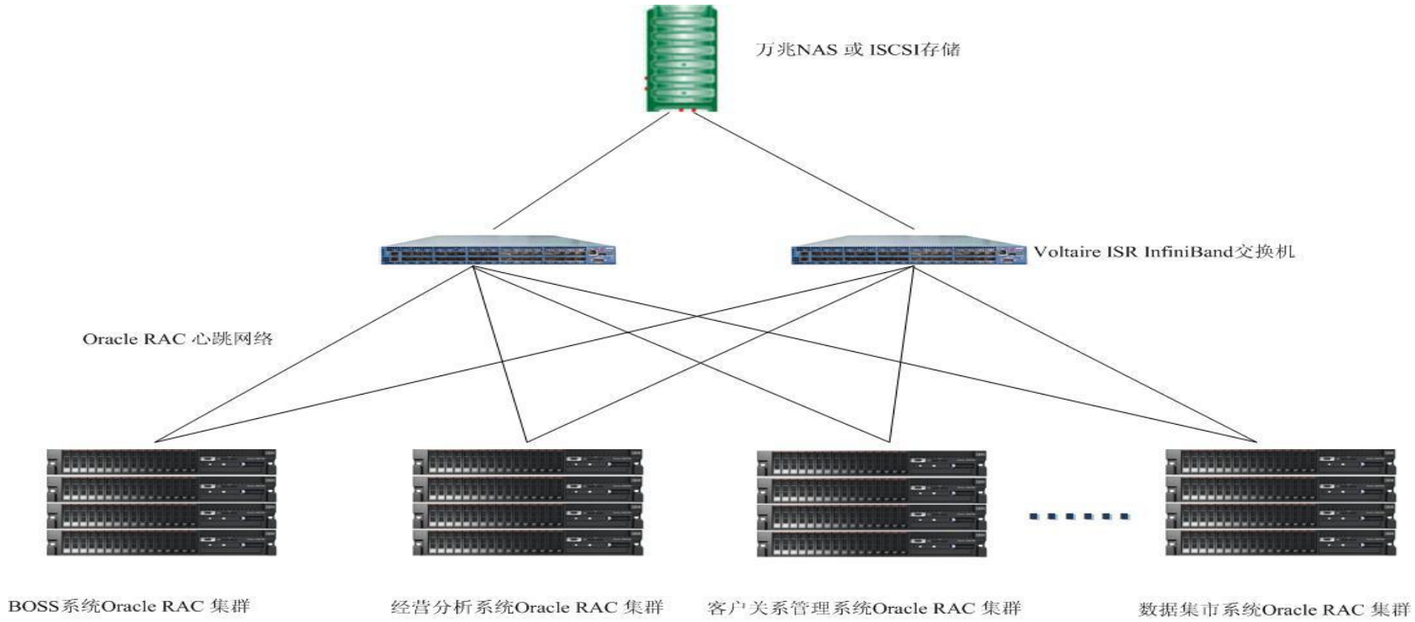
- **IBM DB2 Purescale 数据库:**
  - 需要低延迟高带宽的网络，同时满足高可靠性
  - RDMA 大大降低CPU负荷
  - 实现DB2 Purescale 接近线性的可扩展性

- **微软 SQL Server 数据仓库**
  - 更高性能，更低成本

- **Teradata 数据仓库**
  - 相较以太网，跨机柜SQL查询速度提升2倍
  - 数据加载性能提升4倍

大幅提升性能与可扩展性，降低成本

河南移动Oracle RAC数据库解决方案



- 采用Mellanox InfiniBand交换机作为心跳网络连接设备；
- 全线速无阻塞网络；
- 采用高可用的冗余连接方式，避免单点故障；
- 40Gb/s高通讯带宽、100纳秒超低延迟，全面加速Oracle RAC性能

InfiniBand+PCI-e SSD新架构加速Oracle数据库

生产环境：

处理器：16 CPU Itanium2 1.6GHZ（双核）
内存：192G

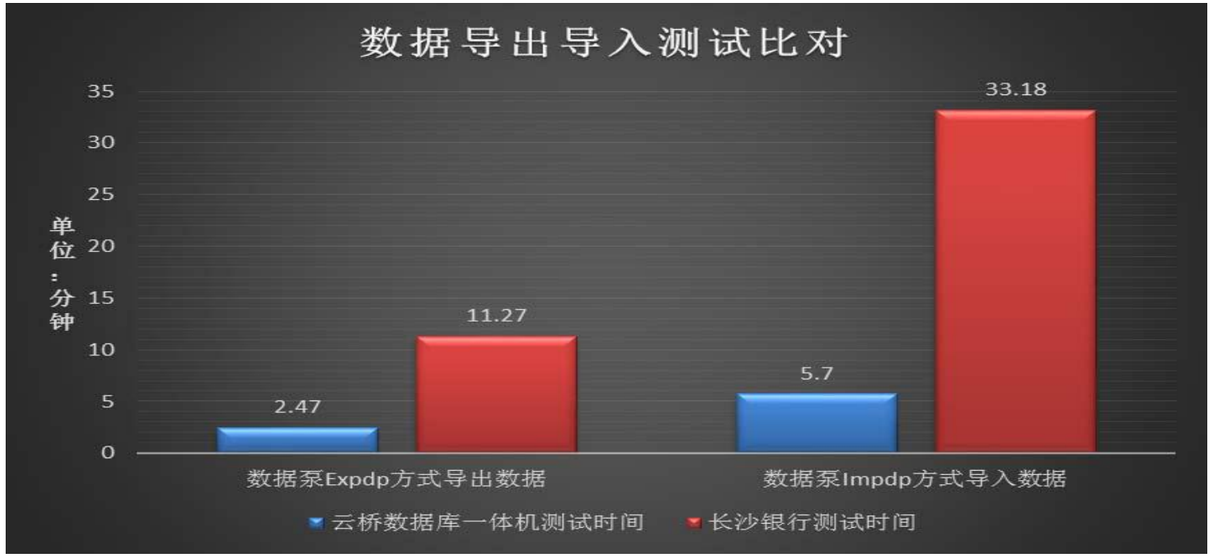
数量：3

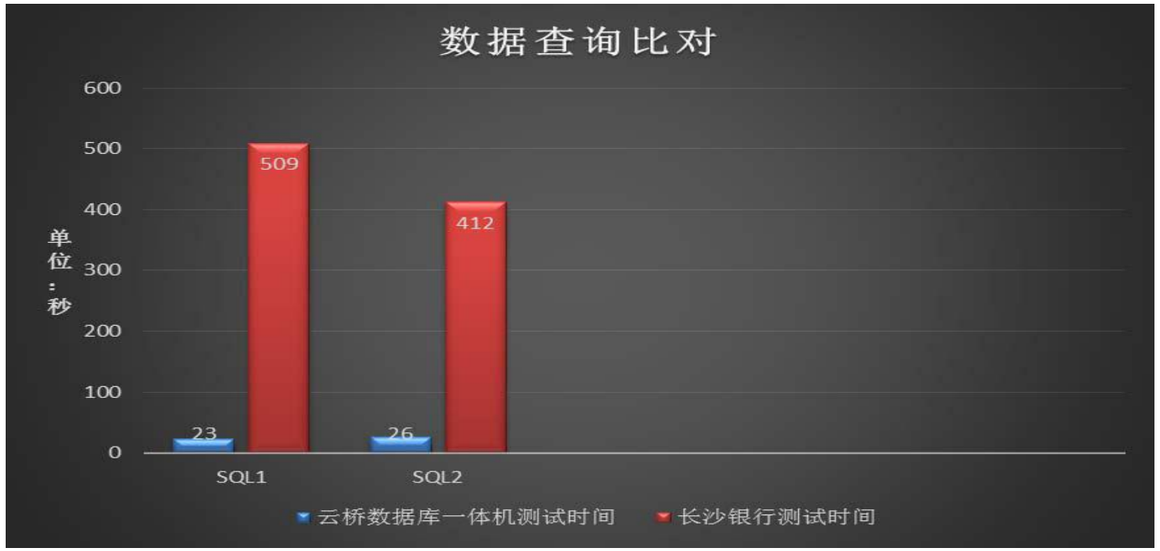
新架构 RAC节点：

AMD Quad-Core 8380 2.5GHZ 4 CPU (4核)
内存：64G

数量：2

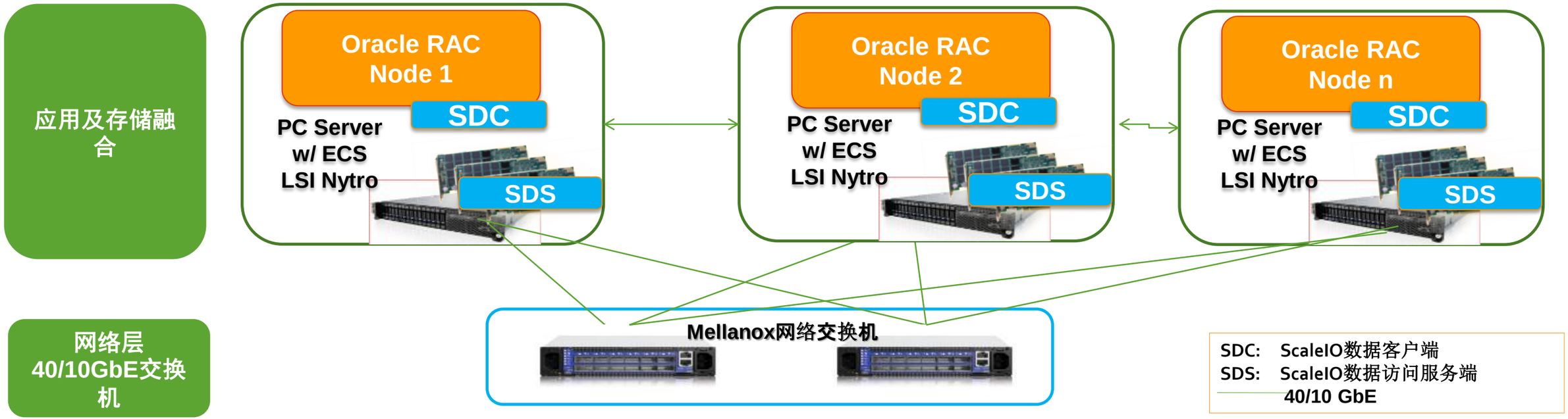
分钟





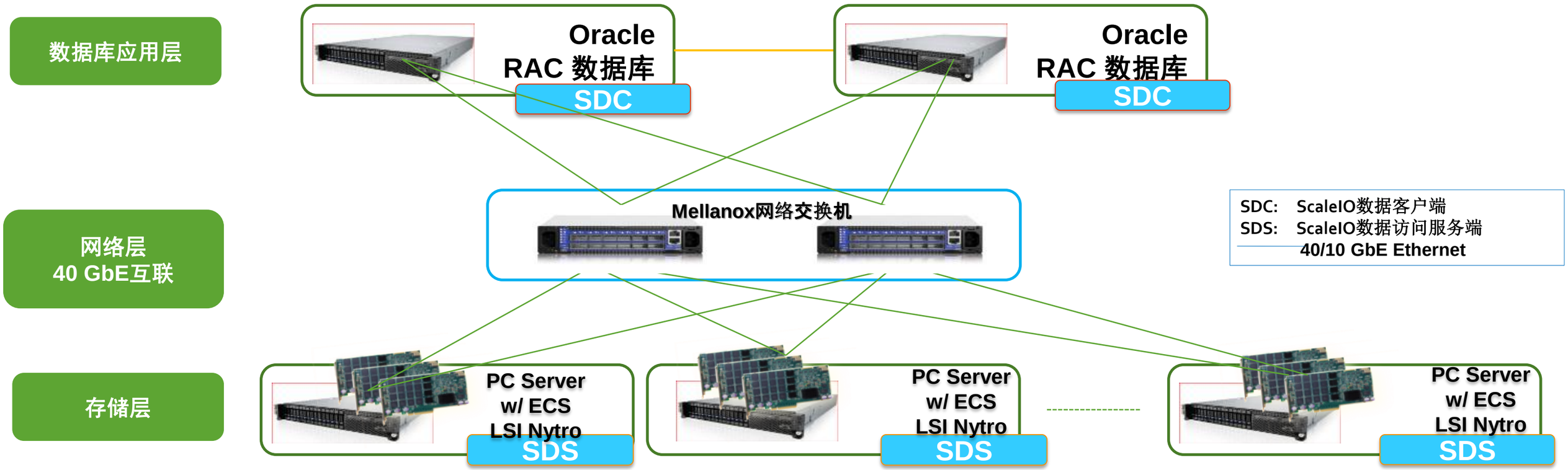
性能提升20倍以上

基于Mellanox以太网的Oracle RAC 方案 1 - 融合架构



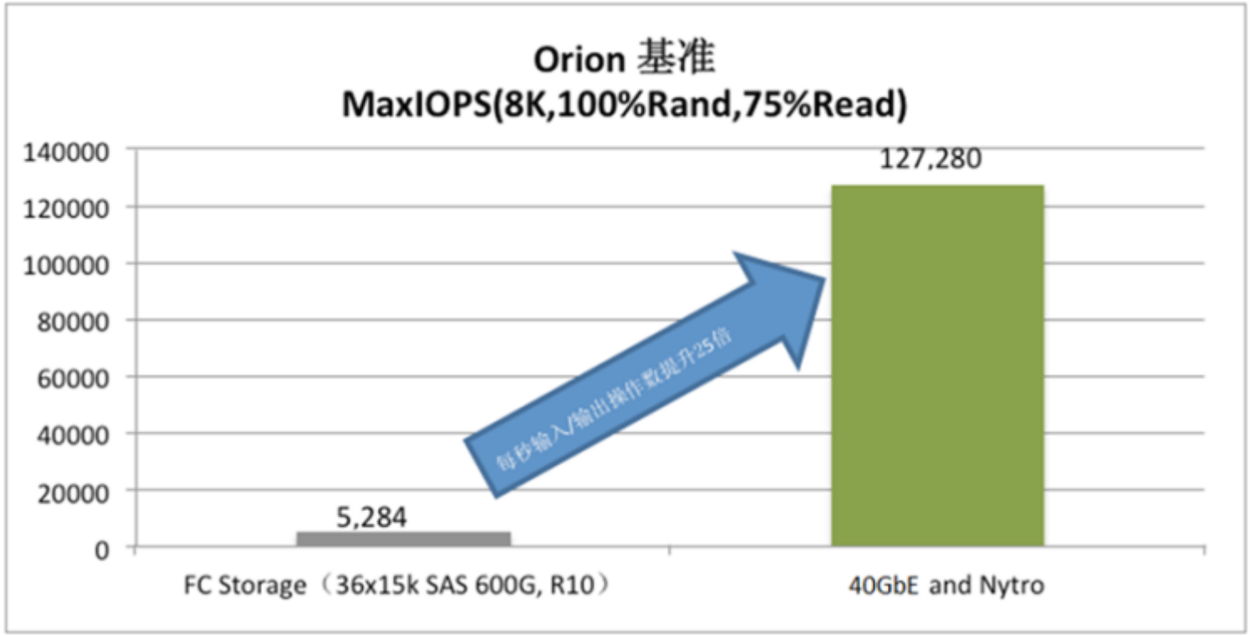
Mellanox 40GbE 交换机+40GbE网卡实现最佳Oracle性能与扩展性

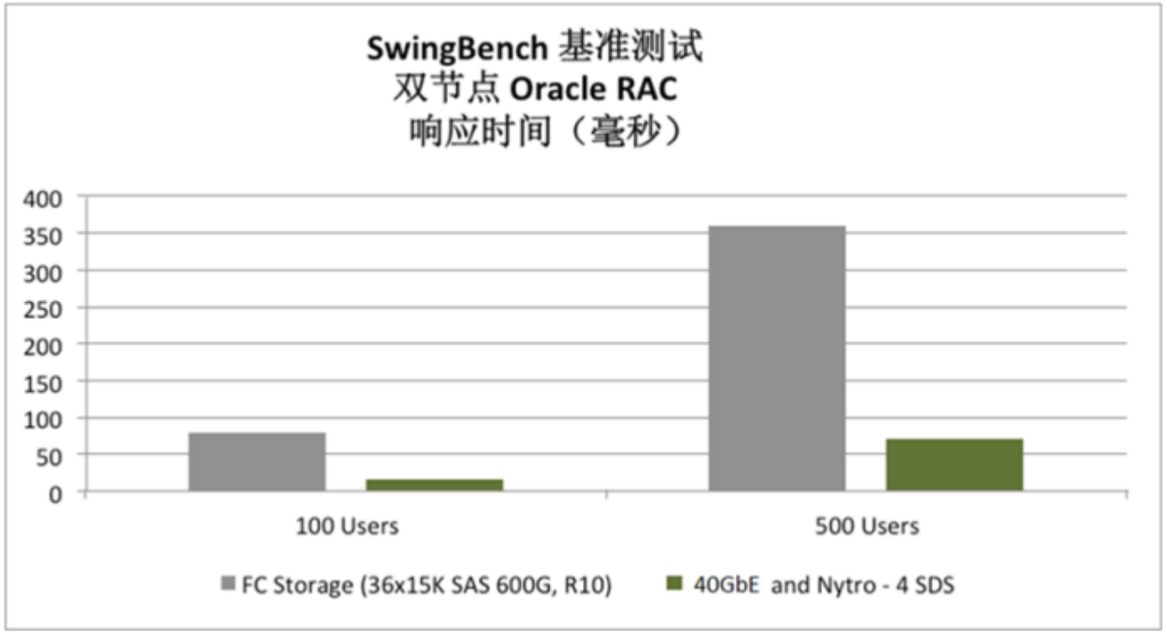
基于Mellanox以太网的Oracle RAC 方案 2 -分层架构



Mellanox 40GbE 交换机+40GbE网卡实现最佳Oracle性能与扩展性

Mellanox加速分布式Oracle RAC性能





Mellanox 40GbE 交换机+40GbE网卡实现最佳Oracle性能与扩展性

加速大数据

Data Intensive Applications Require Fast, Smart Interconnect

hadoop

mongoDB

ceph

redis

Cassandra

End-to-End & Virtual Network Ready InfiniBand and Ethernet Portfolio

ICs

Adapter Cards

Switches/Gateways

Host/Fabric Software

Metro / WAN

metroX

Cables/Modules

Certified Networking Gear

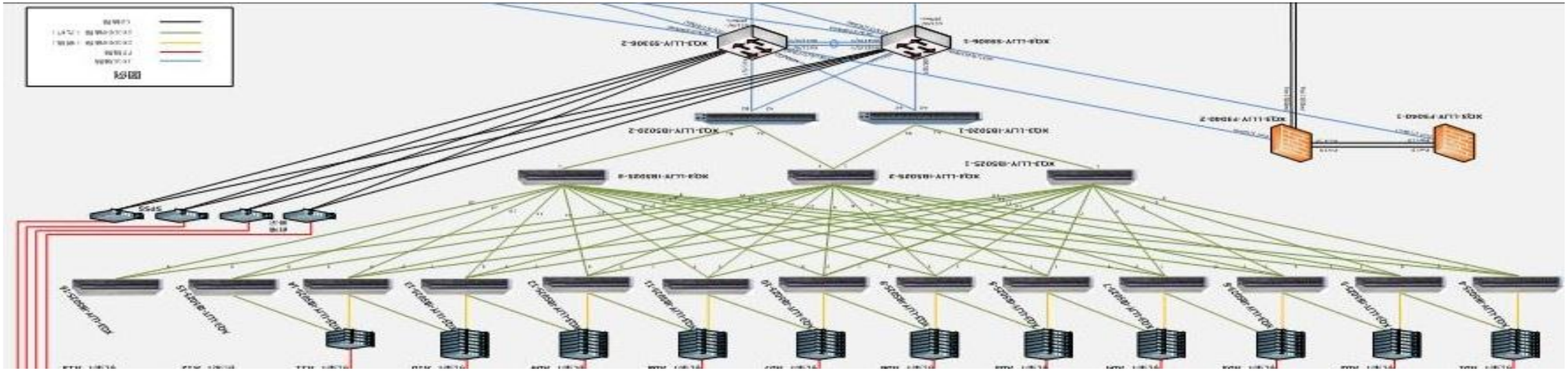
Pivotal™

cloudera CERTIFIED

ConnectX®-3

SwitchX®-2

河南移动大数据部署实例



- 任意服务器之间进行40Gb/s无阻塞通信,消除节点间I/O瓶颈
- 网络采用36口交换机堆叠的Fat-tree架构，最大幅度地降低网络开销，随着节点数量的增加，整体性能线性增加，提供最佳的线性扩展能力
- 集群任意节点均与两个交换机互联，实现系统的高可靠性；
- 全省上网行为数据每天8TB，大数据处理平台（90台）40秒完成忙时数据装载、5小时内完成日报表处理

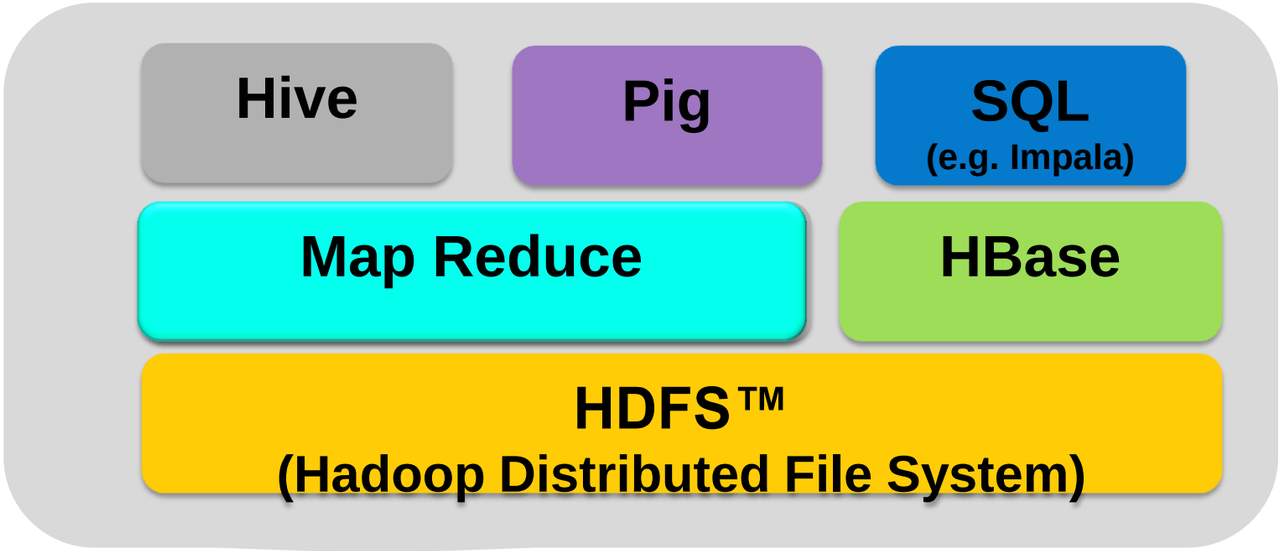
TCO大幅降低高达79.6%

Hadoop缺陷调查

- 管理工具
- 性能
- 可靠性
- SQL支持
- 备份与恢复

451 Research 2013 Hadoop调查

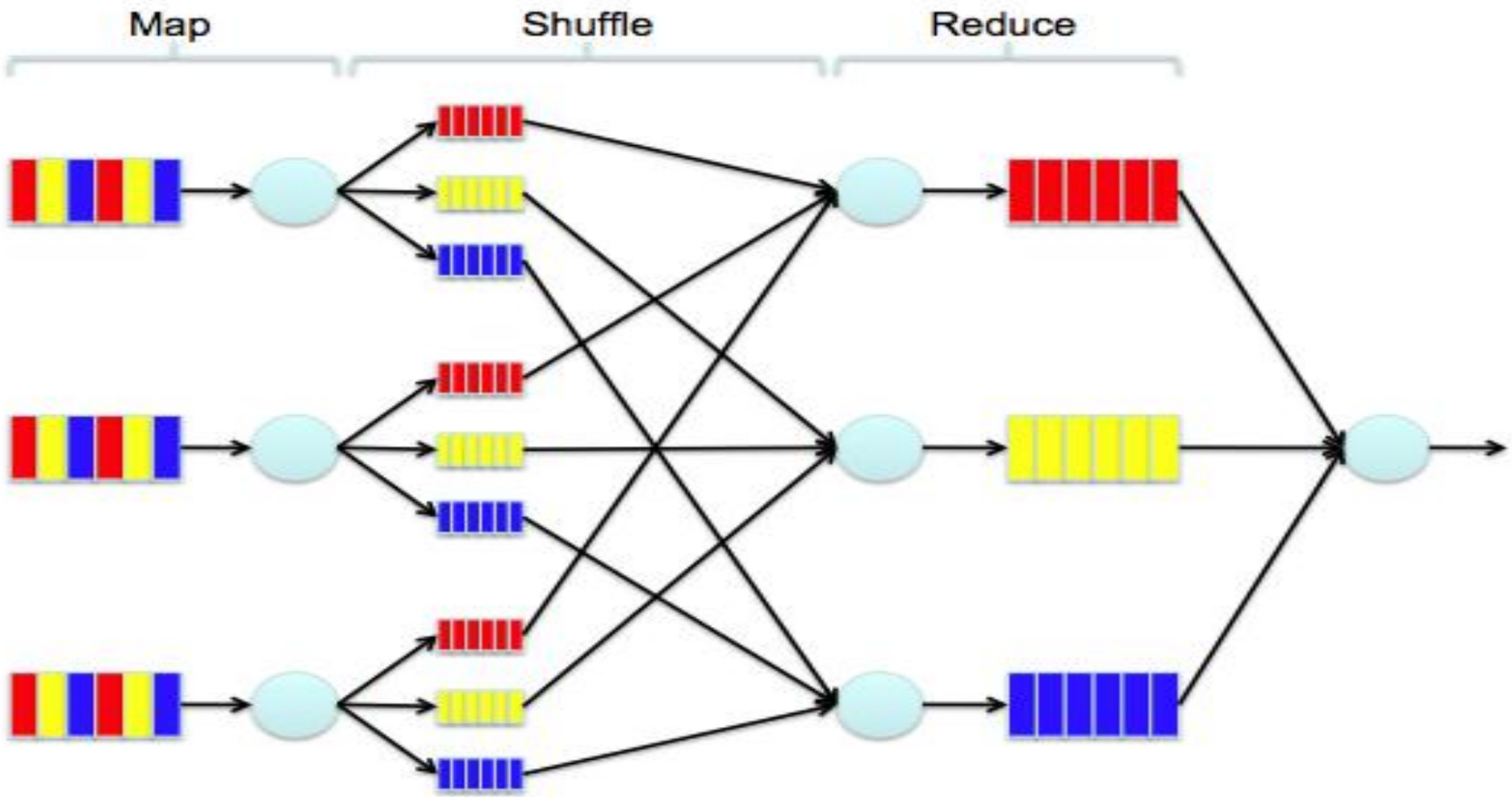
Hadoop性能提升



- **性能提升需求**
  - **实时操作**
  - **更快执行速度**

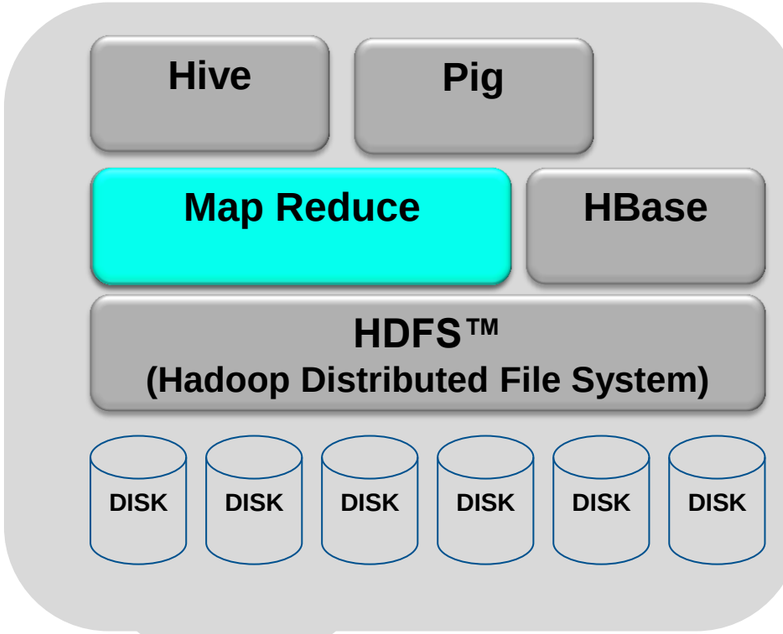
- **挑战**
  - HDFS本事的数据延迟问题
  - 不能支持大量小文件
  - Map Reduce, Hbase, Hive, 等等的效率.

Map Reduce 工作进程



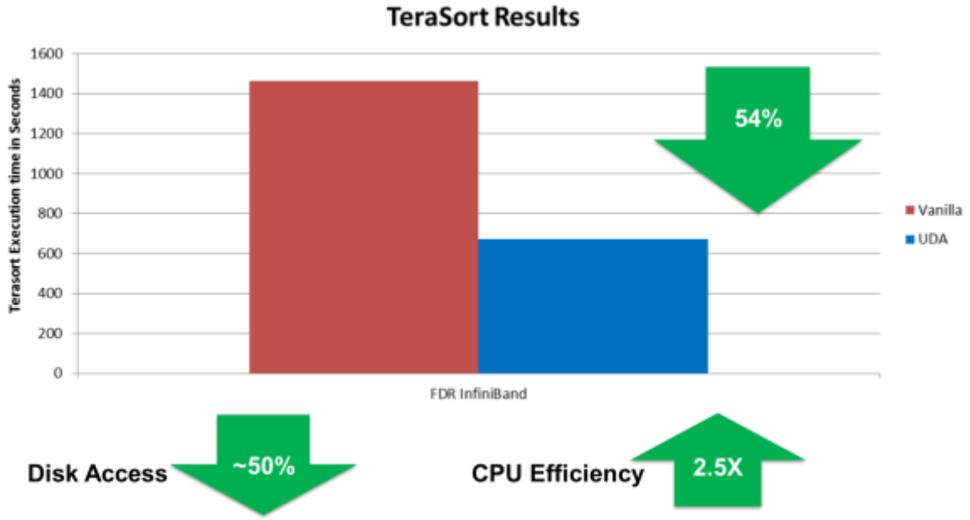
Hadoop MapReduce RDMA优化

- 开源插件
- 支持Hadoop版本
  - Apache 3.0, Apache 2.2.x, Apache 1.3
  - Cloudera Distribution Hadoop 4.4内嵌支持



ConnectX®-3

SwitchX®-2

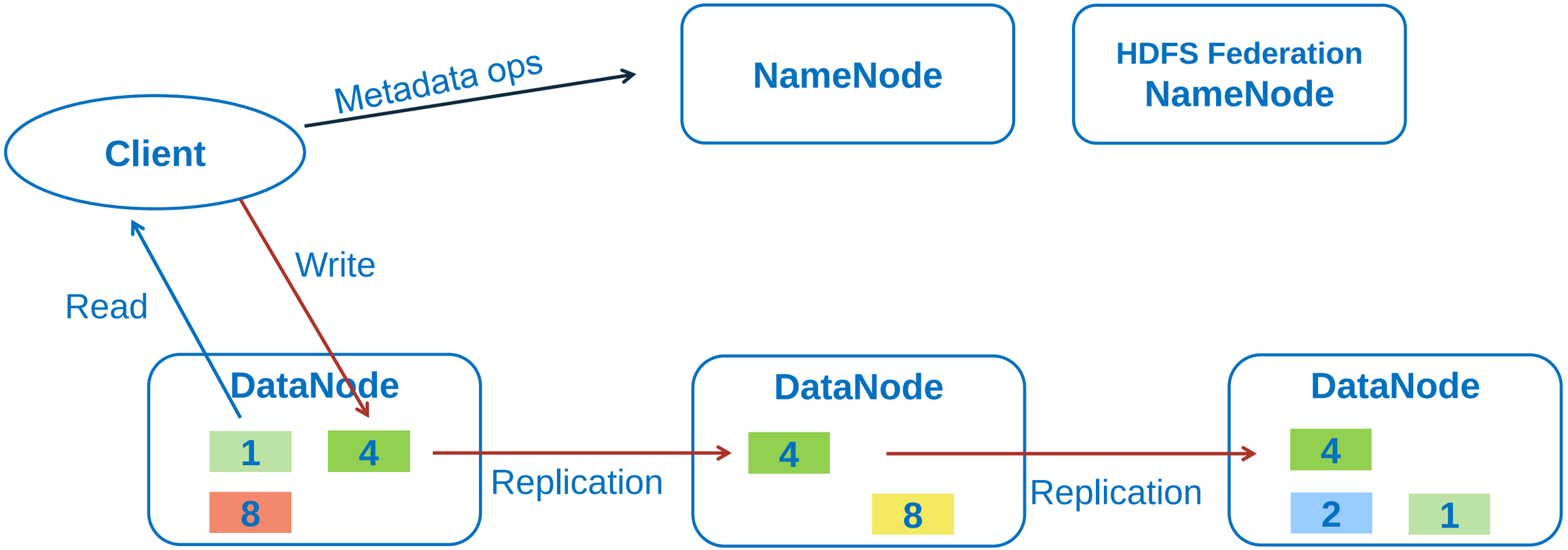


hadoop

AUBURN UNIVERSITY

速度翻倍

HDFS 操作

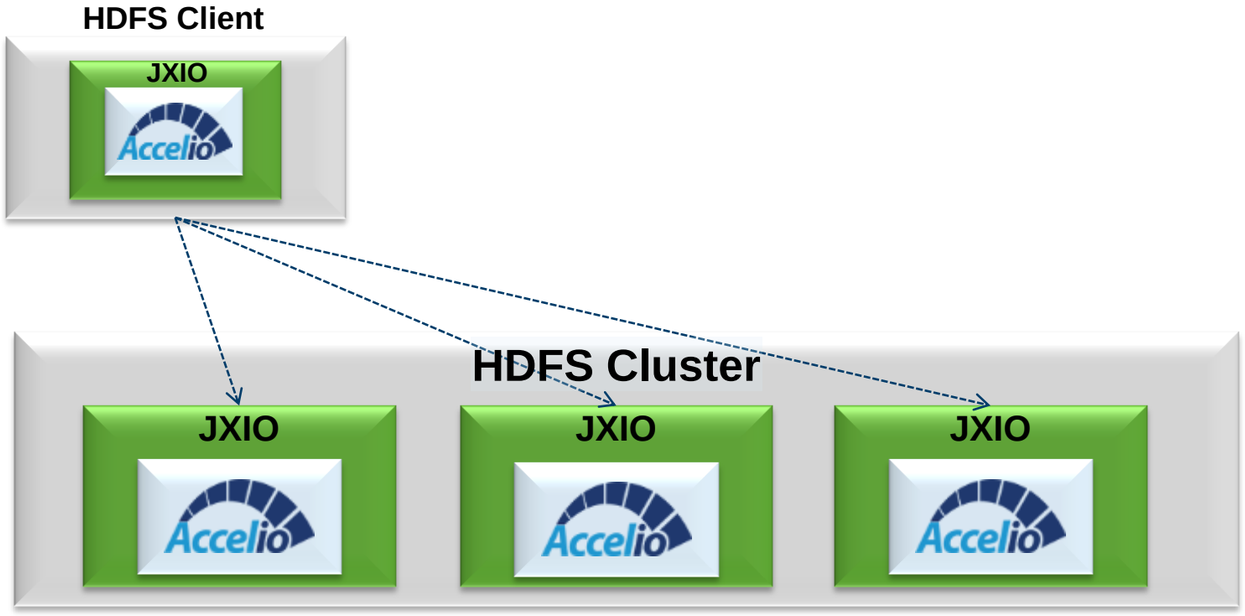


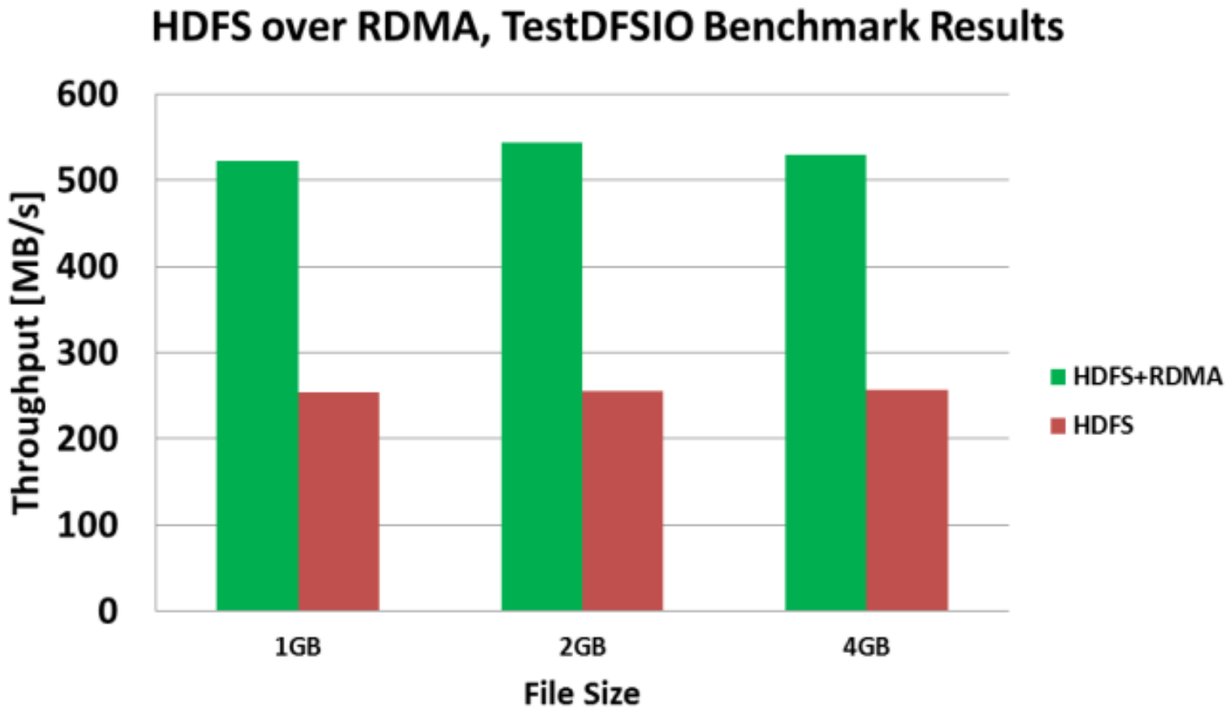
- **HDFS Federation**
- **更快硬盘**
- **更快CPU和内存**

IO成为瓶颈

Hadoop HDFS RDMA优化

- HDFS 基于RDMA进行移植
- 支持CDH5 和 HDP2.1





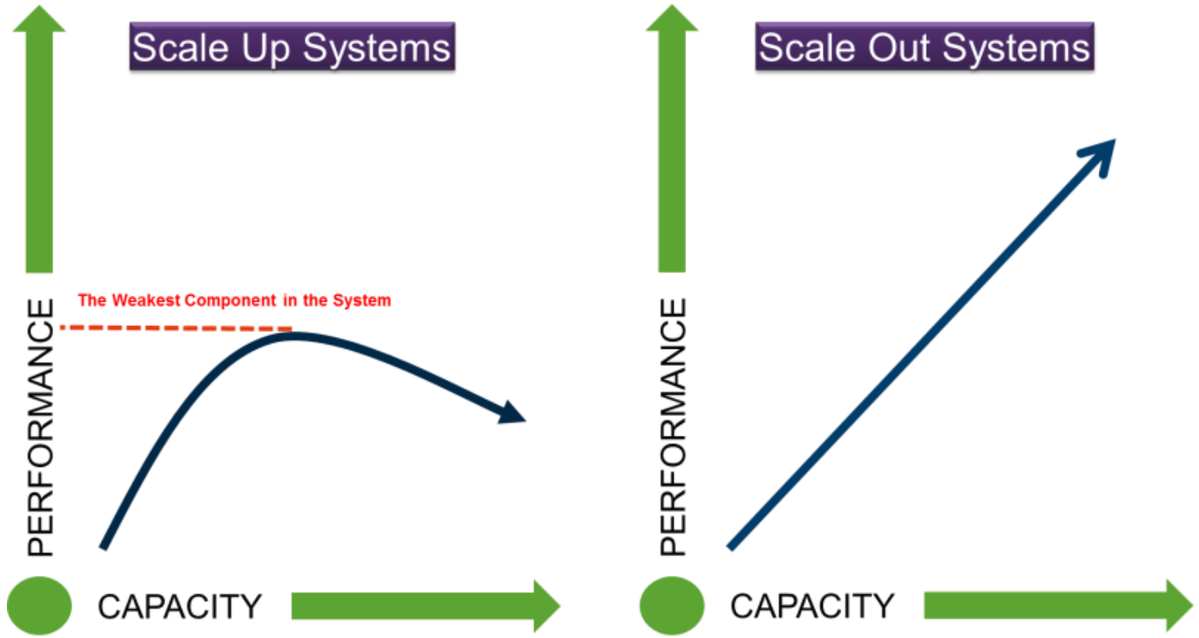
Hadoop存储架构的限制

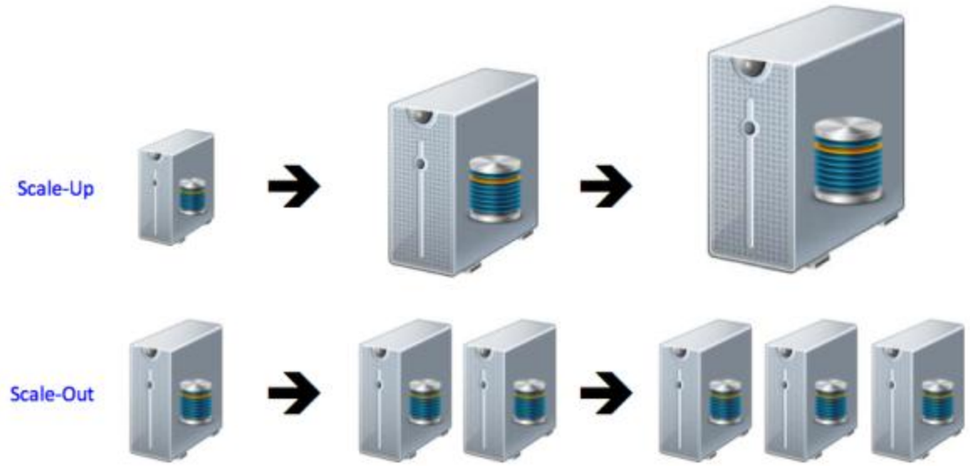
- **Hadoop 使用本地硬盘保持数据本地性和低延迟**
  - 很多高价值数据存在于外置存储
  - 拷贝数据到HDFS, 运行分析, 然后将结果发到另外系统
  - 浪费存储空间
  - 随着数据源的增多，数据管理变成噩梦

- **直接访问外部数据，无需拷贝？**
  - 需要解决性能问题

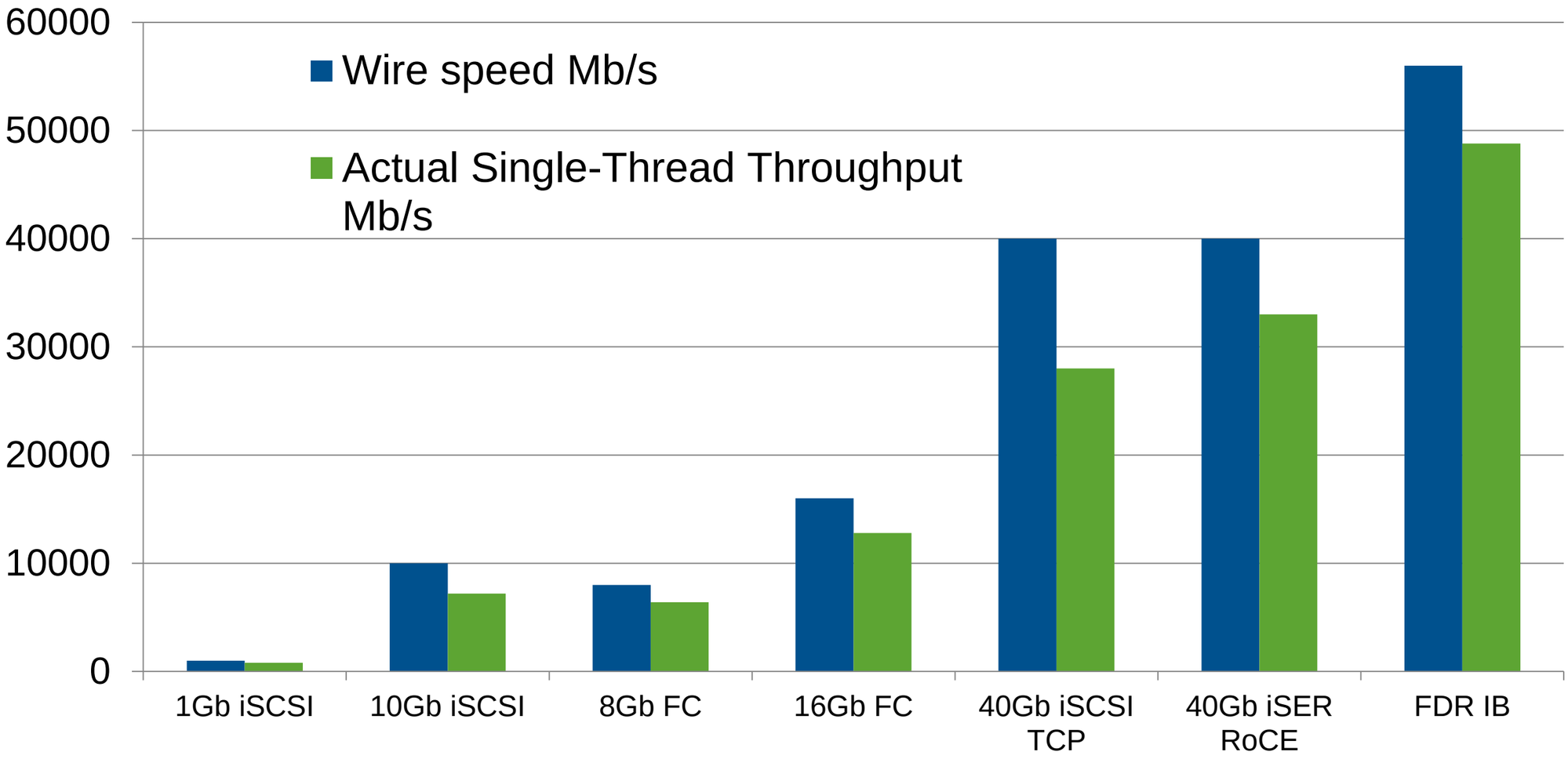
存储: 从Scale-Up 向 Scale-Out 演进

- Scale-out 存储系统采用分布计算架构
  - 可扩展，灵活，高性价比



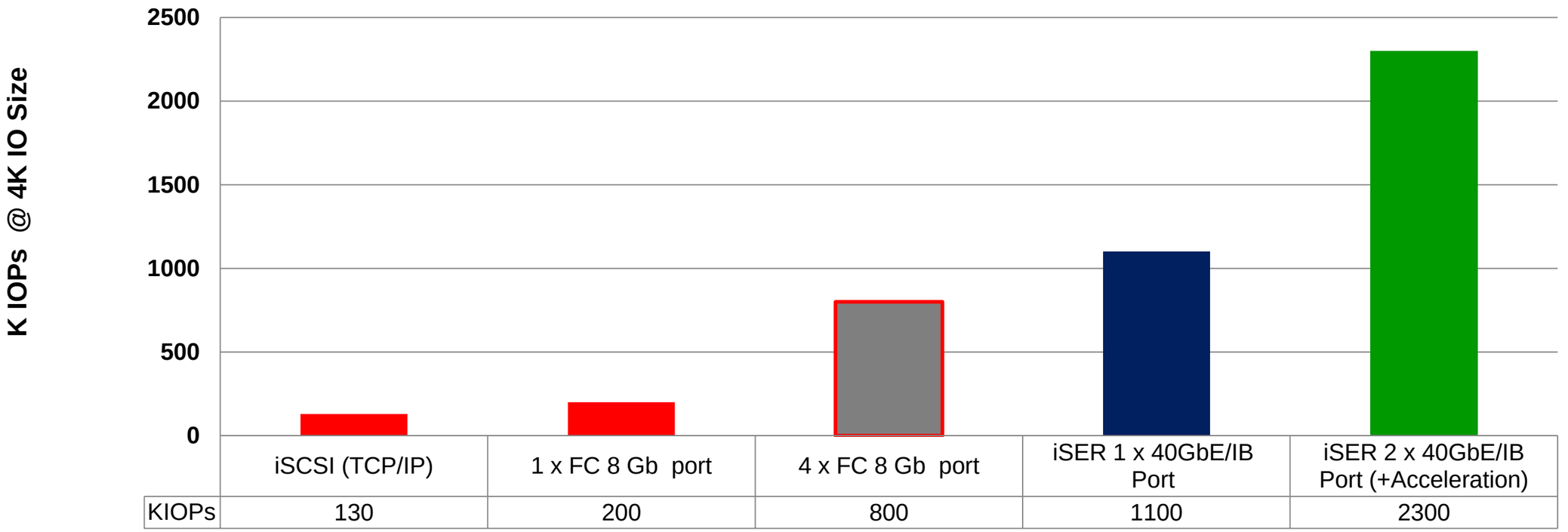


顺序文件读性能 (单端口)



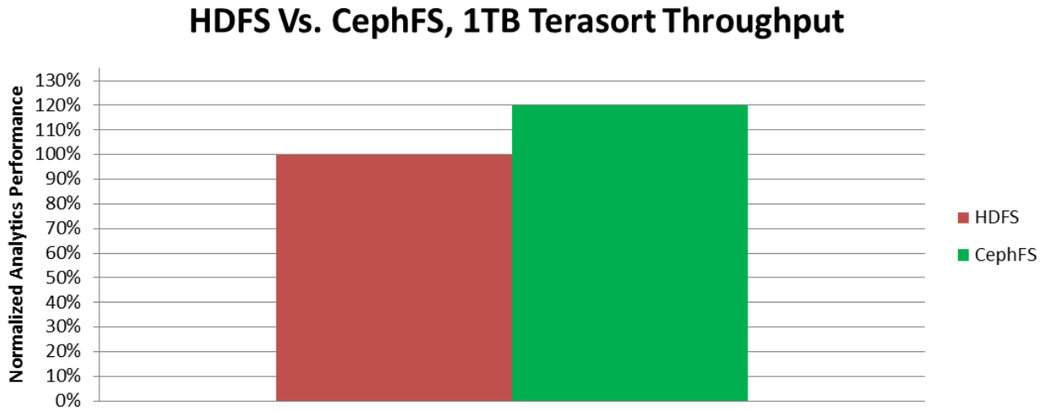
iSER : iSCSI over RDMA

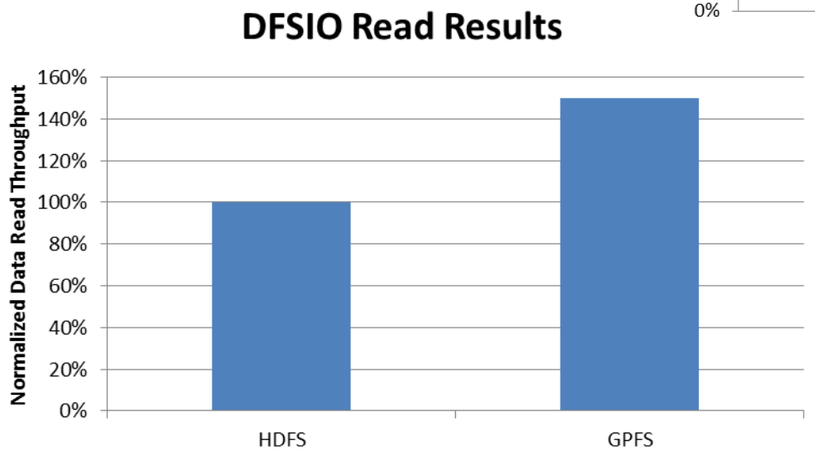
iSER实现最快的存储访问



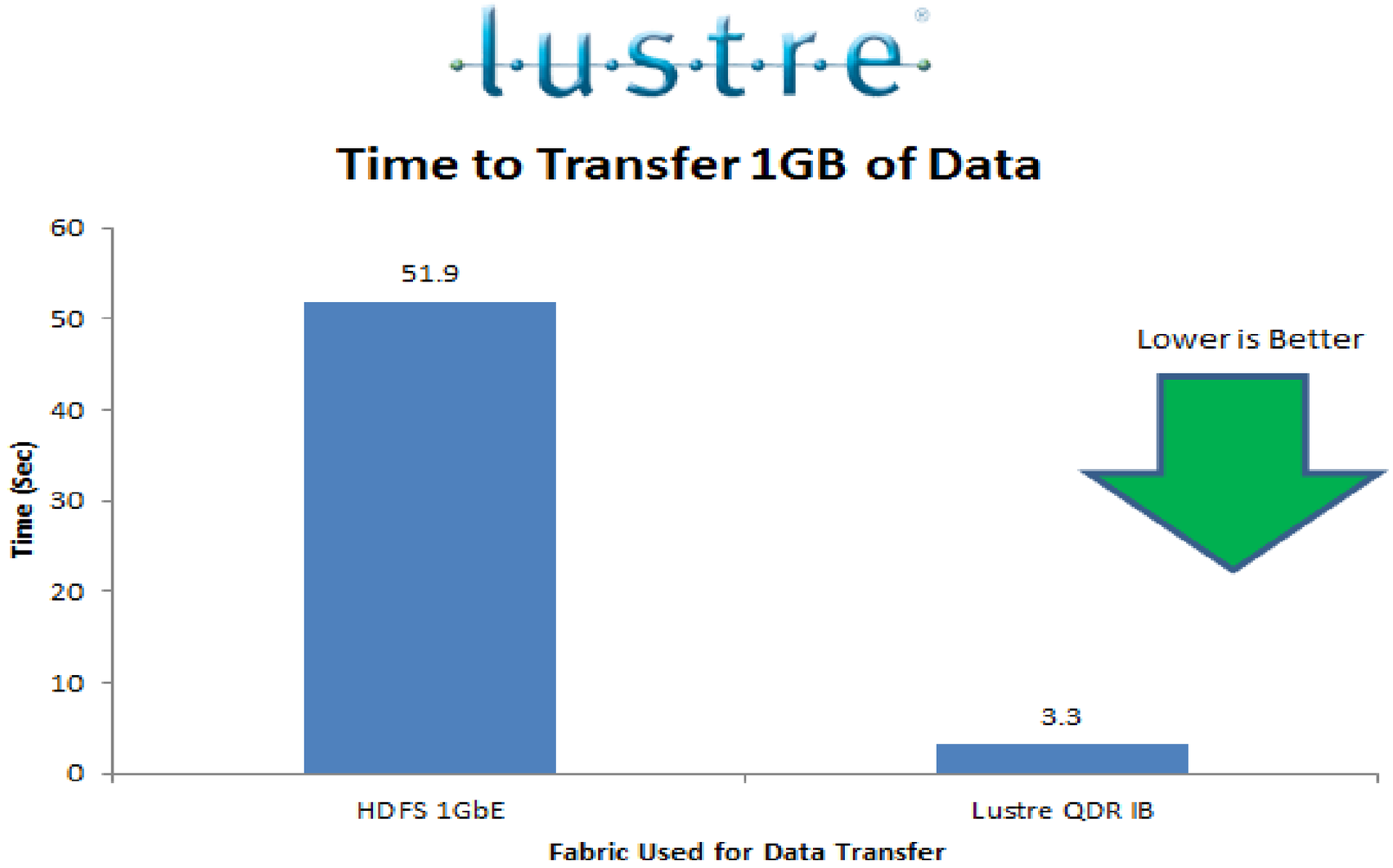
方案1: 使用并行文件系统替换HDFS

- **使用高性能网络和RDMA**
  - 避免性能瓶颈
- **避免单点失败** – HDFS Name Node
- 节省33%磁盘空间!



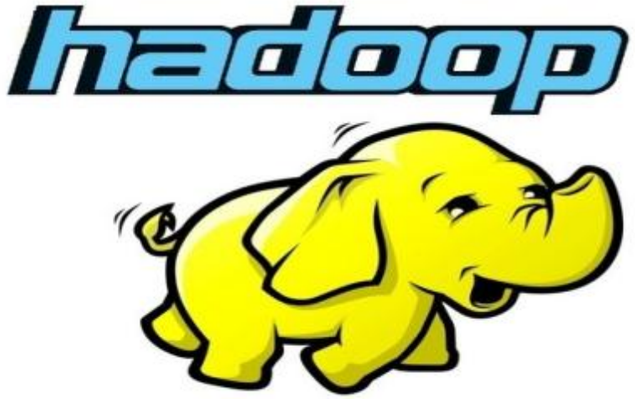


Lustre 作为文件系统方案



Mellanox网络与RDMA技术实现最高 Lustre 性能

Hadoop over Cloud?



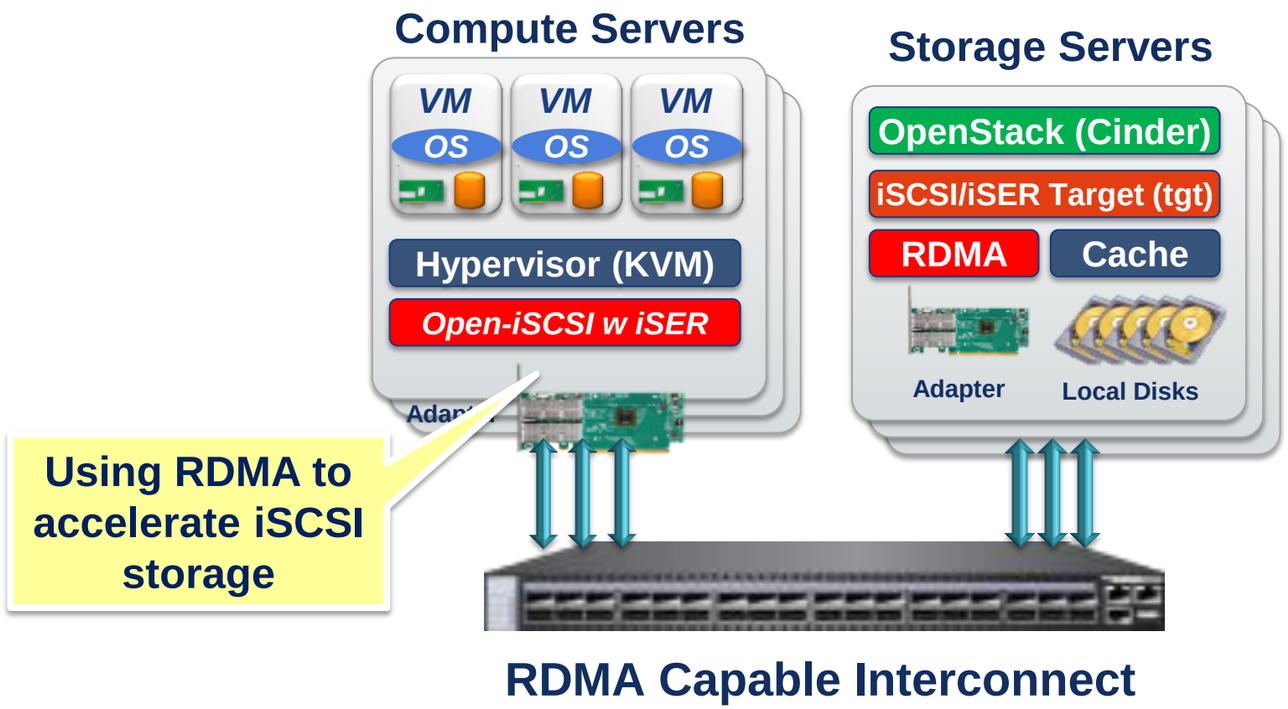
好处:

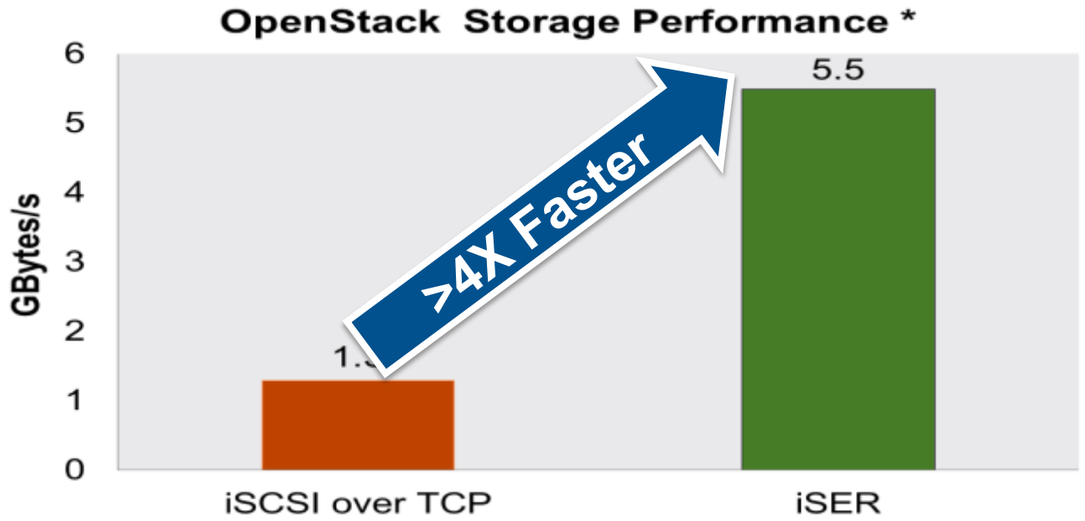
- 降低成本
- 弹性获得大量资源
- 与数据源更近
- 简化Hadoop操作

顾虑:

- 通常满负荷运转，而不是多虚机配置
- 云存储慢且贵

最快的OpenStack 存储速度





- 利用OpenStack 内置组件与管理功能
  - RDMA 已经内置在OpenStack
- RDMA 实现**最快性能, 占用更低CPU负荷**

支持RDMA的高速网络大幅提升大数据应用性能

4倍性能!

Benchmark: TestDFSIO (1TeraByte, 100 files)

2倍性能!

Benchmark: 1M Records Workload (4M Operations)

2X faster run time and 2X higher throughput

2倍性能!

Benchmark: MemCacheD Operations

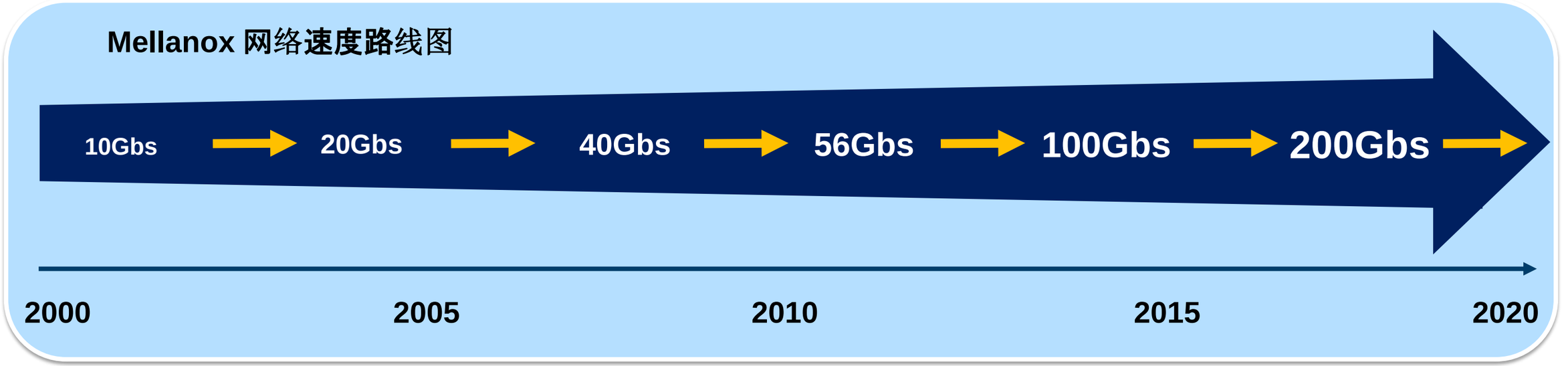
3倍性能!

Benchmark: Redis Operations

步入100G网络时代

通过更快移动数据实现更大数据价值

引领网络速度的发展



Gbs – Gigabit per second

迈向更高网速

进入100G时代

March 2014

LinkX

铜缆(Passive, Active)

光纤 (VCSEL)

硅光

June 2014

SwitchIB

36 EDR (100Gb/s) 端口, <90ns 延迟

吞吐量7.2Tb/s

NEW!

ConnectX®-4

100Gb/s 网卡, 0.7us 延迟

1.5亿消息/秒

(10 / 25 / 40 / 50 / 56 / 100Gb/s)

不止于InfiniBand

端到端高速以太网

Thank You

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™