大数据时代的Intel之Hadoop

系统方案架构师：朱海峰

英特尔®中国云计算创新中心

2013.4　　北京

法律声明

本文所提供之信息均与英特尔® 产品相关。本文不代表英特尔公司或其它机构向任何人明确或隐含地授予任何知识产权。除相关产品的英特尔销售条款与条件中列明之担保条件以外，英特尔公司不对销售和/或使用英特尔产品做出其它任何明确或隐含的担保，包括对适用于特定用途、适销性，或不侵犯任何专利、版权或其它知识产权的担保。

“关键业务应用”是指当英特尔® 产品发生故障时，可能会直接或间接地造成人员伤害或死亡的应用。如果您针对此类关键业务应用购买或使用英特尔产品，您应当对英特尔进行赔偿，保证因使用此类关键业务应用而造成的产品责任、人员伤害或死亡索赔中直接或间接发生的所有索赔成本、损坏、费用以及合理的律师费不会对英特尔及其子公司、分包商和分支机构，以及相关的董事、管理人员和员工造成损害，无论英特尔及其分包商在英特尔产品或其任何部件的设计、制造或警示环节是否出现疏忽大意的情况。

英特尔可以随时在不发布声明的情况下修改规格和产品说明。设计者不应信赖任何英特产品所不具有的特性，设计者亦不应信赖任何标有保留权利摂或未定义摂说明或特性描述。英特尔保留今后对其定义的权利，对于因今后对其进行修改所产生的冲突或不兼容性概不负责。此处提供的信息可随时改变而毋需通知。请勿使用本信息来对某个设计做出最终决定。

文中所述产品可能包含设计缺陷或错误，已在勘误表中注明，这可能会使产品偏离已经发布的技术规范。英特尔提供最新的勘误表备索。

订购产品前，请联系您当地的英特尔销售办事处或分销商，了解最新技术规范。

如欲获得本文或其它英特尔文献中提及的带订单编号的文档副本，可致电 1-800-548-4725，或访问http：//www.intel.com/design/literature.htm 性能测试和等级评定均使用特定的计算机系统和/或组件进行测量，这些测试大致反映了英特尔® 产品的性能。系统硬件、软件设计或配置的任何差异都可能影响实际性能。购买者应进行多方咨询，以评估其考虑购买的系统或组件的性能。如欲了解有关性能测试和英特尔产品性能的更多信息，请访问:英特尔性能指标评测局限

此处涉及的所有产品、计算机系统、日期和数字信息均为依据当前期望得出的初步结果，可随时更改，恕不另行通知。

英特尔、英特尔标识、英特尔酷睿、至强、Core Inside、Xeon Inside、英特尔凌动、英特尔 Flexpipe 和 Thunderbolt 是英特尔公司在美国和/或其他国家或地区的商标。

英特尔® 主动管理技术要求平台采用支持英特尔主动管理技术的芯片组、网络硬件和软件。系统必须接通电源并建立网络连接。就笔记本电脑而言，英特尔主动管理技术可能在基于主机操作系统的虚拟专用网（VPN）上，或者在无线连接、使用电池电源、睡眠、休眠或关机时无法使用或是某些功能受到限制。如欲了解更多信息，请访问：httP：//www.intel.com/technology/iamt。

英特尔® 架构上的 64 位计算要求计算机系统采用支持英特尔® 64 架构的处理器、芯片组、基本输入输出系统（BIOS）、操作系统、设备驱动程序和应用。实际性能会根据您使用的具体软硬件配置的不同而有所差异。如欲了解更多信息£¬请与您的系统厂商联系。

没有任何计算机系统能够在所有情况下提供绝对的安全性。英特尔® 可信执行技术是由英特尔开发的一项安全技术，要求计算机系统具备英特尔® 虚拟化技术、支持英特尔可信执行技术的处理器、芯片组、基本输入输出系统（BIOS）、鉴别码模块，以及英特尔或其它兼容的虚拟机监视器。此外，英特尔可信执行技术要求系统包含可信计算组定义的 TPMv1.2 以及用于某些应用的特定软件。如欲了解更多信息，请访问：httP：//www.intel.com/technology/security/。

†英特尔® 超线程（HT）技术要求计算机系统具备支持英特尔超线程（HT）技术的英特尔® 奔腾® 4 处理器、支持超线程（HT）技术的芯片组、基本输入输出系统、BIOS 和操作系统。实际性能会根据您所使用的具体软硬件配置的不同而有所差异。有关详细信息，包括哪些处理器支持英特尔 HT 技术，请访问 www.intel.com/products/ht/hyperthreading_more.htm。

英特尔® 虚拟化技术要求计算机系统具备支持英特尔虚拟化技术的英特尔® 处理器、基本输入输出系统、BIOS、虚拟机监视器、VMM、以及用于某些应用的特定平台软件、功能、性能或其它优势会根据软硬件配置的不同而有所差异，可能需要对 BIOS 进行更新。相关应用软件可能无法与所有的操作系统兼容。请咨询您的应用厂商以了解具体信息。

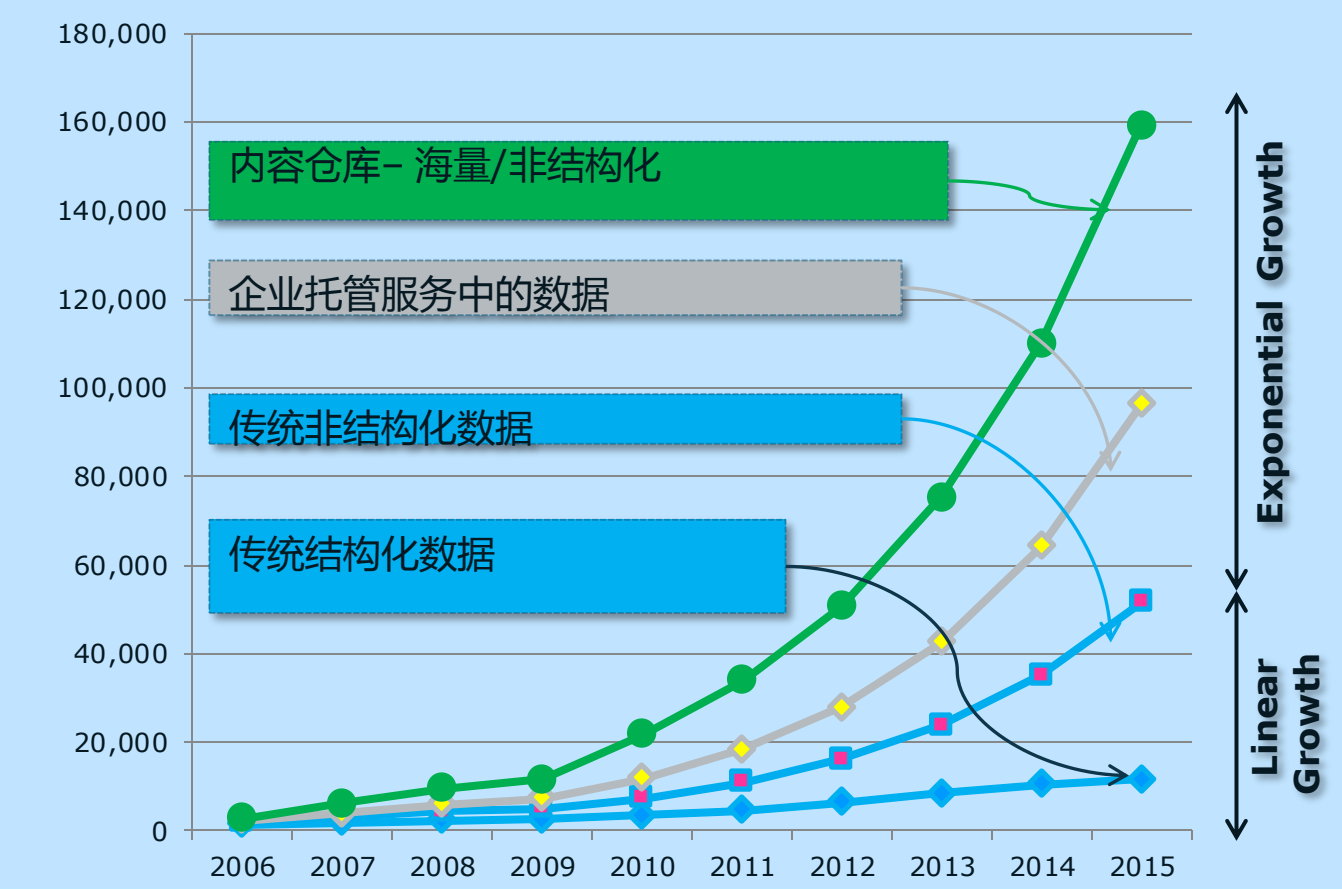
*文中涉及的其它名称及商标属于各自所有者资产。

英特尔所列的厂商仅为方便英特尔客户。但英特尔对于这些设备的质量、可靠性、功能或兼容性不提供任何担保或保证。本列表和/或这些设备可随时更改，恕不另行通知。

提纲

- 大数据时代的新挑战
- 大数据时代的Intel
- 关注产业应用，产研相互促进

大数据时代的数据



从文明诞生到2003年，人类文明产生了5EB的数据；而今天，我们每两天产生5EB的数据。

Eric Schmidt

Google

2011年每天处理的数据超过：

24 PB

2011年6月之前，Facebook平台每天分享资料：

40亿

智慧城市数据
中国某一线城市:

200PB/季度

中国一线城市健康档案数据：

5.5 million

全球 2012 年产生2.7 ZB（1,000,000 PB）数据, 2015 年150亿部接入设备

Source: IDC, 2011 Worldwide Enterprise Storage Systems 2011–2015 Forecast Update.
Worldwide Enterprise Storage Consumption Capacity Shipped by Model, 2006–2015 (PB)

intel

传统的数据处理技术

大数据时代的数据

	传统数据	大数据
数据量	GB -> TB	TB -> PB以上
速度	数据量稳定，增长不快	持续实时产生数据， 年增长率超过60%
多样化	主要为结构化数据	半结构化，非结构化， 多维数据
价值	统计和报表	数据挖掘和预测性分析

“大数据”指**数据集的大小超过了现有典型的数据库软件和工具的处理能力**。与此同时，及时捕捉、存储、聚合、管理这些大数据以及对数据的深度分析的新技术和新能力，正在快速增长，就像预测计算芯片增长速度的摩尔定律一样。

— McKinsey Global Institute

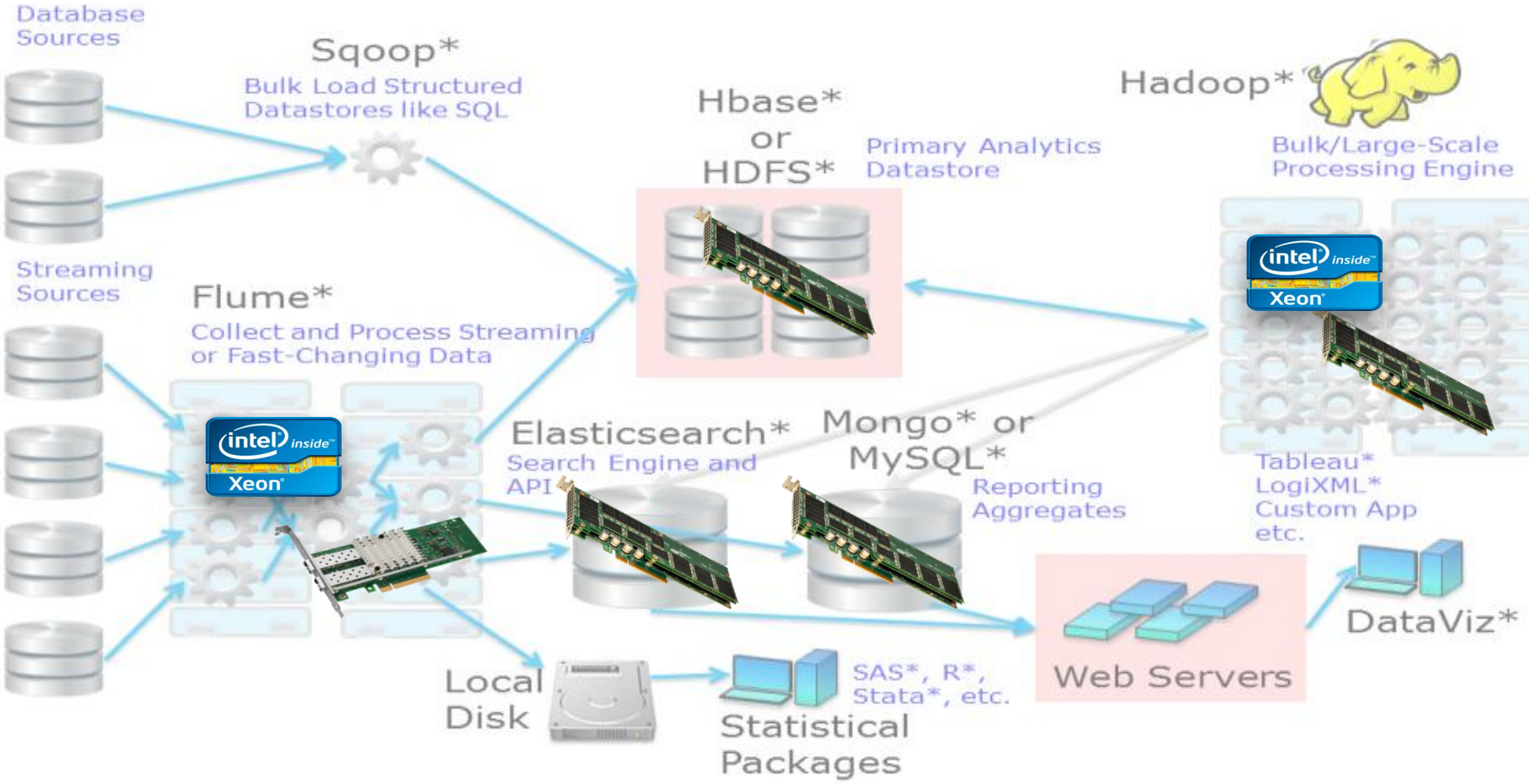
大数据时代的Intel

- Intel的角色
- Intel Hadoop商业发行版
- 对象存储技术

Intel的角色

- 面向大数据应用，在计算、存储和网络方面提供更快更为高效的架构级别的优化方案
- 持续投入大数据应用开发，促进软件系统和服务的不断优化和创新
- 推进终端设备和传感器的智能化，构建互联、可管理的和安全的分布式架构

软硬结合

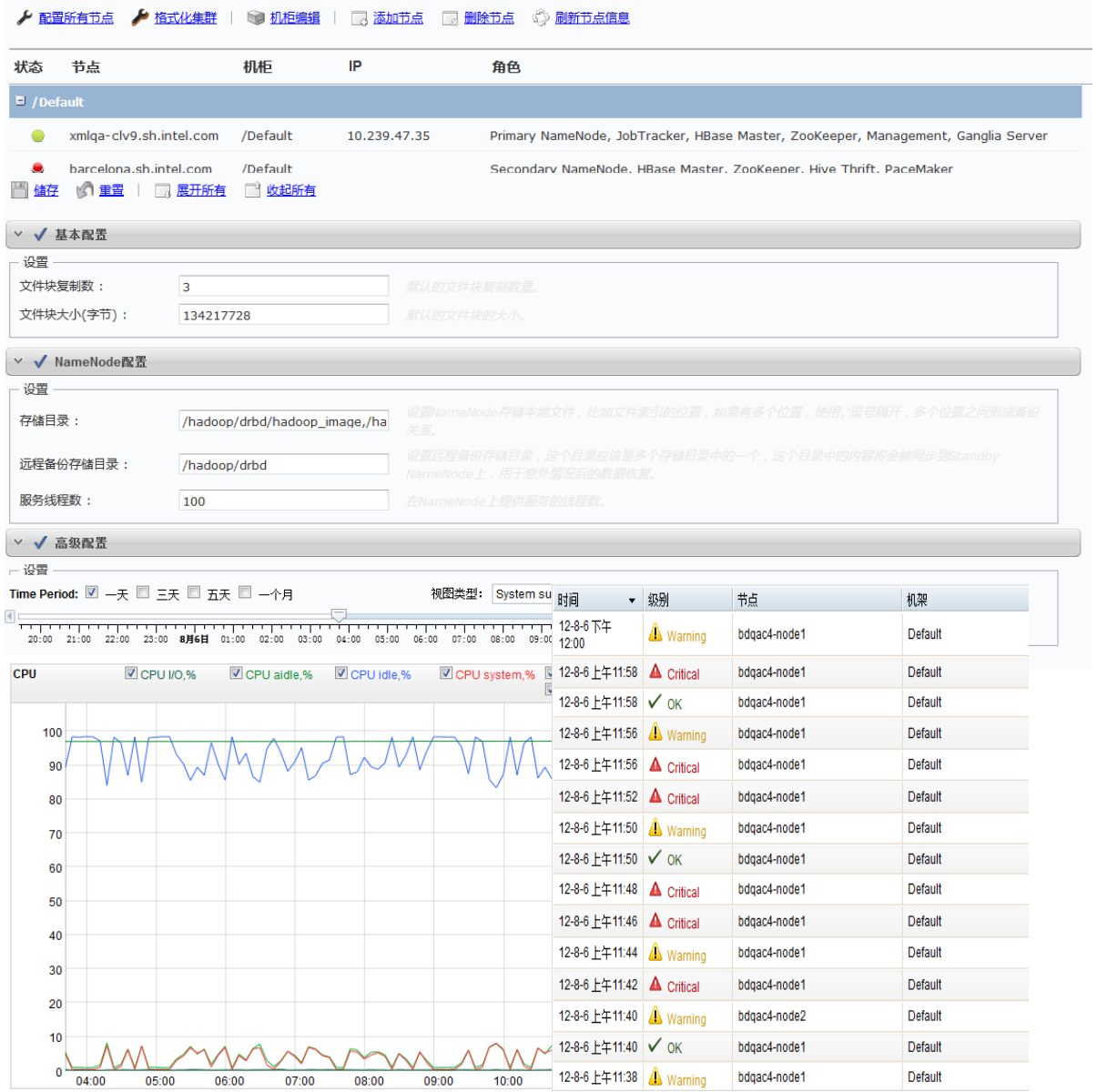

Database Sources
Sqoop*
Bulk Load Structured Datastores like SQL
Hbase* or HDFS*
Primary Analytics Datastore
Hadoop*
Bulk/Large-Scale Processing Engine
Streaming Sources
Flume*
Collect and Process Streaming or Fast-Changing Data
intel inside
Xeon
Elasticsearch*
Search Engine and API
Mongo* or MySQL*
Reporting Aggregates
Tableau*
LogiXML*
Custom App
etc.
DataViz*
Web Servers
Local Disk
SAS*, R*, Stata*, etc.
Statistical Packages

Intel Hadoop商业发行版

优化的大数据处理软件栈

稳定的企业级hadoop发行版	利用硬件新技术进行优化
HBase改进和创新，为Hadoop提供实时数据处理能力	针对行业的功能增强，应对不同行业的大数据挑战

Intel Hadoop Manager – 安装、配置、管理、监控、告警



HBase 统计

状态	:	运行中
集群负载	:	11.00
RegionServer	:	3 (3 运行)
当前主节点	:	bdqac1-node2

ZooKeeper 节点状况

服务器	状态	服务器	状态
bdqac1-node2	运行中	bdqac1-node1	运行中
bdqac1-node3	运行中		

HBase表概览

表名	状态	数据分布
.META.	正常	bdqac1-node4
-ROOT-	正常	bdqac1-node3

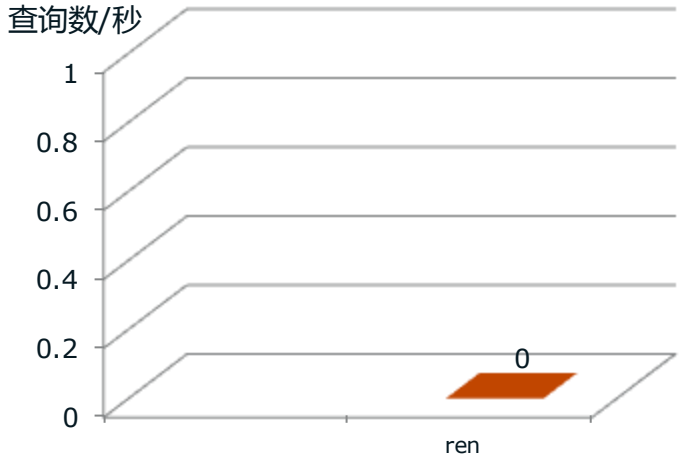
共包含6张表，6张表状态良好，0张表状态异常。Error Table : 0.

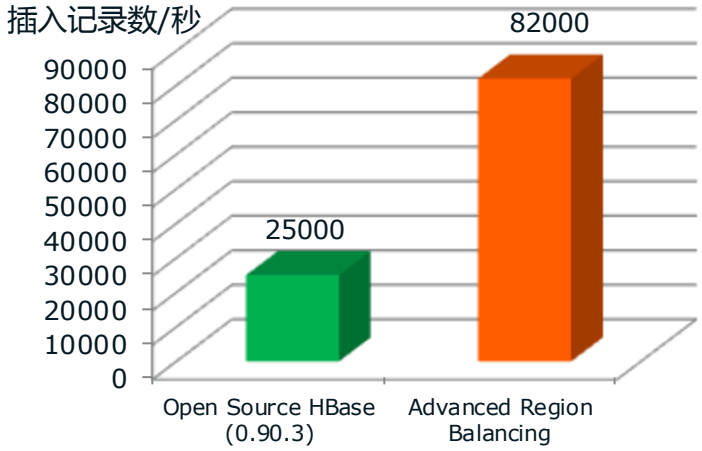
点击此处浏览用户表详细信息...

英特尔Hadoop性能优化

测试配置

- ✓ 性能数据在8台英特尔至强服务器组成的小规模集群上测试得到
- ✓ 服务器配置：6核Intel E5 CPU，48GB内存，8块 7200rpm SATA硬盘，千兆以太网





测试用例和性能

- ✓ 向HBase集群插入1KB大小的记录
- ✓ 每台服务器平均每秒插入1万条记录，峰值在2万条记录
- ✓ 每台服务器，从磁盘扫描数据，每秒完成400个扫描。
  一次扫描从HBase表中获得单个用户一个月内的所有记录（平均100条）

HBase写入性能讨论

写入时的性能瓶颈：

- 客户端
  - 使用Write buffer减少RPC
  - 避免频繁创建HTable对象
  - 如果可以，关闭WAL
- Region负载不均衡：要让写均匀分布到所有的region server上
  - 如果写入的row key是基本单调的（例如时序数据），那么基本上会都落在同一个region上，所以只有一个region server活跃，总体性能会很差
  - “加盐”
- 过多的compaction和compaction不及时
  - 尽量避免：比方说增加compaction thread数，防止阻塞写入
- 过多的split
  - 预分配region

大对象的高效存储（IDH2.3）

在交通、金融等领域，要求存储大量的图片

- 将图片存入HBase，引起大量的compaction
- 将图片存入HDFS，管理使用麻烦

IDH引入了表外存储以解决大对象的高效存储问题

- 类似Oracle的BLOB存储
- 对用户透明
- 2X以上的写入性能，还有进一步提升的空间
- 2X的随机访问性能
- 1.3X的Scan性能
- 接近直接写入HDFS性能

Interactive Hive over HBase

可通过Hive来访问HBase，进行SQL查询

- 使用MapReduce来实现
- 比Hive访问HDFS慢3~5倍

IDH引入了Interactive Hive over HBase

- 完全的Hive支持：常用功能（select, group-by等）用HBase coprocessor实现，其余功能用MapReduce实现，无缝连接
- 去除了MapReduce的overhead，大大减少了数据传输
- 性能有3X~10X的提升

HBase的性能优化

预分配region

启用压缩已减少HDFS数据量，可提高读性能

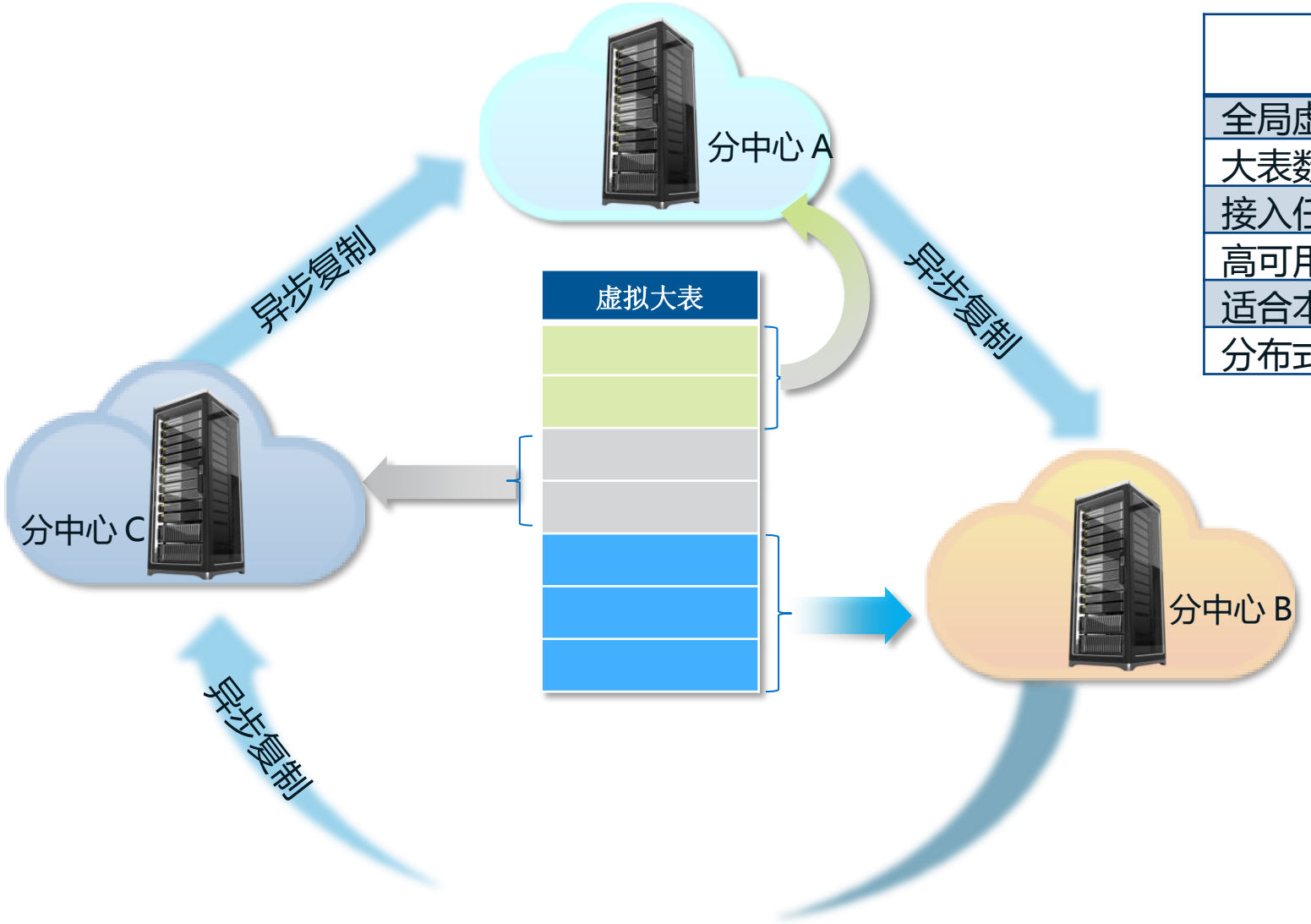
Region Server进程配置大内存（>16G）

每个Region Server拥有的region数量<300

优化表结构设计，防止少数几个region成为瓶颈

- 一个简单的经验公式：每台region server纯写入时高负载应能达到>1万条记录/秒（每记录200字节）

英特尔Hadoop功能增强 - 跨数据中心大表



特点与优势
全局虚拟大表，访问方便
大表数据分区存放在物理分中心
接入任何分中心可访问全局数据
高可用性
适合本地高速写入
分布式聚合计算，避免大数据传输

英特尔Hadoop发行版 – 主要特色

经实际验证的企业级 Hadoop 发行版

•全面测试的企业级发行版，保证长期稳定运行，集成最新开源的和自行开发的补丁，用户可以及时修正漏洞保证各个部件之间的一致性，使应用顺滑运行

实时数据处理的分布式大数据应用平台

•通过对 HBase 进行改进和创新，英特尔 Hadoop 发行版提供实时数据处理功能。为企业对数据的实时监控和即时处理提供有效保障

针对企业用户开发的新的平台功能

•提供企业关键应用程序所需的即时大数据分析，以及其他针对企业用户需要的增强功能，例如：提供跨数据中心的 HBase 数据库虚拟大表功能，实现 HBase 数据库复制和备份功能， 等等。

提供底层 Hadoop 性能优化算法和稳定性增强

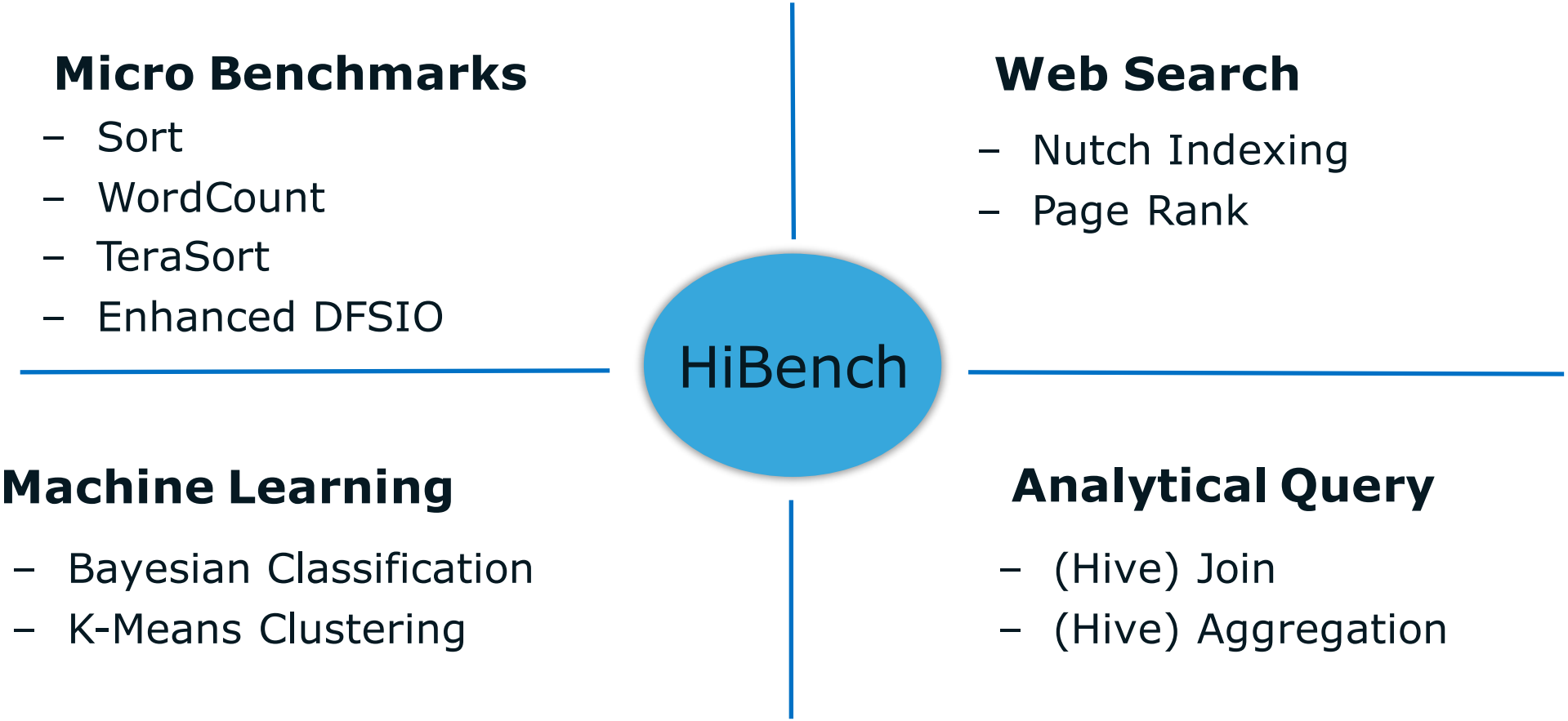
•基于 Hadoop 底层的大量优化算法，配合英特尔优化架构，使应用效率更高、计算存储分布更均衡，系统安装程序计算得出的优化参数配置，适合大多数应用情况，与硬件技术相结合，提高平台性能

提供企业必须的管理和监控功能

•提供独有的基于浏览器的集群安装和管理界面，解决开源版本管理困难的问题，提供网页、邮件方式的系统异常报警

intel

性能评测工具：Intel HiBench



HiBench 1.0 paper ("The HiBench Suite: Characterization of the MapReduce-Based Data Analysis") published in ICDE'10 workshops

HiBench 2.2 released to open source under Apache License 2.0 at https://github.com/intel-hadoop/hibench

HiBench典型测试: Microbenchmarks

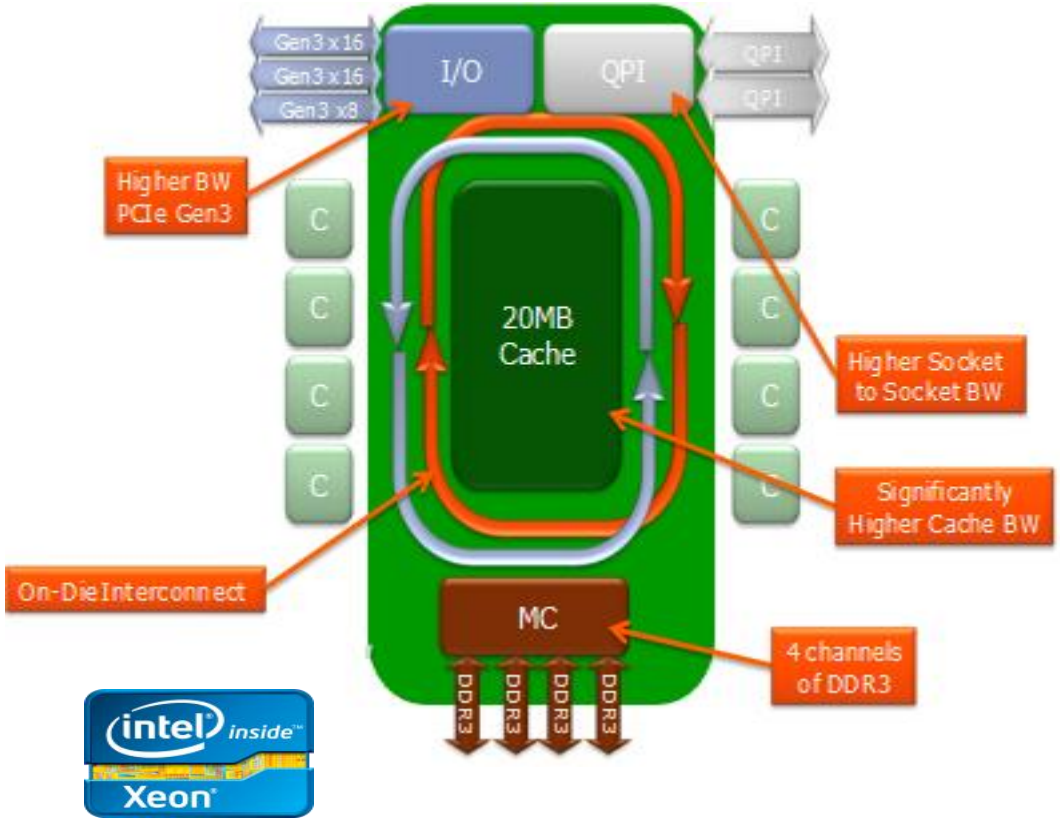
Workload	Description	Rationale
Sort	Example in Apache Hadoop distribution ○ Sort a large amount of text data	Representative of a large subset of real-world MapReduce jobs ○ Transform data from one representation to another
WordCount	Example in Apache Hadoop distribution ○ Count occurrence of each word in input	Representative of a large subset of real-world MapReduce jobs ○ Extract a small amount of interesting data from a large data set
TeraSort	Example in Apache Hadoop distribution ○ Sort 10 billion 100-byte (1TB) records	Standard sorting benchmark started by Jim Gray ○ Used by Google and Yahoo to demonstrate the performance of their MapReduce clusters publicly
Enhanced DFSIO	Enhanced version of *TestDFSIO* in Apache Hadoop distribution ○ Measure aggregate read/write bandwidth of HDFS cluster	MapReduce job measuring HDFS performance ○ The original TestDFSIO program only computes the average I/O rate & throughput of each Map task, instead of aggregate bandwidth of HDFS cluster

HiBench典型测试: Search

Workload	Description	Rationale
Nutch Indexing	The indexing subsystem of Apache Nutch (an open source search engine)	Large scale indexing system is one of the most significant uses of MapReduce (e.g., in Google, Facebook, etc.)
Page Rank	Open source implementation of *page-rank* algorithm (by the CMU Pegasus project)	

Intel® Xeon® =智能数据中心的“核心”

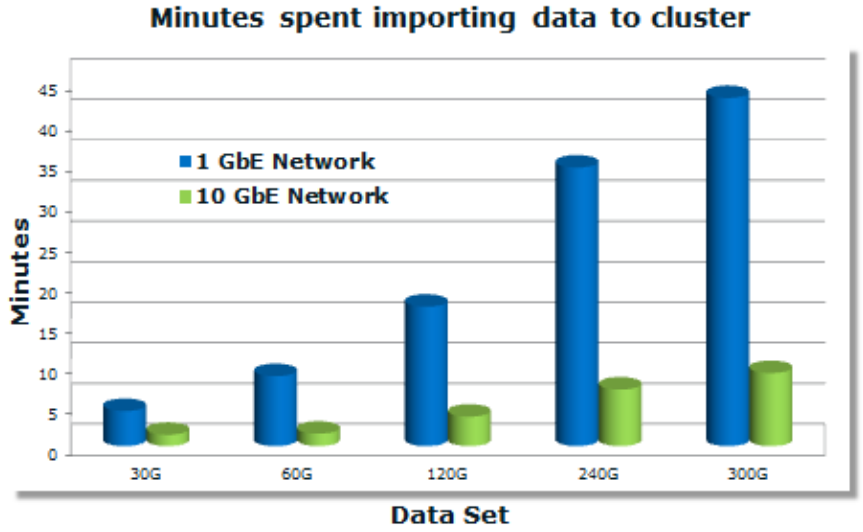
- Integrated PCI Express* Gen 3.0
- Intel® Hyper-Threading Technology, two Threads/Core
- Shared Last Level Cache, 2.5 MB/Core
- Higher memory bandwidth with DDR3
- Integrated Memory Controller
- PCIe Non-Transparent Bridge
- Asynchronous DRAM self-refresh (ADR)
- Intel® QuickData Technology Direct Memory Access

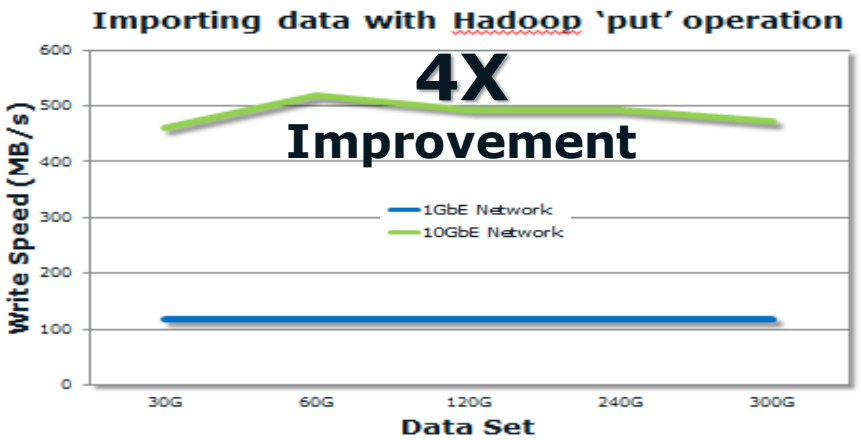


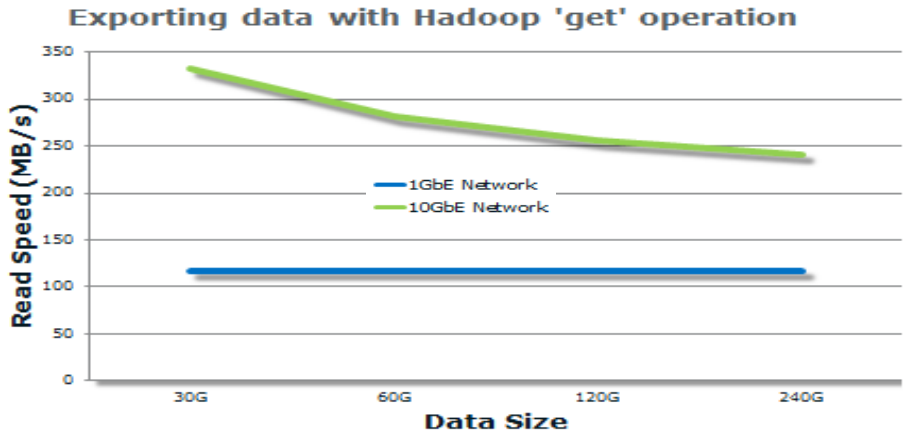
Intel® Xeon®助力大数据计算

高速网络提升大数据平台处理性能

80% less time waiting



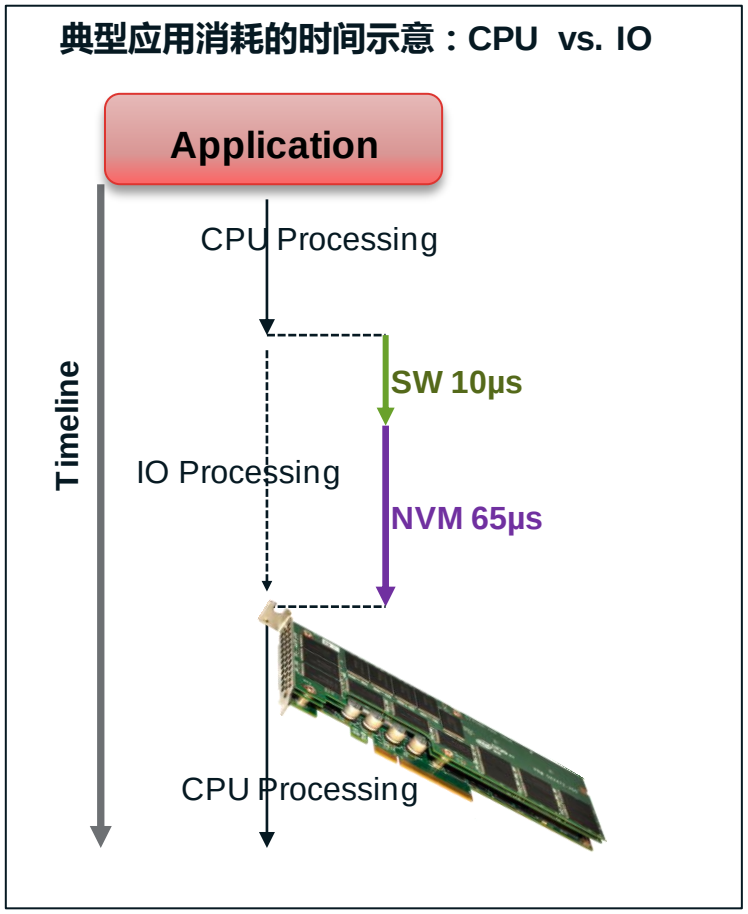




Performance comparison using best submitted/published 2-socket server results on the SPECfp*_rate_base2006 benchmark as of 6 March 2012.

10GbE全面提升系统吞吐，价格也可接受

新的存储架构— NVM



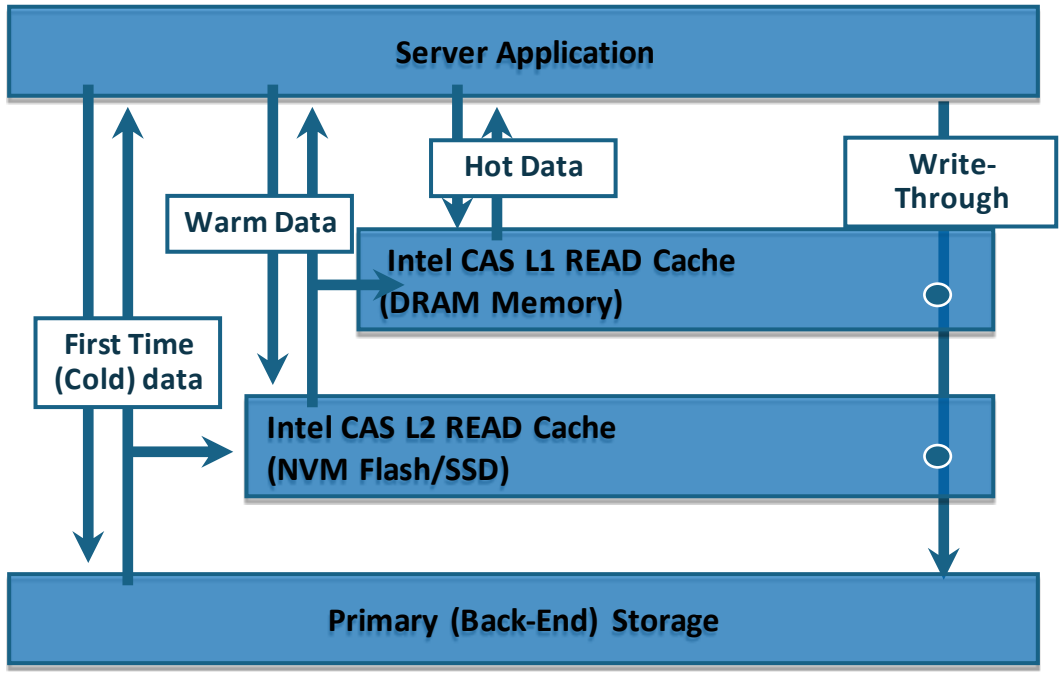
Intel® SSD 910系列

- 性能增强
  - 顺序读/写：2.0/1.0 GB/s
  - 随机读/写: 180/75 KIOPS
  - 读/写延迟：65/65µs
- 高耐久技术(HET)的25nm MLC
  - 写入次数提升10倍
  - 相对传统MLC寿命提升30倍t

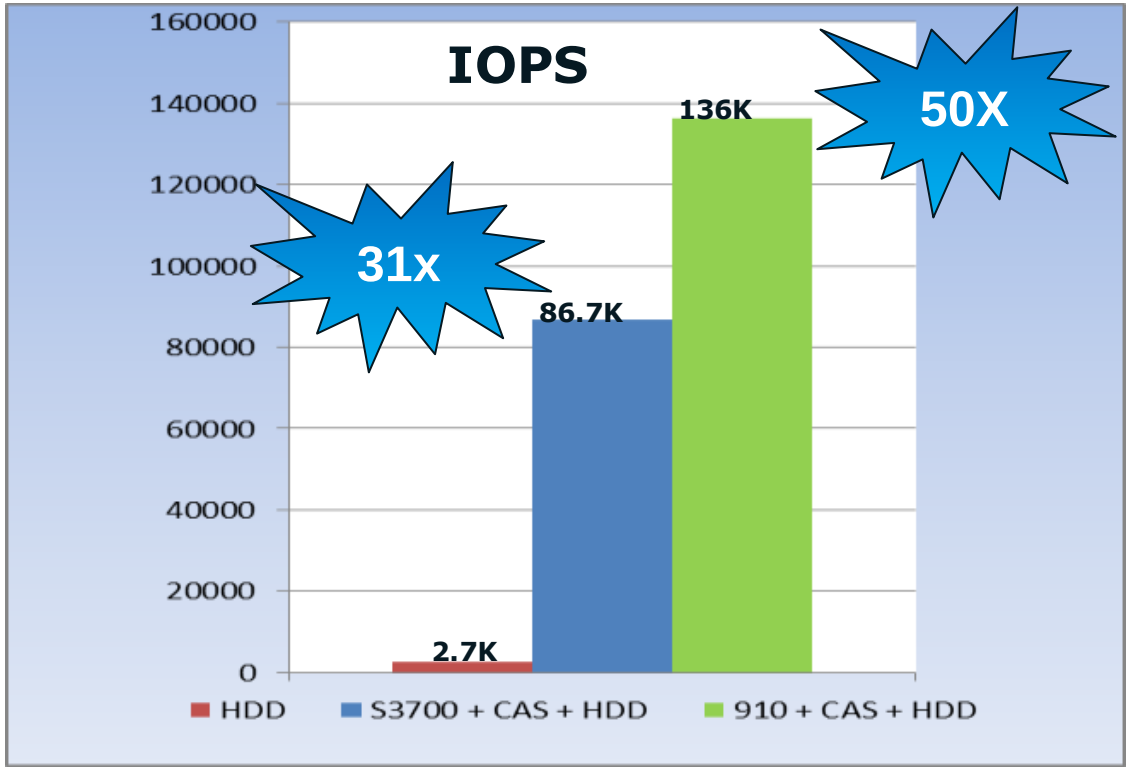
降低延迟，大幅提升IOPS

软件存储加速：Intel® CAS

- Microsoft Windows平台以服务方式运行;Linux上是kernel module
- Multi-Level Cache; 与系统内存整合一起提高性能
- 对应用透明
- 被缓存设备，可以挂载成普通文件系统

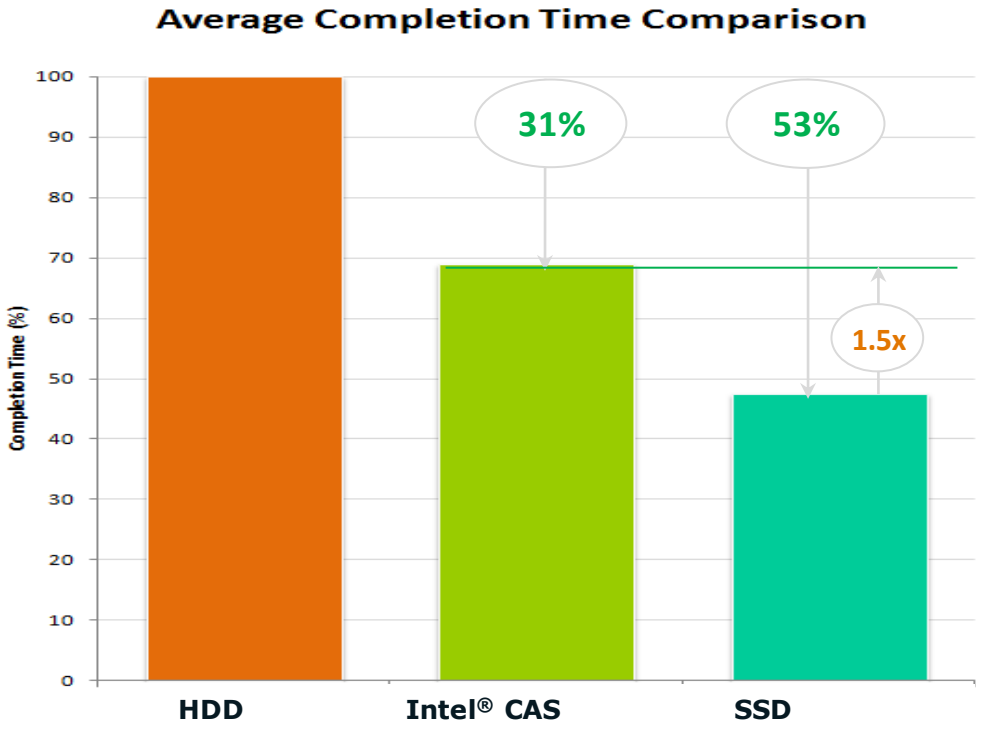


性能指数级提升



100% Random Reads

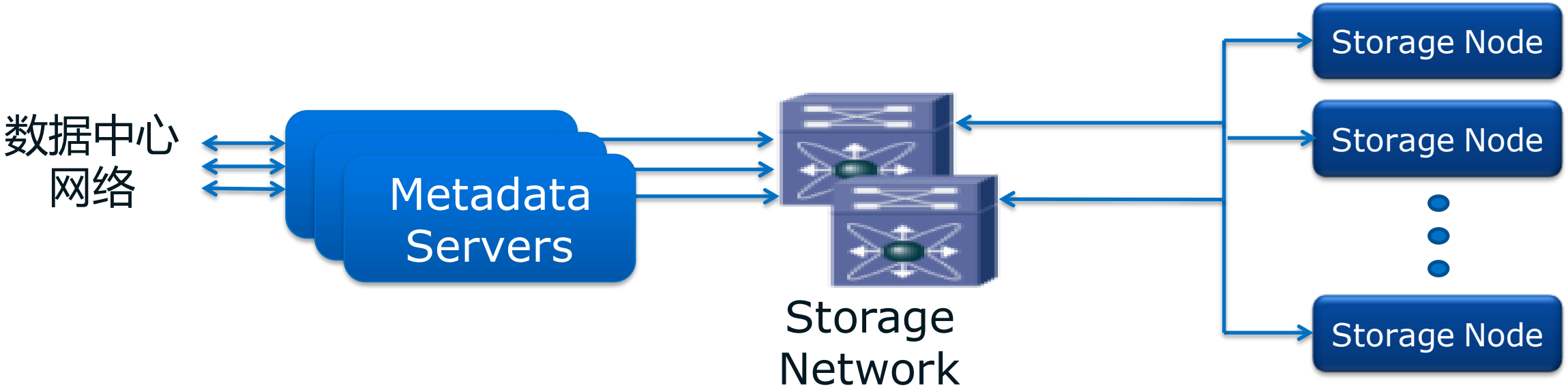
Significant throughput performance increase
for I/O bound, read-intensive workloads[1]



Hadoop* Mixed Workloads

Boosts performance across the Hadoop* Cluster[2]
Solves the primary challenge – I/O bottleneck

对象存储架构



应用接口
- REST
- PUT/GET/DELETE

元数据服务
- Encode/Decode
- Distribution
- Location

存储节点
- Houses data
- Maintains data

高扩展能力的对象存储架构

对象存储架构

数据持久性

Replication

GLUSTER

openstack™

amazon.com web services

Cloudian

HDFS

Erasure Code

AMPLIDATA

Quantum®

ATMOS CLOUD STORAGE PLATFORM

SCALITY

Cleversafe™

Erasure Code 提高存储效能

*文中涉及的其它名称及商标属于各自所有者资产。

intel

关注产业应用，产研相互促进

- 英特尔®中国云计算创新中心
- Intel Hadoop研发团队
- 行业应用

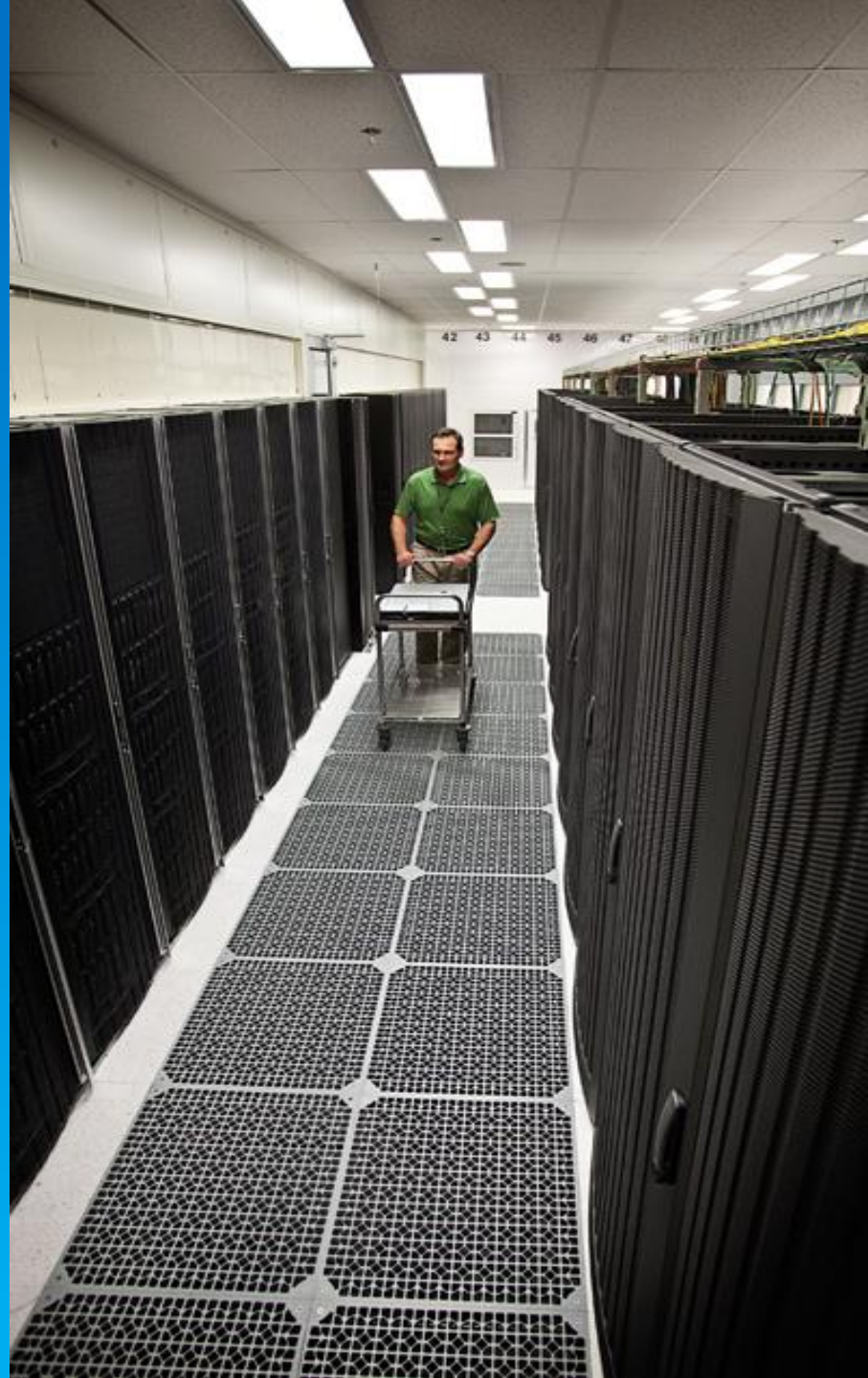
英特尔®中国云计算创新中心

数据中心:

- 11 个机柜，其中网络机柜1个，服务器机柜10个
- 电气容量：6 kW/机柜
- 配电：一路220V AC 市电 + 一路240V DC 直流
- 冷源采用冷冻水系统，末端采用行间送风
- 封闭热走廊

Intel Hadoop研发团队

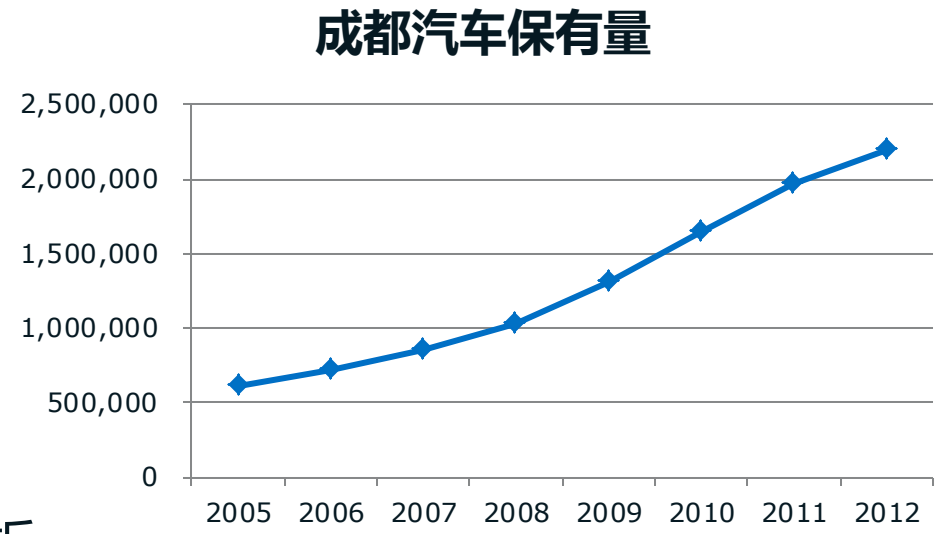
推动产业应用



交通指挥的挑战

——典型中国二线城市

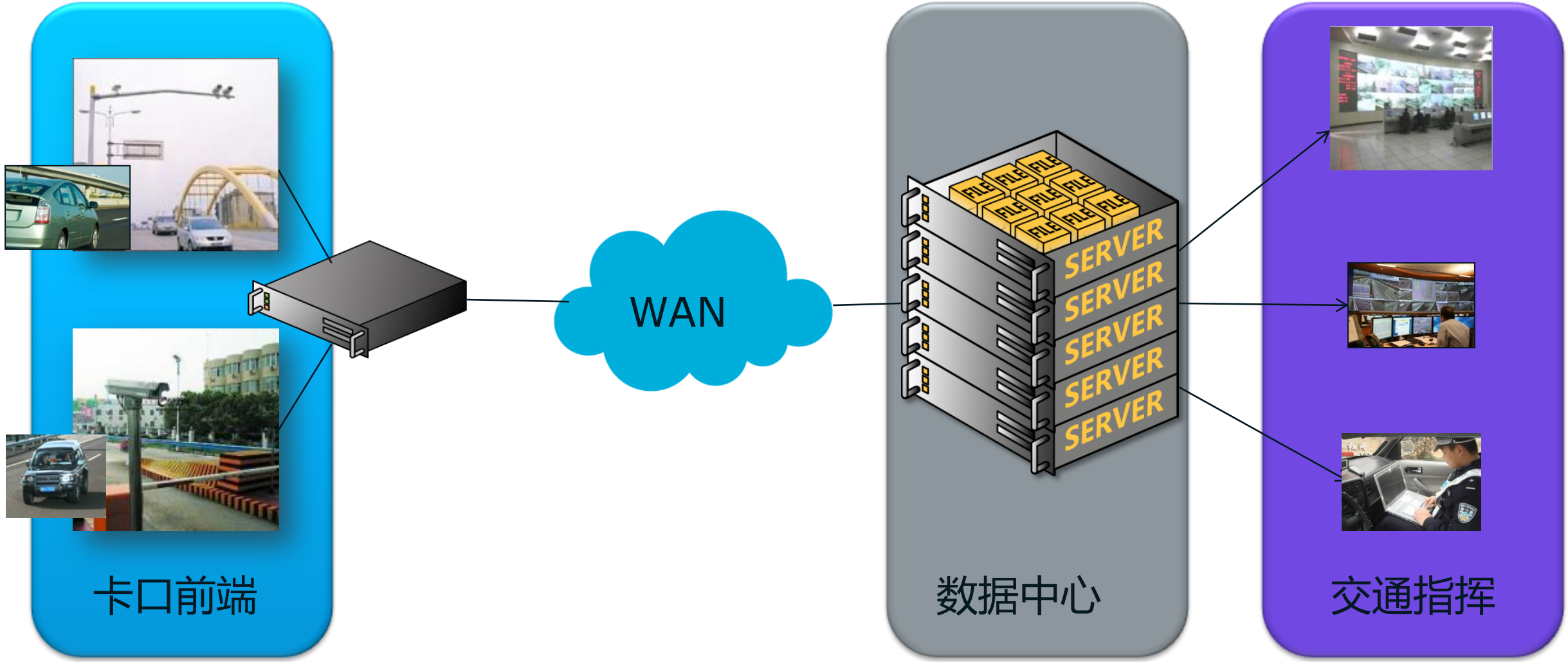
- 机动车的迅速增加



- 复杂数据分析
  - 数据挖掘与预测
  - 突发事件应对
- 公众服务
  - 公众访问高并发
  - 其他系统互连

面对快速增长的数据，如何满足交通指挥要求?

城市交通指挥管理示意



12000个卡口，每年采集超过1000亿条过车的图片和数据信息

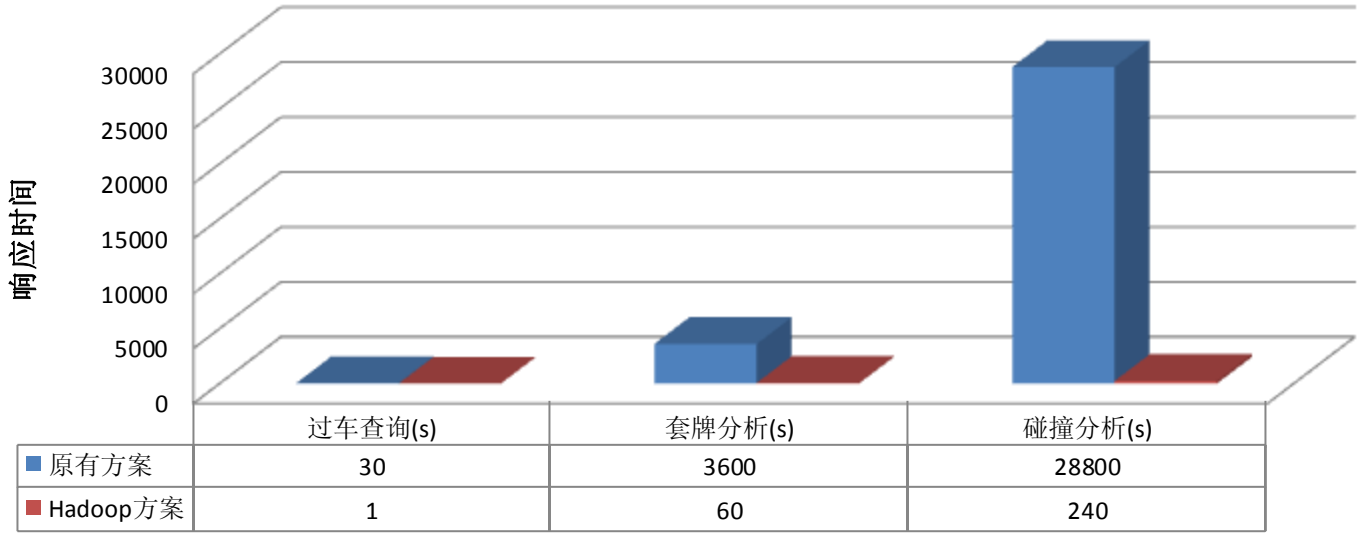
基于Hadoop的新型数据中心方案

原有方案

RDBMS：过车记录
文件系统：过车图片

Hadoop方案

HBase：过车记录
HDFS：过车图片

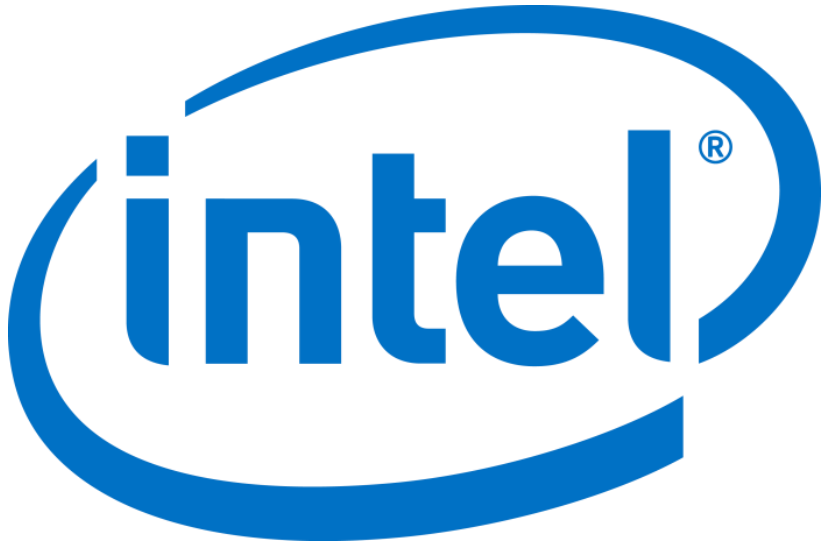


	过车查询(s)	套牌分析(s)	碰撞分析(s)
原有方案	30	3600	28800
Hadoop方案	1	60	240

数据库成本：1PB> 6000万 RMB
数据库维护成本> 1500万RMB

数据库成本：1PB, 1000万RMB
数据库维护成本< 100万RMB

架构灵活，适应业务要求，成本大幅降低


intel