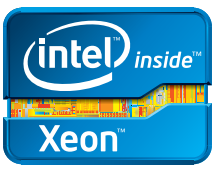
白皮书
英特尔® 万兆位以太网
融合网络适配器
Arista Networks 交换机
Hadoop*

基于万兆位以太网的 Hadoop* 集群

基于经济高效的 10GBASE-T 基础构建均衡系统的通用指南



英特尔™ 至强™

Hadoop* 的应用在处理大数据方面变得日渐流行。主流计算和存储资源的大幅改善使得 Hadoop 集群可以满足大部分企业的需求。但是要提供一个均衡的系统，这些构建模块必须与万兆位以太网（10GbE）配合使用，而不是传统的千兆位以太网（GbE）。基于 10GBASE-T 的基础，并结合采用 Arista 交换机、英特尔® 万兆位以太网融合网络适配器和基于英特尔® 至强® 处理器的服务器，该项研究显示此种构建方法取得了出色的成功。

所有行业的 IT 部门现在都需要处理、存储和分析超大型数据存储，操作 PB 级数据正在变得越来越普遍。与此同时，IDC 最近进行的一项研究显示，这些数据存储正在以 40% 的年复合增长率增长。[1]

由于其中许多数据均为非结构化数据，不适用于预定义模型（或关系型数据库表），处理这些大规模的数据正变得日益复杂。此类数据通常还具有大量文本，例如由社交媒体或搜索功能生成的数据等，传统的关系型数据库则无法满足管理和分析这些数据的需求。

用于处理大量非结构化数据的传统工具具有专有性特点、并且非常昂贵，同时需要专业化的技术去配合，这使得大多数企业难以负担。今天，Hadoop* 集群等常用解决方案为在商用（COTS）服务器中处理大数据提供了卓越的支持。尽管构建 Hadoop 环境可能存在一定的风险，但是这些风险可以被规避。

如果您或您周围的人接到构建 Hadoop 集群的任务，那么无需担心。本白皮书致力于为您提供背景信息与介绍，帮助您轻松开启构建第一步。

减少处理大数据的挑战和成本

企业常常在采集大量数据集方面有大笔支出，但却止步于此、未能采取下一个步骤从信息中提取价值。造成这种中断的原因有二。首先，用于管理和分析非结构化数据的传统工具非常昂贵和复杂，让企业望而生畏。其次，企业不知道应从哪里开始。事实上，在万兆位以太网不断提高成本效益的协助下，强大的 x86 服务器和 Hadoop 可以帮助解决成本和复杂性问题。本白皮书将帮助您制定计划，创建切实可行的 10GBASE-T Hadoop 集群，以此作为成功的基础。

这一方法旨在使初始构建工作尽可能简单、经济且灵活，同时为验证和规划未来投资奠定一个重要基础。Hadoop 是一个卓越的技术选择，可将与大容量数据存储有关的负担转化为可以帮助促进组织取得成功的宝贵资产。目前 Hadoop 通过开源和商业软件包的形式提供。

Hadoop 简介

Hadoop 是一个 Apache 软件框架，能够分析 PB 级非结构化数据，并将其转化为可管理的形式，供应用使用。Hadoop 基于 Google 的 MapReduce 和分布式文件系统构建而成，可部署于通用的商业网络和服务器硬件之上。

目录

***《纽约时报》*的大数据处理**

人类创新的历程源于不断的自我完善，几年前 Derek Gottfrid 在《纽约时报》使用 Hadoop 的示例便证明了这一点。[2] 这份因为“灰色老妇人”的敬称（听起来有些奇怪）而广为人知的报纸决定，将 1851 年至 1980 年间的全部 1,100 万篇公共领域的文章制作成 PDF 文件。完成这项工作需要将每篇文章的多个 TIFF 图片缩放并拼接在一起，通过使用为在 Amazon Simple Storage Service（S3）和 Elastic Compute Cloud（EC2）上运行而构建的代码，Gottfrid 实施了自动化流程。

在合理的时间内处理大量 TIFF 图片数据集（4TB）需要能够在多个设备中并行运行。为了解决这一问题，Gottfrid 构建了他的第一个 Hadoop* 集群，并在此基础之上成功完成了这一冒险举措（用他自己的话来说）。通过在 Amazon 的 EC2 云中使用 100 个 Hadoop 节点，该工作在 24 小时内完成，花费 240 美元（不包括网络带宽成本）。[3]

Hadoop 针对特定任务类别进行了优化，例如大型数据集索引编制和分类、数据挖掘、日志分析和图片处理等。它并非用于不涉及大量数据的实时处理或进程密集型任务。从架构上而言，Hadoop 具有两个主要部分：

- **Hadoop 分布式文件系统（HDFS）**使用一次写入、多次读取模式，能够将数据拆分为多个存储块，分布在多个节点之中，以实现容错和高性能。除了由多个节点构成的大规模汇聚 I/O 之外，约为 128 MB 的大型存储块也进一步提升了性能。与之相比，在 Linux* 实施中更为典型的规模约为 4 KB。
- **MapReduce Engine** 通过其 *Job Tracker* 节点接收来自应用的任务。该节点将工作分解成若干个小型任务，然后指派给 *Task Tracker* 节点。当连接至网络拓扑结构感知交换基础设施时，Job Tracker 节点智能地将工作保存在与其所需数据相邻的位置，性能显著得到提升。

对均衡系统的需求

Hadoop 针对通用硬件进行了特别设计和优化。多年来，服务器持续创新，促使主流系统现在也可以提供大规模处理能力。为了跟上这种能力的发展步伐，必须在专门设计的环境中部署 Hadoop，以达到在计算、存储和网络之间保持均衡的目的。

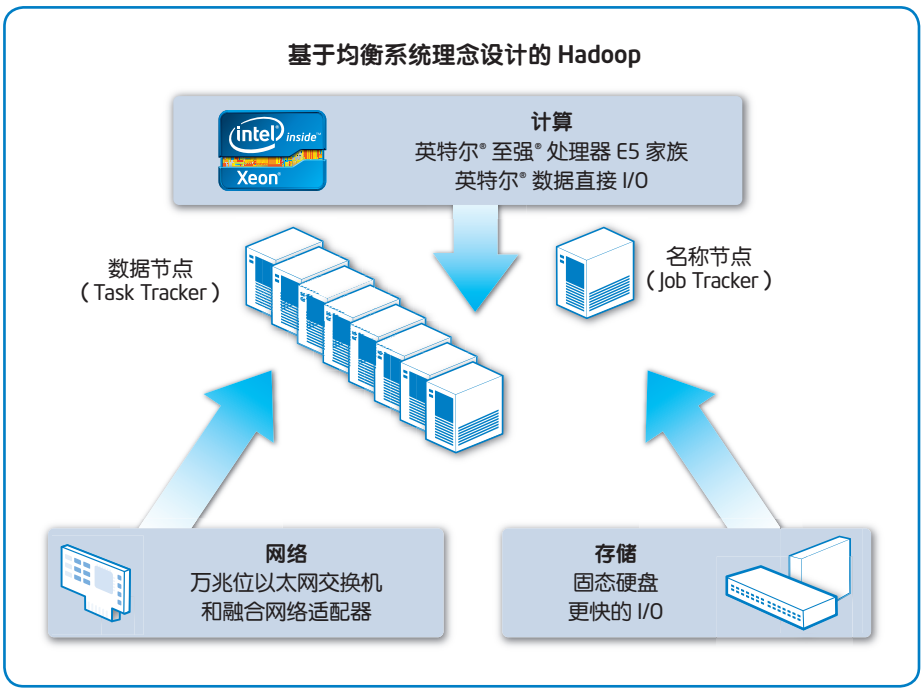


图 1. 均衡问题

图 1 显示了与创建和保持均衡系统有关的关键因素。

- **处理器性能持续攀升。**英特尔® 至强® 处理器 E5-1600/2600 产品家族与上一代产品相比性能提高 80% 。[4,5]
- **存储速度持续上升、延迟速度降低和成本持续下降。**存储 I/O 方面的性能在持续改进，获得固态硬盘（SSD）的性能优势需要的成本也在不断下降。例如，在 2011 年 SSD 存储的每 GB 价格下降至 2.42 美元（2007 年为 40 美元），根据某些人士的预测在 2012 年其价格将下降至 1.00 美元。[6]
- **节点间通信正在加快网络速度。**Hadoop 等分布式服务器和存储架构的大规模 I/O 要求需要高吞吐量。万兆位以太网交换提供了绝佳途径、以经济高效的形式满足了这些需求。

服务器平台的增强不仅限于处理器性能。在 Hadoop 环境中，与英特尔至强处理器 E5 家族架构有关的一个特定因素是英特尔® 数据直接 I/O 技术（英特尔® DDIO）[7]，它对于平台的总体 I/O 性能提升起到了重要作用。英特尔 DDIO 能够智能地将 I/O 数据包直接发送至处理器高速缓存，跳过主系统内存。通过消除不必要的内存步骤，此举可极大地降低延迟，改善总体系统带宽和功效。

在该均衡系统的网络方面，千兆位以太网（GbE）的性能已经严重制约了 Hadoop 的总体性能。例如，使用大型存储块意味着当数据包被丢弃和重发时，系统需要处理大量数据，在千兆位以太网环境中会占用大量网络带宽。万兆位以太网（10GbE）在 Hadoop 集群中实现了较高的网络使用率，为集群带来了出色价值，显示了更高带宽带来的优势。

实现万兆位以太网的全部优势

由无缝组合的构建模块组成的 Hadoop 基础设施包括支持高性能计算和存储资源的万兆位以太网。除了 Hadoop 之外，当前其它技术进步和市场因素也在推动着万兆位以太网的普及。其中最重要的因素当属大规模虚拟化的出现。由于每个物理主机上的大量虚拟机需要极高的网络吞吐率，万兆位以太网成为了一个自然的选择。同样，在万兆位以太网交换网络上整合流量（包括多个千兆位以太网接口、NFS 和 ISCSI 等）同样需要更高的带宽。

随着万兆位以太网在主流环境中的采用不断加快，更多万兆位以太网设备的推出帮助降低了相关网络适配器和交换机的价格。更低的价格促进了进一步的采用，为实现广泛采用和经济高效性创造了一个良性循环的环境。与千兆位以太网相比，万兆位以太网的每 GB 带宽成本现在已大幅下降，成为大多数服务器实施工作的理想选择。

英特尔® 平台 I/O 能力提升

通过在由基于英特尔® 至强® 处理器 E5 家族[5] 和英特尔® 以太网融合网络适配器和控制器的平台组成的全英特尔基础上构建 Hadoop* 服务器，您将可以获得整合工程（Cohesive Engineering）带来的协同优势。

- **英特尔® 集成 I/O** 将 I/O 子系统移至处理器，将服务器平台延迟降低多至 30%[8]，同时将带宽提高多至 2 倍[9]，并支持 PCI Express* 3.0 规范接口。
- **英特尔® 数据直接 I/O 技术（英特尔® DDIO）**支持将 I/O 流量直接发送至处理器高速缓存（而不是通过内存），可有效降低延迟和功耗，让内存能够更长时间保持在低功耗状态。
- **英特尔® 以太网**产品充分利用了英特尔三十多年来在以太网技术领域的领先地位，采用了可信、可靠的硬件和软件，显示出产品的创新特性，并带来卓越性能。

英特尔® 以太网控制器 X540 是英特尔最新的万兆位以太网控制器，也是业界首个完全集成的 10GBASE-T 控制器，在一个芯片中结合了 MAC 和 PHY。这款控制器旨在帮助实现低成本、低功耗的 10GBASE-T 板载局域网（LOM）和融合网络适配器（CNA）产品。英特尔以太网控制器 X540 包括高级 I/O 虚拟化（虚拟机设备队列，支持 PCI-SIG 单根 I/O 虚拟化）和以太网存储（包括 NFS、ISCSI 和以太网光纤通道），并支持在新的英特尔至强处理器 E5 家族上实现 I/O 的性能提升。

为何采用 10GBASE-T?

Arista 7050T 交换机和英特尔® 以太网控制器 X540 等 10GBASE-T 产品的推出正在推动业界向万兆位以太网（10GbE）迁移。这些产品可带来以下优势：

- **较低的部署成本**：支持通过集成的万兆位以太网产品、更低的电缆成本和灵活的电缆长度将万兆位以太网技术带入更广泛的市场。
- **轻松迈进万兆位以太网环境**：向后兼容现有千兆位以太网的交换和布线基础设施。
- **较低的复杂性**：采用行业标准双绞线铜缆，并支持高级 I/O 虚拟化和统一网络功能。

10GBASE-T 交换端口的以太网交换创新和产品成熟度在帮助降低 Hadoop 集群部署成本方面扮演着重要角色，尤其是与光纤电缆和收发器的成本相比较时。此外，10GBASE-T 交换还与许多现有 5 类和 6 类电缆架和电缆槽兼容，无需在采用过程中进行完全重建。

除了带宽容量以外，万兆位以太网交换相比千兆位以太网交换还多项功能得到增强，包括更低的交换延迟，更好的流量均衡和故障切换能力，以及能够随着集群规模的增加进行出色扩展等。交换机的软件功能也可以让 Hadoop 部署直接从中获益，其中包括支持 Hadoop Job Tracker 节点优化任务指派的机架感知功能等（如本白皮书的前面内容所述）。

万兆位以太网创新的另一个实例通过其广泛的采用实现，10GBASE-T 连接能力现在通过板载局域网（LOM）设计集成至主流服务器中。通过利用这种 LOM 方法，最终客户可以充分受益于万兆位以太网技术，并无需在核心服务器平台之外支付额外费用。这些解决方案均采用了英特尔® 以太网控制器 X540。后者是业界首个完全集成（MAC 和 PHY）的 10GBASE-T 控制器，专为 LOM 和融合网络适配器而设计。

来自 ARISTA NETWORKS 的高带宽、低延迟、拓扑结构感知交换机

Arista 最近宣布开始发运在本白皮书的概念验证中使用的 Arista 7050T 交换机。该款交换机是一款 1RU 线速、低功耗 10GBASE-T 交换平台，可以用作 Hadoop* 集群中的架顶交换机，或者在存在多个机架时用作主干交换机（处理第 3 层交换）。

用作主干交换机时，7050T 提供的第 3 层交换（路由）优势可以为 Hadoop 的第 3 层文件系统提供重要补充，能够当在机架之间移动流量时实现更好的可扩展性。因为它具有拓扑结构感知特性，Hadoop 可充分利用 7050T 提供的拓扑结构智能优势，将计算任务置于距离相关存储最近的位置，以提高效率。7050T 交换机还支持深度缓冲，当 Hadoop 出现链路超额利用时，无需丢弃和重发数据包即可显著提高性能。

7050T 交换平台提供了深度数据包缓冲和用于机架感知集成的开放式接口，使其成为从小型 Hadoop 基础设施扩展至中型规模的理想设备。7050T 交换平台与英特尔® 以太网控制器 X540 完全兼容。

我们的系统工程师 Stefanie 正在执行与集群有关的工作。

装配 Hadoop 集群

除了展示性能和可扩展性以外，本概念验证（POC）还介绍了一个在万兆位以太网中构建 Hadoop 集群相对简单的方法。本白皮书介绍了如何一次性取得成功，建立一个具有出色可扩展性的可靠 Hadoop 环境。当然，当集群建立并运行以后还有充足的空间可以进行微调。

概念验证 Hadoop 环境方案

我们的 Hadoop 集群设计演示了万兆位以太网与千兆位以太网相比所具有的价值。它具有中等规模（10 个节点），基于常用组件，如表 1 所示。例如，我们使用传统的硬盘驱动器而不是 SSD，因为传统的磁盘驱动器更常用。同时我们使用了单一客户端设备向集群提交工作任务。

表 1. Hadoop* 测试集群中的组件

硬件配置	
网络交换机	**Arista 7050T 48 端口 1 Gbps/10 Gbps 交换机**
计算节点： ▪ 一个主节点（名称节点，Job Tracker） ▪ 九个从属节点（数据节点，TaskTracker）	10 台 Super Micro Super Server* 1026T-URF 服务器（1U，双插槽）： ▪ 英特尔® 至强® 处理器 5690 ▪ 48 GB 内存 ▪ 五个 700 GB SATA 硬盘驱动器 @ 7200 RPM
网络接口卡	每台服务器具有一个英特尔® 万兆位以太网融合网络适配器（10GBASE-T）和一个英特尔® 千兆以太网服务器适配器（1000BASE-T）。
软件配置	
操作系统	Cent OS 6.2
Hadoop* 版本	Cloudera distribution CDH3（0.20.2）
Java* 开发套件（JDK）	Oracle JDK 1.7.0

在该配置中，我们创造了术语“主节点”来表示同时承载名称节点和 Job Tracker 的主机。在生产部署中，需要为名称节点提供冗余配置，否则存在潜在的单点故障。名称节点冗余可以通过在两个单独的设备中部署，或通过将两个服务器适配器使用双主用配置与交换机连接来实现。注意，在我们的概念验证中，并未使用上述两种方法提供冗余配置。在任何情况下，其他节点都采用故障切换设计，因此无需特定的冗余配置。

为了实现简单化设计目标，我们未对默认配置进行许多更改，仅做了一些必要调整。我们实施了压缩以避免存储瓶颈，同时此举也能够更好地凸显出网络子系统的性能，以用于测试目的和比较万兆位以太网与千兆位以太网技术。集群使用默认 128 MB 存储块和默认 HDFS 副本系数 3，这表示名称节点将输入的数据复制到三个数据节点。

Alternate Cluster Config 是一个简单的文件，允许您通过更改参数对 Hadoop 环境进行更改。这一方法支持对不同产品进行试验，之后可以轻松返回默认配置，从而为探索 Hadoop 的各种功能提供了一种安全且有效的方式。

评估结果及未来规划

构建了概念验证集群后，我们接下来开始执行测试，以便评估万兆位以太网和千兆位以太网之间的性能差异。我们测试的第一个使用案例是在互联网上释放一个假定的 spider 以便采集大量非结构化数据，例如最终客户对产品的评论。当数据被采集并保存在客户机上之后，下一个步骤被称为 PUT 操作。该操作包括将来自客户机上的数据导入至我们的 Hadoop 集群主节点，然后主节点将数据拆分为存储块，并将每个存储块复制到三个独立的数据节点。

图 2 显示了使用千兆位以太网和万兆位以太网时具有各种数据集规模的 PUT 操作的测试结果。注意这些结果主要关注网络性能，不包括通过 MapReduce 操作处理数据花费的时间。使用万兆位以太网时，完成 PUT 操作的时间只有使用千兆位以太网时执行相同操作所需时间的五分之一。另外还需注意，对于万兆位以太网和千兆位以太网，完成操作所需的总时间根据数据集的规模呈现线性扩展趋势，因此与千兆位以太网相比，使用万兆位以太网时所需时间约减少了 80%。鉴于此，对于大小在数 TB 左右的数据集，其差异可能导致多个小时的等待时间。

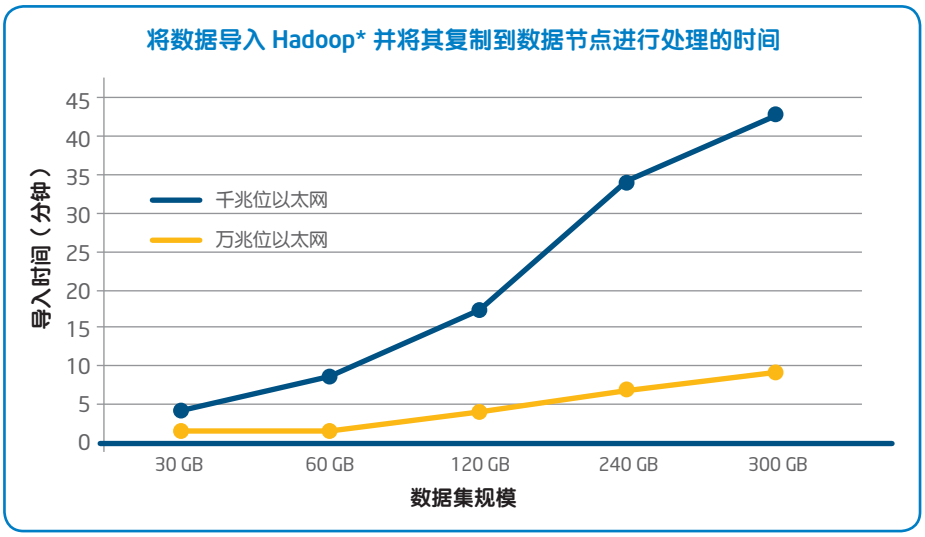


图 2. 与使用千兆位以太网相比，使用万兆位以太网完成 Hadoop* PUT 操作的时间缩短了 80%。

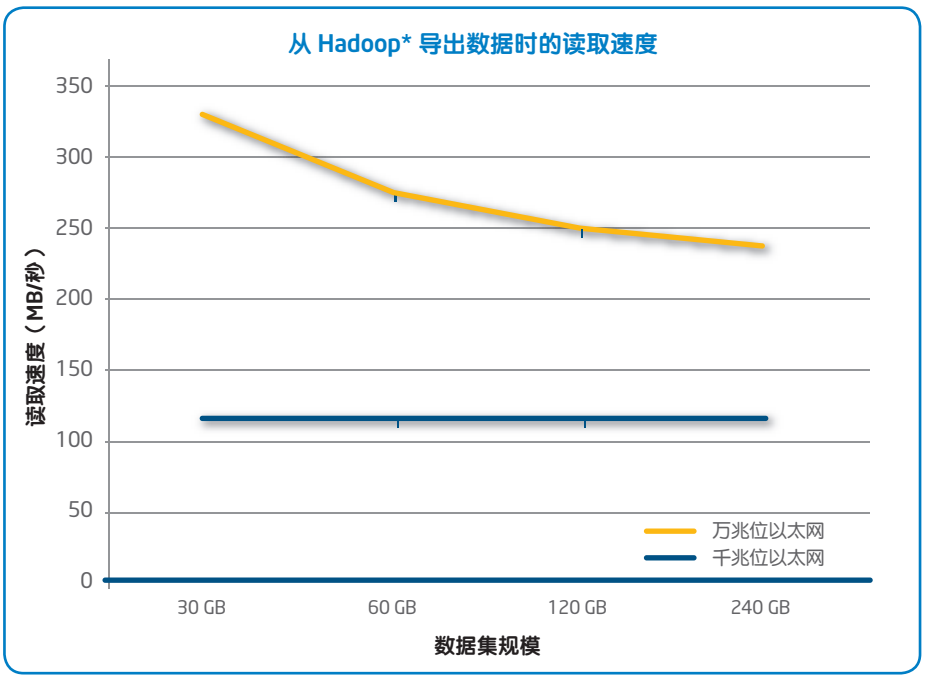


图 3. Hadoop* GET 操作显示万兆位以太网与千兆以太网相比具有明显优势，后者受到存储子系统瓶颈的限制。

与 PUT 操作伴随进行的任务是 GET 操作，即将数据从 Hadoop 集群中提取处理的过程。例如，在上文的侧栏中提到的《*纽约时报*》的示例中，一旦 TIFF 文件拼接在一起创建了 PDF，您将需要再将数据提取出来。图 3 所示为 GET 操作的测试结果。

在执行 GET 操作时，我们再一次清晰地看到在 Hadoop 集群中采用万兆位以太网的优势。受到存储子系统的限制，当数据规模增大时该优势也会稍微下降（由于“本地磁盘空间不足”错误，数据未包括 300 GB 数据集）。我们计划在未来的工作中探索存储技术的发展对该瓶颈产生的积极影响，例如各种类型的非易失性存储器和涉及将非易失性存储器放置在大容量磁盘前端的分层战略。随着这些高级存储技术不断被开发出来，我们预计万兆位以太网将能够为 GET 操作带来更显著的优势。

一旦您按照本文所述构建了简单的测试平台，我们预计您将会立即在可扩展性和处理工作负载方面，享受到万兆位以太网可为 Hadoop 集群带来的优势。同时您也将可以快速掌握各种方法来优化集群，以实现最佳成果。在本问中我们将不提供相关的优化信息。您可以阅读英特尔白皮书“优化 Hadoop 部署”[10]，了解相关信息，尽享 Hadoop 集群之旅！

总结

构建简单的 Hadoop 集群属于日常的 IT 工作范围之内。当然，围绕着对系统环境进行微调有许多详细的信息，但是正如本白皮书所述，基础搭建部分并不复杂。构建一个在计算、存储和网络资源之间保持均衡的环境是项目成功的基础。万兆位以太网在这一方面具有明显的优势，在将数据导入到集群中时，万兆位以太网能够比千兆位以太网节省 80% 的时间。

基于英特尔® 至强处理器的服务器、英特尔® 万兆位以太网融合网络适配器和 Arista 交换机是构建经济高效的 Hadoop 集群的主要硬件元素。具有基于英特尔以太网控制器 X540 的 LOM 连接的服务器和 Arista 的 10GBASE-T 交换机等新型 10GBASE-T 产品正在不断降低万兆位以太网的成本。最新的英特尔处理器专为实现万兆位以太网及更高的速度而设计。Hadoop 的线性可扩展性允许您随着数据集规模的加大扩展环境，将数据集从一项负担转变为宝贵资产。

如欲了解更多信息，请访问
www.intel.com/go/ethernet

www.aristanetworks.com

解决方案提供商：

intel® ARISTA

[1] 数字来源于 IDC 在 2012 年的报告。

[2] http://open.blogs.nytimes.com/2007/11/01/self-service-prorated-super-computing-fun/

[3] www.slideshare.net/dgottfrid/hadoop-world-oct-2009

[4] 使用 SPECint*_rate_base2006、SPECfp*_rate_base2006、STREAM*_MP Triad 和 Linpack* 的性能指标评测结果的几何平均数进行的性能比较。166.75 的基准几何平均数得分在上一代双路英特尔® 至强® 处理器 X5690 平台上得出，采用了在 www.spec.org 上公布的最佳 SPECrate* 得分和截止到 2011 年 12 月 5 日英特尔内部在 STREAM_MP Triad 和 Linpack 上获得的最佳测量结果。306.74 的最新几何平均数得分基于英特尔的内部测量估计值，使用具有两颗英特尔® 至强® E5-2690 处理器的英特尔® Rose City 平台，启用睿频技术技术、EIST 和超线程技术，128 GB 内存，Red Hat Enterprise Linux* Server 6.1 beta for x86_6，英特尔® 编译器 12.1，SPECfp_rate_base2006 中禁用 THP，SPECint_rate_base2006 中启用 THP。

[5] 请注意，该白皮书中报告的测试使用英特尔® 至强® 5600 系列处理器执行，它是英特尔® 至强® E5 处理器产品家族的前代产品。

[6] http://royal.pingdom.com/2011/12/19/would-you-pay-7260-for-a-3-tb-drive-charting-hdd-and-ssd-prices-over-time/

[7] www.intel.com/content/www/us/en/io/direct-data-i-o.html

[8] 英特尔测量的在空闲状态下 I/O 设备读取本地系统内存的平均时间。英特尔® 至强® E5-2600 处理器产品家族（230 纳秒）与英特尔® 至强® 5500 系列处理器（340 纳秒）之间的改进比较。基准配置：具有两颗英特尔® 至强® E5520 处理器（2.26 GHz，4C）的 Green City 系统，12 GB 内存 @ 1333，禁用 C 状态，禁用睿频加速技术，禁用 SMT，Rubicon* PCIe* 2.0 x8。新配置：采用两颗英特尔® 至强® 处理器 E5-2665（C0 步进、2.4GHz、8C）、32GB 内存 @1600 MHz 的 Meridian 系统，启用 C 状态和睿频加速技术。测量结果基于使用英特尔内部 Rubicon（PCIe* 2.0）和 Florin（PCIe* 3.0）测试卡的 LeCroy* PCIe* 协议分析器，运行环境为 Windows* 2008 R2（SP1）。

[9] PCIe* 3.0 规范接口中的 8 GT/秒和 128b/130b 编码特性能够将互联带宽提升为 PCIe* 2.0 规范接口的两倍。资料来源：www.pcisig.com/news_room/November_18_2010_Press_Release

[10] software.intel.com/file/31124

0412/ME/MESH/PDF ♻ 请注意环保 327302-001CN

intel®