



使用 InfiniBand+RDMA 在 GPFS 上推动 IBM BigInsights 性能提升

执行摘要.....	1
背景.....	2
文件系统体系架构.....	2
网络体系架构.....	3
IBM BigInsights.....	5
结果.....	8
结论.....	9

执行摘要

本研究的目的是评估 IBM 通用并行文件系统 (GPFS) 作为 IBM BigInsights Hadoop 部署的文件系统的功能，并使用 GPFS 测试 Mellanox 远程直接内存访问 (RDMA) 用于 BigInsights 应用程序的性能优势。为了提供比较基础，测试运行时比较了 GPFS 与 Apache Hadoop 分布式文件系统 (HDFS) 的使用情况。基准测试结果表明，在分析基准测试（Terasort 基准测试）中，GPFS 使应用程序性能比 HDFS 提升了 35%，而在使用增强 TestDFSIO 基准测试时，写入测试性能提升了 35%，读取测试性能提升了 50%。

本文提供了有关研究过程中的测试体系架构、方法、结果和结论的详细信息。

背景

“大数据”已经成为当代数据存储和分析讨论中的热门词汇。“大数据”可以描述为不容易适应传统关系数据库管理系统 (RDBMS) 的数据，因为数据量很大，或者由于它没有良好定义的行和列结构或明确定义的数据类型。大型 Web 2.0 公司最先遇到这些庞大的数据集，他们的开创性工作造就了 Apache Hadoop 框架。业内其他领域正在进行剧烈的转型，以努力应对不断增长且日趋复杂的数据。公司正在寻找在最短时间内从庞大的数据集中提取可操作情报的方法。这意味着系统架构师需要弄清楚如何应对捕获、策划、管理和处理这些数据的挑战。IBM 通过基于底层 Hadoop 体系架构构建的创新且强大的 IBM BigInsights 产品来解决大数据的挑战。

Hadoop 框架包含两个核心组块。其一是为大量数据提供可扩展存储机制的 Hadoop 分布式文件系统 (HDFS)。其二是提供名为 MapReduce 的分析组件的框架，该组件组织挖掘存储在文件系统上的数据。Hadoop 体系架构正在经历快速采用阶段，给客户带来了新一轮的挑战。本文档中介绍的体系架构可以提供解决方案，解决连续数据引入或写入繁重工作负载问题。



图 1：IBM 大数据平台

文件系统体系架构

虽然 HDFS 是 Apache Hadoop 软件包的一个组件，但它有多项缺点，可通过使用其他文件系统替换 HDFS 来克服这些缺点。IBM 的 BigInsights 也采用了这种方法，即 IBM 通用并行文件系统 (GPFS)。

通用并行文件系统

GPFS 是 1998 年首次发布的 IBM 产品。GPFS 是高性能企业级分布式文件系统。多年来，它已经发展到支持各种工作负载，并且可以扩展到数千个节点。GPFS 在许多企业客户生产环境中部署和使用，以支持机器关键型应用程序。

网络体系架构

GPFS 应用：

- 数字媒体
- 工程设计
- 商业智能
- 财务分析
- 地震数据处理
- 地理信息系统
- 可扩展文件服务

GPFS 功能：

- 无缝容量扩展，可处理数字信息的急剧增长，并通过企业范围的跨部门信息共享来提高效率
- 高可靠性和可用性可以消除生产中断并提供无中断的维护和容量升级
- 性能可满足最苛刻的应用程序
- 策略驱动的自动化工具可简化信息生命周期管理 (ILM)
- 可扩展的管理和监控基础架构，可简化文件系统管理
- 具有成本效益的灾难恢复和业务连续性
- 符合 POSIX（可移植操作系统接口）标准

Hadoop 分布式文件系统

HDFS 是作为 Apache Hadoop 包的一部分提供的固有分布式文件系统。HDFS 数据分布在通过网络连接在一起的多台计算机的本地磁盘上。包含数据的计算机被称为数据节点。数据节点通常是同质的，每个节点具有多个硬盘。写入 HDFS 的每个文件被分割成较小的块，分布在不同的数据节点上。默认情况下，HDFS 配置为 3 路复制，以确保数据可靠性。

HDFS 功能

- 安装简单
- 可在文件创建过程中更改块大小和复制
- 固有的可靠性
- 专为 Hadoop 工作负载一次写入多次读取而设计

InfiniBand 和 RDMA

InfiniBand 提供了消息传递服务，可供应用程序直接访问，而不需要操作系统。与基于 TCP/IP 字节流的传输相比，InfiniBand 不需要在应用程序和网络之间进行复杂交换。直接访问意味着应用程序不需要依赖操作系统来传递消息。这种“以应用程序为中心”的计算方法是 InfiniBand 与 TCP/IP 网络的关键区别。

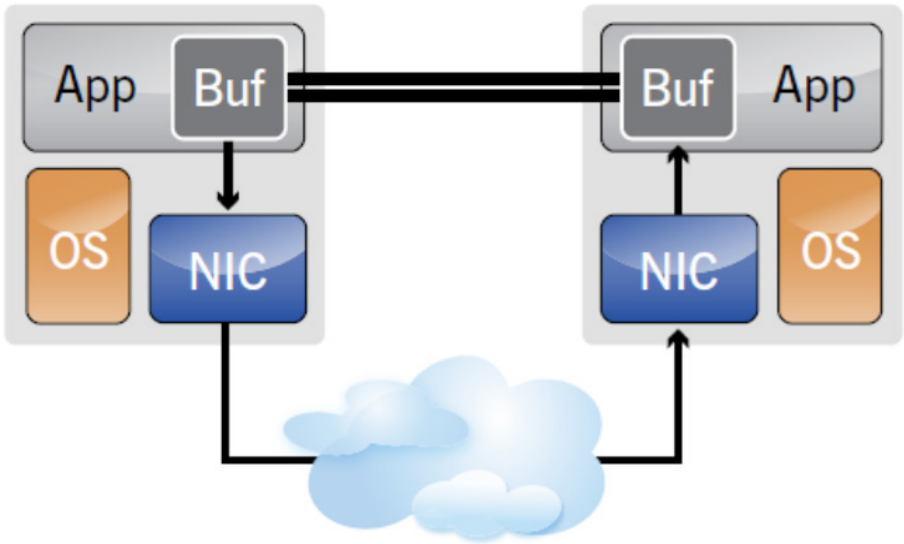


图 2：InfiniBand 架构中的消息传递体系架构

InfiniBand FDR 当前速度为 56Gb/s，应用程序间的通信延迟小于 1 微秒。

InfiniBand 的关键功能是远程直接内存访问 (RDMA)。RDMA 提供从一台计算机内存到另一台计算机内存的应用程序级直接访问，而无需任意计算机上的操作系统的任何服务。这实现了高吞吐量、低延迟和低 CPU 传输开销的节点间通信，这对于大规模并行计算集群至关重要。如图 3 中的 InfiniBand 体系架构所示，软件传输接口就位于传输层之上。软件传输接口定义了应用程序充分利用 RDMA 传输服务所需采取的方法和机制。

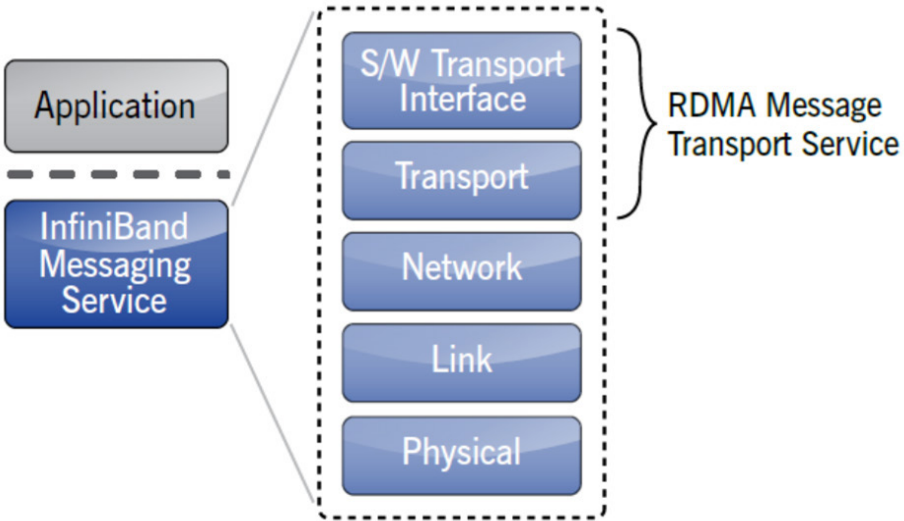


图 3：InfiniBand 体系架构

IBM BigInsights

SX6036 交换机系统概述

SX6036 交换机系统通过为高性能计算和企业数据中心提供 4.032Tb/s 的无阻塞带宽以及 200 纳秒的端口间延迟，以 1RU 外形规格提供性能最高的架构解决方案。SX6036 使用 Mellanox 最新的 SwitchX®-2 InfiniBand 交换设备构建，每端口提供高达 56Gb/s 的全双向带宽。该独立交换机最适合用于架顶式叶片连接或者建立小型至中型集群。它旨在通过有保证带宽和精细服务质量 (QoS) 的组合来承载融合的 LAN 和 SAN 流量。

SX6036 具有支持 InfiniBand 和以太网连接的 Virtual Protocol Interconnect (VPI)，为 PCI Express Gen3 服务器提供性能最高和最灵活的互连解决方案。VPI 用一种硬件设计来服务于多种架构，从而简化系统开发。VPI 通过允许一个平台在同一机箱上运行 InfiniBand 和以太网来简化当今的网络。

具有 Virtual Protocol Interconnect® 的 ConnectX®-3 Pro 双端口适配器概述

ConnectX-3 Pro 适配器卡具有支持 InfiniBand 和以太网连接的 Virtual Protocol Interconnect (VPI)，为企业数据中心、高性能计算和嵌入式环境中使用的 PCI Express Gen3 服务器提供性能最高和最灵活的互连解决方案。集群数据库、并行处理、交易服务和高性能嵌入式 I/O 应用程序会实现显著的性能改进，从而缩短完成时间并降低每操作成本。

IBM BigInsights V2.1 版本引入了对 GPFS 文件放置优化器 (FPO) 的支持。GPFS FPO 是一组功能，允许 GPFS 在没有共享磁盘的集群上支持映射归约应用程序。

GPFS-FPO 优势

- 局部感知能力，因此可以在包含数据的节点上调度计算作业
- 区块允许大块和小块在同一文件系统中共存，以充分利用数据的局部性
- 写入亲和性允许应用程序指定不同节点上的文件布局，以最大限度地提高写入和读取带宽
- 分布式恢复可最大限度地降低故障对进行中计算的影响

下表显示了 GPFS 相对于 HDFS 的优势，从而正确地考虑性能预期。

功能	HDFS	GPFS
不同文件大小处理	仅支持非常大的块。	支持广泛的文件大小，在整个文件大小范围中具有稳定的性能。
是否符合 POSIX 标准	否	是
管理和高可用性	在 2.0 版的临时管理工具之前，不支持高可用性元数据	完善的高可用性支持、标准接口和管理自动化工具

改进 Hadoop 的弹性和性能

如下图 4 所示，在使用 HDFS 时，由一个称为 NameNode 的主节点管理元数据。NameNode 发生故障可能意味着数据完全丢失。然而，像 IBM BigInsights 这样的体系架构提供了高可用性 (HA)，以消除此单点故障 (SPOF)。具体来说，IBM BigInsights 包括对 NameNode HA（自 BigInsights V2.1 版本起实现）的支持。一些实现使用包含备用 NameNode 的网络文件系统 (NFS)，它不提供完整的 HA，但在 NameNode 发生故障的情况下，它可提供相当快速的手动恢复选项。

Hadoop Installation with HDFS (Hadoop Distributed File System)

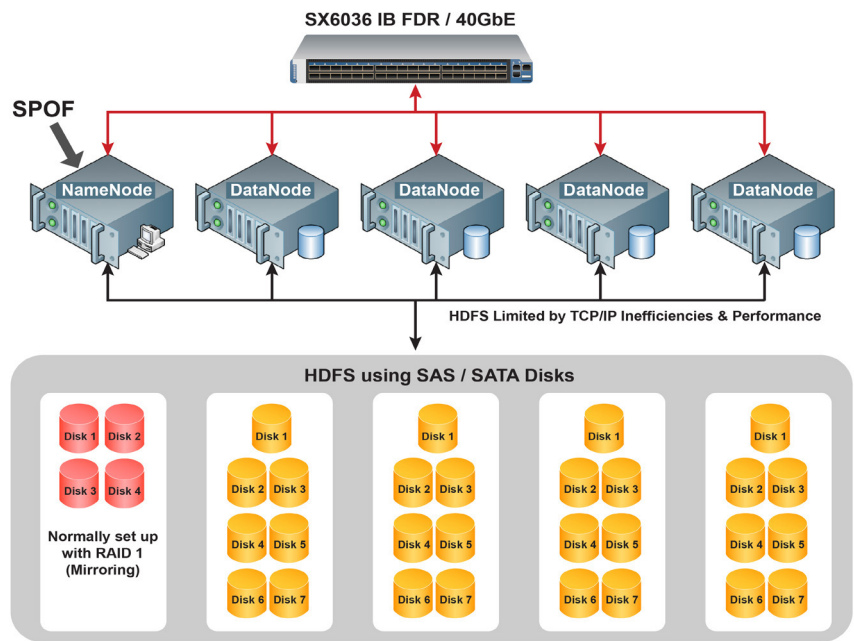


图 4：Hadoop HDFS 体系架构

截至撰写本文时，HDFS 1.2 不支持分层存储（混合不同大小和类型的存储设备）。这意味着文件系统会同样地对待所有存储，并且无法根据所使用的存储类型来优化整体应用程序性能。例如，Hadoop 节点中最慢的组件是旋转磁盘，它为每个设备提供大约 100-200MB/s 的最大吞吐量。在连续数据引入期间，这些磁盘成为瓶颈，并且在存储 PB 数据时使用快速 SSD 可能不具有成本效益。

GPFS 提供了应对这些挑战的解决方案。图 5 显示了当使用 GPFS 进行 Hadoop 映射归约时，一个 NameNode 丢失的精确概念。每个节点对元数据具有相同的访问权限，可以使用与数据相同的方式复制元数据以实现可靠性。而且，GPFS 提供了混合使用各种存储设备的灵活性。现在可以让 SSD 位于第一层，快速 10K RPM SAS 存储位于第二层，7.2K RPM SATA 驱动器位于第三层。

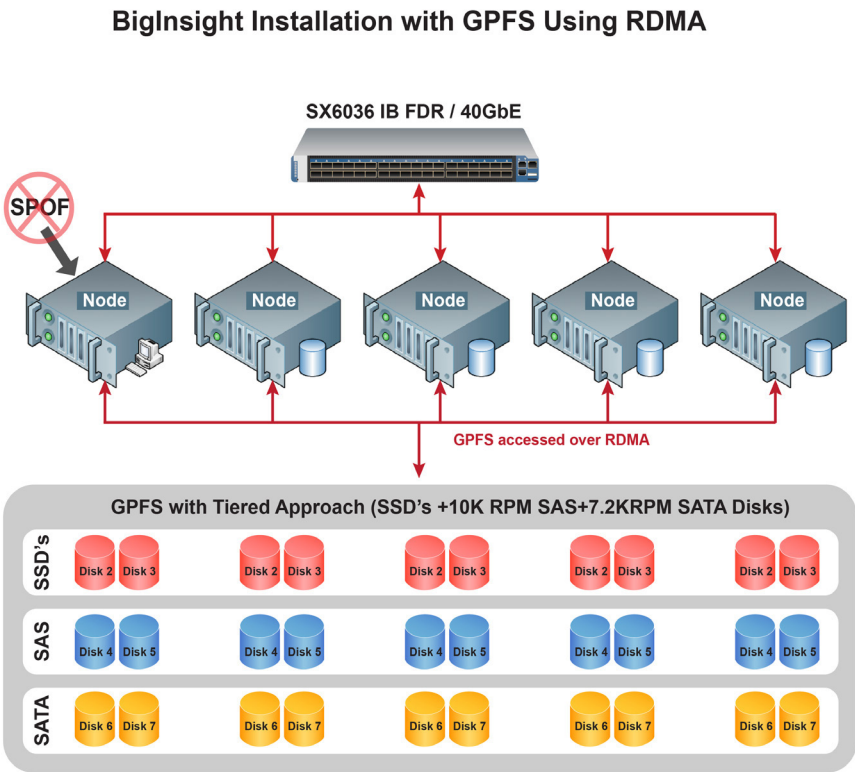


图 5：Hadoop GPFS 体系架构

当用于映射归约 (MapReduce) 工作负载时，GPFS 与 HDFS 相比具有许多优势：

- 消除 NameNode 单点故障，而不需要昂贵的高可用性设计
- 提供分层存储，以便将数据放置在合适类型的存储（SSD、15k SAS、7200 RPM NL/SAS 或任何其他块设备）中
- 使用原生 InfiniBand RDMA，为您的应用程序提供更高的吞吐量、更低的延迟和更多的 CPU 周期

使用 IBM BigInsights 与 GPFS 并利用 SSD、SAS 和 SATA 驱动器功能，让映射归约 (MapReduce) 工作负载获得最佳性价比。

基准测试设置

使用两个运行 IBM BigInsights 且大小相同的五节点集群进行比较，一个集群采用 HDFS，另一个集群采用 GPFS。两个集群都将文件系统元数据存储存储在 SSD 中。目的是比较 HDFS 与 GPFS 上的 Hadoop 映射归约性能。

两个集群使用相同的底层 Infiniband 架构。在进行此测试时，HDFS 不支持 RDMA，因此使用基于 Infiniband 的 TCP/IP (IPoIB) 进行节点间通信。GPFS 集群设置使用文件系统的 RDMA 功能。

部署的完整技术细节和细微差别将在单独的技术文档中公布。

结果

TeraSort 基准测试结果

TeraSort 是著名的 Hadoop 基准。TeraSort 使用映射归约作业以尽可能快的速度对 1TB 数据进行排序。TeraSort 对 Hadoop 的文件系统和映射归约层都会造成压力。图 6 显示了 1TB TeraSort 结果。该测试在 GPFS 上的运行速度比在 HDFS 上快 35%。

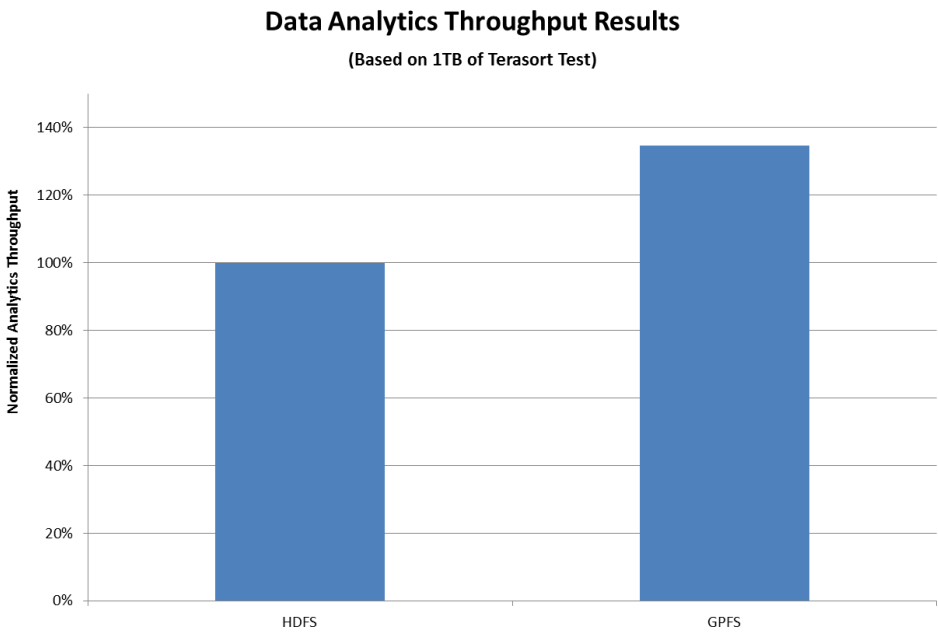


图 6：数据分析性能结果比较，数字越高表示结果越好

HiBench 增强 DFSIO 结果

HiBench 是包含多种 Hadoop 测试的基准套件。来自 HiBench 套件的增强型 DFSIO 测试文件系统连续读写工作负载的性能，以测量文件系统的最大吞吐量。图 7 显示了使用 E-DFSIO 测试的写入吞吐量性能，与 TeraSort 测试中相同，使用 GPFS 的集群提供的吞吐量比 HDFS 集群高 35%。图 8 显示了使用 E-DFSIO 测试的读取吞吐量性能。在此测试中，GPFS 比 HDFS 快 50%。

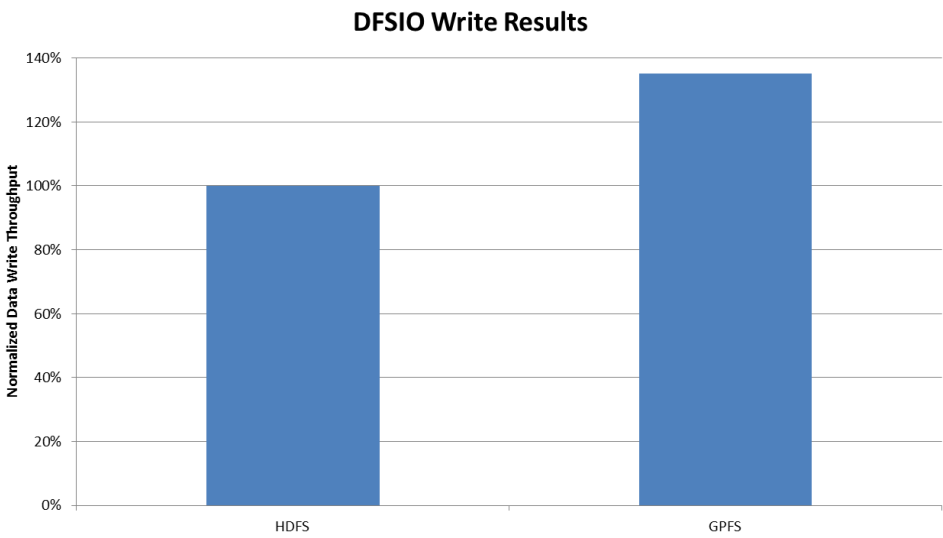


图 7：写入性能，数字越高表示结果越好

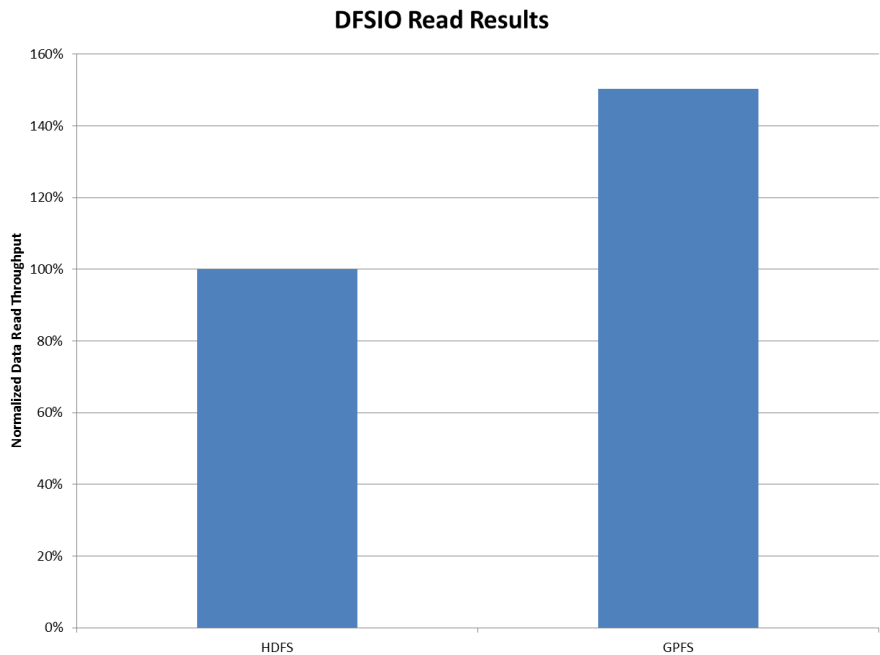


图 8：读取性能，数字越高表示结果越好

结论

E-DFSIO 测试表明，在所使用的基准系统上，使用 GPFS 相对于 HDFS 具有明显的 IO 性能优势：写入性能提升 35%，读取性能提升 50%。此项目的目的是测量利用 InfiniBand RDMA 的 GPFS 在处理连续数据引入（写入）工作负载方面的优势。此外，GPFS 在分析部分（基于 Terasort 基准测试）表现得更好，35% 的提升证明了数据检索、存储和数据分析的综合工作量优势。

在基准系统中，GPFS 表现出的性能比 HDFS 更好。测得的性能提升以及 GPFS 符合 POSIX 标准、包含丰富的管理工具集，以及具有良好的可靠性记录，这一切使其成为映射归约工作负载极具吸引力的文件系统选择。由 Mellanox 互连、IBM Biginsights 的企业级功能和 IBM GPFS 提供的基础架构满足了大数据应用程序的高性能需求。

特别声明

本文档中包含的信息按原样发布，尚未提交以进行任何正式的 IBM 测试。使用此信息或实施其中任何技术由客户承担责任，并且取决于客户评估及将其集成到客户操作环境中的能力。虽然 IBM 可能审查了每个项目在特定情况下的准确性，但不能保证在其他地方获得相同或相似的结果。尝试调整这些技术以用于自己环境的客户自行承担风险。

本文档中包含的任何性能数据均是在各种受控实验室环境中确定的，仅供参考。客户不应针对自己的环境调整这些性能数据以作为系统性能标准。在其他操作环境中获得的结果可能有显著差异。本文档的用户应该验证数据是否适合其特定环境。

了解更多信息

http://www.mellanox.com/related-docs/prod_ib_switch_systems/SX6036.pdf

http://www.mellanox.com/page/products_dyn?product_family=161&mtag=connectx_3_pro_vpi_card

http://www.mellanox.com/pdf/products/SwitchSystem_Brochure.pdf

有关 IBM InfoSphere BigInsights 的更多信息，请访问：

ibm.com/software/data/infosphere/biginsights

有关 IBM GPFS FPO 的更多信息，请访问：

ibm.com/systems/software/gpfs



北京市朝阳区望京东园七区保利国际广场 T1 15 层

Tel: 010-5789 2000

www.mellanox.com