# 千兆和万兆位以太网在HPC领域的中比较

## 利用万兆位以太网构建HPC集群

### 前言

目前高性能计算(HPC)在各行各业的应用迅速升温。HPC的传统客户是那些需要进行计算流体动力学、地质学和航空航天研究的行业组织与政府部门。最近新加入HPC客户群的机构包括金融服务、数据仓储和数据挖掘领域的大型企业。

显然，采用更强大的计算引擎来处理工作负载将为这些应用带来巨大优势，而互连网络在提高吞吐率和降低延迟方面的作用却经常被忽视。目前，大多数高性能计算应用均使用千兆以太网络，延迟敏感型应用则使用InfiniBand (IB)网络。

本文提供了两种著名HPC应用的性能比较数据，并在一个基准测试环境中考察了千兆以太网、万兆以太网和InfiniBand对其性能的影响。

### 比较千兆以太网、万兆以太网和InfiniBand对吞吐率和延迟的影响

我们可以开发网络技术来优化某些参数，如带宽、延迟或服务质量分类等等。如果这些值达到上限，那么技术本身可能能够满足要求，但网络的设置和维护将非常困难。因此，网络的采用者将面临两种选择：要么通过性能优化支持单个案例，要么采用广泛普及的成熟技术来支持其多数案例（但不是全部）。我们利用著名的HPC应用基准评测，比较了千兆以太网、万兆以太网和IB技术，并显示了每种技术在可管理性和易用性方面的相对指标。

图1显示了用于获取基准评测结果的系统配置。测试平台和测试方法在HPC领域一家知名系统供应商的协助下完成实施。

| | |
|---|---|
| 处理器 | 2xIntel Xeon Harpertown 3.00GHz<br>2x6MB L2高速缓存 |
| 内存 | 16x2GB DDR2 FBDIMM 667MHz |
| 磁盘控制器 | 2xSAS磁盘驱动器146GB 10000 RPM, RAID 1<br>6xSAS磁盘驱动器146GB 10000 RPM, RAID 5 |
| 互连 | • 2xBroadcom Corporation NetXtreme II BCM5708千兆以太网卡<br>• ConnectX Infiniband 4x DDR适配器，或<br>• Chelsio S320 E万兆以太网适配器，或<br>• Mellanox万兆以太网卡 |
| 固件 | System BIOS 1.OO |
| 操作系统 | Red Hat Enterprise Linux Server 5.3<br>Kernel 2.6.28.9-smp |
| 软件 | Intel Fortran Compiler 11.0.074/Intel C/C++ Compiler 11.0.074<br>Intel MKL 10.1.1.019<br>Intel MPI 3.2.011<br>OFED 1.3.1 |

**图1:系统配置**

测试采用了FLUENT v6和v12 (一种CFD代码) 以及LS-DYNA MPI v3.2.1 (一种非线性崩溃分析代码)。这两种代码都是HPC领域的常用代码，能够有效标示总体系统性能——FLUENT用于支持计算流体动力学，LS-DYNA则用于模拟复杂的非线性现象。在服务器级别，Fluent主要侧重于CPU和内存，LS-DYNA则主要侧重于CPU。两者在执行繁重的MPI任务方面都具备出色的可扩展性。HPC行业通常在16到64路范围内使用Fluent。LS-DYNA通常用于8路到64路。

**DYNA测试结果：**

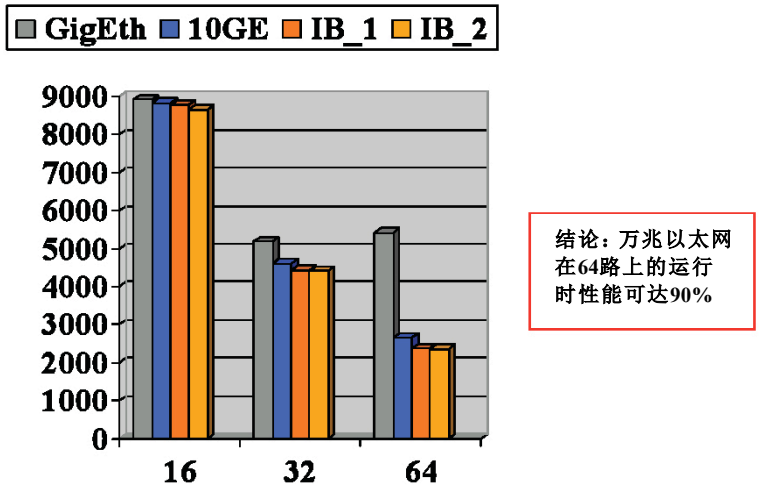


图2: LS--DYNA测试结果–挂钟运行时间（秒）。模式：“车到车(Car to Car)”

**FLUENT测试结果：**

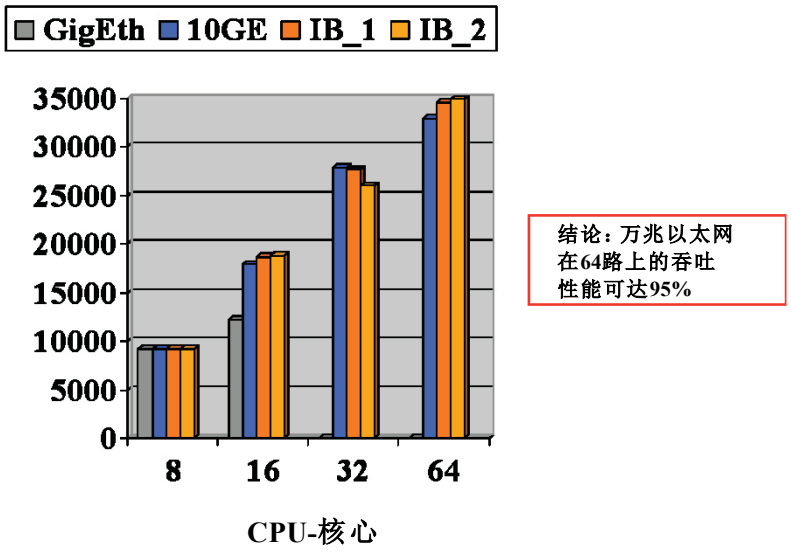


图3：FLUENT测试结果-Fluent吞吐率测量值。模式：“S3”

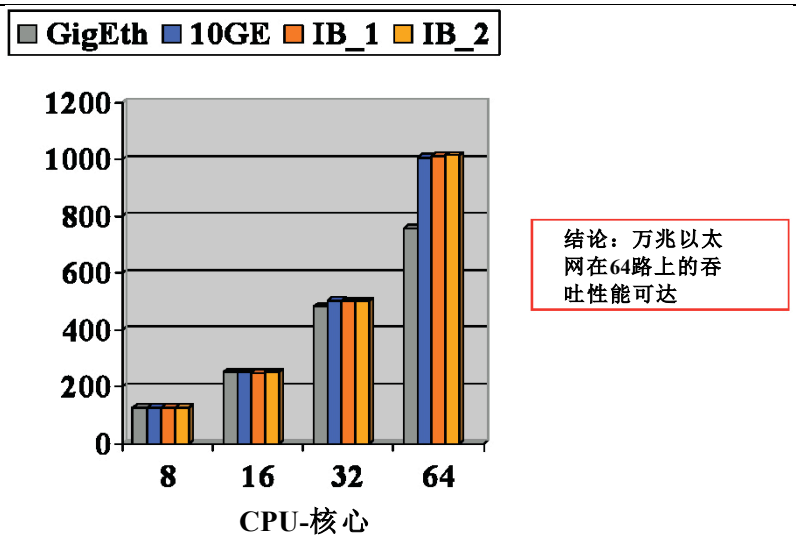


**图4：FLUENT测试结果-Fluent吞吐率测量值。模式：“L3”**

## 测试结果分析

从测试结果中我们得出一个重要结论，在大多数HPC应用中，万兆以太网都具有千兆以太网无法比拟的优势。资料显示，随着CPU核心数量的增加，千兆以太网无法提供所需的低延迟和吞吐率。而万兆以太网在延迟和吞吐率方面具备可扩展的性能，充分显示出其独特的价值，此外它还保留了以太网的易用性优势。

应用在万兆以太网与4x InfiniBand下的性能比较显示，在大多数情况下万兆以太网能够实现比IB高10%的应用性能。这充分表明，随着万兆以太网的来临，IB的应用机会将大大减少。

以太网在易于管理方面具有明显的优势，，安装过程中，可使用ping和traceroute等工具。这样就无需重新培训员工，从而缩短了万兆以太网在HPC环境中的部署周期。

## 总结

少数对于延迟极为敏感的HPC应用。InfiniBand网络始终是面向这些应用的最佳解决方案。因为以太网非常容易配置、管理和排障，从而当前大多数HPC应用都是在千兆以太网上运行。在很大程度上，这种可用性的实现，得益于各种各样工具的使用以及精通IT和以太网技术的训练有素的网络人员的帮助。

以下测试结果显示，通过将基础千兆以太网升级到万兆以太网，HPC应用的性能将获得显著提升。升级后的性能将达到千兆以太网的数倍，而与难于管理的IB网络相比，IB仅高出近10%。面向HPC应用的千兆以太网采用的是专为通用企业应用开发的以太网交换机。这些产品无法提供新型万兆以太网交换机所具备的低时延和无阻塞带宽。

总部

5470 Great America Parkway
Santa Clara, California
USA95054
408-547-5500

支持

support@aristanetworks.com
408-547-5502
866-476-0000

www.aristanetworks.com

销售

sales@aristanetworks.com
408-547-5501
866-497-0000



ARISTA