采用PCI Express 2.0的BONET性能评估

隐中山，岩泽健太郎，陈池武，拓富富雄

AKIBSYSTEMS Inc.台湾

{hide,kentaro,timothy\_wu,takuo}@akibsystems.com

**摘要**：在本文中，AKIBSYSTEMS提出了一种新型的互连网络系统：BONET系统架构，目的是为云计算应用提供足够的带宽。我们开发了基于PCI Express（第2代）的BONET系统架构作为以太网上的网络接口。通过使用PCI Express 2.0（第2代），它不仅可以提高吞吐量和延迟，还可以为并行计算提供动力。本文将不涉及延迟测量。我们的评估结果表明，使用AKIBSYSTEMS的BONET系统可以实现总体的带宽性能提升。

**关键词：**BONET，Hermite，HPC。

1介绍

在HPC（高性能计算）和分布式并行计算方面，瓶颈主要出现在网络接口，处理从节点到节点的数据传输中。假设有几个计算节点具有强大的计算能力，并且这些节点在短时间内完成了指定的计算数据段。但是，所有计算的数据都在Gbe（千兆比特以太网）网络接口卡中排队，等待数据传输。由于网络接口是性能瓶颈，这限制了分布式并行计算的能力。在BONET系统中，我们通过提供更高的带宽能力而不是其他现有互连体系结构来释放瓶颈的局限性。而且，BONET系统与所有TCP / IP和UDP应用程序完全兼容。在第3节中，我们将通过使用三种基准评估应用来说明在具有不同CPU的不同平台上测量的带​​宽性能结果。我们在AKIBSYSTEMS的HermiteⅠ 4通道网络适配器中获得了超过12Gbps的带宽，在AKIBSYSTEMS的HermiteⅡ 8通道网络适配器中获得了近21Gbps的带宽。评估结果揭示了BONET系统的优异能力，打破了目前的网络瓶颈。

2背景

在PCI Express已被改编为标准之前，PCI被广泛用于PC（个人计算机）架构。 PCI是连接诸如声卡，网卡和视频卡等外围设备的一种方式。但是，PCI也有一些缺点。随着CPU功率的提高，网卡和显卡速度的提高，PCI仍然停留在33MHz或66MHZ的总线频率上，仅提供32位宽的长指令访问。 2002年，PCI-SIG和Arapahoe Work Group完成了3GIO草案规范，正式将其重命名为“PCI Express”，其缩写为PCIe。 PCI Express的发展如下所述。

对应的作者。

R.-S. Chang，T.-h. Kim和S.-L. Peng（编辑）：SUComS 2011，CCIS 223，pp.125-130，2011。

©Springer-Verlag柏林海德堡2011

PCI-SIG推出了PCIe 1.0a，其数据速率为250 MB/s，传输数据速率为2.5 GT/s。

2005年，PCI-SIG推出PCIe 1.1。这完全与PCI Express 1.0a兼容，并且更新了一些规范，包括说明和一些改进。数据速率与PCIe 1.0a相同。

2007年，PCI-SIG宣布推出PCI Express Base 2.0规格。PCIe 2.0标准将PCIe 1.0标准的250 MB/s的每通道吞吐量翻倍至500 MB/s。这意味着32通道PCI连接器（x32）可以支持高达16 GB/s的总带宽吞吐量。基础时钟也从2.5GHz增加到5GHz。

另外，基于PCI Express的系统I/O设备可以通过I/O桥连接到内存控制器，命名为“Root Complex”。 PCI Express仍在开发和改进之中。目前的PCI Express实施是3.0版。我们不会详细描述PCI Express。更多信息可以在[1]中找到。

2.1 BONET系统

BONET系统架构是2009年由AKIBSYSTEM提出的。在目前的以太网PCIe插槽网络互连系统中，实现它的常用方法是开发可在特定载体上将PCIe 2.0协议转换为以太网协议进行传输的芯片组如1千兆位以太网和10千兆位以太网。所以我们（AKIBSYSTEMS）认为在发送和接收数据包路径中的数据包转换都需要额外的协议转换开销。为了消除这种开销，我们提出了一种简单地使用PCIe 2.0协议的网络互连系统来进行传输，而不是转换为以太网或任何其他协议。与万兆以太网相比，通过BONET系统从PCIe 2.0插槽传输的数据包仍然是PCIe数据包。我们在没有芯片组管理协议转换的情况下实现了BONET系统中的以太网PCIe。通过BONET系统，我们可以充分利用PCIe 2.0带来的带宽。下一节将中进一步的说明评估的实际性能结果。

3绩效评估

3.1评估应用程序

在本节中，我们使用三种不同的评估应用程序评估AKIBSYSTEMS的BONET系统。首先是Netperf版本2.4.5 [2]，可用于评估网络系统时使用不同模式来衡量带宽性能。 Netperf不仅可以测量TCP和流的带宽，还可以测量流的延迟。其次是iperf 2.0.5 [3]，它是由NLANR / DAST开发的，用于发现TCP和UDP的最大带宽。最新NetPIPE版本的是3.7.1 [4]。 NetPIPE有几种协议评估工具。它可以用来测量TCP，MPI等。在本文中，我们只关注NetPIPE类别下的使用NPtcp的TCP性能。在使用这些应用程序时，我们没有设置任何特定参数。 netperf，iperf和NetPIPE的参数设置几乎设置为默认设置。以下是我们为应用程序设置的参数和命令。

1）iperf

#iperf -c <IP地址> -t 60 -i 30

2）netperf

#netperf -H <IP地址>

3）NetPIPE

#NPtcp -h <IP地址>

3.2先决条件测试环境

图1显示了连接方法。 我们通过特定的PCI Express电缆直接连接客户端（CN1）和客户端2（CN2）。 每个客户配备AKIBSYSTEMS HermiteⅠ4通道（x4）或HermiteⅡ 8通道（x8）网络适配器。 所有由应用程序生成的数据流直接从点对点传输。 本文没有应用控制开关。

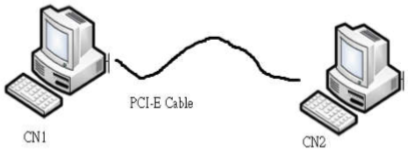


图1.连接方法

表1和表2显示了有关我们的硬件和软件环境设置的详细信息。

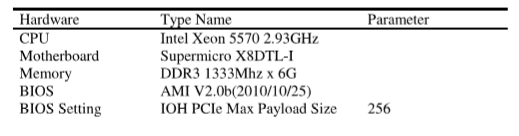


表1.硬件环境设置

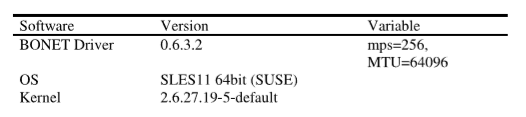


表2.软件环境设置

3.3评估结果

我们已经成功地测试了一对Hermitel网络适配器，它们与PCI Express 4通道宽度直接相连，HermiteⅡ网络适配器与PCI Express 8通道宽度在表1和表2中所述的相同环境下进行测试。结果如图2所示HermiteⅠ的NetPIPE和HermiteⅡ网络适配器在上述测试环境中的应用。 NetPIPE开始从0字节的消息大小发送8388611字节的消息大小，试图探测最大带宽。 在NetPIPE中也测量了每种模式的延迟。

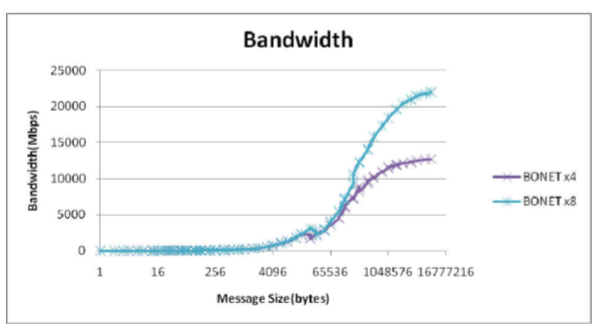


图2. BONETx4（HermiteⅠ）和BONETx8（Hermitell）的NetPIPE

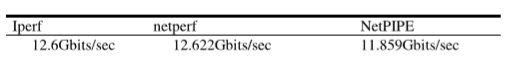


表3. Supermicro X8DT-L上的Hermitel网络适配器

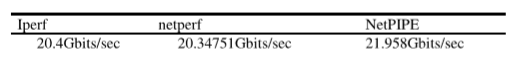


表4. Supermicro X8DT-L上的HermiteⅡ网络适配器

在表3和表4中，我们将评估结果总结为包括iperf，netperf在内的相同测试环境（如NetPIPE）中的列表。表3中列出的测量数据显示iperf和netperf与测量结果几乎相同的NetPIPE。 HermiteⅡ网络适配器的带宽性能比HermiteⅠ网络适配器提高了66％，特别是在大型邮件中，因为HermiteⅡ可以充分利用PCI Express带来的资源。

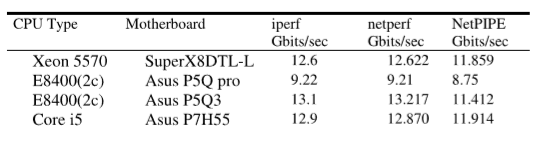


表5.不同平台上的HermiteⅠ 4通道（x4）网络适配器

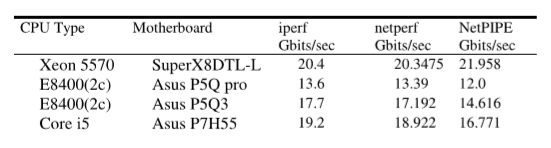


表6.不同平台上的HermiteⅡ 8通道（x8）网络适配器

在表5和表6中，我们列出了带有PCI Express 4通道宽度的HermiteⅠ网络适配器和在不同平台上使用PCI Express 8通道宽度的HermiteⅡ网络适配器以及我们测试过的不同CPU的带宽性能。我们发现CPU的能力会影响带宽性能。表6显示，使用Intel Xeon 5570 CPU的数量比Intel Core i5和Intel E8400的数量更多。采用Supermicro X8DTL-I的Intel Xeon 5570 CPU在评估中具有最佳性能结果。英特尔酷睿i5 CPU比英特尔酷睿E8400有13％的提升。根据实验数据，我们假设如果我们应用更强大的CPU，那么我们可以获得更好的性能结果。通过以上评估，性能瓶颈不再是网络接口。这取决于你的CPU有多强大。

4结论和未来的作品

在这项工作中，AKIBSYSTEMS推出了一种新的互连网络系统，命名为“BONET”系统。我们为评估应用了一些流行的带宽评估应用程序。对于大消息，性能结果表明使用HermiteⅡ 8通道（x8）网络适配器可以实现出色的带宽性能。在高性能计算领域，带宽不是唯一需要考虑的因素。

延迟也是一个重要因素。 未来，我们计划通过使用更多评估应用程序继续我们的评估以寻求低延迟。 MPI和BONET系统将成为另一个有趣的研究方向。

参考

1. PCI-SIG，http://vm7w.pcisig.com/specifications/pciexpress

2. Netperf，http://www.netperf.org/netperf

3.iperf，HTTP：//sourceforge.net/projects/iperf

4. NetPlPE（网络协议独立性能评估器），http://Www.sc1.ameslab.gov/netpipe