TESLA® CASE STUDY

GPUDirect RDMA, 让加速更深入

背景

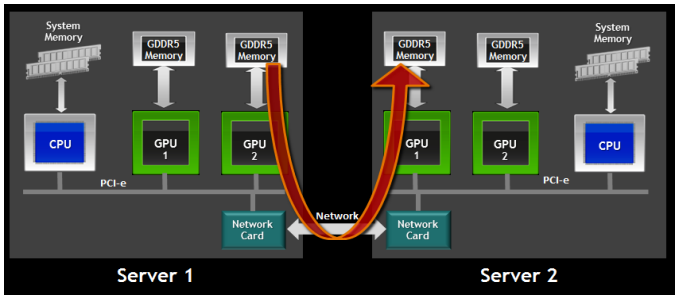
多年来，通用处理器CPU一直试图在追赶着摩尔定律的步伐。但如今，无论是在频率，内存带宽，多核乃至制程和指令集上的优化，都遇到了前所未有的困难。恰逢其时，由NVIDIA带来的GPGPU在异构计算领域打开了一扇门，愈来愈多的超算中心，企业和研究机构正在构建以协处理器为核心的计算资源池，并在异构平台上发展和优化出适配的应用层，以满足日益增长的计算能力需求。

而GPUDirect RDMA，则是在GPU加速领域的又一个技术深入，结合Infiniband技术的原生支持，多个GPU协处理器之间的带宽和延迟性能会得到大幅度的提升，从而使CUDA编程环境下的应用加速有了更大的发挥空间。

针对此技术，曙光公司推出了完全适配的集成方案，并对其性能做了深入全面的测试，以期给用户应用的快速构建和性能提升带来更多的价值。

挑战

NVIDIA 的 GPUDirect 技术始于 2010 年，并不断得到优化和完善，目前已相当成熟。从 CUDA3.1 开始，GPU 支持与网卡，存储等设备共享内存页，简化了 CPU 对 GPU 显存的读写路径，进而获得性能上的提升。随 CUDA 4.0 发布的 GPUDirect 2.0 则开始支持同一 PCIe 总线上的 GPU 之间的内存访问，即所谓的 P2P（Peer-to-Peer）。而直到 2012 年底，GPUDirect RDMA 才完美解决了计算集群节点间 GPU 卡跨 PCIe 总线的通信的问题。



即便如此，底层的技术革新仍然需要时间来完成和其他硬件及应用的适配。主导Infiniband市场的Mellanox在与NVIDIA的合作和推动下，于2014年初发布了支持GPUDirect RDMA的网卡驱动程序，藉此机会，中科曙光在Mellanox的支持下，完成了基于GPUDirect RDMA的异构计算优化方案，并进行了详尽的兼容性和性能测试。

从结果可以看出，在小包通信的传输带宽和网络延迟方面，GPUDirect RDMA可以带来3倍甚至更多的性能提升，对并行应用的扩展性提供了更好的保证。

方案

作为 GPU 异构计算领域坚定的支持者和践行者，曙光公司从国内第一套 GPU 异构集群开始到 HC2000 异构计算方案的推出，一直在积极推进国内 HPC 领域的异构计算加速技术。而 GPUDirect RDMA 则作为方案中的重要技术得到了广泛的推广。

无论是 GPU 工作站，服务器，还是高密度计算刀片，曙光都完全保证其方案的兼容性和性能，并对最新的 NVIDIA Tesla K80 加速卡和 ConnectX-4 EDR 100Gb/s InfiniBand 网络提供加速支持，结合 GPUDirect RDMA 的深度加速技术，给用户带来完美的异构计算加速体验。

很多用户都对曙光公司的方案设计、性能优化和集群运维能力感到满意。如中科院理论物理所的基于 W580I-G10 的 GPU 工作站集群，其上运行的材料计算软件 Ultra-Mat 已成为 GPU 加速领域的应用典范。再如中科院网络中心“元”超算系统基于 I620-G15 的高密度 GPU 服务器集群，其一期系统即已经在为各个科研院所的工作者们提供高达每秒 80 万亿次以上的异构加速能力。

同时，像石油和天然气勘探，地球环境模拟，工业制造仿真等领域，也存在着大量 GPU 异构计算的用户和计算资源需求。在和 NVIDIA，Mellanox 等知名厂商的合作中，曙光对 HC2000 异构计算方案和 GPUDirect RDMA 给用户带来的价值充满信心。

影响

通过对 GPUDirect RDMA 的研究，我们了解到，异构计算的加速可以藉此更深入，更高效。而用户在此过程中收获的价值仅仅在技术上是无法估量的，也是曙光公司努力地方向。新的技术不断地产生和发展，需要我们将其融合和贯通后展现给需要的人。就如同近来得到关注的 cuDNN 的推出，也是给机器学习（Machine Learning）在异构计算领域的发展铺平了道路。

GPUDirect RDMA 的使用在日趋广泛，异构计算领域的革新和竞争也在不断深入，相信曙光公司和 NVIDIA 的进一步合作将能够给用户带来全新的价值和更好的计算体验。