# GPUDirect技术需要的软件和硬件

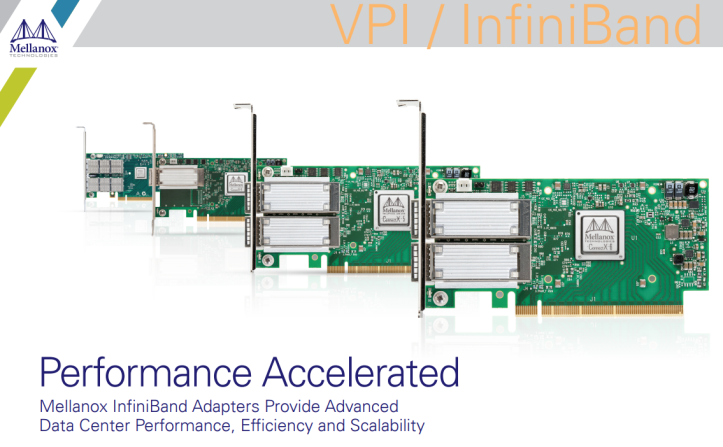
## GPUDirect的优势

* 通过直接往返固定的GPU内存拷贝数据，避免不需要的系统内存拷贝和CPU利用率
* GPU设备和Mellanox RDMA设备间的点对点传输
* 使用高速DMA传输，在P2P设备之间拷贝数据
* 使用直接内存访问（DMA）消除CPU带宽和延迟瓶颈
* 通过GPUDirect RDMA，GPU内存可以用于远程直接内存访问（RDMA），从而带来更高效的应用
* 通过零拷贝支持，提升消息传递接口（MPI）应用

## 硬件

mellanox网卡（必须的）

https://www.mellanox.com/files/doc-2020/ib-adapter-card-brochure.pdf



NVIDIA显卡

NVIDIA® Tesla™ / Quadro K系列或Tesla™ / Quadro™ P系列GPU

* [Quadro RTX 8000, 6000, 5000, RTX 4000](http://www.nvidia.com/object/quadro-desktop-gpus.html)
* [Quadro GV100, GP100, P6000, P5000, P4000](http://www.nvidia.com/object/quadro-desktop-gpus.html)
* [Quadro M6000, M5000, M4000](http://www.nvidia.com/object/quadro-desktop-gpus.html)
* [Quadro K4000, K4200, K5000, K5200 and K6000](http://www.nvidia.com/object/workstation-solutions.html)
* [Quadro 4000, 5000, and 6000](http://www.nvidia.com/object/workstation-solutions.html)
* [Tesla T4](http://www.nvidia.com/object/tesla-servers.html)
* [Tesla V100, P100, K10, K20, K20X, K40](http://www.nvidia.com/object/tesla-servers.html)
* [Tesla C2075 and M2070Q](http://www.nvidia.com/object/tesla-servers.html)

数据交换机

Mellanox数据交换机

TP-LINK交换机？ （10GB/s）

华为交换机？

## 软件

<https://www.mellanox.com/products/GPUDirect-RDMA>

|  |  |  |
| --- | --- | --- |
| 驱动程序 | 平台 | 系统要求 |
| [nvidia-peer-memory\_1.1.tar.gz](https://www.mellanox.com/sites/default/files/downloads/ofed/nvidia-peer-memory_1.1.tar.gz) | HCA | * [ConnectX-6 Lx](https://cn.mellanox.com/products/ethernet-adapters/connectx-6-lx) * [ConnectX-6 Dx](https://cn.mellanox.com/products/ethernet-adapters/connectx-6-dx) * [ConnectX-6](https://cn.mellanox.com/page/products_dyn?product_family=265&mtag=connectx_6_vpi_card) * [ConnectX-5](https://cn.mellanox.com/page/products_dyn?product_family=258&mtag=connectx_5_vpi_card) * [ConnectX-4 Lx](https://cn.mellanox.com/products/ethernet-adapters/connectx-4-lx-en) * [ConnectX-4](https://cn.mellanox.com/page/products_dyn?product_family=201&mtag=connectx_4_vpi_card) |
| GPUs | * NVIDIA® Tesla™ / Quadro K系列或Tesla™ / Quadro™ P系列GPU |
| 软件/插件 | * MLNX\_OFED v2.1-x.x.x or later: [www.mellanox.com](http://cn.mellanox.com/)-> Products -> Software - > InfiniBand/VPI Drivers -> Linux SW/ Drivers * Plugin module to enable GPUDirect RDMA: [www.mellanox.com](http://cn.mellanox.com/)-> Products -> Software - > InfiniBand/VPI Drivers -> GPUDirect RDMA (on the left navigation pane) * NVIDIA Driver: <http://www.nvidia.com/Download/index.aspx?lang=en-us> * NVIDIA CUDA Runtime and Toolkit: <https://developer.nvidia.com/cuda-downloads> * NVIDIA Documentation: <http://docs.nvidia.com/cuda/index.html#getting-started-guides> |

## FAQ

0、What is GPUDirect, and GPUDirect RDMA?

GPUDirect is a term for improving interoperability with NVIDIA GPUs and third-party devices, such as Mellanox ConnectX-3 or Connect-IB devices. GPUDirect RDMA is a feature introduced in Kepler-class GPUs and CUDA 5.0 that enables a direct path for communication between the GPU and a peer device using standard features of PCI Express. The devices must share the same upstream root complex. A few straightforward changes must be made to device drivers to enable this functionality with a wide range of hardware devices. This document introduces the technology and describes the steps necessary to enable an RDMA for GPUDirect connection to NVIDIA GPUs on Linux.

1、Is GPUDirect RDMA supported on Windows platforms?

At this time, GPUDirect RDMA is only supported on Linux.

2、What is the difference between the GPUDirect RDMA Alpha, Alpha 2.0 and the BETA releases?

GPUDirect RDMA was initially developed with the community OFED 1.5.4 release and a set of patches provided by Mellanox that was only targeted with RHEL 6.2. This implementation was a development vehicle only. The GPUDirect RDMA Alpha 2.0 patches were provided to support capabilities of MLNX\_OFED 2.0 and additional support was provided for RHEL 6.3 and RHEL 6.4 development platforms. The latest BETA release represents a significant change in driver architecture between supported devices and is more of generic and standard “plug-in” implementation. This basically means there is the potential for more supported devices to take advantage of the RDMA network to move data directly between different co-processors, storage or other accelerators in the future.

3、Will GPUDirect RDMA run over any generic Ethernet or any network adapter?

No. Mellanox GPUDirect RDMA is supported by Mellanox VPI, Ethernet and InfiniBand devices which provide native hardware-based RDMA offload capabilities.

4、Will GPUDirect RDMA run over RoCE supported adapters?

Yes, GPUDirect RDMA is supported by Mellanox VPI devices in Ethernet mode and will also take advantage of RDMA over Converged Ethernet (RoCE) protocol.

5、Will previous versions of GPUDirect 1.0 developed software still run?

GPUDirect 1.0 allowed GPU and HCA to register and share data buffers on the host memory (i.e. no need to copy the data buffers when switching ownership between HCA and GPU). Initially it was supported as a kernel patch (as well as special MLNX\_OFED version). Later on, it was supported into the kernel (w/o the need for special MLNX\_OFED). Most software developed on MPI should work seamlessly with the latest version of GPUDirect RDMA. There are however, cases where some software may have been developed directly to the GPUDirect RDMA alpha patches and take advantage of lower level verbs level implementation; such as supported MPIs. These applications and MPIs would need to provide slight changes to take advantage of the new plug-in architecture of GPUDirect RDMA BETA. Please ensure you are using the latest available MPIs released after 30 December 2013.

5、What MPIs are currently providing support for GPUDirect RDMA?

Currently, MPIs that provide support for GPUDirect RDMA are Mellanox HPC-X, MVAPICH2-GDR, OpenMPI, and SpectrumMPI.

6、Where can I find out more information about GPU Direct RDMA support in Open MPI?

This link (<http://www.open-mpi.org/faq/?category=building#build-cuda>) has information about configuring Open MPI

9. I am having performance issues after installing the software.

We can distinguish between three situations, depending on what is on the path between the GPU and the third-party device:

PCIe switches only, single CPU/IOH, CPU/IOH <-> QPI/HT <-> CPU/IOH

The first situation, where there are only PCIe switches on the path, is optimal and yields the best performance. The second one, where a single CPU/IOH is involved, works, but yields worse performance. Finally, the third situation, where the path traverses a QPI/HT link, doesn't work reliably.Tip: lspci can be used to check the PCI topology（查看设备上的PCIe端口的拓扑结构）:

$ lspci -t