The evolution of rCUDA towards GPUDirect RDMA

rCUDA is designed according to a client-server architecture. The client side is the computer hosting the application requesting GPGPU services. The server side consists of the remote computer with the real GPU. In order to decouple the virtualization features from the communication characteristics, rCUDA leverages a modular layered architecture, where each layer evolves independently from each other. The communication between rCUDA clients and servers has evolved with time as shown below:

**2010-2012**

**rCUDA versions 1, 2, and 3**

In the initial versions, the focus was on providing the CUDA virtualization services. Therefore, communications were not elaborated, just supporting only the TCP/IP protocol.

**2012**

**rCUDA version 4beta**

In the next version, support for the InfiniBand fabric was included. In addition, memory copies were improved by pipelining the data transfers among communicating peers.

**2012-2013**

**rCUDA version 4**

A further improvement in the communication approach enhanced the pipelined transfers with the use of GPUDirect for the InfiniBand fabric.

**2013**

**rCUDA version 5**

rCUDA currently leverages the new Mellanox OFED GPUDirect RDMA feature. The new rCUDA version will be publicly available at the end of Q4 2013.

**About rCUDA**

In the context of HPC and datacenter clusters, the rCUDA framework grants CUDA-accelerated applications being executed in a server transparent access to GPUs installed in other server of the cluster. In this way, applications are not aware of being accessing an external device, as the rCUDA remote GPU virtualization framework hides all the details, while maintaining application performance.