Abstract English

With the rise of GPUs as hardware accelerators for general purpose computing comes the need for efficient distributed GPU virtualization. Typical communication methods employ CPU buffers for inter-server data transfers. This causes overhead because the server providing the real GPUs for virtualization needs to maintain an additional buffer to temporarily store GPU data.
This thesis investigates the use of GPUDirect RDMA to solve this problem. GPUDirect RDMA enables network adapters to directly access GPU memory. All data transfers on the server providing the real GPUs for virtualization bypass CPU involvement and no further memory copy process is necessary. Integrating GPUDirect RDMA into an already existing GPU virtualization layer led to an overall performance increase of up to 5.96%. Using GPUDirect RDMA is therefore profitable for GPU virtualization software, especially if it virtualizes an application containing a large amount of data allocations and data transfers. The work presented in this thesis represent a step closer to performance transparent GPU virtualization.

Keywords: GPU Virtualization, GPUDirect RDMA, InfiniBand

Abstract Deutsch

Mit dem Aufstieg der GPUs als Hardware-Beschleuniger für allgemeine Berechnungen entsteht der Bedarf an effizienter verteilter GPU-Virtualisierung. Typische Kommunikationsmethoden verwenden CPU-Puffer für Datenübertragungen zwischen Servern. Dies verursacht Overhead, da der Server, der die realen GPUs für die Virtualisierung bereitstellt, einen zusätzlichen Puffer für die Zwischenspeicherung der GPU-Daten unterhalten muss.
Diese Arbeit untersucht die Verwendung von GPUDirect RDMA zur Lösung dieses Problems. GPUDirect RDMA erlaubt es Netzwerkadaptern direkt auf den GPU-Speicher zugreifen können. Alle Datenübertragungen auf dem Server, der die realen GPUs für die Virtualisierung bereitstellt, umgehen die CPU-Beteiligung und es ist kein weiterer Speicherkopierprozess notwendig. Die Integration von GPUDirect RDMA in eine bereits bestehende GPU-Virtualisierungsschicht führte zu einer Gesamtleistungssteigerung von bis zu 5,96%. Der Einsatz von GPUDirect RDMA lohnt sich also für GPU-Virtualisierungssoftware, insbesondere dann, wenn sie eine Anwendung mit einer großen Anzahl von Datenallokationen und Datentransfers virtualisiert. Die in dieser Arbeit vorgestellten Arbeiten stellen einen Schritt näher an eine leistungsfähige transparente GPU-Virtualisierung dar.

Stichwörter: GPU Virtualisierung, GPUDirect RDMA, InfiniBand