On the Enhancement of Remote GPU Virtualization in High Performance Clusters

Las unidades de procesamiento gráfico (Graphics Processing Units, GPUs) están siendo utilizadas en muchas instalaciones de computación dada su extraordinaria capacidad de cálculo, la cual hace posible acelerar muchas aplicaciones de propósito general de diferentes dominios. Sin embargo, las GPUs también presentan algunas desventajas, como el aumento de los costos de adquisición, así como mayores requerimientos de espacio. Asimismo, también requieren un suministro de energía más potente. Además, las GPUs consumen una cierta cantidad de energía aún estando inactivas, y su utilización suele ser baja para la mayoría de las cargas de trabajo.

De manera similar a las máquinas virtuales, el uso de GPUs virtuales podría hacer frente a los inconvenientes mencionados. En este sentido, el mecanismo de virtualización remota de GPUs permite que una aplicación que se ejecuta en un nodo de un clúster utilice de forma transparente las GPUs instaladas en otros nodos de dicho clúster. Además, esta técnica permite compartir las GPUs presentes en el clúster entre las aplicaciones que se ejecutan en el mismo. De esta manera, varias aplicaciones que se ejecutan en diferentes nodos de clúster (o los mismos) pueden compartir una o más GPUs ubicadas en otros nodos del clúster. Compartir GPUs aumenta la utilización general de la GPU, reduciendo así el impacto negativo de las desventajas anteriormente mencionadas. De igual forma, este mecanismo también permite reducir la cantidad total de GPUs instaladas en el clúster.

En esta tesis mejoramos un entorno de trabajo llamado rCUDA, el cual ofrece funcionalidades de virtualización remota de GPUs para su uso en clusters de altas prestaciones. Si bien la versión inicial del prototipo de rCUDA demostró su funcionalidad, también reveló dificultades con respecto a la usabilidad, el rendimiento y el soporte para nuevas características de las GPUs, lo cual impedía su uso en entornos de producción. Estas consideraciones motivaron la presente tesis, en la que toda la investigación llevada a cabo tiene como objetivo principal convertir rCUDA en una solución lista para su uso entornos de producción, con la finalidad de transferirla eventualmente a la industria. La nueva versión de rCUDA resultante de este trabajo presenta una reducción de hasta el 35% en el tiempo de ejecución de las aplicaciones analizadas con respecto a la versión inicial. En comparación con el uso de GPUs locales, la sobrecarga de esta nueva versión de rCUDA es inferior al 5% para las aplicaciones estudiadas cuando se utilizan las últimas redes de computación de altas prestaciones disponibles.