Resumen

En la última década la utilización de la GPGPU (General Purpose computing in Graphics Processing Units; Computación de Propósito General en Unidades de Procesamiento Gráfico) se ha vuelto tremendamente popular en los centros de datos de todo el mundo. Las GPUs (Graphics Processing Units; Unidades de Procesamiento Gráfico) se han establecido como elementos aceleradores de cómputo que son usados junto a las CPUs formando sistemas heterogéneos. La naturaleza masivamente paralela de las GPUs, destinadas tradicionalmente al cómputo de gráficos, permite realizar operaciones numéricas con matrices de datos a gran velocidad debido al gran número de núcleos que integran y al gran ancho de banda de acceso a memoria que poseen. En consecuencia, aplicaciones de todo tipo de campos, tales como química, física, ingeniería, inteligencia artificial, ciencia de materiales, etc. que presentan este tipo de patrones de cómputo se ven beneficiadas, reduciendo drásticamente su tiempo de ejecución.

En general, el uso de la aceleración del cómputo en GPUs ha significado un paso adelante y una revolución. Sin embargo, no está exento de problemas, tales como problemas de eficiencia energética, baja utilización de las GPUs, altos costes de adquisición y mantenimiento, etc.

En esta tesis pretendemos analizar las principales carencias que presentan estos sistemas heterogéneos y proponer soluciones basadas en el uso de la virtualización remota de GPUs. Para ello hemos utilizado la herramienta rCUDA, desarrollada en la Universitat Politècnica de València, ya que multitud de publicaciones la avalan como el framework de virtualización remota de GPUs más avanzado de la actualidad.

Los resutados obtenidos en esta tesis muestran que el uso de rCUDA en entornos de Cloud Computing incrementa el grado de libertad del sistema, ya que permite crear instancias virtuales de las GPUs físicas totalmente a medida de las necesidades de cada una de las máquinas virtuales. En entornos HPC (High Performance Computing; Computación de Altas Prestaciones), rCUDA también proporciona un mayor grado de flexibilidad de uso de las GPUs de todo el clúster de cómputo, ya que permite desacoplar

totalmente la parte CPU de la parte GPU de las aplicaciones. Además, las GPUs pueden estar en cualquier nodo del clúster, independientemente del nodo en el que se está ejecutando la parte CPU de la aplicación. En general, tanto para Cloud Computing como en el caso de HPC, este mayor grado de flexibilidad se traduce en un aumento hasta 2x de la productividad de todo el sistema al mismo tiempo que se reduce el consumo energético en un 15 %.

Finalmente, también hemos desarrollado un mecanismo de migración de trabajos de la parte GPU de las aplicaciones que ha sido integrado dentro del framework rCUDA. Este mecanismo de migración ha sido evaluado y los resultados muestran claramente que, a cambio de una pequeña sobrecarga, alrededor de 400 milisegundos, en el tiempo de ejecución de las aplicaciones, es una potente herramienta con la que, de nuevo, aumentar la productividad y reducir el gasto energético del sistema.

En resumen, en esta tesis se analizan los principales problemas derivados del uso de las GPUs como aceleradores de cómputo, tanto en entornos HPC como de Cloud Computing, y se demuestra cómo a través del uso del framework rCUDA, estos problemas pueden solucionarse. Además se desarrolla un potente mecanismo de migración de trabajos GPU, que integrado dentro del framework rCUDA, se convierte en una herramienta clave para los futuros planificadores de trabajos en clusters heterogéneos.