Las nuevas arquitecturas de aceleradores de procesamiento y sus aplicaciones

El Dr. Francisco D. Igual Peña y el Dr. Carlos García Sánchez son Profesores de la Universidad Complutense de Madrid, España. Ambos profesionales son referentes para profundizar acerca de las arquitecturas de aceleradores.

- En la actualidad, ¿cuáles son las arquitecturas de aceleradores dominantes y hacia qué aplicaciones están orientadas?

Desde mediados de la pasada década, el avance en prestaciones de los procesadores gráficos (GPUs) supuso un crecimiento considerable en su utilización para cómputo de propósito general. Con la aparición de entornos software que facilitaron su uso (CUDA y/u OpenCL), su éxito se consolidó hasta el punto de establecerse como un estándar de facto en la construcción de arquitecturas de alto rendimiento al día de hoy. Con ello, la cantidad y tipología de aplicaciones que potencialmente pueden hacer uso de estas plataformas se ha extendido a prácticamente cualquier ámbito, y el impacto en rendimiento es más que considerable siempre que dichas aplicaciones cumplan ciertos requisitos en cuanto a cantidad de paralelismo expuesto. Otras plataformas han intentado ocupar el puesto alcanzado por las GPUs sin tanto éxito (principalmente FPGAs o el recientemente desaparecido Intel Xeon Phi), de momento sin conseguir destronarlas como líderes del mercado en el ámbito de la aceleración para Computación de Altas Prestaciones.

-La Inteligencia Artificial (y en particular el Deep Learning) se ha vuelto uno de los temas más relevantes de la Informática actual. ¿Cómo ha impactado éste tema en los procesadores y aceleradores?

La Inteligencia Artificial en general, y el aprendizaje profundo en particular, son campos maduros que pertenecen a ámbitos con gran trabajo científico subyacente, y que han sido desarrollados durante décadas. Sin embargo, la eclosión y popularización de grandes facilidades de cálculo (incluso a nivel doméstico) ha hecho resurgir el interés por este tipo de algoritmos aplicados sobre cantidades ingentes de datos. La respuesta a este creciente interés por parte de los desarrolladores de arquitecturas ha sido doble: primero, se han introducido modificaciones en los procesadores para dar soporte específico a primitivas únicamente útiles en este tipo de algoritmos (un ejemplo claro es la introducción de Tensor Cores en las GPUs Nvidia de última generación); segundo, ha emergido una enorme familia de procesadores de propósito específico, diseñados y desarrollados con un único objetivo: acelerar los procesos computacionales básicos asociados a las implementaciones para deep learning.

-Google ha desarrollado su propio procesador para Deep Learning, el Tensor Processing Unit (TPU). Por su parte, Nvidia incorpora hardware específico para Deep Learning en su última generación de GPUs (Volta), convirtiéndola en una arquitectura heterogénea. ¿Cree que en el futuro se volverá más popular el

uso de arquitecturas específicas para cada clase de problema?

Sí, la especialización es la respuesta que la industria ha dado a la ralentización en el escalado de Dennard, y todo apunta a que las futuras arquitecturas darán soporte de forma nativa a ciertas aplicaciones de interés (llevando la heterogeneidad dentro del chip), o bien diseñándose desde el inicio como exclusivamente específicas para una aplicación dada. Esto supone un reto, ya que el correcto uso de estas unidades de propósito específico para la aceleración de otro tipo de algoritmos abre la puerta a oportunidades similares a las que se plantearon con el nacimiento de las GPUs y con el advenimiento del término GPGPU (Procesamiento de Propósito General en GPUs).

-Al día de hoy, no existe un sistema en el ranking TOP500 o en el GREEN500 que incluya FPGAs. Sin embargo, en los últimos años, su uso para diferentes tipos de problema viene aumentando. La incorporación en 2015 de esta clase de acelerador a los servidores de Microsoft para su buscador Bing es un claro ejemplo de ello. ¿Cree posible ver un sistema basado en FPGAs en el TOP500 en la próxima década? ¿y en el GREEN500?

Si bien es cierto que no existe en la actualidad un sistema basado en FPGA en cualquiera de las listas mencionadas, hay que mostrarse cauto sobre su evolución en los próximos años. ¿Quién nos iba a decir hace 10 años que las primeras posiciones de estos rankings estarían copadas por sistemas basados en aceleradores cuando por aquel entonces solo unos pocos incorporaban el denostado Cell? Entre las principales razones por las que no existe ningún sistema con FPGAs en las listas mencionadas se encuentra la metodología usada para ordenar ambas listas. Tanto el TOP500 y como el GREEN500 se basan en la ejecución del conocido benchmark LINPACK consistente en la resolución de un sistema de ecuaciones denso con aritmética en punto flotante. Mientras que las FPGAs tienen un comportamiento excelente en aritmética en entero, su rendimiento decae para aplicaciones en punto flotante. Además, el ámbito de aplicación de un supercomputador de propósito general como los de la lista del TOP500 difiere del campo de aplicación de las FPGAs. Sin embargo, nos gustaría indicar que las FPGAs tienen su nicho de interés en aquellos sistemas encargados en realizar tareas repetitivas. Este aspecto es su punto fuerte, permitiendo especializar una FPGA para realizar ese trabajo repetitivo, lo que unido a su mayor eficiencia energética lo hacen altamente interesante en aplicaciones como el mencionado ejemplo del buscador. Por último, nos gustaría resaltar dos hitos del último año en la industria que pueden suponer un cambio de tendencia: (1) el anuncio de Intel del nuevo procesador "Skylake" Xeon SP con un FPGA integrada en el mismo chip y (2) la incorporación de instancias F1 con FPGAs de Xilinx en los servicios en la nube de Amazon.

- En la última década, ha crecido significativamente el uso de procesadores basados en la arquitectura ARM, tanto en el segmento de móviles como de embebidos. ¿Cuál cree que es el futuro de esta clase de procesadores?

Es evidente que este tipo de procesadores estarán muy presentes en el sector de sistemas empotrados donde su cuota de mercado actual es enorme. A modo de ejemplo, la compañía Samsung ha desbancado a Intel en volumen de ventas durante el 2018. Con la irrupción del Internet de las Cosas con más dispositivos conectados a la red no es descabellado pensar que la supremacía de ARM en este sector será aún mayor si cabe. Sin embargo, es cierto que en el contexto de HPC, pese a iniciativas como el sistema Mont-Blanc basado en ARM, no han existido muchos casos éxito motivado principalmente por la diferencia de rendimiento entre un procesador con arquitectura x86 frente a un ARM.

-¿La importancia de estas nuevas arquitecturas requiere incluir temas en las currículas de Informática? Que consideran Uds. que es necesario que conozca un egresado de Licenciatura/Ingeniería sobre estas arquitecturas con vista a su labor profesional?

Tal y como se ha comentado anteriormente, la tendencia actual de la industria de los procesadores es hacia la especialización. El primer hito de especialización lo podemos datar en la inclusión de pequeñas unidades vectoriales para incrementar el rendimiento de las aplicaciones multimedia en los procesadores Intel Pentium MMX a finales de los 90. Desde entonces esta tendencia no se ha frenado y un claro ejemplo es la popularización de las GPUs desde sistemas de escritorio hasta teléfonos móviles. Desde nuestro punto de vista es muy recomendable que los egresados conozcan la heterogeneidad presente en el hardware actual, y aquellas herramientas informáticas que facilitan el desarrollo de aplicaciones en estos sistemas. Este enfoque de heterogeneidad debería de ser transversal a todo el currículum. Afecta a gran parte de las asignaturas de los grados, desde aquellas más relacionadas con el hardware como el diseño de un computador y la arquitectura interna del procesador, pasando por las necesidades de adaptación de un sistema operativo y la programación paralela de estos dispositivos. Por último nos gustaría hacer notar el número de oportunidades laborales que se abren para los estudiantes en Informática con el auge del Internet de las Cosas donde la heterogeneidad estará muy presente. La demanda de empleo en este sector es mucho mayor que la oferta actual y las previsiones hablan de una demanda aún mayor por lo que adaptación de los planes de estudio no solo es imprescindible, es además urgente.