Una Experiencia en GPU Computing entre FaMAF e INVAP

Fabio Bustos (INVAP SE), Nicolás Wolovick (FaMAF-UNC)

Resumen

Mostramos la experiencia realizada entre INVAP SE y FaMAF-UNC en desarrollo de software de procesamiento de imágenes sobre placas de procesamiento gráfico de propósitos generales (GPGPU por sus siglas en inglés).
En el marco de un sistema de adquisición de imágenes desarrollado por INVAP para uno de sus clientes, resulta necesario incluir un módulo de software capaz de realizar seguimiento automático de puntos de interés identificados en un video. Se encomienda a un grupo de investigación de la UTN un primer desarrollo de este módulo sobre procesadores convencionales CPU, y, a partir de la integración del sistema, surge la necesidad de optimizar el desempeño de dicho módulo en términos de tiempos de ejecución y uso de recursos computacionales a fin de que se obtenga un procesamiento en tiempo real (30 cuadros por segundo). Para satisfacer esa necesidad se decide migrar el módulo a una GPU presente en el sistema, para lo cual se recurre al grupo de investigación GPGPU Computing de FaMAF, reconocido a nivel nacional por su experiencia en esa tecnología. Se establece una metodología de trabajo en conjunto a través de un Enunciado de Trabajo en el cual se definen el alcance, los entregables, los plazos de ejecución de los trabajos y los interlocutores por ambas partes. En términos generales, el equipo de investigación toma a su cargo el diseño, la implementación y la verificación del módulo desarrollado en ambiente de laboratorio, mientras que el equipo de INVAP se responsabiliza por la verificación y validación en el ambiente real de ejecución. La interacción se caracteriza por la comunicación fluida y la rigurosidad en la metodología de medición del desempeño de la aplicación, su verificación y validación. Con motivo de esta experiencia se ejercitaron también los canales administrativos de ambas entidades para generar el marco formal de trabajo mediante los correspondientes convenios. En resumen, se presenta un caso de aplicación exitosa de los saberes del sistema científico tecnológico nacional para la resolución de un problema de la industria que tiene impacto concreto en un proyecto de desarrollo de un sistema de alta complejidad.

Palabras Clave

GPU Computing, Visión de Computadoras, Tiempo Real

Introducción

El Grupo de GPGPU Computing de FaMAF-UNC nace en 2008 a través de un Subsidio a Profesores por parte de la empresa NVIDIA, Este subsidio reúne un grupo interdisciplinario de docentes-investigadores en Física y Computación. Con este subsidio y la donación de dos placas de cómputo GPU se contratan pasantes de I+D para realizar tareas de paralelización de códigos de simulaciones físicas que tienen aplicación en investigaciones de ciencia básica dentro de la FaMAF. Dado que la tecnología brinda la oportunidad de tener una Supercomputadora en el Escritorio, muchos docentes investigadores se acercan a plantear sus problemas que en la mayoría de los casos fueron acelerados de manera dramática, mejorando de manera cualitativa y cuantitativa la producción en ciencia.

Otro hito importante es la creación de la "Escuela Argentina de GPGPU Computing para Aplicaciones Científicas" (EAGPGPU) que tiene su primer edición en 2011.
El Grupo de GPGPU Computing recibe galardones año a año por parte de empresas y organismos internacionales (NSF, NVIDIA), así como donaciones de hardware y software (NVIDIA, Intel, PGI, Adapteva), que reconocen la calidad del trabajo que se realiza en divulgación, docencia, investigación y transferencia en estas tecnologías particulares, así como en Computación de Alto Desempeño en general (High Performance Computing - HPC por sus siglas en inglés).

INVAP SE es una empresa dedicada al desarrollo de sistemas tecnológicos complejos con más de treinta años de trayectoria en el país y veinte en el exterior. Su experiencia se centra en el desarrollo de proyectos multidisciplinarios de alta complejidad tales como reactores nucleares, satélites y radares, cubriendo desde la concepción inicial hasta la entrega de los productos finales al cliente así como el soporte y la actualización tecnológica a lo largo de todo el ciclo de vida de los sistemas. La empresa mantiene una estrecha relación con la Comisión Nacional de Energía Atómica y la Comisión Nacional de Actividades Espaciales (entidades con las que ha llevado a cabo proyectos de gran envergadura), así como con las instituciones del sistema científico-tecnológico nacional, dentro del cual ocupan un lugar relevante las universidades nacionales. INVAP se encuentra organizada según sus áreas de actividad, entre las que se cuentan las áreas Nuclear, Espacial y Gobierno, Industrial y Energías Alternativas, TICs y Servicios Tecnológicos. El 85% de su personal está formado por profesionales y técnicos altamente especializados, organizados en estructuras dinámicas adecuadas para la realización de los proyectos y distribuidas geográficamente en el país, ubicándose la Sede Central en Bariloche, con oficinas en Córdoba y Buenos Aires.

Ambos actores se conocen previamente. Personal de INVAP ha formado parte de muchas de las instancias de formación brindadas por el Grupo de GPGPU Computing tanto en Córdoba como en Bariloche. Los equipos de trabajo de INVAP se nutren a su vez de egresados de la FaMAF tanto de grado como de posgrado de las carreras de Física, Matemática y Ciencias de la Computación.

Para uno de los proyectos desarrollados por INVAP se requiere disponer de la capacidad de seguir objetos a partir de imágenes de alta resolución en movimiento. La solución técnica de dicho requerimiento incluye sistemas mecánicos, módulos electrónicos y diversos componentes de software. A partir de un análisis de las posibles alternativas y la mejor combinación de ellas para resolver el requerimiento, se decide recurrir al Laboratorio de Técnicas Avanzadas (LTA) de la UTN-FRC, para desarrollar un prototipo del módulo de software de seguimiento, el cual es implementado utilizando técnicas y algoritmos en el estado del arte. Dicho software resulta correcto respecto a la precisión obtenida para el seguimiento, pero con la potencia de cálculo disponible en la CPU del equipo, presenta un desempeño insuficiente para lograr un seguimiento en tiempo real, es decir con una tasa de refresco de imágenes cercana a 30 cuadros por segundo (frames per second, FPS por sus siglas en inglés), considerando que además se requiere utilizar capacidad de la misma CPU para otras tareas propias del sistema.
Frente a esta situación se resuelve incorporar al sistema un procesador gráfico (GPU) y migrar el módulo de software de seguimiento a esta placa, encomendándose esta tarea al Grupo de GPGPU Computing de la FaMAF-UNC.

Objetivos

El objetivo era entonces la descarga de los cálculos del software de seguimiento a la GPU que fue agregada al equipo de cómputo, a fin de liberar la CPU para atender a los otros procesos que se necesitaban y obtener una tasa de refresco cercana a los 30 FPS dentro de un flujo de imágenes en resolución FullHD, manteniendo la precisión del algoritmo respecto al seguimiento de objetos.

Metodología

La placa GPU agregada a la configuración de cómputo posee una potencia pico de cálculo de 600 GFLOPS (miles de millones de operaciones aritméticas por segundo), 3 veces más que los 200 GFLOPS que entrega de potencia pico la CPU que se estaba utilizando. Se estimó que con esta potencia de cálculo era posible reemplazar la CPU en la tarea de seguimiento cumpliendo con las restricciones de un flujo de imágenes FullHD con una tasa de refresco de 30 FPS.
El software de seguimiento tiene dos módulos principales, el Módulo de Rastreo y el Módulo Detector. Se plantea una primera etapa donde el objetivo es paralelizar en GPU el Módulo Detector, pieza clave en el seguimiento de objetos y que tiene la mayor carga computacional.
La descarga a la GPU del Módulo Detector se realiza en los meses de Noviembre y Diciembre de 2013, con un periodo de descanso intermedio y finalizando a fines de Febrero de 2014, con un tiempo total de 10 semanas de trabajo.
En esta etapa participan del lado de FaMAF un coordinador general, un coordinador técnico, un consultor en Visión de Computadoras y tres programadores, dos de ellos senior y uno junior.
La descarga o paralelización en GPU del Módulo Detector se lleva a cabo dentro de las instalaciones de FaMAF, en una oficina asignada a tal efecto, con todos los servicios necesarios: mobiliario, red, servidores de versiones, servidores de cómputo, seguridad 24 horas, y personal específico de apoyo para llevar adelante los acuerdos y contratos.
El servidor de cómputo, réplica del equipo de cómputo que llevará el producto, llega a préstamo a FaMAF por parte de INVAP.
Se utiliza tecnología NVIDIA CUDA para la programación ya que la plataforma seleccionada para GPU es una NVIDIA Quadro K2000. La validación de la descarga a la GPU se realiza comparando el comportamiento entre los submódulos, y aceptando sólo lo que pasaba los tests de unidad y generaba el mismo resultado comparando CPU vs. GPU.
El desarrollo contó con dos fases bien diferenciadas, la obtención de un programa GPU correcto respecto a la versión CPU y luego la optimización del código GPU para obtener el máximo desempeño en la plataforma de cómputo. La primera fase fue de 6 semanas y la segunda de 4 semanas.

Luego de entregado el producto, éste se integra con el resto de los componentes tanto de software como de hardware que forman el producto final. En Junio de 2014 se detecta que aunque el desempeño del software de seguimiento era satisfactorio, la integración introducía demasiada carga a la CPU lo que generó la necesidad de descargar también el Módulo de Rastreo en la GPU y esto generó una segunda fase del proyecto.
La segunda fase abarcó 8 semanas en los meses de Septiembre y Octubre de 2014. Gracias a la experiencia anterior, el proceso fue más fluido en sus partes administrativas y técnicas.

El grupo de coordinación y consultoría se repitió, pero los programadores cambiaron y fueron un desarrollador y una desarrolladora con experiencia senior en la tecnología CUDA.
El proceso de desarrollo fue similar, 4 semanas para obtener un producto funcional corriendo en la GPU y 4 semanas para obtener un producto optimizado para que tenga alto desempeño.
A principios de Noviembre de 2014 se entregó el software de seguimiento completamente descargado en la GPU.

Análisis e Interpretación de los Resultados

Al finalizar la primera etapa a fines de Febrero de 2014 se entregó el Módulo Detector paralelizado de manera completa en GPU, el cual logró acelerar el seguimiento de objetos de 15 cuadros por segundo a 35 cuadros por segundo. La CPU quedó descargada desde un 100% de uso de sus 4 núcleos a solo el 50%, es decir, 2 núcleos.

La segunda etapa se entregó a principios de Noviembre de 2014 con el software de seguimiento completamente descargado en la GPU, obteniendo una tasa de refresco de 43 cuadros por segundo y una CPU con un 25% de carga (1 solo núcleo ocupado).

La integración con el resto de los subsistemas de software fue satisfactoria y se obtuvieron, dependiendo de las condiciones externas, una tasa de refresco que va de los 25 a los 34 cuadros por segundo.

Es importante destacar que la GPU utilizada es modesta en prestaciones, con una potencia pico de cómputo de 600 GFLOPS, cuando las placas tope de gama brindan hasta 4000 GFLOPS de potencia de cómputo. La CPU en cambio no era modesta, el equipo contaba con un Xeon E3 1275 v3 con 224 GFLOPS de potencia pico, y este es uno de los procesadores más potentes del mercado encapsulado en una sola pastilla.
El software resultó óptimo respecto a la utilización de los recursos de la GPU, ya que teniendo 3 veces más potencia de cómputo que la GPU, se obtuvo 3 veces la tasa de refresco.

Se realizaron pruebas de escalado del software en placas de mayor potencia y se obtuvieron desempeños de hasta 110 FPS en GPUs de gama alta.

Conclusiones

La empresa INVAP contrató un desarrollo de I+D en un área técnicamente compleja y la FaMAF-UNC pudo llegar a los objetivos en tiempo y forma, teniendo en cuenta los ajustados plazos de entrega.
La relación Empresa-Universidad funcionó muy bien, en parte porque los actores tenían conocimiento de las necesidades y capacidades del otro, y además porque existía una valoración positiva entre las partes. Por lo tanto, en la práctica, funcionó como un proyecto conjunto.
Jugó un papel fundamental el conocimiento y el prestigio del Grupo de GPGPU Computing, que tiene dos miembros actualmente trabajando en la oficina de INVAP en Córdoba. También fue importante que durante los 7 años de existencia del Grupo de GPGPU Computing se realizaron varias instancias de formación (Escuelas, Cursos,

Workshops) tanto en Córdoba como en Bariloche donde siempre asistieron empleados de INVAP.
La empresa obtuvo un producto que le permitió cumplir los requerimientos del cliente y el Grupo de GPGPU Computing comprobó que sus conocimientos pueden ser aplicados de manera inmediata en la industria de alta tecnología local.
Además de este trabajo, se va a presentar un poster en una conferencia internacional en tecnología de GPU Computing.

Creemos que las claves que permitieron cumplir los objetivos planteados, con tecnologías complejas y en un tiempo muy acotado son las siguientes.

- Se planteó un desarrollo incremental evolutivo con entregas parciales.
- Ambas partes demostraron solvencia técnica, lo que permite interacción fluida a través de dominios diferentes.
- Se establecieron criterios de aceptación claros y compartidos.
- Se ejerció flexibilidad en la gestión administrativa, la cual sirvió para dar el marco correspondiente y de ninguna forma obstaculizó el trabajo.
- Se ejerció una comunicación fluida y se exhibió disponibilidad para evaluar las situaciones y tomar decisiones en plazos cortos.
- Se mantuvo transparencia en las estimaciones de esfuerzo y hubo mutua disponibilidad para discutirlas y acordar alcance, costo y plazo, respetando los criterios de calidad establecidos para el desarrollo.

Trabajos Futuros
El vínculo de colaboración está funcionando perfectamente, tanto en la parte administrativa como en la parte técnica.
El Grupo de GPGPU Computing está investigando la *embedded supercomputer* (supercomputadora portátil) Jetson TK1 de NVIDIA para evaluar su funcionamiento en este tipo de sistemas. Con el software corriendo en esta plataforma, el campo de aplicaciones aumenta considerablemente y el producto obtenido puede ser reutilizado en otros proyectos actualmente en proceso de ingeniería dentro de INVAP.
INVAP está creando un grupo de GPGPU Computing para atender a sus crecientes necesidades respecto al uso de estas tecnologías. Se esperan colaboraciones en un futuro cercano, tanto en consultoría como en desarrollos y optimizaciones de código dentro de las áreas de HPC y GPU Computing.

Referencias Bibliográficas

1. Wu, Y., Lim, J., & Yang, M. H., Online object tracking: A benchmark. *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.