Abstract

The increasing demand of decoding high-quality videos can lead to a challenging computational requirement for conventional Central Processing Unit (CPU) architectures. Graphics Processing Units (GPUs) have emerged as another important type of general-purpose computing architecture, which in general provide higher computational power than CPUs. Efficient GPU execution, however, requires massive parallelism and little executing divergence, two criteria are not fully met by all video decoding kernels. This thesis exploits how GPUs can be effectively used in video decoding applications. The challenges include proper workload distribution between the CPU and GPU, task optimizations on two heterogeneous devices, and efficient communication between them.

We first analyzed the suitability of decoding kernels in H.264/MPEG-4 Advanced Video Coding (H.264) and High Efficiency Video Coding (HEVC) standards for the GPU architecture. The entropy decoding kernel shows strong data dependency and hence is not suitable for GPU computation. The other kernels can be adapted for GPUs as they characterize with reasonable degree of data parallelism and with fair executing divergences.

In the H.264 standard, the inverse transform and motion compensation were parallelized and optimized for GPUs. We compacted and separated input coefficients of inverse transform to remove unnecessary computations caused by zero coefficients. In motion compensation, we decomposed kernel computation into multiple sub-stages that can be shared between different computation modes, in order to mitigate the potential divergence.

In the newer HEVC standard, its in-loop filters were parallelized and optimized for more efficient GPU execution. Compared to the state-of-the-art GPU approach, the proposed parallelization provided an on-demand workflow and a more efficient thread mapping.

Finally, a complete parallel HEVC decoder was proposed for heterogeneous CPU+GPU systems. We exploited available decoding parallelism on the CPU, GPU, and between the two devices simultaneously. On top of the parallel design, two workload balancing schemes were implemented, in order to adapt computation resource variation on CPU and GPU. In addition, an energy measurement module was developed for energy efficiency analysis.

Evaluated results showed that suitable decoding kernels can be accelerated substantially (up to 28.2×) on GPUs at the kernel level. At the application level, using GPU architecture can provide significant acceleration when only a low number (1 to 8) of CPU cores are available. On a system consisting of an NVIDIA Titan X Maxwell GPU and an Intel Xeon E5-2699v3 CPU, with four CPU cores, the proposed HEVC decoder delivers 167 frames per second for 4K videos, corresponding to a speedup of 2.2× over the state-of-the-art CPU decoder using four CPU cores. When more CPU cores (¿8) are employed, the benefit of using GPU vanishes and the performance is eventually outperformed by the CPU decoder due to GPU overloading. With respect to energy, because of its high power consumption GPU architecture is not as efficient as the CPU for HEVC decoding.

Zusammenfassung

Der steigende Bedarf an qualitativ hochwertigen Videos stellt eine zunehmend grere Herausforderung fr die Rechenleistung konventioneller Prozessoren (CPUs) dar. Grafikprozessoren (GPUs) haben sich zu einer weiteren Computerarchitektur entwickelt, die im Allgemeinen eine hhere Rechenleistung als CPUs bietet. Allerdings erfordert eine effiziente Programmausfhrung auf GPUs ein Hchstma an Parallelitt bei geringer Divergenz. Beide Kriterien werden von verschiedenen Kerneln aktueller Videodekoder nicht vollstndig erfllt. Diese Doktorarbeit untersucht, wie GPUs effizient fr Videodekodierungsanwendungen eingesetzt werden knnen. Die wesentlichen Herausforderungen bestehen dabei in der geeigneten Aufgabenverteilung zwischen CPU und GPU, der Optimierung auf zwei heterogenen Architekturen, sowie der effizienten Kommunikation zwischen diesen.

Zuerst wurden die verschiedenen Kernel der Standards H.264/MPEG-4 Advanced Video Coding (H.264) und High Efficiency Video Coding (HEVC) auf ihre Eignung fr eine GPU-Implementierung untersucht. Die Entropiedekodierung ist von starken Daten- und Steuerabhngigkeiten geprgt, weshalb sie auf einer GPU-Architektur nicht effizient implementiert werden kann. Alle anderen Kernel sind aufgrund eines hohen Anteils an Datenparallelitt bei moderater Divergenz sehr gut fr die Ausfhrung auf GPUs geeignet.

Die inverse Transformation und Motion Compensation des H.264-Standards wurden fr die Ausfhrung auf GPUs parallelisiert und optimiert. Die Koeffizienten der inversen Transformation wurden separiert, um unntige Berechnungen mit Nullkoeffizienten zu vermeiden. Fr die Motion Compensation wurden die Berechnungen in mehrere Stufen unterteilt, die bei verschiedenen Berechnungsmodi gemeinsam verwendet werden knnen, um so die Divergenz zu verringern.

In dem neueren HEVC-Standard wurden die Loop Filter parallelisiert und fr eine effizientere Ausfhrung auf GPUs optimiert. Im Vergleich mit dem aktuellen Stand der Technik konnte eine effizientere Zuordnung der Aufgaben zu den Recheneinheiten erreicht werden.

Abschlieend wurde ein vollstndig paralleler HEVC-Dekoder fr ein heterogenes CPU +GPU-System entwickelt. Es wurden sowohl die parallelen Mglichkeiten beider Gerte, als auch die Parallelitt zwischen ihnen ausgenutzt. Zustzlich wurden zwei Mechanismen implementiert, welche die Aufgaben anhand der Rechenressourcen auf beiden Gerte verteilen. Auerdem wurde ein Gert zur Messung der Energieeffizienz entwickelt.

Die Ergebnisse zeigen, dass einzelne geeignete Videodekoderkernel auf GPUs massiv beschleunigt werden knnen (bis zu 28,2x). Auf der Anwendungsebene ist eine Verbesserung fr ein CPU+GPU-System nur bei einer geringen Anzahl an CPU-Kernen (1 bis 8) mglich. Der vorgestellte heterogene HEVC-Dekoder kann auf einem System mit einer NVIDIA Titan X Maxwell GPU und einer Intel Xeon E5-2699v3 CPU mit vier Kernen ein 4k-Video mit 167 Bildern pro Sekunde dekodieren. Dies entspricht einer Verbesserung um den Faktor 2,2 gegenber einem reinen CPU-Dekoder. Werden mehr als acht CPU-Kerne verwendet, wird der Vorteil des heterogenen Systems geringer. Aufgrund einer berlast der GPU kann das reine CPU-System sogar schneller sein. Die GPU-Architektur hat wegen ihres hohen Energieverbrauchs eine niedrigere Energieeffizienz bei der HEVC-Dekodierung als die CPU-Architektur.