运用 GPU 计算面向非规则应用的非合并内存访问优化

郑然[1,2,3,4]，刘元栋[1,2,3,4]，金海[1,2,3,4]

[1]华中科技大学大数据技术与系统国家地方联合工程研究中心，中国武汉市，430074

[2]华中科技大学服务计算技术与系统实验室，中国武汉市，430074

[3]华中科技大学集群与网格计算实验室，中国武汉市，430074

[4]华中科技大学计算机科学与技术学院，中国武汉市，430074

摘要：通用图形处理器（GPGPU）可大大提升规则应用的计算性能。然而，很多应用中存在非规则内存访问模式，大大限制了GPU的性能优势。近年来，一些研究提出解决方案来移除静态非规则内存访问。然而，利用软件消除动态非规则内存访问仍然面临严峻挑战。本文提出一种纯软件解决方案用于消除动态非规则内存访问，尤其是间接内存访问，无需硬件扩展和离线分析。提出数据重组和索引重定向以减少内存访问次数，从而提高GPU内核性能。为提高数据重组效率，卸载重组数据操作至GPU以降低开销并传输数据。通过并发执行数据重组和数据处理内核的统一计算设备架构（CUDA）流，可降低数据重组开销。完成这些优化后，相比于CUSPARSE基准测试，使用该方法GPU内核的内存数据传输减少了16.7%–50%；同时，NVIDIA Tesla P4 GPU上的内核性能提高了9.64%–34.9%。

关键词：通用图形处理器；内存合并；非合并内存访问；数据重组

https://doi.org/10.1631/FITEE.1900262