doi:10.1631/FITEE.1700059

题目：CWLP：一种在 GPU 中协同的线程束调度和局部性保护的高速缓存分配策略

概要：随着我们正在接近百亿亿次超级计算机的时代，一个拥有强大运算能力和低能耗的均衡的计算机系统变得越来越重要。GPUs 是在最近投入运营的超级计算机中被广泛使用的加速器。它采用大规模多块程来隐藏长访存延迟，同时它拥有高能效。相对于其强大的运算能力，GPUs 的每个流多核处理器只有几兆的片上资源。面向吞吐率的执行模型与它的高速缓存层次结构设计不匹配，使得 GPUs 缓存表现出较差的运行效率。由于片上存储器的严重缺少，受较差的缓存性能影响，GPU 的计算能力急剧下降，限制了系统性能和能效。提出一种协同的线程束调度和局部性保护的缓存分配策略（CWLP），以充分利用数据局部性和隐藏延迟。首先，设计了一种基于指令 PC 的局部性保护方法（LPC）以提升 GPU 性能。使用一个基于 PC 的收集器收集每个高速缓存块的重用信息。在获取缓存块的动态重用信息后，采用一个智能缓存分配单元（PCAU），它结合了重用信息和 LRU（最近最少使用）替换策略，以找到拥有最少局部性的缓存块并将其逐出。此外，局部性信息被线程束调度器用来实现一个智能的重排序策略，用以获取局部性和隐藏延迟。实验结果表明，CWLP 能够提供高达 19.8%的性能加速比和超过基准策略平均 8.8%的性能提升。

关键词：局部性 GPU；cache 分配；线程束调度