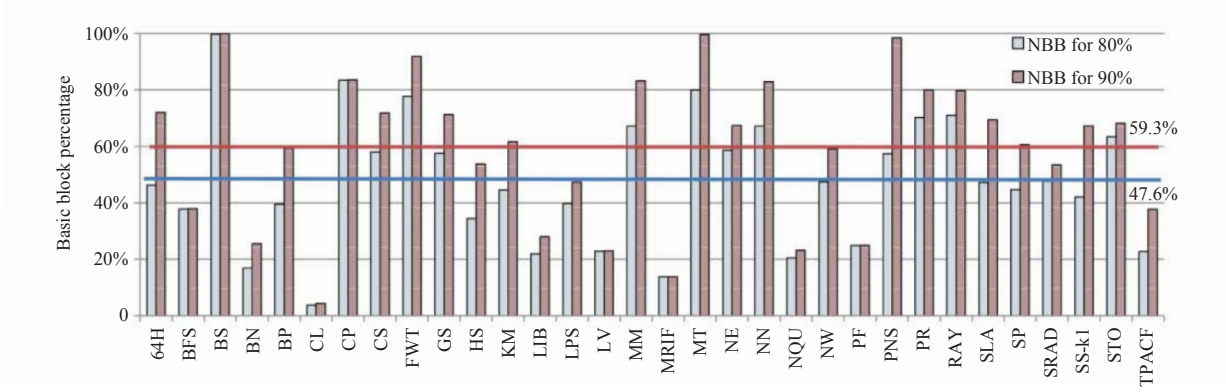
喻之斌研究员团队提出一种对 GPGPU 内核程序多级特征分析和优化的工具——MIC

中国科学院深圳先进技术研究院异构智能计算体系结构与系统研究中心喻之斌研究员团队主导的研究在 GPGPU 内核程序的多级特征分析和优化取得进展。相应成果为“Liu QX, Chen ZF, Yu ZB. MiC: multi-level characterization and optimization of GPGPU kernels [J]. ACM Journal on Emerging Technologies in Computing Systems (JETC), 2019, 15(3): 25(MIC: GPGPU 内核程序的多级特征分析和优化)”。

图形处理器通用计算(GPGPU)在并行程序和新计算模式下的应用使得 GPUs(图形处理器)越来越受欢迎，且随着实时处理大数据需求的不断增加，GPU 在为大数据分析提供有效解决方案的方面具有很大的潜力。然而，由于 GPGPU 集成了大量的处理器阵列和成千上万的执行线程，使得移动 GPU 的研制及其应用面临着巨大的挑战，并且目前还没有方法能够揭示 GPGPU 程序性能损失的根本原因。为此，该文提出了一种框架——MIC，它能够在指令级、基本块级和线程级对 GPGPU 内核进行全面的特征分析。该文通过分别设计了指令向量和基本块向量、线程相似矩阵和发流统计图来对各个层次的信息进行分析，并通过采用流行 GPGPU 测试程序集(如 CUDA SDK、Rodinia 和 Parboll)中的 34 个内核的描述对 GPGPU 内核进行了深入的研究。

结果发现：(1)与中央处理器(CPU)工作负载相比，GPGPU 内核程序有相当的指令级并行性；(2)GPGPU 基本块的数目明显小于 CPU 工作负载，平均仅为 22.8；(3)每个线程的动态指令数从几十条到几万条不等，与 CPU 测试程序相比非常小；(4)在 CPU 程序中普遍存在的 Pareto 原理(二八定律)不适用于 GPGPU 内核程序；(5)GPGPU 内核程序中的循环模式明显与 CPU 工作负载下的不同；(6)GPGPU 内核程序的分支比低于 CPU 程序但高于纯 GPU 工作负载。此外，该研究还通过对一个 GPGPU 内核的特征分析，给出了一个 GPGPU 内核的优化实例，结果显示性能提升了 16.8%。该研究所提出的 MIC 可为 GPGPU 架构师、编译器设计人员和开发人员提供一个简便的工具，同时该文研究成果对在移动设备上部署 GPGPU 并优化其性能具有重要意义。



在程序执行为 80% 和 90% 时的基本块级的占比图