Graphical Processing Units: multithreading and bottlenecks

The current programming of the GPUs consists of two distinct programming models: one for the host CPU, which is written in C/C++ with restricted C-like API for memory management; and the other for the GPU device, which is written using a device-dependent, explicit parallel programming model such as CUDA. The programming processing flow starts by copying input data from CPU memory to GPU DRAM memory. Then, the GPU program is loaded from the CPU and executed. Finally, the results are copied from the GPU DRAM memory to the CPU memory. Therefore, when an application running on a GPU does not meet the latency requirements, then it is consequence of both CPU and GPU performance. The reasons for not meeting those constraints may be that the CPU is bottlenecking the GPU or simply the GPU does not have enough processing power. If the number of cores increases on the GPU, then it is required that the size of the internal DRAM of the new GPU also increases to fetch more data to be processed in parallel. Otherwise, the latency requirements are not going to be fulfilled. Since the programming flow of the GPU depends heavily on the CPU, it is also required that the CPU increases its processing power with higher number of cores and bigger on-Chip memories with high bandwidths to satisfy the timing requirements of the application. As a consequence, only increasing the number of cores in the GPU is not a sufficient condition to reduce the latency of a running application.

Before acquiring a new GPU, it is necessary to test if the CPU/GPU configuration produces bottlenecks. To do so, there are commercial tools such as HWiNFO which quantifies how good the match between a CPU and GPU is. On the other hand, it is possible to estimate how is the GPU and CPU usage when the former is performing a task. If it is found that the CPU is using more than 90% of is resources, while the GPU does not reach that value of utilization, then it is clear that the CPU is bottlenecking the GPU.