Lecture 18 Summary

This lecture talked about Intel GPU performance debunking paper. Throughput computing applications is defined as an application with process deadline and a feature that it has plenty of data level parallelism and the data can be processed independently and in any order on different processing elements for a similar set of operations such as filtering, aggregating, ranking, and so on. Two major computing platforms are deemed suitable for this new class of applications. The first one is the general-purpose CPU (central processing unit). The second one is the GPU (graphics processing unit) that is designed for graphics processing with many small processing elements. Fundamentally, CPUs and GPUs are built based on very different philosophies: CPUs are designed for a wide variety of applications and to provide fast response times to a single task. Architectural advances, such as branch prediction, come at the price of increasing complexity/area and power consumption, and as a result, main stream CPUS can pack only a small number of processing cores on the same die to stay within the power and thermal envelops; GPUs are built specifically for rendering and other graphics applications that have a large degree of parallelism, and it trade off single-thread performance for increased parallel processing. Two platform used in this paper were Intel Core i7-960 CPU and Nvidia GTX 280 GPU. And in the performance analysis, there were 6 aspects to be considered. First, for the Bandwidth aspect, the performance ratio between GPU and CPU used in SAXPY, LBM, and SpMV were 5.3X, 5X and 1.9X. Second, in the compute flops aspect, the ratio in SGEMM, Conv and FFT were in the 2.8-4X range. MC has a ratio of 1.8X, and Bilat is better than 5X. In the Cache aspect, since the local buffer on a GPU is small, it required more time to sort on a GPU than a CPU. In the gather and scatter aspect, since Core i7 did not provide hardware gather/scatter support, GPU were better in this aspect. In the reduction and synchronization, GPU did not achieve 5X performance compared with CPU because of overhead on GPU. In the fixed function aspect, GPU was much better than CPU.