Heterogeneous computing with GPUs

Heterogeneous systems combining GPUs and CPUs are becoming mainstream. The performance potential of these processors working together is significant, yet most applications choose to use such systems as homogeneous, exploiting either GPUs or CPUs exclusively. This choice is due to the perceived imbalance between the high difficulty of deploying heterogeneous applications and the unknown performance gain.

The main goal of this tutorial is to demonstrate that implementing and deploying heterogeneous applications is simpler than expected, is supported by easy-to-use tools, and leads to significant performance gain for many applications.

Therefore, our tutorial has two main parts: (1) introduce the principles of heterogeneous computing models to sketch the current landscape, and (2) present in more detail the representative models that can efficiently cover most applications.

When discussing the principles and design of modern heterogeneous computing, we cover static workload partitioning, dynamic run-time based models, and hierarchical multi-node models. We further give representative examples from each class, focusing on their strengths and weaknesses. Further, we show how one can use workload characterization to match applications with the right heterogeneous computing model. Finally, we provide several examples to illustrate the usability and practical differences between tools.