Heterogeneous Computing

Kushagra Gupta, *2012MT50599*
Department of Mathematics
Indian Institute of Technology, Delhi.

Abstract— **Heterogeneous Computing refers to using different type of processors to compute the given problem. GPUs along with CPUs form the basis of heterogeneous architecture. It provides huge speedup. It is the most common architecture used nowadays. From smartphones to supercomputers, almost everything is using it. This paper discusses the basic of heterogeneous computing, role of GPUs, their energy efficiency and future prospects.**

Keywords— **CPU, GPU, Heterogeneous Computing, Parallel Computing.**

I. Introduction

WITH the uniprocessors hitting the power wall, we have now entered an age of low power multicore processors. An age where many of us are carrying devices with multiple cores and GPUs. The age of parallel computing.

Parallel computing refers to dividing the given large problem into many sub problems and computing them simultaneously on different processors thus reducing the time of computation. It is the foundation for modern day supercomputers, workstations and even the cell phones. The common type of parallel computing involves processors of same type and is called homogeneous computing. We all have computers with multicore processors. One problem with the homogeneity is that if the tasks at hand could not provide the required amount of homogeneous parallelism then the estimated speedup could not be met and in fact we are not utilizing the hardware to its full potential. Moreover increase in number of cores means bigger processors and increased energy consumption. So to increase the speed up we use heterogeneous computing. It involves separate (completely different architecture) processors for different type of tasks. For example GPUs are used in modern day machines for floating point computations.

If we need to increase the computational powers of our systems without compromising with the energy issues, heterogeneous computing up as one of the most feasible solution.

II. Components of Heterogeneous Systems

Processors nowadays use both CPUs and GPUs for effectively exploiting heterogeneous computing. So to understand how they facilitate the computation we must understand their basic structures.

The main aim of the CPU is to reduce the latency i.e. reduce the total processing time. For achieving this CPU have powerful ALUs (and hence more energy consuming), larger caches (so that time taken to access memory is as less as possible) and sophisticated controls for branch prediction, forwarding and error handling.

On the other hand a GPU focuses on increasing the throughput i.e. the number of tasks completed per unit time. For this it has a large number of cores (hundreds of cores) that perform multiple tasks together. It has many energy efficient ALUs that may have a higher latency but are heavily pipelined for increased throughput. It has small caches to boost the memory throughput and has small control units. Moreover GPUs operate at almost 10 times the memory bandwidth of a normal CPU. GPUs when used for general applications other than graphics are called GPGPUs (General Purpose GPUs).

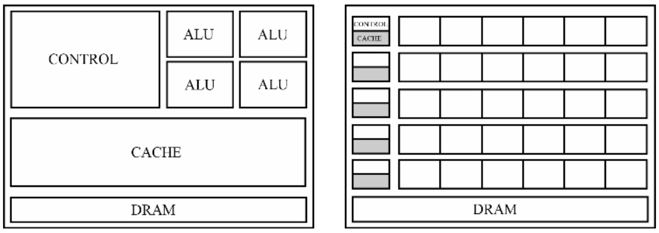


Fig. 1 Basic Architecture of CPU and GPU

For an application to computed at the maximum efficiency it needs to use both CPU and GPU as CPU could be efficient for sequential parts where latency matters and GPU could be beneficial for parallel parts where throughput matters.

Apart from CPU and GPU, other important processors used in devices are Digital Signal Processors (Microprocessors that are specially designed for signal processing. They are used for audio processing, object detection etc.), FPGA (Field Gate Programmable Array), Video Decoders etc.

III. Energy and Computing Efficiency

GPUs have enabled us to compute the same amount of data at very less energy. According to the TOP500 [4] and GREEN500 [5] the most efficient outputs i.e. performance per unit watt are all provided by computers using both CPUs and GPUs. When computing parallel instructions GPU could do it in a very less no of processors due to its throughput as compared to CPU and hence less cooling required. Currently sixth fastest computer in the world according to TOP500 is “Piz Daint” which is also the fourth most energy efficient supercomputer according to GREEN500. It is a modified version of its predecessor “Monte Roza” which didn’t used GPUs. Piz Daint has 20 times more computational power and uses 2.5 times less energy as compared to its predecessor.

Another example of energy efficiency of GPUs is when Bloomberg shifted its bond pricing application running on 2,000 CPUs to a 48 GPU rack of NVIDIA Tesla GPUs. The CPU system cost $4 million and $1.2 million in annual energy bills; the GPU one cost under $150,000, with about $30,000 yearly in energy as mentioned in [6].

GPU have provided a great boost in the computation power of the processors as could be seen from the figure 2. So heterogeneous computing could play a big role in reducing costs and at the same time increasing the computational power.

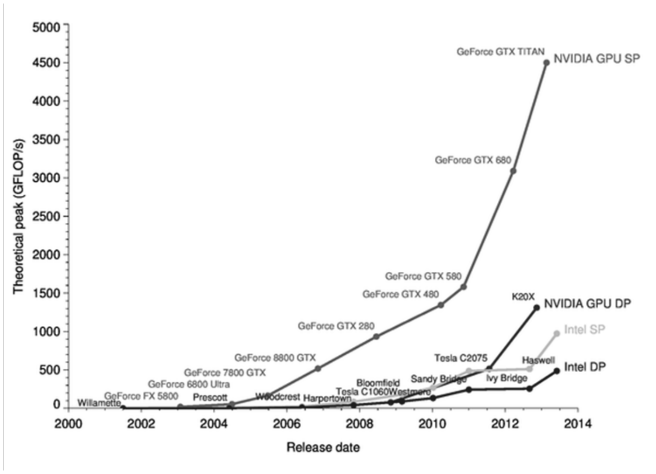
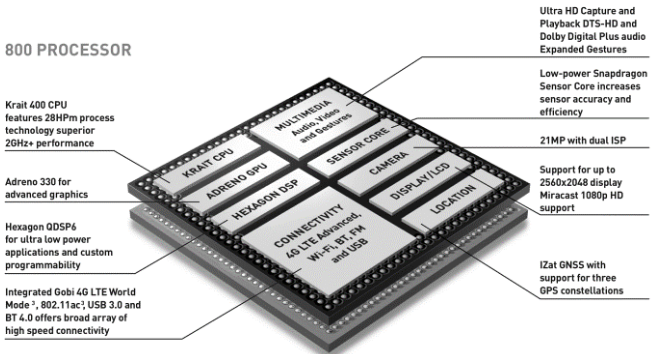


Fig. 2 Graph showing comparison between computation power of CPUs and GPUs [7]

IV. Heterogeneous Computing in a Smartphone : Qualcomm Snapdragon 800

Size of smartphones in current era is getting smaller and their abilities are growing. Continuous improvisation in the computing abilities without compromising with battery life is entirely possible due to heterogeneous computing. Due to the diverse needs of a smartphone such as realistic physics, computational photography, gestures, computer vision etc., a typical processor of a smartphone needs various computing units namely CPU, GPU, DSP, FPGA and separate special units for specialised tasks such as image processing etc.

Consider the Qualcomm Snapdragon 800 processor which is used in Google Nexus 5. It is one of the best processors in the smartphone market currently. It has a 28nm quad core Krait CPU (has 11 stage pipelining), Adreno 330 GPU (has 32 ALUs), Hexagon DSP and separate cores for Camera, GPS Sensors, Multimedia (Audio, Video and Gestures), Connectivity and Display/LCD. It enables multitasking and moreover energy efficiency as tasks such as face detection could be more efficiently computed using GPU rather than CPU as each pixel could be manipulated independently.



800 PROCESSOR
Krait 400 CPU features 28HPm process technology superior 2GHz+ performance
Adreno 330 for advanced graphics
Hexagon QDSP6 for ultra low power applications and custom programmability
Integrated Gobi 4G LTE World Mode ³, 802.11ac³, USB 3.0 and BT 4.0 offers broad array of high speed connectivity
KRAIT CPU
ADRENO GPU
HEXAGON DSP
MULTIMEDIA Audio, Video and Gestures
SENSOR CORE
CAMERA
DISPLAY/LCD
LOCATION
CONNECTIVITY 4G LTE Advanced, Wi-Fi, BT, FM and USB
Ultra HD Capture and Playback DTS-HD and Dolby Digital Plus audio Expanded Gestures
Low-power Snapdragon Sensor Core increases sensor accuracy and efficiency
21MP with dual ISP
Support for up to 2560x2048 display Miracast 1080p HD support
IZat GNSS with support for three GPS constellations

Fig. 3 Qualcomm Snapdragon 800 basic architecture [8]

V. Challenges

So far we have seen that heterogeneous computing is efficient and energy saving but it has its own challenges too.

The first challenge is to identify what kind of different processors we want to incorporate for given task and to what level we could exploit the heterogeneity. This is called machine selection.

When we have decided the components, we need a system so that the given task is properly divided and distributed among the various available processors. So we need new algorithms that could exploit the heterogeneity. But if some components are changed or new ones are added the entire algorithms may have to be changed or updated. We need portable and scalable applications (Portability is defined as the ability of an application to be run on different type of processors. Scalability is defined as the ability of a program to run on different versions of same type of processor) so that the software costs are reduced. Another challenge is the communication overheads. The amount of time taken for communication between various components is needed to be small. Other than these we need mechanisms for synchronisation between various units.

The problems discussed above are not really big problems and many of them have already been tackled by the HSA (Heterogeneous System Architecture)

VI. Conclusion and Future Prospects

So far in this paper we have discussed that the heterogeneous computing is increasing energy efficiency and improving computational abilities at every level, from smartphones to middle size computing to supercomputing. It is being facilitated by highly throughput oriented GPUs and specialized units. We have discussed the important role of GPUs and how are they different from traditional CPUs. But we saw there are some challenges faced

Heterogeneous Computing forms the basis for the next generation applications such as augmented reality, scientific simulations, ambient computing, speech recognition etc. So we conclude that the heterogeneous computing is the only energy-efficient answer to the increasing needs and versatility of applications.

References

[1] Khokhar, Ashfaq A., et al. "Heterogeneous computing: Challenges and opportunities." *IEEE Computer 26*.6 (1993): 18-27.

[2] “Heterogeneous System Architecture” http://developer.amd.com/resources/heterogeneous-computing/what-is-heterogeneous-system-architecture-hsa/

[3] “Heterogeneous Parallel Programming” https://class.coursera.org/hetero-002

[4] “TOP500 Project”, www.top500.org.

[5] “GREEN500 Project”, www.green500.org.

[6] “Carbon Disclosure Project – 2012 Investor Request” http://www.nvidia.com/docs/IO/124731/nvidia-2012-cdp-final.pdf

[7] Michael Galloy, “CPU vs GPU performance” http://michaelgalloy.com/wp-content/uploads/2013/06/cpu-vs-gpu-thumbnail.png

[8] “Qualcomm Snapdragon 800” http://www.qualcomm.com/sites/default/files/pods/tout/800-diagram-1152013.jpg

[9] David B. Kirk and Wen-mei W. Hwu, “*Programming Massively Parallel Processors*”. Elsevier 2010.