Third KAUST-NVIDIA workshop on Accelerating Scientific Applications using GPUs Tuesday, February 23rd, 2016

Program

08:30 - **Registration**

08:45 - 08:50 **Welcome!**
Dr. Saber Feki, *Workshop Chair, KSL &* Frederic Pariente, *NVIDIA*

08:50 - 09:10 **Opening Talk: Preparing for the Grand Convergence**
Prof. David Keyes, Professor, Applied Mathematics and Computational Science Director, Extreme Computing Research Center

09:10 - 10:10 **Keynote Talk: Deep Learning – the Killer App for GPUs**
Dr. Timothy Lanfear, *Manager, Solution Architecture and Engineering EMEA at NVIDIA: The World Leader in Visual Computing*

Although not a new idea, the interest in deep learning has exploded in the past few years because three key ingredients have come together: huge amounts of data is gathered by internet service providers; researchers are rapidly advancing the techniques used; and the GPU has been recognised as the ideal computing platform to deliver the enormous processing resources needed. We will explain some of the underlying concepts of deep learning, look at a few application areas, and review the hardware and software solutions that NVIDIA offers.

10:10 - 10:30 **Coffee Break**

10:30 - 10:55 **Detection Algorithms for Massive MIMO on a GPU Platform: A Performance-Complexity Tradeoff**
Dr. Zouheir Rezki, *Senior Research Scientist*, Computer, Electrical and Mathematical Sciences and Engineering Division, KAUST

The problem of detecting a large vector whose components are depicted from a lattice, from a noisy observation, in an efficient way is of crucial interest in communication. Applications of this problem include uplink communication in a cellular network where few tens of cellular users send their message to a base station equipped with a large number of antennas, two base stations with a large number of antennas, each, communicating in a backhaul mode, a set of sensors communicating with a multi-antenna data center (DC), to cite only few. While linear decoders are relatively simple to implement, they provide only a suboptimal error performance. On the other hand, nonlinear decoders such as maximum likelihood decoders (MLDs) or sphere decoders (SDs) are known for their high error performance, but their high complexity, and hence their high latency, prevents their deployment in real time applications.

In this talk, leveraging GPUs computation capabilities, we show that a SD, that outperforms the best linear decoder, can be implemented in real time complexity. Our approach provides a tremendous 12 dBs transmit power gain.

10:55 - 11:20 **Adaptive Optics Simulation for the World's Biggest Eye on Multicore Architectures with Multiple GPUs**

Dr. Hatem Ltaief, *Senior Research Scientist*, Extreme Computing Research Center, KAUST

We present a high performance comprehensive implementation of a multi-object adaptive optics (MOAO) simulation on multicore architectures with hardware accelerators in the context of computational astronomy. This implementation will be used as an operational testbed for simulating the design of new instruments for the European Extremely Large Telescope project (E-ELT), the world's biggest eye and one of Europe's highest priorities in ground-based astronomy.
The simulation corresponds to a multi-step multi-stage procedure, which is fed, near real-time, by system and turbulence data coming from the telescope environment. Using modern multicore architectures associated with the enormous computing power of GPUs, the resulting data-driven compute-intensive simulation of the entire MOAO application, composed of the tomographic reconstructor and the observing sequence, is capable of coping with the aforementioned real-time challenge and stands as a reference implementation for the computational astronomy community.

11:20 - 11:45 **KBLAS: Redesigning Triangular Dense Matrix Computations on GPUs**

Ali Charara, *PhD Student*, Extreme Computing Research Center, KAUST

A new implementation of the BLAS3 triangular matrix kernels are described on GPU hardware accelerators. We propose adopting a recursive formulation, which enriches these kernels (TRMM and TRSM) inner structures with GEMM calls. The new implementation enables efficient use of the GPU memory hierarchy and mitigates the latency overhead, to run at the speed of the higher cache levels. Performance comparisons show up to eightfold and twofold speedups for large dense matrix sizes, against the existing state-of-the-art respective implementations from NVIDIA cuBLAS, across various GPU generations. The new kernel implementations are part of the open-source KBLAS software library.

11:45 - 12:10 **High Performance Hierarchical Matrix-Vector Multiplication using Hardware Accelerators**

Wajih Boukaram, *PhD Student*, Extreme Computing Research Center, KAUST

We present a high performance hierarchical matrix vector multiplication using hardware accelerators. By properly mapping the tree structures to the GPU and overlapping the phases of the computation using streams, we greatly outperform the CPU implementations and achieve up to 80% of the sustained bandwidth of the GPU.

12:10 - 12:20 ***Group Photo***

12:20 - 13:30 **Lunch/Prayer Break**

13:30 - 15:00 **Tutorial: Introduction to OpenACC**
Brent Leback, *Service and Support Manager*, PGI/NVIDIA

15:00 - 15:20 **Coffee Break**

15:20 - 15:45 **Introduction to GPU Resources on Campus**
Niall O'Byrnes, *Research Applications Specialist,* IT Research Computing

In this session, we will introduce the GPU capabilities that are available and accessible on campus. We will describe the GPU hardware and the GPU-enabled applications supported by IT Research Computing. We will also go through scenarios of running MATLAB, VASP, ADF, GROMACS and LAMMPS on GPUs and we will explain how to best utilize these resources. After the presentation, the audience should have sufficient information to get started with GPUs. The available GPU hardware is based on NVIDIA Tesla K20 and K40 and are part of the Noor cluster.

15:45 - 16:10 **Enablement and use of Library device routines from within OpenACC compute regions**
Anas Almousa, *PhD Student,* King Fahd University of Petroleum and Minerals

Many Libraries have been written to port many of the common scientific methods into GPU devices. Moreover, automated parallelization technologies are catching up rapidly to make performance on these devices more accessible. Interoperability between automated parallelization technologies such as OpenACC and scientific libraries such as CuBLAS is still underrated by researchers in terms of easiness of use and performance. In this presentation, we shed the light more on the interoperability between OpenACC and libraries that provide the ability to be called from Device code (code regions that would be executing on device) . We show how to overcome some obstacles in the way of mixing OpenACC with libraries that hide parts of their implementation. Our tests achieved a speedup that reached 2.5 in some cases over the plain use of CuBLAS host based interface, while the speed up reached about 34 with respect to purely based OpenACC solution in some cases. Moreover, a decrease in code size of about 50% with respect to purely using OpenACC was noted.

16:10 - 16:35 **Chemistry Codes on GPUs: Case study of VASP**
Dr. Zhiyong Zhu, *Computational Scientist,* KAUST Supercomputing Laboratory

VASP is a code for atomic scale materials modeling from first principles, and is widely used in the areas of physics, chemistry, and materials science. The development of its GPU version has already entered the beta testing phase, in which several KAUST

research groups have participated. In this talk, I will present some preliminary results on the efficiency of the GPU version of VASP.

16:35 - 16:45 **Closing Remarks** -- Dr. Saber Feki, *Workshop Chair*