# Fourth KAUST-NVIDIA workshop on Accelerating Scientific Applications using GPUs
# Sunday, February 5th, 2017

## Program

*08:30* - **Registration**

*08:40 - 08:50* **Welcome!**

Dr. Saber Feki, *Workshop Chair, KSL* & Frederic Pariente, *NVIDIA*

*08:50 - 09:20* **Keynote Talk: State-of-the-art Machine Learning Algorithms and How They Are Affected By Near-Term Technology Trends**

Rob Farber, *CEO,* TechEnablement.com, USA

Industry and Wall Street projections indicate that Machine Learning will touch every piece of data in the data center by 2020. This has created a technology arms race and algorithmic competition as IBM, NVIDIA, Intel, and ARM strive to dominate the retooling of the computer industry to support ubiquitous machine learning workloads over the next 3-4 years. Similarly, algorithm designers compete to create faster and more accurate training and inference techniques that can address complex problems spanning speech, image recognition, image tagging, self-driving cars, data analytics and more. The challenges for researchers and technology providers encompass big data, massive parallelism, distributed processing, and real-time processing. Deep-learning and low-precision inference (based on INT8 and FP16 arithmetic) are current hot topics. This talk will merge two state-of-the-art briefings.

1) Massive scale and state- of -the art algorithm mappings for both machine learning and unstructured data analytics including how they are affected by current and forthcoming hard ware.

2) The technology trends at Intel (the Intel® Scalable Systems Framework including both Intel Xeon Phi Knights Landing and Knights Mill plus the Skylake Purely uArch), NVIDIA (Pascal GPUs including P100 training and the 8-bit arithmetic inference optimized GPUs), IBM (Power8/9 plus TrueNorth bee-brain on a chip), ARM and OpenPower that will affect algorithm developments.

The goal is to give attendees a sense of the fast-track algorithm + technology combinations for both research and commercial success as well as an overview of the state- of -the-industry and near-term industry directions

*09:20 - 9:50* **Unleashing GPU Computational Power to Tackle Multidisciplinary Research Projects @ECRC**

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

Come and learn how GPU technology helps enabling massive MIMO simulations, accelerating climate/weather model prediction, deploying the largest ever-built ground telescope, developing the fastest dense SVD solver, and promotes a new set of batched high performance linear algebra operations for machine learning. These various in-house and international projects are fostered at the Extreme Computing Research Center @KAUST in the context of the HiCMA and KBLAS projects.

*09:50 - 10:15* **ExaGeoStat: A High Performance Unified Framework for Geostatistics on Manycore Architectures**

*Dr*. Sameh Abduallah, *Post-Doc*, Extreme Computing Research Center, KAUST

The ExaGeoStat project is a high performance framework for geostatistics on manycore architectures. Based on a geospatial statistics model, ExaGeoStat allows to estimate and eventually improve the prediction of soil moisture, sea-surface temperature, etc., in the context of climate / weather applications. Such models engender large dense covariance matrices from multivariate spatial data sets. This approach may be considered as an alternative computational scenario to WRF simulations, while providing a much more efficient parallel implementation as opposed to the state-of-the-art computational statistics software R, for which performance issues exist for large-scale simulations.

*10:15 - 10:30* **Coffee Break**

*10:30 - 10:55* **Scaling Computer Vision and Numerical Optimization with GPUs**

*Victor Escorcia, PhD Student,* Visual Computing Center, KAUST

In the last couple of years, there has been an explosion in the use of GPUs for certain computer vision and numerical optimization tasks. At KAUST, IVUL has adopted GPUs to enable some demanding computational tasks. In this talk, we will dive into how we take advantage of the computational power of GPUs to successfully recognize activities in videos from Youtube as well as localize and identify faces on consumer photographs. Moreover, we are going to appreciate how classical optimization problems such as LASSO can benefit tremendously from GPU acceleration without traditional BLAS routines or linear system solvers.

*10:55 - 11:20* **Introduction to KSL**

Dr. Jysoo Lee, *Director*, Supercomputing Core Laboratory, KAUST

*11:20 - 12:00* **Keynote Talk: NVIDIA — The AI Computing Company**

Dr. Timothy Lanfear, *Director, Solution Architecture and Engineering, EMEA,* NVIDIA

Artificial intelligence is the use of computers to simulate human intelligence. Learning from data — a computer's version of life experience — is how AI evolves. GPU deep learning is a new computing model in which deep neural networks are trained to recognize patterns from massive amounts of data. This new model has set off a string of "superhuman" achievements in image and speech recognition and sparked the era of AI computing. NVIDIA offers a complete solution to the AI computing problem comprising an end-to-end hardware product family deployable in any situation; software optimised to run on NVIDIA platforms; and in-depth expertise in deep learning and artificial intelligence.

*12:00 - 12:10* ***Group Photo***

*12:10 - 13:30* **Lunch/Prayer Break**

*13:30 - 14:00* **Keynote Talk: PGI Compilers for Heterogeneous Supercomputing - 2017 Update**

Brent Leback, PGI Customer Service Manager

PGI has been providing the HPC market with compilers and tools for over 25 years. In the last year PGI has delivered a production release of our compilers for OpenPOWER + Tesla, made a no-cost, full-featured PGI Community Edition available for free download, and enabled Pascal P100 efficiency gains from OpenACC and CUDA Fortran. In addition we've added a multicore target to OpenACC and added features and improvements enabling a number of new ISVs and community applications both portability and performance when targeting GPUs.

*14:00 - 14:25* **Asynchronous Task-Based Polar Decomposition on Manycore Architectures**

Dalal Sukkhari, *PhD Student*, Extreme Computing Research Center, KAUST

This talk introduces the first asynchronous, task-based formulation of the polar decomposition and its corresponding implementation on manycore architectures. Based on a new formulation of the iterative QR dynamically-weighted Halley algorithm (QDWH) for the calculation of the polar decomposition, the proposed implementation replaces the original and hostile LU factorization for the condition number estimator by the more adequate QR factorization to enable software portability across various architectures. Relying on fine-grained computations, the novel task-based implementation is also capable of taking advantage of the identity structure of the matrix involved during the QDWH iterations,which decreases the overall algorithmic complexity. Furthermore, the artifactual synchronization points have been weakened compared to previous implementations, unveiling look-ahead opportunities for better hardware occupancy. The overall QDWH-based polar decomposition can then be represented as a directed acyclic graph (DAG), where nodes represent computational tasks and edges define the inter-task data dependencies. The StarPU dynamic runtime system is employed to traverse the DAG, to track the various data dependencies and to asynchronously schedule the computational tasks on the underlying hardware resources, resulting in an out-of-order task scheduling. Benchmarking experiments show significant improvements against existing state-of-the-art high performance implementations (i.e., Intel MKL and Elemental) for the polar decomposition on latest shared-memory vendors'systems (i.e., Intel Haswell/Broadwell/Knights Landing, NVIDIA K80/P100 GPUs and IBM Power8), while maintaining high numerical accuracy.

*14:25 - 14:50* **Batched Triangular DLA for Very Small Matrices on GPU**

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

Batched dense linear algebra (DLA) kernels are ubiquitous in scientific applications, like tensor contractions in deep learning and data compression in hierarchical low rank matrix approximation. Batch calls remove the expensive overhead of multiple API calls while increasing the occupancy of the underlying hardware. We describe the design and performance of a new class of batched triangular DLA kernels on very small data sizes using NVIDIA GPUs. By deploying recursive formulations, stressing the register usage, maintaining data locality, and reducing threads synchronization, thanks to CUDA shuffle instructions, the new batched kernels outperform existing state-of-the-art implementations on both x86 CPU and GPU architectures.

*14:50 - 15:05* **Coffee Break**

*15:05 - 15:30* **Manycore Implementations of a Real-Time Tomographic Reconstructor for the European Extremely Large Telescope**

Nicolas Doucet, *PhD student* Observatoire de Paris, France, in collaboration with Extreme Computing Research Center, KAUST

Multi-object adaptive optics (MOAO) is a novel adaptive optics employing deformable mirrors that applies dedicated wavefront corrections to numerous separated tiny patches spread over a field of view as large as that of the telescope, itself. Each mirror is controlled individually using a tomographic reconstruction of the phase based on measurements from a number of wavefront sensors (WFS) pointing at natural and artificial guide stars in the field. A key step of the tomographic reconstruction consists of a pseudo-inversion of a large (10^5 to 10^6 dimensional) dense symmetric matrix, which is the rate-limiting step in the control. For geographically remote telescope applications, a conventional power-hungry supercomputers is not a solution, but contemporary energy-efficient manycore accelerators can be optimized for this regular and recurring task. The OdP-KAUST implementation employs a task-based programming model with a scheduler and hardware accelerators, such as KNLs or GPUs including DGX-1. We examine different implementations and their trade-offs, including the use of hierarchically low-rank approximation methods.

*15:30 - 15:55* **Fast Batched SVD on GPU**
Wajih Boukaram, *PhD Student*, Extreme Computing Research Center, KAUST

Hierarchical matrices are an efficient way for storing the dense matrices of very large dimension that appear in the discretization of integral operators associated with elliptic PDEs, in Schur complement methods exploiting dimension reduction, in spatial statistics and computational astronomy when describing pairwise relations between data points, etc. They exploit the fact that the underlying matrices, while formally dense, are data sparse. They have a structure consisting of blocks, many of which can be well-approximated by low rank factorizations, resulting in compressing the dense matrix in an accuracy-controlled manner. Singular value decomposition of these very small hierarchical blocks is a crucial operation and should be optimized by ensuring high occupancy on throughput-oriented manycore architectures. This representation can avoid superlinear growth in memory requirements to store $n \times n$ dense matrices in a scalable manner requiring $O(n)$ units of storage with a constant depending on the representative rank $k$ for the low rank blocks.

*15:55 - 16:10* **Introduction to GPU Resources on Campus**
Antonio Arena, *Research Computing Lead,* IT Research Computing, KAUST

*16:10 - 16:20* **Closing Remarks** -- Dr. Saber Feki, *Workshop Chair*