

# THIRD PARADIGM AT KAUST

## النموذج الثالث في جامعة الملك عبد الله للعلوم والتقنية

كانت النظرية و التجربة يعتبران الركبتين الأساسيين للعلم، ثم أصبحت المحاكاة الآن ركنًا ثالثًا حيث نجد أن الحوسبة العالية الأداء تُسرّع من عجلة الاكتشاف و الابتكار. كما يقدم كل من شاهين «الحاسوب العملاق» وفريق مختبر الحوسبة الفائقة من باحثي الحوسبة ومديري النظم للباحثين والمتعاونين الفرصة لتشغيل عمليات المحاكاة التي لم تكن حتى في الخيال قبل عقدين من الزمن.

Centuries ago scientists worked independently; they made observations, calculated, theorized and documented their observations and theories in notebooks. Over the next several hundred years very little changed. But scientific discovery has undergone profound transformation over the last four decades – from a culture of science with minimal data to science that works with increasingly large amounts of data. This data is shared and enriched by collaboration among scientists from multiple disciplines. The cross-fertilization of ideas and expertise often leads to a better understanding of complex physical phenomena scientists are attempting to analyze today.

High performance computing is accelerating discovery and innovation – where theory and experimentation were once the two pillars of science – simulation is now the third. It complements, informs, and when experiments prove too dangerous, expensive, or are otherwise prohibitive, computational science and engineering research can replace experimentation. Therefore, the third paradigm is the use of computer simulation for new scientific research. After proper verification and validation, simulation in high fidelity can achieve predictive abilities and help us in global design optimization with huge economic impact.

Thanks to relatively inexpensive data storage systems, broadband networks, and sensors that are

continuously monitoring the earth from high up in space to deep under the sea; a virtual avalanche of data, constantly growing in volume and diversity, is both a blessing and a curse. It holds the promise of significant breakthroughs in scientific discovery for years to come and yet, in order for it to be effectively tapped, it must be properly managed and efficiently manipulated.

Researchers now have the power to look at correlations in existing data and consider the relationships between data in multiple domains. The big challenge is how to analyze the enormous amount of data available within a practical time frame (even with a supercomputer).

Researchers in KAUST's Mathematical and Computer Sciences and Engineering (MCSE) Division mostly create theoretical techniques or software implementations that add capabilities to the computational ecosystem – they are enablers. The computational researchers in KAUST's other two divisions complete the ecosystem as users with software applications. They run ensembles of large simulations or process large batches of data.

KAUST enablers include people like Dean of MCSE, Prof. David Keyes and Research Scientist Xiangliang Zhang. For the past two decades, Prof. Keyes' research has responded to the growing gap between the complexity of the applications that scientists would like

to run and the complexity of the architecture of high-performance computers. Application codes typically have requirements that can be expressed as mathematical abstractions, such as "solve a linear system with this matrix" or "find the normal modes of this system". Scientists who make these demands on the hardware should be spared the details that their matrix needs to be spread out over tens or hundreds of thousands of individual computer memories, connected to similar numbers of processors at the ends of a complex network that routes information through the computer, and ultimately on and off of storage units and display peripherals. Prof. Keyes has developed solution algorithms for the distributed-hierarchical memory message-passing supercomputers that have dominated simulation for the past two decades, during which computational capabilities have been improved by a cumulative factor of about one million, at fixed cost.

Dr. Zhang's work is motivated by data streaming applications, which are characterized by data that typically must be processed in real time, and cannot be stored in its entirety or revisited, as it is too voluminous. She employs statistical methods to look for patterns in such data, to detect events of interest. Such events could be scientific in nature, such as the analysis of observations from spectrometers, telescopes, sensors, etc., but the same tools apply to many

domains outside of science, such as intrusion detection in computer networks, or the supply of electricity in response to the time-varying demands of millions of users and hundreds of suppliers.

According to Prof. Keyes, the software that he and Dr. Zhang develop allows them "to partner with a large variety of scientists and engineers – an interdisciplinary richness that we and others in the MCSE Division treasure about our work at KAUST!"

Whether collaborating across divisions or across the KAUST Global Collaborative Research network, the one constant is the team of computational scientists within the KAUST Supercomputing Laboratory who provide the expertise to assist domain experts to optimize their software for Shaheen and support projects from start to finish. Their reputation is quickly spreading worldwide as the popularity of their exhibit at the SC10 Supercomputing Conference in New Orleans last November can attest. Not only were those watching demonstrations and lectures interested in possible faculty, staff, or student openings, but many were also considering collaborative projects on Shaheen. □



A synthetic velocity model of an oil and gas reservoir under a salt body

## SEISMOLOGY

Seismic imaging estimates the earth's rock properties by deciphering seismic data recorded on the earth's surface. For example, earthquake seismologists tomographically decipher traveltimes from earthquakes to understand the geologic history of our planet, helioseismologists invert sunquake measurements to understand the internal physics of the sun, and exploration seismologists use controlled source seismic data to identify oil, gas, and mineral deposits. Without such technology, there would be more than an order-of-magnitude fewer natural resources available for today's advanced civilization.

Recently, KAUST researchers in the earth science program have co-pioneered multisource seismic imaging to significantly expedite the inversion of seismic data. Instead of sequentially inverting one seismic record at a time, they encode and combine all seismic data

together and invert the combined data in one step. This is analogous to listening to simultaneous conversations at a cocktail party, and quickly being able to decipher all conversations at the same time. Dr. Chaiwoot Boonyasiriwat, and Ph.D. students Wei Dai, Yunsong Huang, Xin Wang, and Ge Zhan, are leading the KAUST efforts to further develop and broaden this technology with the help of KAUST Prof. Gerard Schuster and their Academic Excellence Alliance (AEA) partner Prof. Paul Stoffa at University of Texas at Austin. In particular, the KAUST and AEA researchers have developed parallel modeling and inversion codes on Shaheen to validate and refine the multisource inversion technologies. These codes have been successfully tested for 2D and 3D synthetic data, with current efforts aimed at inverting land data and marine data from the Red Sea. These research efforts would not be possible without the computational power of Shaheen. □

## BIOSCIENCE

Biology is generating massive complex data sets and computation is increasingly important to understand and extract useful information from them. Increased size and capability of high-end computer systems make it possible, for instance, to quickly compare new molecular sequences against all known DNA and protein data so far generated.

The new field of metagenomics, in which genome sequence segments from an entire environmental sample are determined, has revolutionized our approach to studying ecology and population genetics. Recently, the 2010 KAUST Red Sea expedition was carried out. This, in collaboration with the American University of Cairo, generated 12% of all current metagenomic data in one experiment. The metagenomic analysis of these seven million new sequences would have taken over a year to

run on a standard computer server. When Senior Research Scientist, Dr. Heikki Lehvaslaiho, and Research Scientist, Dr. Intikhab Alam in the Computational Bioscience Research Center needed to analyze these data, they turned to the computational capabilities of Shaheen which allowed them to complete this task in just weeks.

Dr. Lehvaslaiho and Dr. Alam were able to take advantage of work done at Virginia Tech (VT) on a special version of the well known Basic Local Alignment Search Tool (BLAST) which is one of the standard tools for identifying regions of local similarity between biological sequences. The group at VT, together with others, has developed a "parallelized" version of the BLAST code called mpiBLAST that could run on BlueGene/P systems. At KAUST, they adapted this code to run efficiently on Shaheen and made a pipeline which allows mpiBLAST to automatically undertake many different tasks in the future. □

