DOI:10.1145/3447737

BY DAVID KEYES

The Arab World Prepares the Exascale Workforce

THE ARAB WORLD is currently host to eight supercomputers in the Top500 globally, including the current #10 and a former #7. Hardware can become a honeypot for talent attraction—senior talent from abroad, and rising talent from within. Good return on investment from leading-edge hardware motivates forging collaborative ties to global supercomputing leaders, which leads to integration into the global campaigns that supercomputing excels in, such as predicting climate change and developing sustainable energy resources for its mitigation, positing properties of new materials and catalysis by design, repurposing already-certified drugs and discovering new ones, and big data analytics and machine learning applied to science and to society. While the petroleum industry has been the historical motivation for supercomputing in the Arab World, with its workloads of seismic imaging and reservoir modeling, the attraction today is universal.

However, it is not sufficient to install and boot supercomputers. Their purpose is performance, and their acquisition and operating costs are too high to use them any other way. In each phase of computation, the limiting resource must be identified and computation reorganized to push the bottleneck further away, ideally guided by a performance model. The soul of the machine is the software: the distributed shared memory data structures, the task graphs, the communication patterns. The software is generally not performance-portable; it must be re-tuned in each application-architecture context. As applications become more ambitious and architectures become more austere, algorithms and software must bridge the growing gap.

Hands-on opportunities to resolve this application-architecture tension are a lure to students who had not previously considered supercomputing careers. In some cases, they must surmount significant hurdles in mathematical or computational preparation to enlist, but enlist they do, and they often wind up in globally leading institutions upon graduation. The stories in this article grew up around a university-operated supercomputer, but they largely can be replicated with much less investment because the main challenges today are not in coordinating tens of thousands of nodes across a low-latency, high-bandwidth network. Rather, the challenges lie in extracting performance *from within* increasingly heterogeneous nodes. Furthermore, the cloud now provides high-performance computing (HPC) environments. It is estimated that the percentage of HPC jobs run in the public cloud nearly doubled from 10%–12% in 2018 to 20% in 2019.[4] Many supercomputers, including the currently #1-ranked Fugaku (featuring ARM-based Fujitsu A64fx) and #2-ranked Summit (featuring

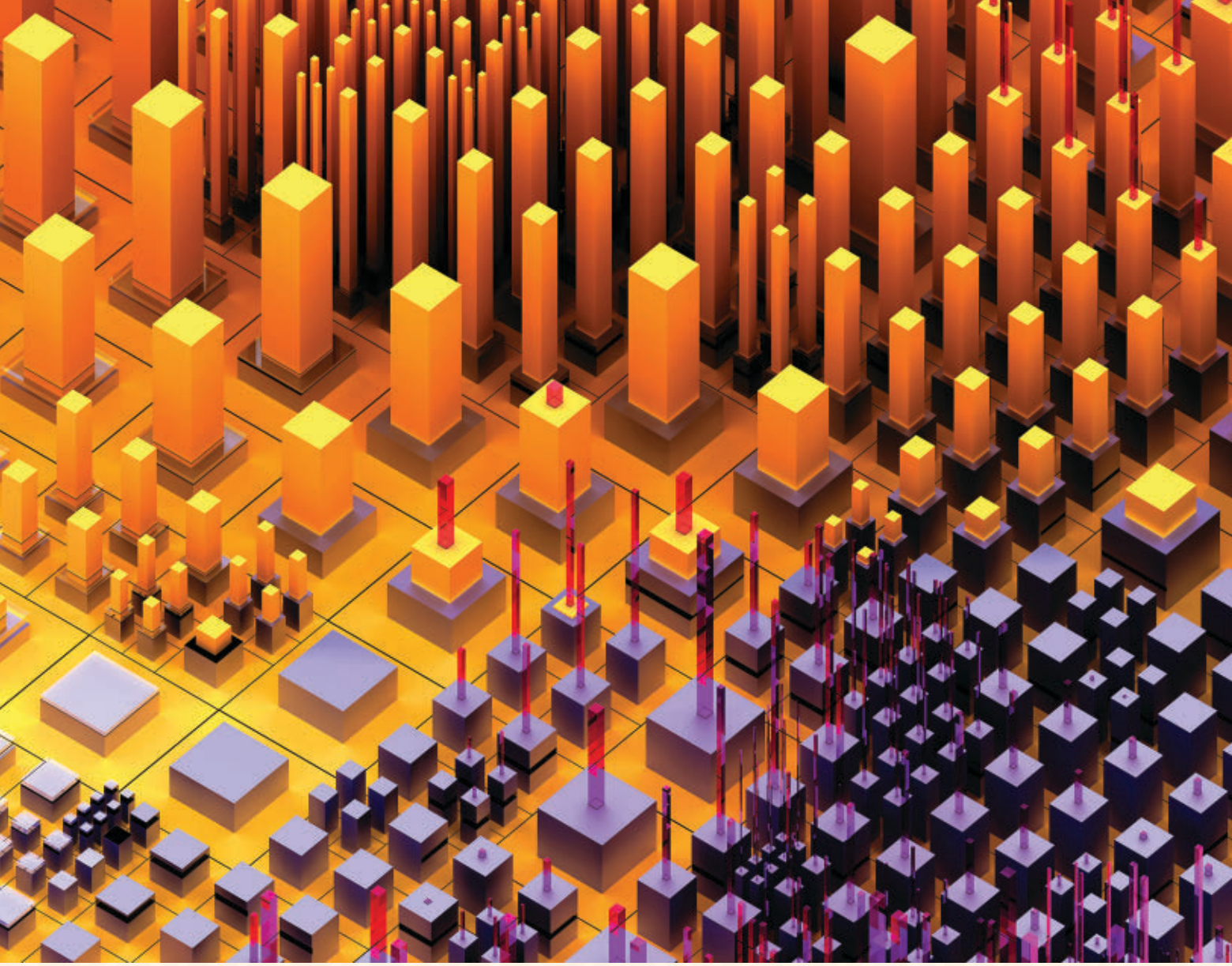


IMAGE BY PHILIPP TUR

NVIDIA V100 and IBM Power9), offer small competitively awarded research accounts at no cost. (Disclosure: the author's team presently employs accounts on each.) We argue that a menu of *Awareness*, *Examples*, *Instruction*, *Opportunity*, and *Utilization* will *Yield* members of the exascale workforce, mnemonically: *A*, *E*, *I*, *O*, *U* and sometimes—hopefully often(!)—*Y*.

From the Middle East to the Best of the West

The basis for confidence in this five-fold agenda for preparing students for the heterogeneous environments of exascale computing is empirical: nine of the author's earliest Ph.D. students from the King Abdullah University of Science and Technology (KAUST, founded in 2009 in Saudi Arabia) received as their first job offer a U.S. DOE-funded post-doc, either in NERSC's NESAP, NNSA's PSAAP, or the agency-wide ECP. All of them held their U.S. DOE-based job offer before they defended their dissertations. Three Saudi Ph.D.s beyond the nine sought by the U.S. DOE elected to stay at home and grow their careers with their increasingly information-based national economy. One joined Boeing and another joined NEOM, the futuristic green city along the Gulf of Aqaba billed as "an accelerator of human progress" that has 10 times the land area of Hong Kong. The last joined a digital start-up he had co-founded on the side as a student, already with Series A financing of over USD $2 million. Eleven of these 12 students hail from the Arab World: from Egypt, Jordan, Lebanon, Saudi Arabia, and Syria. Some students had their software integrated into NVIDIA's cuBLAS or Cray's LibSci before they graduated, and one had their software integrated into a prototype of Saudi Aramco's next-generation seismic inversion code; integration into NEC's Numeric Library Collection is pending.

Of these 12 HPC Ph.D.s, four are women. All but two completed their bachelor's degrees in Arab World universities in departments of computer science, computer engineering, or information science. Only one was proactively recruited into an HPC-oriented fellowship at KAUST. The others came from a globally cast net offering doctoral fellowships to study computer science or applied mathematics more generally and were lured to supercomputing by its opportunity.

Two of the 2020 graduates won major conference awards for papers based on their thesis work: Noha Alharthi lead-authored *Solving Acoustic Boundary Integral Equations using High Performance Tile Low-Rank LU Factorization*, which was awarded the Gauss Center for Supercomputing Award in

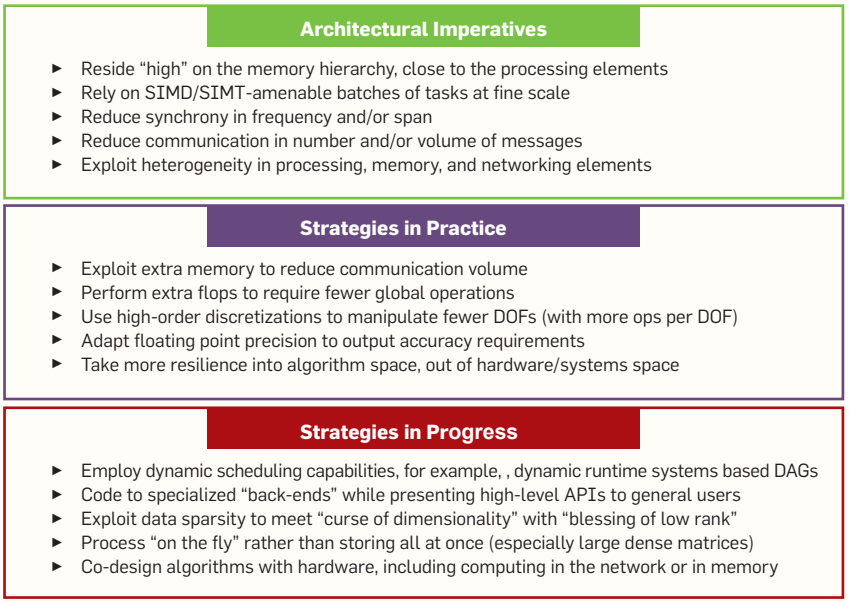
June 2020 at the (virtualized) 35th International Supercomputing Conference (ISC'20), and Tariq Alturkestani lead-authored *Maximizing I/O Bandwidth for Reverse Time Migration on Heterogeneous Large-Scale Systems*, which was awarded Best Paper at the (virtualized) 26th Euro-Par Conference in September 2020. Each paper is interdisciplinary: Alharthi's spans the discretization of singular integral equations, massively distributed data-sparse linear algebra, and acoustic scattering from irregularly shaped bodies, while Alturkestani's generalizes 2-level cache protocols to *N*-level memory hierarchies for hiding pre-fetching and write-back times of the huge datasets of reverse-time migration on massively distributed systems (including globally #2-ranked supercomputer Summit) for seismic imaging of petroleum deposits. Topics pursued by the graduates who were recruited to U.S. DOE-funded post-docs include: dense and hierarchically low-rank linear algebra kernels implemented on graphics processing units (GPUs), singular value decomposition (SVD) and eigensolvers implemented on massively distributed memory systems; high-order stencil update protocols for Cartesian lattices implemented on many-core central processing units (CPUs) with shared caches; a many-core implementation of an unstructured-grid implicit external aerodynamics code, and a fast-multipole method preconditioned boundary-integral equation solver.

Following their DOE-sponsored post-docs, two of the U.S.-based alumni moved to Intel, one to NVIDIA, and one to a machine learning start-up in the Bay Area; the rest remain in DOE Exascale Computing Project (ECP)-funded research positions. Not all of them dealt with heterogeneity as a first-class consideration for their theses, but many implementations were GPU or hybrid. While all scaled in a performance-oriented way to distributed memory, the main contributions took place within a node. They were immersed in a roofline modeling culture and became fluent with DAG-based dynamic runtime programming. Hands-on, they compared as many vendor platforms as were available locally and then gained access to guest accounts abroad by sharing tantalizing locally generated results. They also compared against the prior state of the art on a given platform, whether libraries like MKL, CuBLAS, PLASMA, or MAGMA, or runtimes like PaRSEC, QUARK, StarPU, or OpenMP-LLVM. All but a couple released their codes at github.com/ecrc, where the visibility of one code calls attention to another.

Facing the 'Universals' of Exascale Computing

As part of their inauguration into research, each student was presented with a list of "universals" for exascale computing that cuts across most applications. The list has been growing over the past few years to include the 15 grouped in Figure 1 into five architectural imperatives, five strategies already widely in practice, and five strategies in progress. Each student was asked to identify a research contribution among these "universals" and to adopt a particular demanding application to keep the work practically motivated, typically through a co-advisor. From our experience with this cadre of students, we advocate the following five principles for equipping the next-generation exascale workforce.

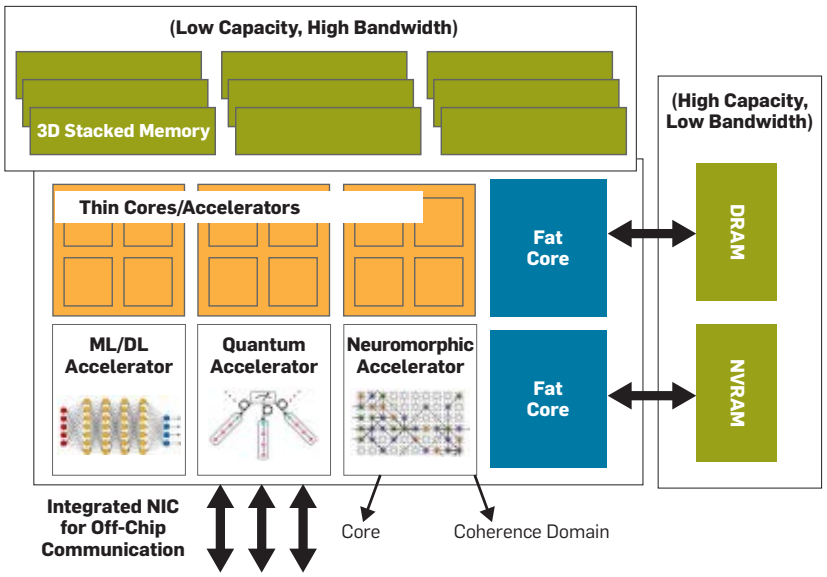
Figure 1. Fifteen "universals" of exascale computing.



Awareness of heterogeneity should be emphasized from the beginning. Heterogeneity of processing, memory, and network elements is now the norm, driven by opportunities for energy efficiency for specialized instructions, such as the $D <- A*B + C$ 4x4 matrix-multiply-and-add in convolutional neural networks that does 64 FMADD operations in one instruction. Increasingly, performance-oriented programmers will make choices about in which memories to stash their data structures, and how to route their data, possibly doing operations like transposes or reductions of the data *en route*. In Figure 2, we adapt a figure from a DOE report on exascale architecture[1] by adding deep learning, quantum, and neuromorphic elements. While a quantum device likely will need to be off-board for cryogenic engineering purposes, students should, for example, recognize that an unconstrained optimization step in a large scientific code may be ideal to offload to such a device in the future to quickly examine billions of random possibilities and return one that, with high probability, is within a tight tolerance of the optimum. Programmers of the exascale era have to think about *what* runs best *where*.

Examples should be provided. Students should read success stories about applications that profit from heterogeneity and how; for example, ACM Gordon Bell Prize finalist papers, like the 2019 OMEN code with its data-centric DaCE programming model,[5] or the 2020 DeePMD-kit with its use of machine learning to replace inner loops of expensive floating-

Figure 2. Augmented example of a heterogeneous node from Ang et al.[1]



point function evaluations of *ab initio* electronic structure calculations.[3] Today, the examples employ vector units, GPUs, tensor processing units (TPUs), and field-programmable gate arrays (FPGAs); in mainstream scientific campaigns, more neuromorphic and quantum devices may soon be relevant. Memory systems stretch from registers, through multiple levels of cache with varying nestedness, to HBM, DRAM, NVRAM, local disk, and federated data bases. Communication channels range from direct optical, through copper, to optical fiber. For petabyte datasets, users need to consider whether they should use a courier service, or leave data globally distributed and manage it as a federated entity.

Instruction should be given on two levels. High-level multidisciplinary thinking estimates thresholds for using a technique that amortize the overheads, and how to recognize amenable kernels in applications. Low-level training in how to express scheduling, data placement, and heterogeneous targets such as vector extensions, CUDA, and libraries for remote operations, is also important. Syntax often can be taught outside of credit-bearing courses, such as through hosting vendors for weekend tutorial/hackathons, while the conceptual parts belong in proper courses.

Opportunity to experience development at the cutting edge motivates and equips. Sometimes, this is best accomplished in a three-month to six-month internship. The students described here interned at mission-oriented research labs like Argonne; academic institutions with performance expertise we needed like Erlangen, home of the LikWid performance tools,[a] HPC vendors like NVIDIA, and HPC customers like Saudi Aramco. In some cases, the thesis topic arose from the internship advisor. In other cases, the application that motivated the algorithmic innovation or implementation of the thesis was mastered in the internship.

Utilization is the ultimate goal: hands-on code development as part of the thesis, ideally plugging into a multidisciplinary team so the specialist effort is part of something bigger and real-world, like clean combustion, vehicle aerodynamics, or geospatial statistical inference of the weather. The application motivates, may lead to sponsorship, and brings visibility beyond the algorithmic and software accomplishments.

Fashioning a Computational Mecca

Without question, the lure of top Arab World talent to the opportunities of exascale computing that are being established beyond *and within* the

The soul of the machine is the software: the distributed shared memory data structures, the task graphs, the communication patterns.

a https://github.com/RRZE-HPC/likwid.

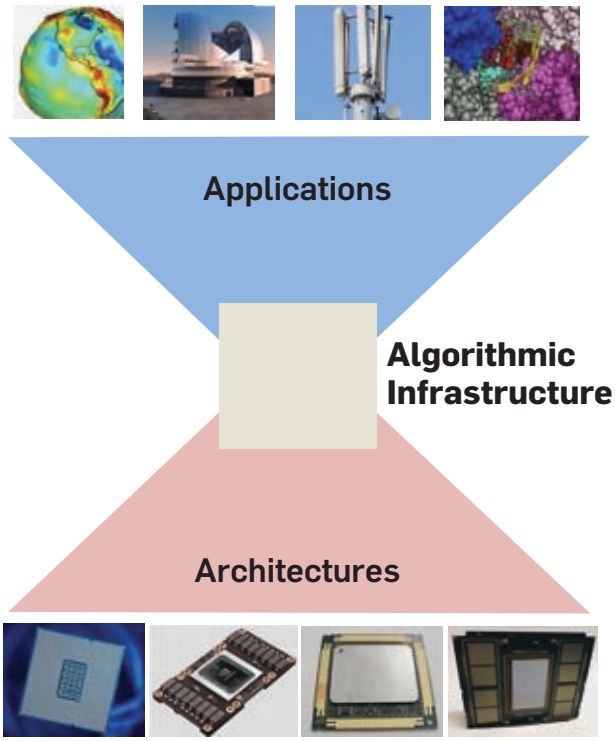
The function of the ECRC is best understood in terms of the 'hourglass model' of software, a concept borrowed from the TCP-IP philosophy.

Arab World itself is due in part to access to a petascale supercomputer as a technological stepping stone and a sheer source of inspiration. However, universities need not have the means to bring a supercomputer to their campus to participate. Besides access to HPC in the commercial cloud (which shifts the expense from capital to operating), many of the more than 500 petascale supercomputers in the world in the hands of universities or national laboratories offer exploratory grants at no cost, including to offshore collaborators, and some also offer summer training programs in hopes of attracting a future workforce. More importantly, as mentioned earlier, the most significant bottlenecks to performance scalability now lie within the individual (often heterogeneous) nodes, meaning that a modest collection of the latest processor offerings from providers such as AMD, ARM, Fujitsu, IBM, Intel, and NVIDIA put experimental proofs of concept within reach.

The Extreme Computing Research Center (ECRC), which sponsored the students discussed earlier, was created outside of KAUST's 16 degree programs as one of 14 mission-oriented research centers. The Centers create critical mass beyond the capacity of individual faculty members to encourage translating basic research into translational ends. They induce faculty and students from the degree programs to their missions with expertise, centrally supplied competitively awarded funding, space, research facilities, and reputation. The Centers support a small number of experienced research scientists. In the case of the ECRC, these are professional software engineers who, together with the faculty, contribute beyond the university to such industry-standard open-source libraries as PETSc, MFEM, SPECFEM3D, ug4, CLAWPACK and pyCLAW, mpi4py, and OpenFOAM.

The function of a center dedicated to software infrastructure is best understood in terms of the "hourglass model" of software, in Figure 3, a concept borrowed from the TCP-IP philosophy:[2] many diverse scientific applications (the top of the hourglass) are enabled to run with high performance on many diverse computer architectures (the bottom of the hourglass)

Figure 3. An hourglass model for scientific software.



through a standard interface (the neck of the hourglass) implemented as callable software libraries whose purpose is to partially hide the complexity and diversity of the architecture. The role of such a center becomes more interesting as architectures evolve under the premium of energy efficiency to become more specialized to certain tasks, thus presenting a host of heterogeneous resources in processor, memory, and network components. The diversity of applications here refers both to domain subject matter, from seismic imaging to genome-wide association studies, and to technique, from simulation based on first-principles models to machine learning, where first-principles models are not readily constructed but input-output maps can be learned from data.

There are manifold ways to improve a scientific computation, such as: increase its *accuracy* (computational resolution of an underlying continuum); increase its *fidelity* (inclusion of a system's full features in a computational model); tighten its *uncertainty* (bound the error of a model's output in terms of errors in its inputs); and reduce its *complexity* (computational costs, in terms of storage and operations) to achieve a sought accuracy, fidelity, and confidence. Modelers generally customize the first three to their application and are happy to hand off the fourth, complexity reduction and architectural tuning, as a productive separation of concerns. KAUST's ECRC focuses its resources on complexity reduction and architectural tuning for widely used computational kernels in simulation and data analytics.

Ph.D. students can become a source of widely distributed software for the implementation of efficient algorithms for simulation and data analytics on high-performance hardware by facilitating the transition to algorithms that exploit a hierarchy of scales, such as multigrid, fast multipole, hierarchical low-rank matrices, and hierarchical coarsenings of graphs. Hierarchical algorithms are much more efficient than their traditional "brute force" counterparts, but also more complex to implement because of their nonuniformity of scales. These algorithms exploit the mathematics of "data sparsity," architecture-specific instruction-level concurrency, and they aim to reduce communication and synchronization, the latter much more expensive than operations on cached data. As a further step, algorithms that exploit randomization, such as stochastic gradient descent in machine learning and algorithms based on randomized subspace selection in linear algebra, can to *high probability* deliver a *highly accurate* answer at a much *lower cost* than their deterministic equivalents.

Technology translation for software includes both computer vendors (for example, Cray, NVIDIA) and commercial users (such as Aramco, McLaren). Translation efforts train post-docs and master's students, as well as the Ph.D. students emphasized herein, for the rapidly expanding workforce in simulation and big data analytics, for placement in the world's leading computing establishments for the future of the national economy, such as at Saudi Aramco, which operates three of the world's Top500 systems. Some ECRC members also carry out computational science and engineering campaigns of their own.

The ECRC vision fits in the "digital pillar" of KAUST's 2020–2025 strategic plan (see the article by Elmootazbellah Elnozahy in this special section), especially with respect to climate prediction and artificial intelligence, with also a recent foray into smart health. Traditionally, it has supported other institutional pillars, especially energy and environment. More than half (90 as of the time of this writing) of KAUST's faculty have accounts on KAUST's supercomputer Shaheen-2, currently the third most-powerful system in the Arab World and one of the few most powerful operated by any university on behalf of its own researchers. Eighteen research organizations beyond KAUST in Saudi Arabia have accounts on Shaheen-2. Tellingly, many of these users are KAUST alumni who now work in ministries or other universities. They were the first to bring expectations from supercomputing into their organizations. This illustrates the "flying embers" effect of a supercomputer.

Student theses in such innovations as data sparsity in linear algebra on GPUs through its HiCMA and H2Opus software—a critical technology in spatial statistics and engineering optimization, and rapid mesh traversal on many-core shared-memory accelerators through its GIRIH software—a critical technology in seismic wave propagation, have ultimately attracted collaborations with major vendors to the Arab World. Today, the majority of the members of the exascale workforce trained in the Arab World find their best opportunities in countries already possessing a fully developed supercomputing ecosystem, including more capable supercomputer hardware. With the UAE and Morocco recently joining Saudi Arabia in operating a Top 100 supercomputer and with increasing trends in simulation and analytics/machine learning in all disciplines of science and engineering, we expect an increasing fraction of the workforce trained in Saudi Arabia will remain in Saudi Arabia and become the human core of the regional ecosystem. For a decade now, Saudi Arabia has hosted a High-Performance Computing Symposium that attracts researchers from around the region and allows students to mix in poster sessions and sense the regional spirit of HPC. In becoming a source of students and a source of software, a university can aspire to become a "mecca" for high-performance computing. C

References

1. Ang, J.A. (Ed) et al. Abstract machine models and proxy architectures for exascale computing. In *Proceedings of Hardware-Software Co-Design for High-Performance Computing*, (New Orleans, LA, 2014), 25–32; doi: 10.1109/Co-HPC.2014.4.
2. Beck, M. On the hourglass model. *Commun. ACM 62*, 7 (June 2019), 48–57.
3. Jia, W., Wangy, H., Chenz, M., Luz, D., Lin, L., Car, R., Weinan E. and Zhang, L. Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning. In *Proceedings of the Intern. Conf. High-Performance Computing, Networking, Storage and Analysis*, (Nov. 2020).
4. Norton, A., Conway, S. and Joseph, E. *Bringing HPC Expertise to Cloud Computing*. Opinion Whitepaper, Hyperion Research, Apr. 2020.
5. Ziogas, A.N., Ben-Nun, T., Fernández, G.I., Schneider, T. Luisier, M. and Hoefler, T. A data-centric approach to extreme-scale Ab initio dissipative quantum transport simulations. In *Proceedings of the Intern. Conf. High-Performance Computing, Networking, Storage and Analysis*, (Nov. 2019).

David Keyes is a professor of applied mathematics and computational science and director of the Extreme Computing Research Center at the King Abdullah University of Science and Technology, Saudi Arabia.