A NEW DECADE OF HIGH PERFORMANCE COMPUTING

Sandia National Laboratories

Printed in March 2021

TECHNOLOGY *THAT CONNECTS US*

Watch innovation come to life using SNLSimMagic, an augmented reality application developed at Sandia National Laboratories.

Download SNLSimMagic on your iOS device.

Use your phone to scan images with an icon to watch content come to life.

CONTENTS

SANDIA'S HIGH PERFORMANCE COMPUTING

TACKLES THE NATION'S TOUGHEST CHALLENGES

James Peery

Laboratories Director
Sandia National Laboratories

Sandia National Laboratories is known around the world for its expertise and leadership in High Performance Computing (HPC), a capability critical to its national security missions. Our commitment to HPC includes advances in numerical algorithms, pioneering approaches to software development, and the deployment and application of the next-generation of supercomputing hardware. HPC is a foundation of Sandia's Nuclear Deterrence (ND) stockpile stewardship mission and has far-ranging impact in areas including state-of-the-art energy systems, advanced materials, earth system modeling, and most recently our efforts to combat COVID-19.

Sandia is performing ND work on a level not seen in over 30 years. Our HPC capabilities support advanced physics simulation codes developed at Sandia that allow scientists and engineers to explore a variety of engineering options to identify low-cost, high-performing designs. These simulations are critical to understanding how systems behave in the most challenging environments and help ensure weapons will perform as intended.

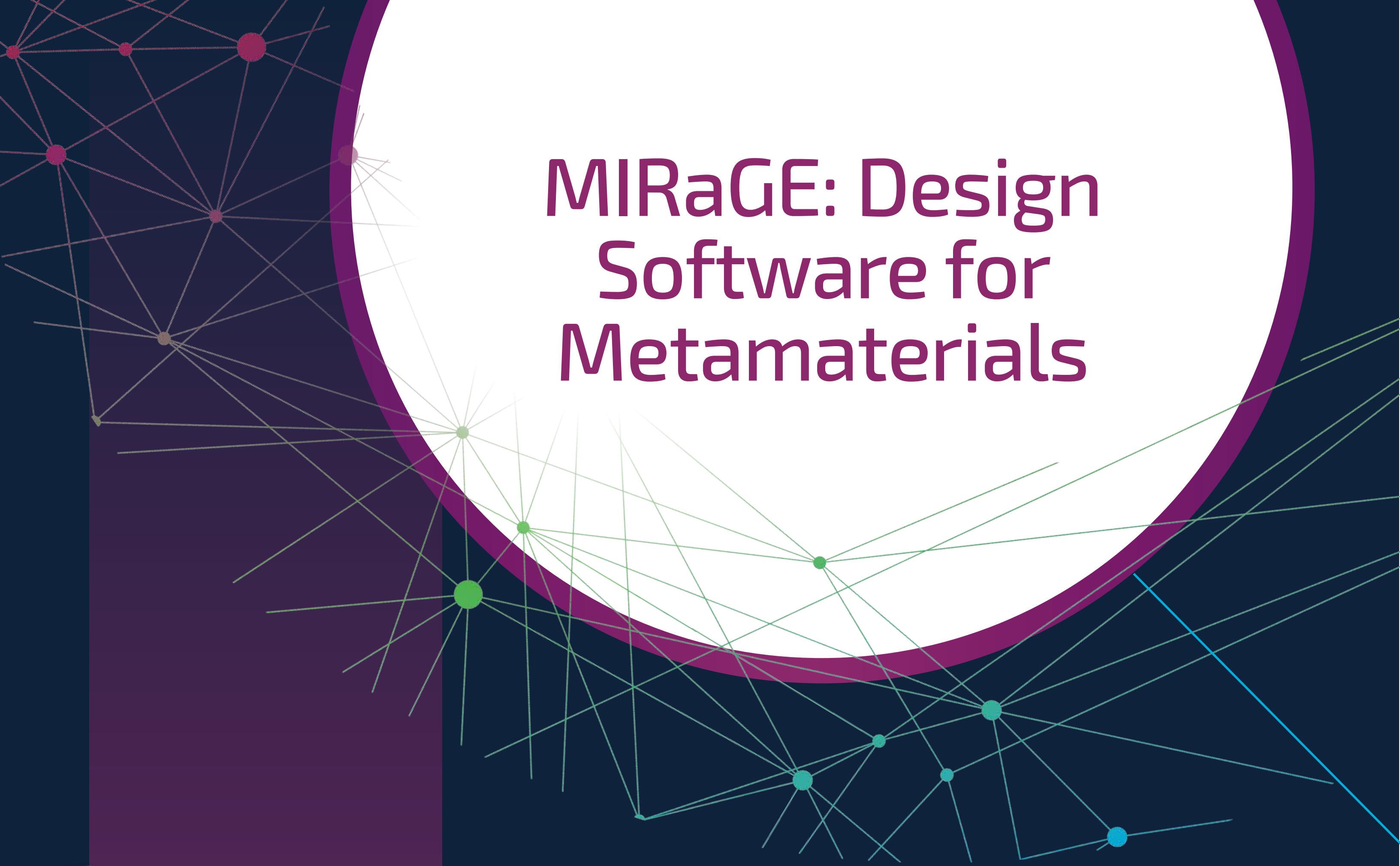
At Sandia, the high-consequence nature of our national security work requires different, more exacting levels of simulation fidelity and confidence from those of other engineering endeavors. Design cycles and qualification turnaround times are steadily shortening, leading to an increased use of models as opposed to field tests and experiments. Consequently the demands on Sandia computer architectures and software has steadily increased and will enhance the sophistication and performance of HPC.

Harnessing the power of emerging, next-generation platforms requires new approaches to software engineering. A core software capability is the Kokkos library, which allows science and engineering application developers to achieve performance portability across multiple hardware architectures. Kokkos has been developed under Laboratory Directed Research and Development (LDRD) projects, the National Nuclear Security Administration's (NNSA) Advanced Simulation and Computing (ASC) program, and the Department of Energy's (DOE) Advanced Scientific Computing Research program including the Exascale Computing Project. Sandia recently developed two Kokkos-based exascale applications, EMPIRE and SPARC, that are the culmination of a multi-year effort supported by ASC. EMPIRE provides new simulation capabilities for a wider range of plasma regimes while SPARC is a hypersonic flow code that offers important new simulation capabilities for atmosphere re-entry environments.

Our high-fidelity modeling and simulation capabilities require computing platforms at the forefront of state-of-the-art. Great examples are the Advanced Architecture Testbed systems procured and managed by Sandia's Computer Science Research and Operations centers that have kept the Labs at the leading edge of HPC technologies. Recent highlights include a new Fujitsu system with A64FX processors, acquisition of the large-scale neuromorphic Loihi system, and our upcoming Graphcore system targeting research and development in advanced machine learning algorithms. Sandia also uses its computing expertise to select emerging technologies as part of the Vanguard advanced technology prototype program. Following a successful demonstration of Arm-based processors on Astra, the first peta-scale class ARM platform, staff on Vanguard are selecting the next technology to evaluate in support of stockpile stewardship.

Importantly, Sandia scientists have recently harnessed the power of HPC to study and understand the spread of COVID-19. As the disease moved rapidly across the country, groundbreaking research using HPC gave decision-makers data to better understand how the virus might spread and mitigation options. That work continues as of this writing.

Our path is challenging but also exciting with emerging technologies and increased partnerships with colleagues, sponsors, and stakeholders. We are inspired and energized by these challenges and opportunities in the realm of HPC. ●



MIRaGE: Design Software for Metamaterials

TEAM

Ihab El-Kady,
Denis Ridzal
Timothy Wildey

Contributing Writer:
Johann Snyder

Metamaterials are artificial optical structures that allow control of light in ways not found in, or offered by, naturally occurring materials. Sandia's Multiscale Inverse Rapid Group-theory for Engineered-metamaterials (MIRaGE) software, which won an R&D100 award in 2019, allows researchers to deterministically design and produce metamaterials with unique characteristics. MIRaGE also provides powerful autonomous optimization techniques for real-world performance in a rigorous, robust, and accurate manner.

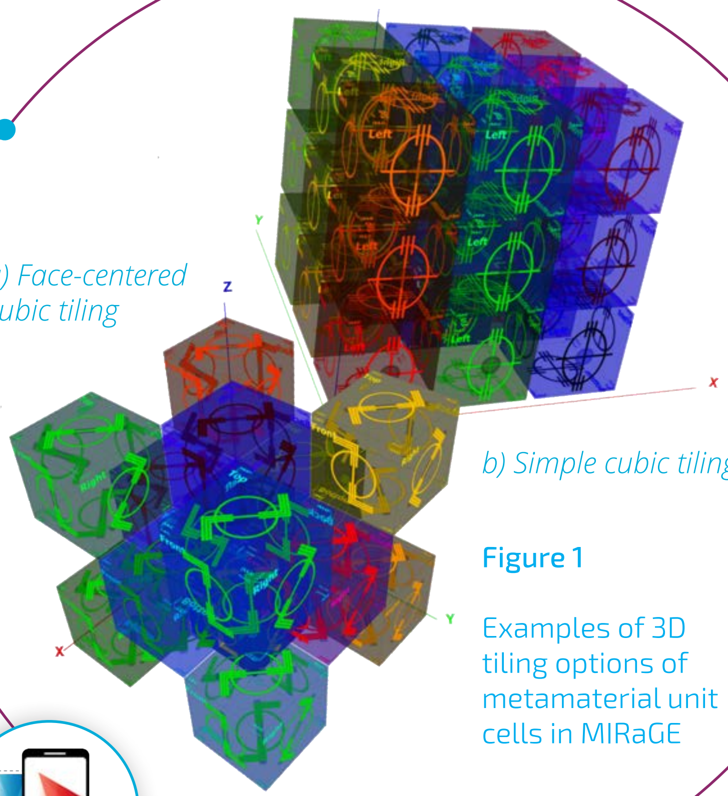
Metamaterials can be better understood by considering the example of elemental carbon. Both diamonds and graphite are made of this element, but the difference is that one has carbon atoms arranged on a staggered face-centered lattice and the other in a hexagonal lattice. Diamonds are optically transparent and one of the hardest known materials known, while graphite is pliable and opaque. The critical feature is the 3D geometric arrangement. Using the same rationale, through the metamaterial approach, it's possible to engineer graded properties between diamonds and graphite. A third material is not needed for mixing and tuning. Instead by geometrically assembling artificial carbon structures in unique 3D arrangements, semi-transparency can be realized. Scale is important, as this is not an atomistic approach.

Of particular interest to researchers are metamaterials with unusual electromagnetic behaviors appearing as effective optical properties. Since its inception, the field of optical metamaterials has lagged in its attempts to achieve full potential because of its dependence on trial and error. Prior to MIRaGE, only a handful of the eighteen possible electromagnetic classes (tensors) of behaviors had been realized – an enormous number of possibilities had to be navigated by intuition and trial and error, no matter the optimization techniques. MIRaGE overcomes challenges in design, while ensuring the desired qualitative behavior.

Metamaterials are designed on a length scale ~10-times smaller than the wavelength of interest. Unlike natural materials, the incoming wave does not see the micro-structure of the metamaterial, but instead an aggregate-effective behavior. When designing metamaterials, resonances are engineered and tuned through the 3D geometrical arrangement, shape, and topology. These act to give an effective behavior that is not a simple average of the underlaying materials. With metamaterials, the outcome can be designed to yield effective values that are larger or smaller than that of the individual components and can be tuned to a spectrum of values by changing the sub-structure.

One of the goals of MIRaGE is to create optically invisible metamaterials by making the refractive index (n) of a solid material in an optical band of interest match that of air/vacuum (n=1), or alternatively, make a flat surface focus and concentrate light like a curved lens. This can be done by engineering n< 0. Since there are no transparent natural materials with a negative refractive index (n<0), no rule of mixtures can be applied. This can only be achieved by engineering the optical structural resonances and their behaviors in response to an incoming optical wave. The result is a huge savings in SWaP (Size, Weight and Power), a factor critical in the size of telescopes and satellites and achieving lighter payloads.

Unique to MIRaGE is its inverse approach, allowing a user to start with a desired optical outcome, and through a series of sequential steps guided by the tool, get the result as an optimized metamaterial. The inverse design software relates desired properties to groups of molecular symmetries that possess those properties. By using those symmetries to predict behavior, a metamaterial can be designed that is guaranteed to exhibit the desired properties (such as semi-transparency). MIRaGE allows the researcher to explore various configurations, simulate the system, and validate the behavior. It enables researchers to optimize the design by tuning it precisely to the requirements without guesswork. MIRaGE retains its speed across a variety of computing platforms and can provide support at various levels of design proficiency.



a) Face-centered cubic tiling

b) Simple cubic tiling

Figure 1

Examples of 3D tiling options of metamaterial unit cells in MIRaGE

Scan using SNLSimMagicApp

Another unusual feature of MIRaGE is its Group-theory foundation and its ability to simulate and tile heterogeneous unit metamaterial cells into various configurations. This allows the user to combine different metamaterial properties to achieve a new functionality, hierarchical functionality, or even directional functionality where metamaterials are added together to yield different behaviors in different directions. The user can take full advantage of the 3D space, creating compact and agile multi-functional designs. Traditional electromagnetic simulation tools usually limit simulation to a single repeating unit cell (to reduce the computational overhead and time needed to produce an outcome), but with MIRaGE's powerful computational engine, as many as eight tiled heterogeneous (non-identical) unit cells can be simulated on a standard laptop using MIRaGE-Lite. This number can be greatly expanded if deployed on a sizable workstation to hundreds or thousands of heterogeneously tiled unit cells using MIRaGE-Elite. "Lite" and "Elite" are MIRaGE computational modes that use frequency domain and time domain solvers respectively. The former is geared to rapid prototyping, and the latter towards exascale-level computations.

MIRaGE software can be used to design and produce metamaterials with unique characteristics.

Categorization of designs based on symmetry groups has generated a card catalog of metamaterial building blocks. Modules can now be arbitrarily combined, and performance predicted by simply referring to a lookup table correlating each building block to its expected behavior. The library function draws on an extensive and expanding experimental base to suggest viable starting points for the design.

MIRaGE's metamaterials are used in a variety of specialized optics, such as advanced lasers, cloaking materials, and thin, flat lenses. The MIRaGE approach represents a more computationally efficient technique for metamaterial design than other contemporary options because it delivers the objective behavior faster, or more accurately, or both. ●

Materials by Design

High-Throughput Optimization of Advanced Alloys

The use of High Performance Computing recently led to a remarkable breakthrough in the optimization of thermo-mechanical properties for an alloy of platinum and gold, an elemental combination that is ideal for a wide range of electrical applications due to the oxidation resistance of these constituents. Sandia researchers showed that the structure and composition of this noble metal alloy can be tailored in a way that produces extraordinary resistance to deformation, fatigue, and wear. This work triggered the development of high-throughput computational and experimental methods for material design that researchers are now using to optimize far more complex alloys for mechanical strength and resistance to extreme temperature.

HPC tools enable Sandia to test many alloy combinations in a short amount of time, spanning a divide that covers millions of combinations of alloys with diverse characteristics.

By coupling cutting-edge theory and testing, including high-throughput, machine learning-optimized molecular dynamics simulations, additive manufacturing synthesis methods, and mechanical property testing, it is possible to rapidly scan the vast compositional space of advanced alloys and find optimal configurations for high-strength, melting temperature, etc.

TEAM

Nic Argibay
Michael Chandross
Andrew Kustas

Contributing Writer:
Sarah Johnson

HPC enables optimization of thermo-mechanical properties of a vast range of alloys, producing extraordinary resistance to deformation, fatigue and wear.

Researchers developed these novel techniques from a confluence of practical questions collected over decades from a wide range of customer needs. A common theme identified was metal electrical contacts, that depend on shearing metal junctions with minimal wear and without destroying the electrically conductive pathway. Many materials with properties ordinarily superior to metals, e.g., diamond-like carbon coatings, with typically far higher wear resistance, are obviated due to a lack of conductivity. The physics of metal interfaces is complicated not only because of the details of metals' deformation, but also due to influences from the environment that can compound problems through oxidation or corrosion. Distilling the basics of the problem—structure-dependent deformation mechanisms, led to new fundamental insights about the aspects of alloys that could be tuned to better resist failure and wear, even in complicated environments.

Researchers initially worked on core concepts to more rigorously understand the fundamental physics (e.g., the mechanisms of deformation) and how practical considerations such as the environment can influence these in real-world applications. They then pursued cutting-edge applied R&D in parallel with systematic, targeted fundamental investigations to identify opportunities to tweak materials in a way that can 'shut down' or greatly mitigate failure modes. This combination of fundamental concepts and cutting-edge theory focused research on materials optimization and investigations of failure mechanisms in three ways.

The first research focus developed a theoretical model that predicts the peak strength of polycrystalline metals based on the activation energy (or stress) required to cause deformation. Building on extensive earlier work, this model is based purely on materials' properties, requires no adjustable parameters, and is shown to accurately predict the strength of four exemplar metals. This framework reveals new routes for design of more complex high-strength materials' systems, such as compositionally complex alloys, multiphase systems, nonmetals, and composite structures.

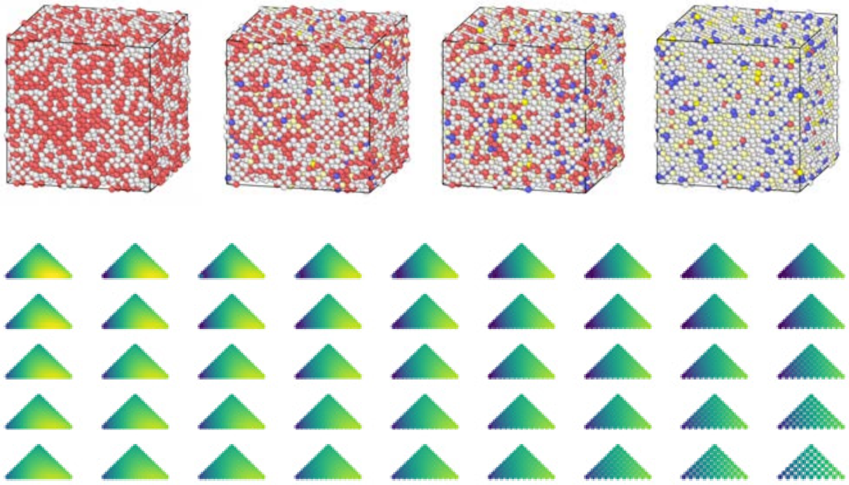
The success of this model in predicting the strength of metals leads to two major conclusions. The first is that in the absence or suppression of dislocation slip, the maximum strength of any metal is determined by the heat of fusion, and thus is a value that can be relatively easily determined. The second conclusion is that it is possible to directly design alloys for maximum strength by calculating and optimizing this property. This has strong implications in a variety of fields but seems particularly of value in the cases of high-entropy or compositionally complex alloys, where the design space is highly multidimensional.

The second research focus, based off of recent work from the Massachusetts Institute of Technology, suggests that thermally stable nanocrystallinity in metals is achievable in several binary alloys by modifying grain boundary energies via solute segregation. Researchers at Sandia fabricated such an alloy out of platinum and gold and found that this alloy exhibited extraordinarily low wear. The potential practical impacts of ultralow wear noble (i.e., oxidation resistant) alloys are significant in many applications, though perhaps most notably in electrical contacts, where bare metal contacts remain an intrinsic requirement for maintaining electrical conductivity across sliding and rolling interfaces. A highly stable nanocrystalline noble metal alloy could address persistent roadblocks to widespread adoption of this technology. The demonstration of fatigue resistance and high strength of a stable nanocrystalline alloy also suggests promising opportunities for this class of alloys as structural materials.

The third research focus lies in the optimization of metal alloys, specifically complex concentrated alloys that are typically made of four or more elements in nearly equal concentrations. A high-throughput simulation approach allows researchers to rapidly scan potential alloys and identify what may be ideal or optimal properties. HPC allows a streamlined approach to this complicated problem. Implications of this work suggest that validating the properties of the compositions predicted by the simulations with high-throughput processing (via additive manufacturing) and characterization (via hardness/scratch testing) can enable rapid screening of complex concentrated alloys, and thus lead to novel materials with optimal properties for structural applications.

Sandia researchers may use this technology to produce unusual and practical results that industry partners can use to address emerging needs for DOE applications, including aerospace systems, power generation, and gas turbines for high efficiency.

The search for optimal alloys grows increasingly complicated as the number of their constituents grows. In the near future, researchers at Sandia hope to add additional properties, such as determination of heat of fusion, as a way to further accelerate the HPC-enabled iterative high-throughput optimization process. Additional capabilities, such as extreme temperature mechanical testing, will also concurrently accelerate this process and enable Sandia to achieve the ambitious goal of rapidly developing tailored alloys for any unique application.



High-entropy alloys with various compositions of FeNiCrCoMn. The triangles represent properties across the 5D composition space.

A New Approach to Simulate Nuclear Waste Container Compaction at the Waste Isolation Pilot Plant

TEAM

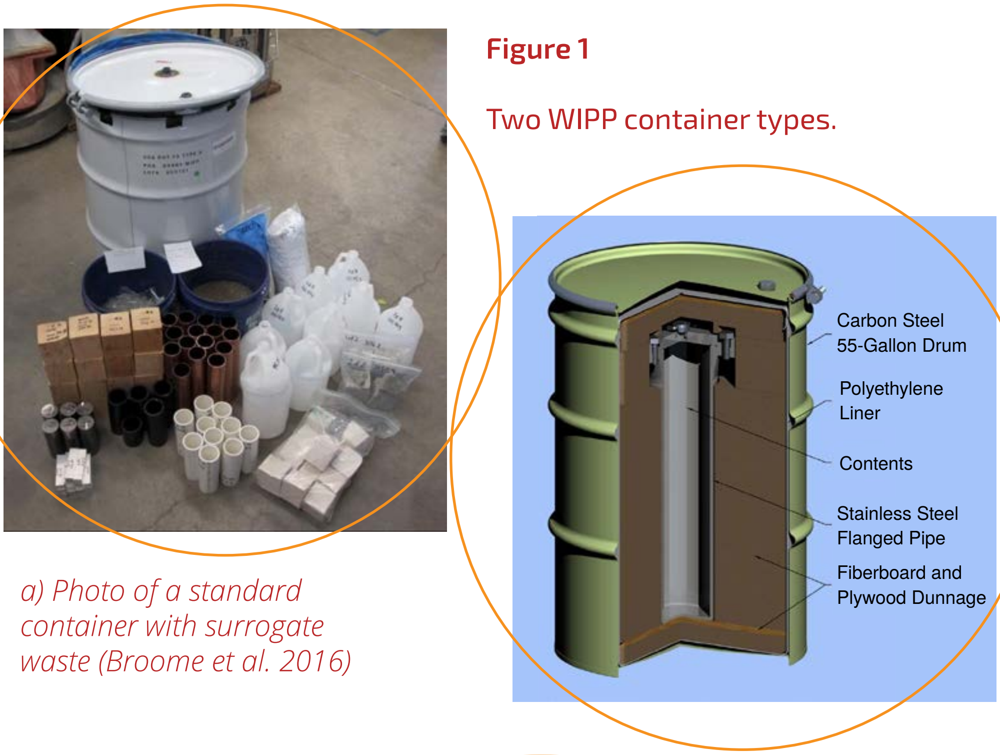
Benjamin Reedlunn
James Bean

The Waste Isolation Pilot Plant (WIPP) is an operating geologic repository in southeastern New Mexico for transuranic (TRU) waste from nuclear defense activities. Nuclear waste containers are placed underground in a disposal room at the WIPP, and then the surrounding rock salt creeps inward to compress the containers over several centuries, thereby isolating the waste from the biosphere.

Past simulations of container compaction at the WIPP have homogenized the containers into a solid, spatially uniform material. As shown in Figure 1a, standard waste containers are 55-gallon drums filled with all types of contaminated debris: cellulosics, metals, sorbants, and polymers. These containers are typically bundled together in hexagon shaped seven-packs: six containers surround one container in the center. Upon emplacement in a WIPP disposal room, the seven-packs are stacked up to three-high and arranged in a hexagonal configuration (see Figure 2a). Geomechanical models developed in the 1980s and 1990s, however, replaced the complexity in Figure 2a with the idealized geometry in Figure 2b. Instead of including every individual container, the models represented all the containers by a single, isotropic material (see, for example, Stone (1997).

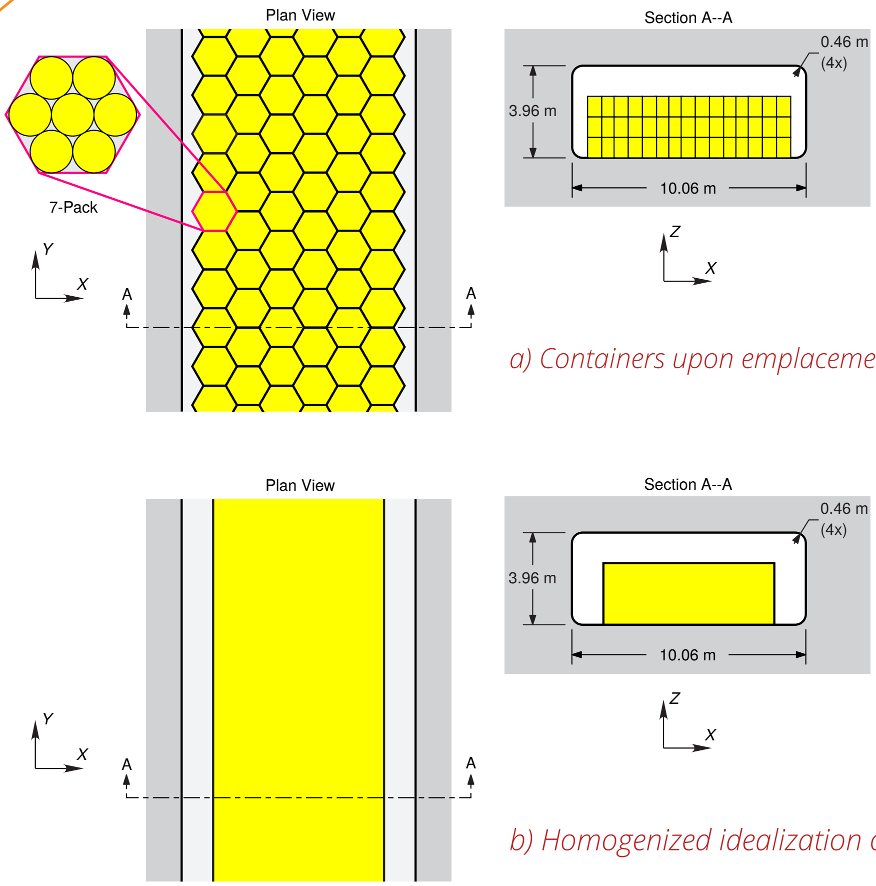
Figure 1

Two WIPP container types.



a) Photo of a standard container with surrogate waste (Broome et al. 2016)

b) Schematic of a 6-inch Pipe Overpack Container (modified from Porter (2013)



a) Containers upon emplacement

b) Homogenized idealization of (a)

Figure 2

Container configuration immediately after disposal and the corresponding homogenized idealization.

This modeling approach may be sufficient for standard waste containers, but WIPP disposes of other container types that may compact quite differently than standard waste containers, such as Pipe Overpack Containers (POCs). As shown in Figure 1b, a POC consists of a thin-walled stainless steel pipe filled with waste, surrounded by fiberboard and plywood, within a 55-gallon drum. These containers are structurally anisotropic, susceptible to buckling, and likely to interact with one another during compaction, which prompted Sandia researchers to discretely model the individual POCs, including the pipe and plywood within each drum. Discrete container modeling comes with a significant increase in computational cost, but processing power has also advanced significantly since studies performed in the 1990s. Even if a homogenized POC model is developed in the future, the higher fidelity model will still be useful as a reference for the homogenization error.

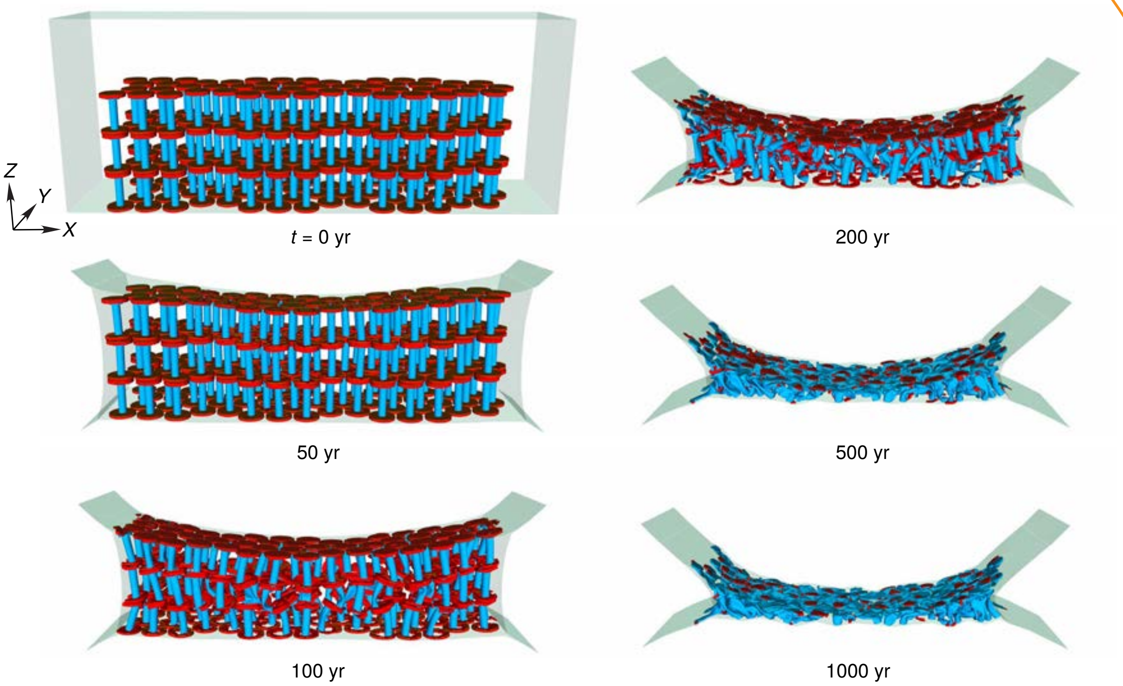


Figure 3

Compaction of several rows of 6-inch Pipe Overpack Containers with stainless steel pipes in blue and top/bottom fiberboard in red. (The surrounding rock salt, lateral fiberboard, and 55-gallon drums are not shown for visualization purposes.) (Reedlunn and Bean (2019)[4])

Through crucial high-fidelity simulations enabled by HPCs, it is possible to evaluate the nuclear waste isolation process, and determine the risk of an adverse event.

Images of 6-inch standard container gradual compaction predictions from Sandia's Sierra/Solid Mechanics are shown in Figure 3 at selected time instances. The plywood, lateral fiberboard, and 55-gallon drums were included in the simulation, but they are hidden in Figure 3 to reveal the flanged pipes. The geology is also hidden, leaving only the semi-transparent surface of the room, and the perspective view is tilted in order to show the 3D nature of the compaction. While actual disposal rooms are 91 m long in the *Y*-direction, a representative *Y* length of 2.15 m was considered sufficient for modeling purposes.

The compaction sequence is as follows: The roof first contacts the containers at room mid-width at $t = 41$ yr and contacts the entire upper layer of containers by $t = 70$ yr. By $t = 100$ yr, many 6-inch flanged pipes have buckled under the axial load supplied by the roof and floor. By $t = 200$ yr, numerous pipes originally stacked in the upper and middle container layers have been pushed down into the bottom layer. Further room closure buckles the remaining flanged pipes and pushes the upper/middle container layers down into the bottom container layer. Nearly all of the fiberboard finite elements disappear as they were deleted upon inversion. At $t = 1000$ yr, the compacted shape of the container array is essentially stabilized to something resembling a tall and slender "I" turned on its side.

The containers clearly deform in a non-uniform and highly complex manner as they slide past one another in Figure 3. Interactions between containers are realistically captured by modeling each container discretely rather than homogenizing them into a single material. These high-fidelity simulations will help Sandia assess the nuclear waste isolation process and evaluate whether the risks of an adverse event are acceptably low. ●

High-resolution 32-year Wave Hindcasts for the U.S. Coastal Waters

TEAM

Sandia	*Vincent Neary* *Seongho Ahn* *Chris Chartrand* *Budi Gunawan*
North Carolina State University	*Nabi Allahdadi* *Ruoying He*

High Performance Computing at Sandia plays a key role in the DOE Water Power Technology Office's mission of advancing the commercialization of wave energy by generating an assortment of geospatial statistics on wave energy resources in the U.S. at an unprecedented level of spatial and temporal resolution. This was recently highlighted in a collaborative project between Sandia's water-power technologies department and North Carolina State University (NCSU).

Ocean wave energy is renewable, has a high energy density, is close to high coastal population centers around the globe, and has limited environmental impacts. In the U.S., wave energy resources make up approximately 80% of the ocean hydrokinetic energy resources (wave, ocean currents, and tidal currents). A wide spectrum of wave energy conversion (WEC) technologies designed to capture, absorb, and convert the energy transferred by ocean waves to electricity, or some other useful form of energy, are under development, but the costs of these technologies are high and industry is still in its pre-commercial phase. Sandia is one of DOE's main national laboratories conducting foundational research since the conception of the Water Power R&D program to improve the techno-economic performance of these technologies, leveraging decades of experience in water power technology engineering, controls and materials research, large-scale testing, and HPC.

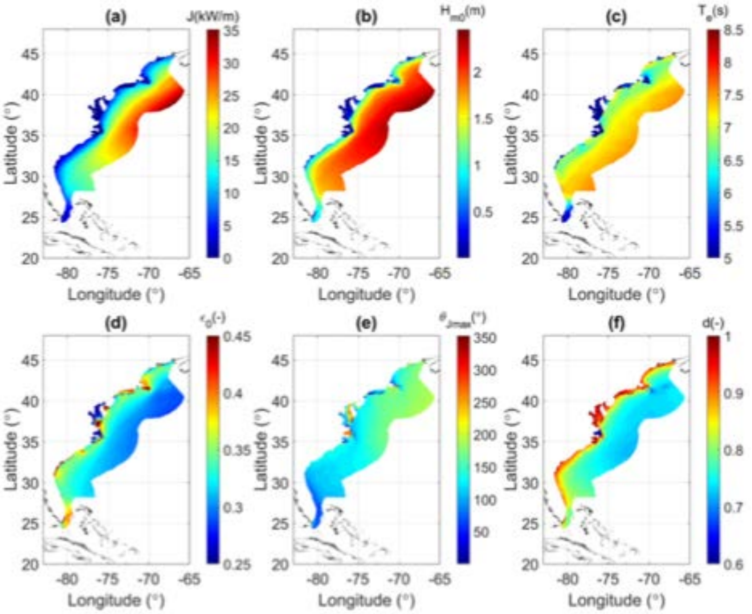
Wave energy resource characterization and assessment is a key thrust area of the DOE's Water Power R&D program strategy to advance the wave energy industry by providing high-resolution data and information on wave energy resource attributes and wave conditions that inform WEC design, and that characterize opportunities, constraints, and risks to wave energy projects. These data and information support project siting, permitting, and development.

The Sandia WaveBot test rig for investigating advanced control systems and hydrodynamic performance of wave energy converters (WEC).

The Sandia-NCSU team applied the Simulating WAves Nearshore (SWAN) model to simulate a 32-year wave hindcast along the U.S. East Coast, which encompassed the entire economic exclusion zone (EEZ) 200 nautical miles offshore with a computational mesh of over 4.3 million grid points. The model generates time series for a variety of wave energy resource statistics, including wave power at a 200 m resolution up to 20 km off the coastline. These statistics are invaluable for characterizing, classifying, and assessing the U.S. wave energy resources, including wave energy project opportunities, constraints, and risks. The model was extensively verified and validated, and efforts are now underway to generate a 32-year hindcast for the Gulf of Mexico and Caribbean Sea with a computational mesh of over 5.7 million grid points.

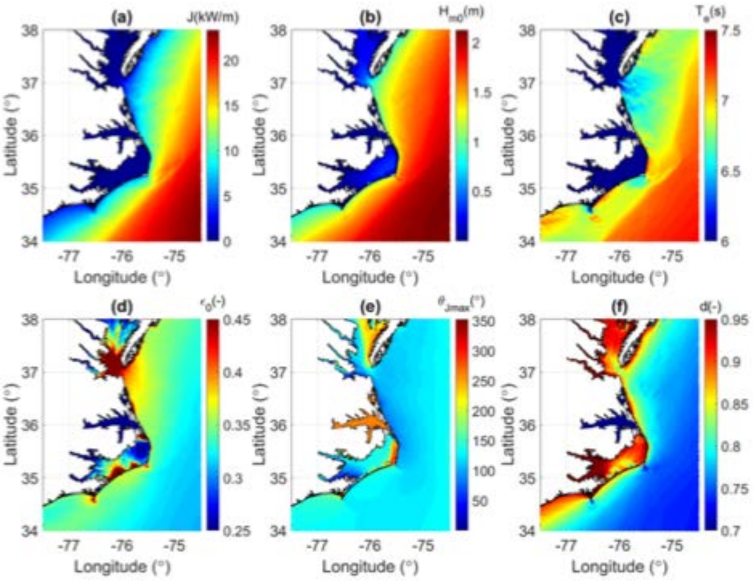
This study demonstrates the value of Sandia's HPC resources for generating modeled data at high spatial and temporal resolutions. The size of the modeled dataset for the U.S. East Coast is over sixty terabytes, and for the Gulf of Mexico and Caribbean Sea it's expected to be over one hundred terabytes. Wave energy resource statistics generated from this data set will be used to upgrade the marine and hydrokinetic (MHK) ATLAS maintained by the National Renewable Energy Laboratory, which provides valuable information to designers of WEC technologies, WEC project developers, and regional energy planners. In addition, the full model hindcast dataset, which includes 32-year 3-hour time series of these resource statistics and over one hundred virtual buoy sites with hourly 2D wave spectra, will be disseminated as an open source database on an Amazon Web Server to promote further research characterizing and assessing the U.S. wave energy resource. ●

Simulated 32-year averages of six wave energy resource statistics for U.S. East Coast



(a) omnidirectional wave power, *J*;
(b) significant wave height, H_{m0};
(c) energy period, T_e;
(d) spectral width, E_0;
(e) direction of maximum directionally resolved wave power, 0_{Jmax};
(f) directionality coefficient, *d*.

Simulated 32-year averages of six wave energy resource statistics in the Chesapeake Region of the U.S. East Coast



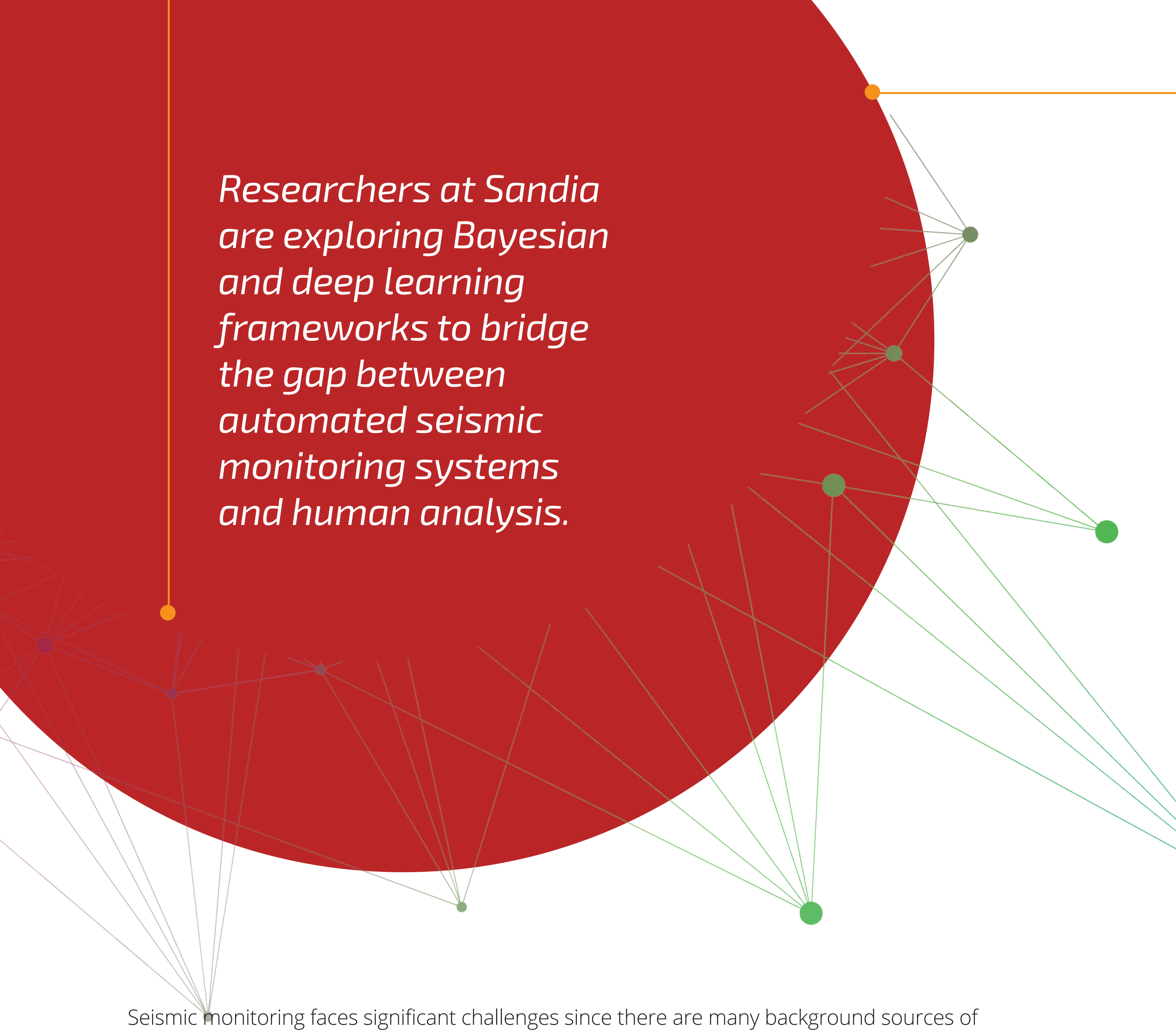
Advanced Machine Learning for Seismic Monitoring

TEAM

Thomas Catanach
Lisa Linville

Contributing Writer:
Sarah Johnson

Sandia is part of a multi-lab Defense Nuclear Nonproliferation Research and Development project designed to improve U.S. capabilities to detect and characterize low yield underground nuclear explosions. Monitoring at low detection thresholds with dynamic monitoring networks requires innovative approaches to leverage and merge new and legacy physics-based knowledge with new data quantities, qualities, and types currently available. Sandia is working to address these challenges through the development of new predictive capabilities, network design, and advanced algorithms for the detection, identification, location, and characterization of underground nuclear explosions. Seismic monitoring provides an important improved capability for detection of nuclear tests around the world, especially those conducted deeper underground in an effort to conceal a nuclear explosion. To contribute to this effort, Sandia is developing new automated tools to analyze seismic data that increase the ability to detect and characterize evasive explosions – understanding the inherent uncertainty – while still maintaining the credibility that human scrutiny provides.



Researchers at Sandia are exploring Bayesian and deep learning frameworks to bridge the gap between automated seismic monitoring systems and human analysis.

Seismic monitoring faces significant challenges since there are many background sources of seismicity, such as natural seismic events like earthquakes or anthropogenic sources like mining. Analysis must differentiate these background events from events that could represent nuclear tests. Automated preprocessing of sensor data is the first step in helping analysts identify potential events of interest. Automation serves as a filter to categorize unique signatures from background noise when there is abundant data from many sensor phenomenologies. Scientists at Sandia are exploring tools capable of bridging the gap between current automated seismic monitoring systems and human analysis. These tools rely on advances in Bayesian statistical inference, computer simulation, and machine learning. Researchers at multiple labs, including Sandia, are adapting datasets, generated by complex and interdependent physical systems, to interact with these emerging algorithms in ways that inform high-consequence decisions. For this reason, physics-based knowledge, data-driven observation, simulation, and novel learning objectives play important roles in advancing seismic monitoring algorithms.

One of the tools Sandia is using to improve automated processing is deep learning. After automated seismic processing is completed, additional analysis determines attributes such as event source, type, and size. Deep learning builds predictive models that can assign attributes, such as whether an event comes from natural or anthropogenic seismicity, with accuracies significantly exceeding existing methods. Deep learning synthesizes years' worth of curated data from monitoring networks to extract predictive features. Researchers developing these models have to be concerned not only with high predictive accuracy, but also with prediction credibility. This is because safe and actionable decision support for high-consequence situations requires capturing uncertainty in deep learning predictions. By providing both accurate and credible predictions, deep learning will increase the amount of automated seismic monitoring and reduce the burden on human expertise.

Sandia is also advancing seismic monitoring capabilities by developing novel methods to precisely locate seismic events. Precise locations are critical not only for seismic monitoring but also for providing data that can improve computer simulations used to model the propagation of seismic waves in the earth. To achieve this goal, researchers at Sandia are developing a novel Bayesian framework for event location which allows scientists to assess the fidelity of inferred locations with higher confidence. A statistical model of a seismic waveform's features learned from offline synthetic simulations of seismic events gives scientists the ability to integrate meaningful physical properties from the simulations into the location algorithm. Building the statistical model offline avoids costly simulations during online monitoring that previously made this approach impractical. Leveraging physics-based waveform features will decrease monitoring thresholds and increase location reliability and accuracy.

Locating seismic events using Bayesian methods and learning the feature model from seismic waveform simulations are computationally intensive; however, they are highly parallelizable on HPC resources. The waveform synthetics are expensive to compute and require detailed physics models to attain the resolution and realism analysts require to replicate real seismic events. Therefore, researchers rely on GPUs in combination with HPC to build predictive models in an efficient, lower-cost manner. Likewise, data-driven models require GPU and HPC resources for efficient and practical development. Widespread access to GPU and HPC has opened up new doors for seismic monitoring algorithm development.

Beyond improving automated seismic pipelines, Sandia is also using HPC to design better sensor networks through Bayesian optimal experimental design. Within this framework, the position and type of seismic sensors are optimized to maximize the expected information gained about possible events of interest. Therefore, the performance of each potential network configuration is tested for many hypothetical events using Monte Carlo sampling. HPC is crucial for speeding up sampling since many events must be considered. By optimizing the sensor configuration, seismic monitoring networks can be designed to improve detection accuracy and reduce location uncertainty, particularly for small seismic events.

Sandia researchers hope to continue to develop ideas that contribute directly to an ability to detect and characterize low yield underground nuclear explosions. The advances they make to U.S. monitoring capabilities may also provide far-reaching solutions to challenges shared across other mission spaces. ●

TEAM

Andrew Baczewski
Ojas Parekh
Mohan Sarovar
John Aidun

Contributing Writer:
Sarah Johnson

Quantum computing (QC) leverages quantum mechanics to enable a vastly different mode of computation than computers based on classical physics, including conventional von Neumann systems. A quantum bit (qubit), like a classical bit, takes a binary 0 or 1 value when measured, usually at the end of a quantum computation. However, the value of a qubit is not deterministic. A quantum state of n interacting qubits is parameterized by 2^n complex numbers, which are called amplitudes and cannot be accessed directly; measuring such a state produces a single random n-bit classical string with probability dictated by a corresponding amplitude.

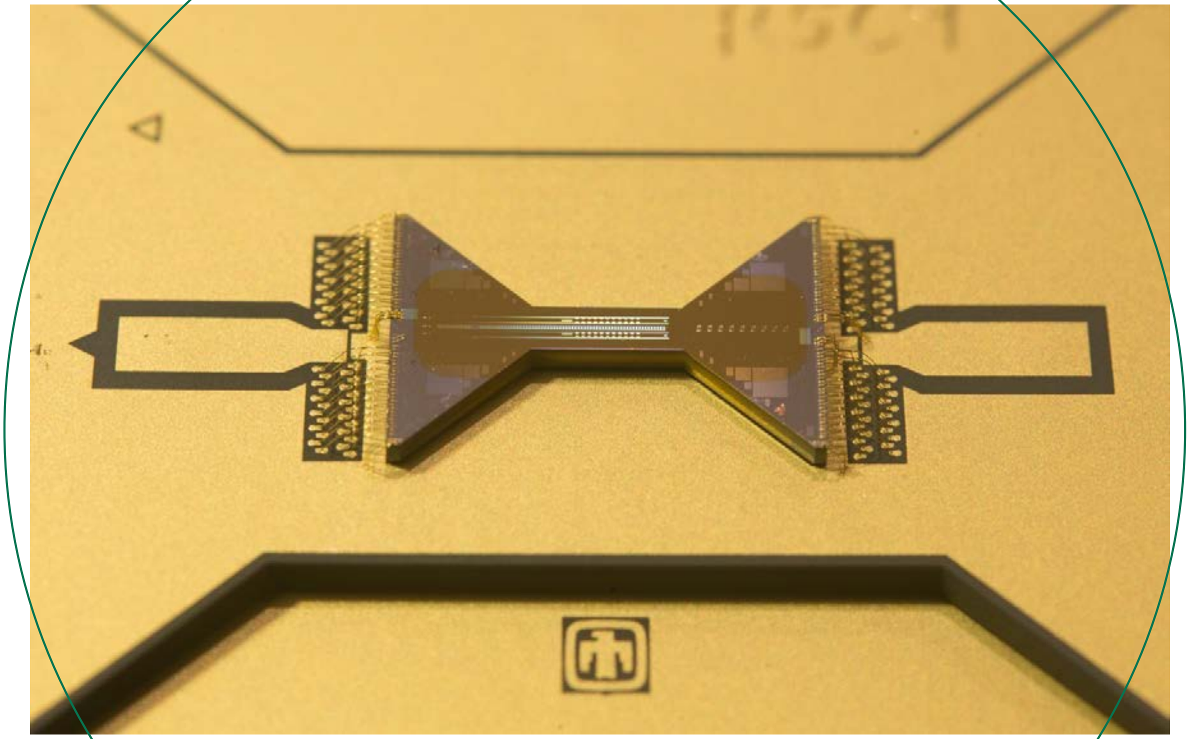
Sandia Brings Its Deep Expertise in Quantum Information to Bear on DOE Missions

A powerful feature of quantum computation is that manipulating *n* qubits allows users to sample from an exponentially larger probability distribution over 2^n outcomes. However, an analogous claim can be made for randomized classical algorithms operating on n probabilistic bits (e.g., flipping *n* coins). A key difference between the two is that quantum algorithms seem to be able to sample from certain kinds of probability distributions that may take exponentially longer for randomized classical algorithms to mimic. For example, Shor's seminal 25-year-old quantum algorithm for factoring integers requires exponentially fewer steps than the best-known classical counterparts. Exponential quantum advantages are also known for other fundamental scientific problems such as solving a certain kind of linear systems of equations and simulating quantum-mechanical systems, currently a critical bottleneck in many physical and chemical applications. The precise source of quantum computational advantage is not well understood; however, it is attributed in part to quantum computation's ability to efficiently generate entanglement among qubits, yielding probability distributions with correlations that in some cases overstep the reach of efficient classical algorithms.

Successes in designing theoretical quantum algorithms have fueled the hope that other quantum advantages can be discovered and exploited. Ideal quantum advantages would provide: (i) an exponential (or at least super-polynomial) computational speedup, (ii) practical applications, and (iii) implementation on a physically realizable quantum system (ideally scalable). A foremost open question in quantum computing is whether all three of these can be simultaneously achieved. A significant hurdle for (iii) is that prepared quantum states are fragile and highly susceptible to environmental noise and rapid entropic decay. Contemporary quantum information science (QIS) research addresses (i) and (ii) by developing novel quantum algorithms and applications and (iii) through scientific and engineering efforts to develop noise-resilient and scalable quantum infrastructure.

After decades of steady progress, mainly in academia, the past five years have seen an explosion of interest and effort in QIS. The fifteen years of QC research at Sandia spans the Labs' expertise from theoretical computer science and physics to microelectronic fabrication, laboratory demonstrations, and systems engineering. Hardware platforms developed at Sandia include a variety of efforts in trapped-ion, neutral atom, and semiconductor spin qubits. Complementary theoretical efforts have created unique capabilities, from quantum characterization verification and validation protocols to multi-scale qubit device modeling tools. Even efforts that are ostensibly purely theoretical, such as quantum algorithms development, are tied to applications of interest ranging from optimization and machine learning to materials simulation The breadth of current Sandia research activities coupled with the longevity of Sandia's program have established Sandia as a leading U.S. National Laboratory in QC and broader QIS research.

Most recently, Sandia has been successful in securing a number of quantum computing projects funded by the recent push from DOE Office of Science and the National Nuclear Security Administration. Among these projects, closest to the hardware, are the Advanced Scientific Computing Research (ASCR)-funded Quantum Scientific Open User Testbed (QSCOUT) and Quantum Performance Assessment (QPerformance) projects. In just over a year, the first edition of the QSCOUT testbed with three trapped-ion qubits was stood up. While this will be increased to thirty-two qubits in time, the testbed is most significant for providing researchers complete access to generation of the control signals that specify how gates are operated so they can further investigate the quantum computer itself. A critical component of this effort is the Sandia-developed Jaqal quantum assembly language which will be used to specify programs executed on QSCOUT. The QPerformance project is aimed at creating techniques for evaluating every aspect of a testbed QC's performance and understanding and tracking how these change with improvements to the QC hardware and software. The effort isn't limited to the QSCOUT testbed and it will invent and deploy platform-independent holistic benchmarks that will capture high-level characteristics that will be predictive in evaluating the suitability of QC platforms for DOE mission-relevant applications.

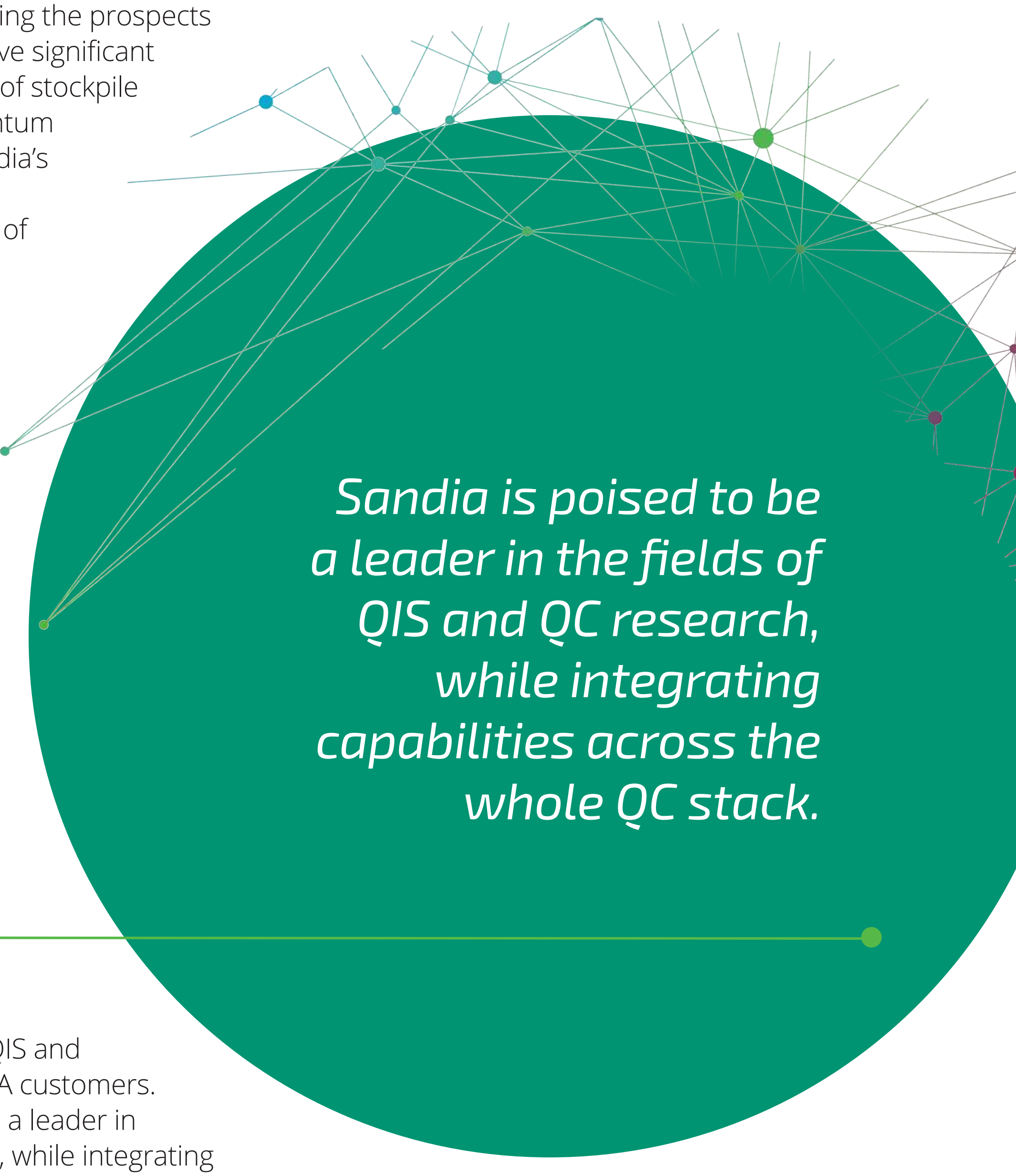


The Peregrine ion trap that is being used in the ASCR-funded QSCOUT testbed.

At the next level of the computing hierarchy sits the ASCR-funded "Optimization, verification and engineered reliability of quantum computers" (OVER-QC). Led by Sandia, this project aims to develop tools that get the most out of near-term QC hardware, which will be noisy and imperfect. By developing specialized techniques to interpret the output, and to increase the reliability of such noisy hardware, OVER-QC aims to understand and push the limits of QC hardware.

Sandia complements these efforts driven by near-term QC hardware with ASCR-funded efforts focusing on developing fundamental hardware-agnostic quantum algorithms for future fault-tolerant quantum computers. These Sandia-led projects, "Quantum Optimization and Learning and Simulation" (QOALAS) and "Fundamental Algorithmic Research for Quantum Computing" (FAR-QC), are multi-institutional interdisciplinary efforts leveraging world-class computer science, physics, and applied mathematics expertise at Sandia and more than ten partner institutions. QOALAS seeks to develop novel quantum algorithms enabling new applications in optimization, machine learning, and quantum simulation. FAR-QC expands upon the scope of QOALAS to identify problems and domains in which quantum resources may offer significant advantages over classical counterparts. Some of the achievements of these projects include new quantum algorithms offering significant advantages for solving linear systems, convex optimization, machine learning kernels, and rigorous simulation of physical systems.

Among the key mission priorities of Sandia are those related to stockpile stewardship. The Advance Simulation and Computing (ASC)-funded Gate-Based Quantum Computing (GBQC) project is focused on understanding the prospects for QC platforms to eventually have significant impacts on the unique problems of stockpile stewardship. In this context, quantum simulation is a key capability. Sandia's stockpile stewardship mission requires models for the behavior of materials in extreme conditions that are both challenging and expensive to evaluate experimentally. GBQC is focused on understanding what will be required to realize a simulation capability that would be exceptionally impactful to ASC and the broader DOE. Recent research directions have broadened the scope of this work to understand the impacts that QCs might have on numerical linear algebra, which is a key capability for not only ASC applications, but most computational science.



Sandia is poised to be a leader in the fields of QIS and QC research, while integrating capabilities across the whole QC stack.

Sandia has spent fifteen years developing a strong program in QIS and QC to better serve DOE and NNSA customers. As a result, Sandia is poised to be a leader in the fields of QIS and QC research, while integrating capabilities across the whole QC stack. ●

Z Machine – An Engine of Discovery

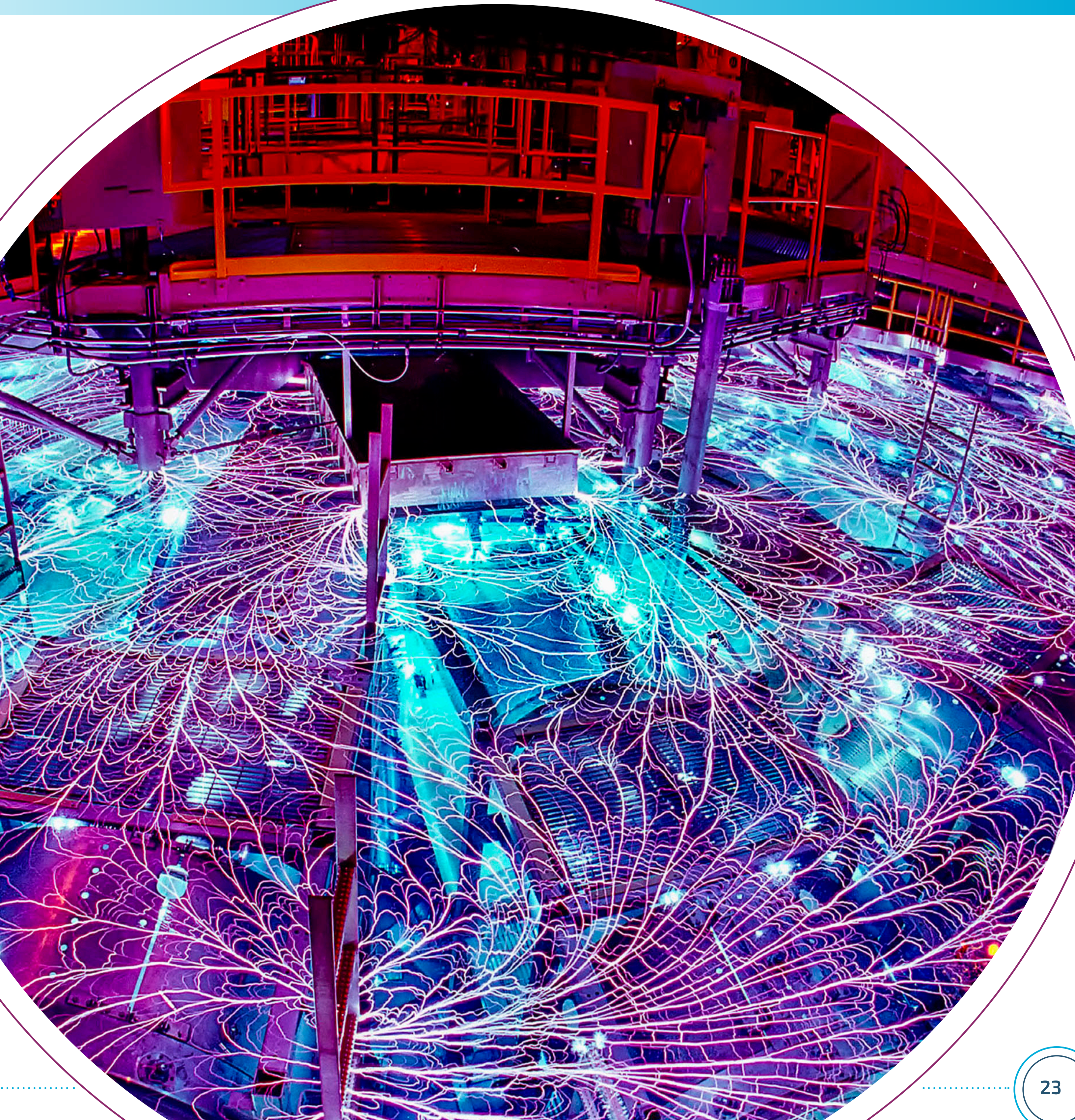
The world's largest pulsed-powered accelerator—the Z Machine housed at Sandia—is an engine of discovery. Z Machine stores up to 22 megajoules of electrical energy in a series of capacitor banks and can release this energy in roughly 100 nanoseconds. The electrical power (P = Current*Voltage) delivered to an experiment on Z can reach 80 TW, or roughly 12 times the total continuous electrical power generation from the world's largest power plants.

Z Machine's primary missions range from studying materials under high pressure, to the development of intense x-ray sources, to inertial confinement fusion (ICF) and magneto-inertial-fusion (MIF). Conventional ICF relies on high velocity spherical implosions to compress and heat fusion fuel to thermonuclear conditions. In MIF, an applied magnetic field is used to relax the requirements to achieve fusion conditions for ICF. Over the past 10 years the Magnetized Liner Inertial Fusion (MagLIF) platform has been developed to study MIF. In MagLIF, the electrical current supplied by Z implodes a 1 cm tall metal tube (cylindrical liner) of beryllium onto fusion fuel via the J x B force (Lorentz force), where J is the axial current density and B is the azimuthal magnetic field produced by the current through Ampere's Law. Before the liner implodes, the fuel is also initially heated by the Z-Beamlet laser, which helps to reduce the necessary compression to achieve fusion conditions.

TEAM

Matthew Weis

Contributing Writer:
Whitney Lacy



As the fuel compresses and heats, it can lose energy radially through thermal conduction to the liner wall, cooling the fuel. The applied axial magnetic field is the key MIF feature of MagLIF and inhibits this heat loss. Heat is primarily transported by electrons, but with the addition of an axial magnetic field, the electrons instead tend to gyrate around the magnetic field azimuthally, as they try to move across the magnetic field, which slows the transport of heat. The combination of these effects has successfully produced thermonuclear fusion conditions in over 70 MagLIF experiments.

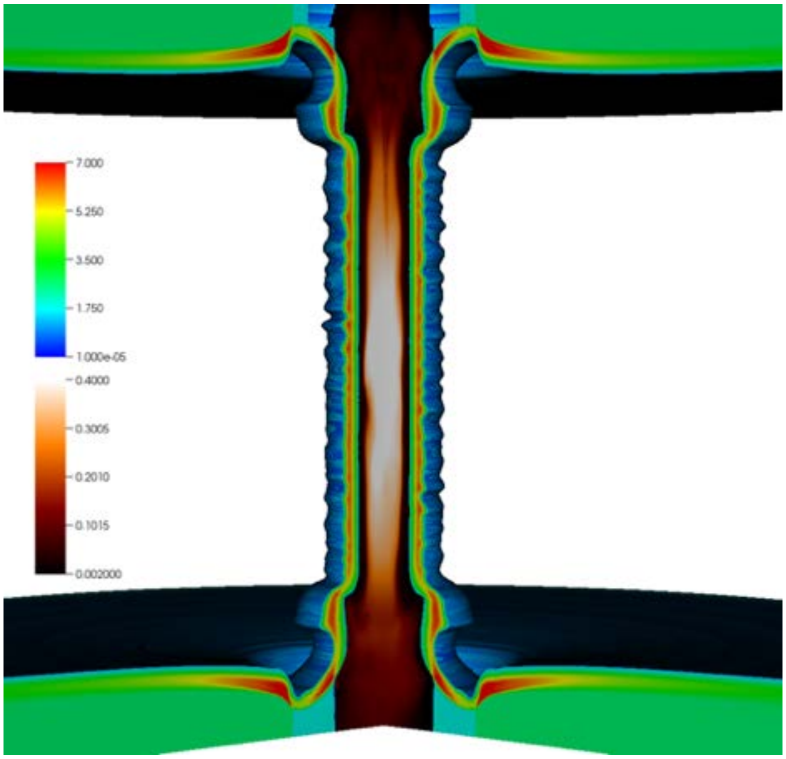
MagLIF, originally conceived in 2009 through 1D and 2D simulations, had its first experiments begin in late 2013. While the experiments successfully produced fusion conditions, they revealed striking 3D structures in the fusion plasma. Like any inertial confinement fusion concept, MagLIF is susceptible to the Rayleigh-Taylor instability (RTI), which tends to break up the liner during implosion and thereby reduce the confinement of the fuel at stagnation.

Radiographic images taken of the liner during implosion show that the RTI structure changes with the axial magnetic field. Instead of azimuthal-correlated structures, which are approximately 2D, the instability becomes helical. Such 3D instabilities cannot be produced by a 2D computer code but are necessary to understand the experimental data.

This is where Sandia's HPC power comes into play. The Advanced Simulation & Computing (ASC) Nuclear Deterrence Program Clusters at Sandia, in coordination with the New Mexico Alliance for Computing at Extreme Scale (ACES), has allowed unprecedented modeling of MagLIF in 3D. It does this by using the HYDRA code. HYDRA is a massively parallel, multi-physics code used on HPC computers for modeling high-energy density plasmas such as those found in ICF. HYDRA was originally developed by Lawrence Livermore National Laboratory (LLNL) for modeling spherical implosions on the National Ignition Facility (NIF). In collaboration with Sandia, HYDRA now includes a full resistive magneto-hydrodynamics (MHD) package for modeling experiments on Z Machine.

HPC simulations enhance Z machine experiments and provide the basis for designing future research on Z.

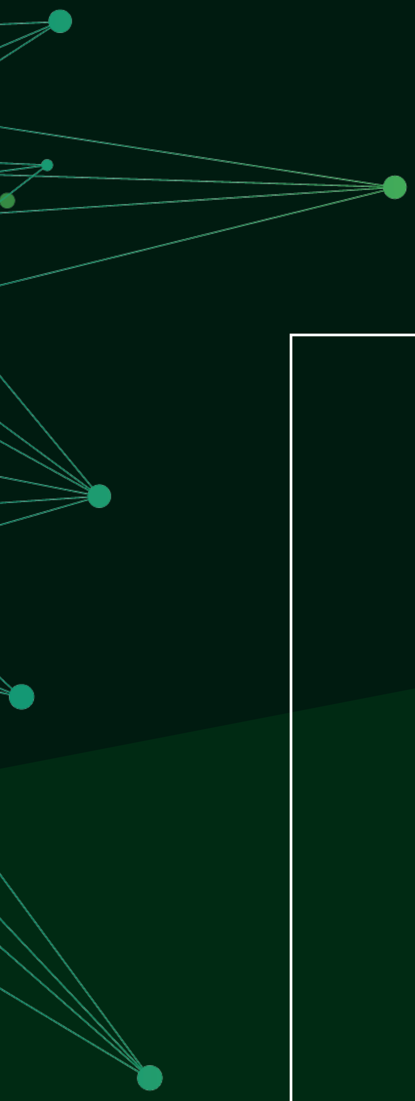
Scan using SNLSimMagicApp



A typical full 3D model of MagLIF with modest spatial resolution requires a simulation volume of roughly ~π*(1 cm)^2 *(2 cm) and 25 million computational cells. To model the laser heating, ~ 1 million rays are used by the ray-tracing package, and the implicit Monte Carlo algorithm models the transport of radiation in the target. HYDRA also incorporates detailed models for material properties. For example, beryllium uses equation of state as well as electrical and thermal conductivity tables developed at Sandia to model the material from solid room temperature, through melt, to the warm-dense matter conditions at stagnation. These detailed 3D simulations produce much better agreement with MagLIF experiments on the Z Machine and provide better insight into the physics of MIF. As well, these HPC simulations enhance Sandia's confidence in using computational tools to both understand current experiments on Z and design meaningful future experiments both on Z and future pulsed-power facilities.

Currently, 3D calculations are for very targeted situations, but as computational resources continue to expand and improve, more routine 3D simulations will enable even more rapid progress in these areas. ●

NEXT-GENERATION COMPUTING



The advent of on-node accelerators such as GPUs is driving a revolutionary shift in next-generation computing, similar to the advent of distributed memory systems and MPI in the 1990s. Single parallelization strategies like MPI are no longer sufficient; instead, to ensure scalability, a hierarchal approach of developing an interface for the threading layer is necessary. The Kokkos library and programming model provide that layer through an abstraction of the different hardware mechanisms and custom languages like OpenMP and CUDA into a uniform interface. This layer can be integrated into Sandia codes and solver libraries to provide performance portability across a wide range of next-generation architectures, including GPUs. A number of utilities have been co-designed and evolved from the lowest levels of Kokkos to the highest levels within two Sandia engineering analysis codes, EMPIRE and SPARC. These applications, with their varied approaches and numerical techniques, ensure that nextgeneration algorithms implemented within software libraries such as Trilinos (a collection of reusable scientific software libraries containing tools for linear algebra, discretizations, embedded analysis, and more) and Pressio (a projection-based model reduction library) will be applicable to a wide range of different problem spaces.

EMPIRE: A Revolutionary Modeling Tool for Agile Design

TEAM

Matt Bettencourt
Keith Cartwright
George Laity

ElectroMagnetic Plasma In Realistic Environments (or EMPIRE) is a modeling and design tool for plasma environments. Plasma is one of the four fundamental states of matter. In the case of water, it is created when water transforms from ice to liquid, liquid to steam, and then break down of the molecules into hydrogen and oxygen with the addition of continually added energy. Eventually, orbital electrons are stripped off, and ions (atoms which have some of their orbital electrons removed) and free electrons are left moving through space. This gas of ions and free electrons is called plasma. Plasma can be generated in several ways, such as by heating a neutral gas or subjecting it to a strong electromagnetic field to the point where naturally occurring free electrons ionize neutral particles in an avalanche (mechanism for lightning). As the ionization fraction of a gaseous substance becomes increasingly high, it becomes more electrically conductive. What makes plasma different than a gas like steam is plasma has a charge and thus it both creates and is modified by electromagnetic fields. The resulting charged ions and electrons become influenced by long-range electromagnetic fields, making the plasma dynamics more sensitive to these fields than a neutral gas.

EMPIRE, which is funded by the Grand Challenge-Laboratory Directed Research and Development (GC-LDRD) and ASC-Advanced Technology Development and Mitigation, has a unique ability to design plasma environments, which is particularly useful for Sandia's pulsed power research. Accelerators like the Z Machine, or Saturn or HERMES (High-Energy Radiation Megavolt Electron Source), take energy

For example, the Z Machine uses a z-pinch to compress material to perform research on inertial fusion for clean power (see *Z Machine--An Engine of Discovery* article on page 23). The z-pinch, also known as zeta pinch, is a type of plasma confinement system that uses an electrical current in the plasma to create enormous magnetic fields which compresses a target to pressures seen in the sun. Energy is delivered rapidly from the capacitors and into the z-pinch, creating a flow of electrical power. In Figure 1, the Marx generators (top) are on the outside, which are discharged into the coaxial intermediate storage capacitor (represented in blue anode and red cathode). From there, there are the laser-triggered gas switch, then pulse forming lines, and then the transmission lines to the vacuum insulator stack in the center. As power flows along the stages, the power density (energy/volume) increases as the physical domain gets smaller and smaller. As the power flows close to the load (red cathode and blue anode, lower right), the voltage gets so high (~2M volts) that it rips electrons out of the metal in the red area. As power flows along the stages, the power density (energy/volume) increases as the physical domain gets smaller and smaller. As the power flows down the red electrodes into the blue electrodes, the voltage gets so high (~2M volts) that it rips electrons out of the metal in the red area, combined with desorbed impurities in the metal which are ionized very quickly, creating a plasma in the gap. In some experiments, the high voltages can carry the plasma between the electrodes, reducing the efficiency for electrical power to reach the target. Fortunately, in addition to the high voltages, there is a very high current which creates large magnetic fields. These magnetic fields prevent the flow of plasma from the cathode to the anode, keeping the plasma tight to the cathode. The combination of the high self-generated magnetic fields which keeps the plasma from crossing the gap is what defines a Magnetically Insulated Transmission Line (MITL), which looks like two cones, where the power flows on the surface.

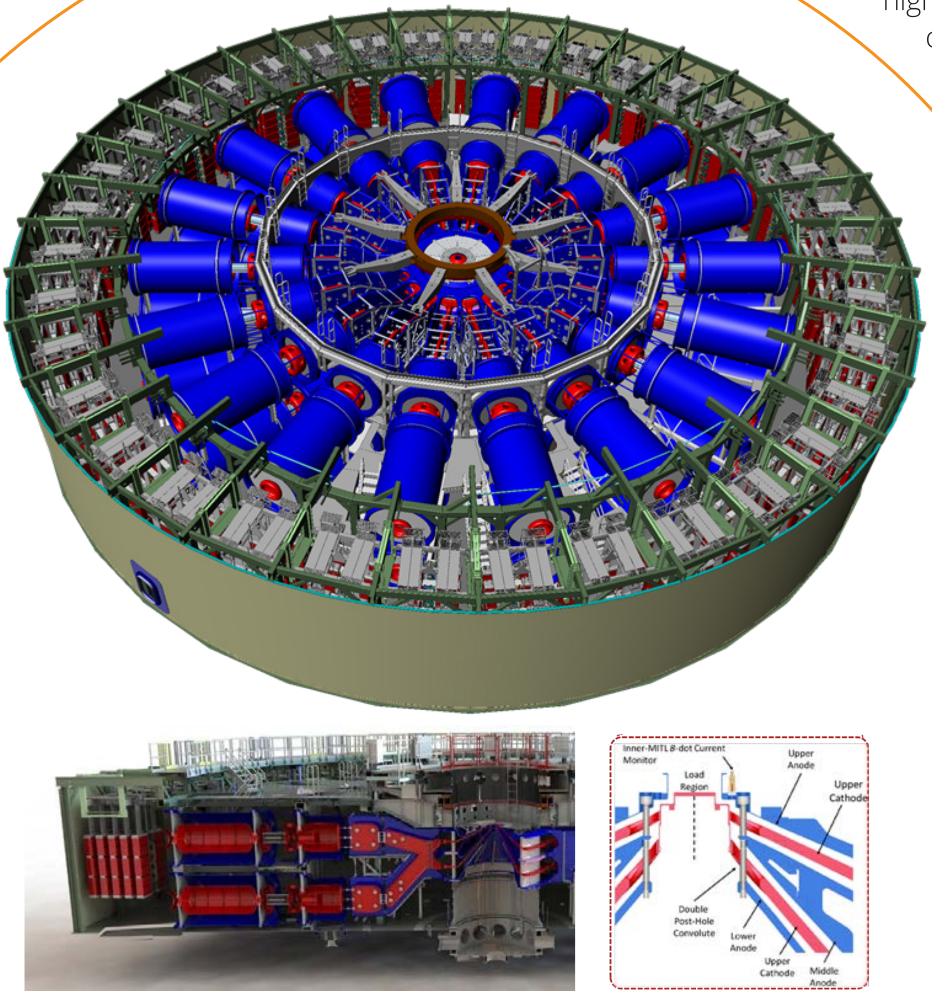


Figure 1.

Internal schematic of Z Machine, a pulsed power machine

Scan using SNLSimMagicApp

Any current crossing the gap is a loss, the goal is to minimize the loss of energy in the MITL and maximize the energy delivered to a target.

EMPIRE's unique ability to predict plasma environments has a wide range of applications including pulsed power, accelerator, oscillators used in national security, commercial and medical devices.

EMPIRE takes a transmission line design and predicts how much energy is delivered to the target vs. how much is lost in the various locations of the accelerator architecture. The modeling is split at the point where the blue area begins, with everything on the left modeled as a circuit and the area to the right modeled as the self-consistent evolution of the plasma and the power flow. EMPIRE models Maxwell's equations for electromagnetics, Newton's Laws with special relativity corrections for the evolution of the plasma, and tracks the temperature on the surface of the MITLs, modeling desorption physics which allows gas and water trapped on the surface and in the bulk of the metal to enter the gap and be turned into plasma.

EMPIRE uses a particle-in-cell formulation for plasma modeling. This is a dual mesh formulation where the electromagnetics are computed on a traditional tetrahedral finite element mesh allowing for local refinement to have resolution where needed and also a particle Lagrangian mesh for particles which are allowed to move as the physics dictates. The particles represent electrons and ions, and since there are too many electrons and ions to represent individually, EMPIRE models them as movements of sections (say one billion) as a unit.

EMPIRE has a wide range of applications, including pulsed power and other plasma devices such as microwave generators (magnetrons, gyrotrons) used in radar, accelerators like the Stanford Linear Accelerator Laboratory (SLAC), laser plasma interactions like National Ignition Facility (NIF), and even for X-ray devices used in medicine. ●

Sandia Parallel Aerodynamics Reentry Code (SPARC)

the Future of Production and Research Aerodynamics

TEAM *Justin Smith*

Contributing Writer: Johann Snyder

At hypersonic speeds, an atmospheric flight vehicle must survive the intense aerodynamic heating environments that are transferred from the hypersonic flowfield to the vehicle surface. The vehicle configuration and external conditions define the heating distribution on the vehicle and the type of thermal protection system necessary to protect its internal components. The Sandia Parallel Aerodynamics and Reentry Code (SPARC) simulates the aerodynamic environment and material thermal response for atmospheric flight vehicles from subsonic to hypersonic speeds—enabling the design, development, and analysis of such vehicles with higher fidelity and more quickly than previously possible.

SPARC solves the compressible Navier-Stokes and Reynolds-Averaged Navier-Stokes (RANS) equations for both laminar and turbulent flow. To do so, SPARC employs several working discretization schemes, with a cell-centered finite volume discretization as the primary scheme. High-Mach number flowfields are achievable through finite rate multi-species reacting gas and two- or three-temperature thermal non-equilibrium models. SPARC's cell-centered discretization scheme allows for solutions on both structured and unstructured meshes, which enables the simulation of complex and diverse vehicle geometries to meet evolving mission needs.

SPARC's principal applications are in computational fluid dynamics (CFD) analysis of transonic flowfields for gravity bomb analyses and hypersonic flowfields for reentry vehicle analyses. SPARC is also used to simulate the material thermal response and ablation of thermal protection materials (TPS). One-way and two-way multiphysics couplings exist between the CFD and ablation solvers within the code, enabling a coupled "virtual flight test" capability to simulate vehicle response from release to impact.

For reentry applications, a sacrificial thermal protection system (TPS) is often employed, which mitigates the heat load by radiating it back into space, conducting it into the body, or through internal decomposition of the TPS material. SPARC's ablation module solves the transient heat equation and associated thermal response and ablation equations for both non-decomposing ablators, such as carbon-carbon, and decomposing ablators, such as carbon-phenolic. SPARC's ablation module can be run for both full 3D problems or more rapidly for 1D problems using a specialized version that incorporates an internal 1D automatic mesh generation capability.

A differentiating capability of SPARC is its Multi-Fidelity Toolkit (MFTK), which employs three different levels of fidelity to produce aerodynamics and aeroheating environments.

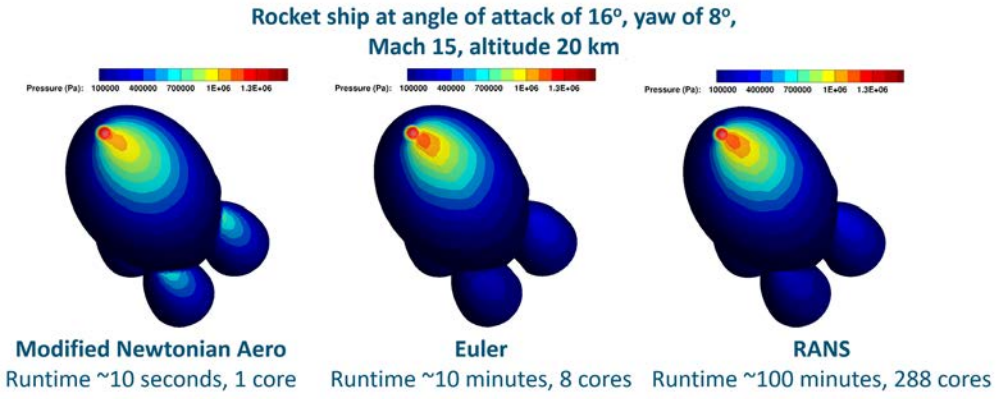

Rocket ship at angle of attack of 16º, yaw of 8º,
Mach 15, altitude 20 km
Pressure (Pa): 100000 400000 700000 1E+06 1.3E+06
Pressure (Pa): 100000 400000 700000 1E+06 1.3E+06
Pressure (Pa): 100000 400000 700000 1E+06 1.3E+06
Modified Newtonian Aero
Runtime ~10 seconds, 1 core
Euler
Runtime ~10 minutes, 8 cores
RANS
Runtime ~100 minutes, 288 cores

Figure 1
Multi-fidelity rocket ship evaluation

> *The new capability will modernize/parallelize/automate the simulation process and generate a common interface for each of the software tools.*

These can be tailored to the level of rigor required by a problem. The lowest fidelity is a Modified Newtonian Aerodynamics (MNA) method with boundary layer corrections. Solutions are achievable on complex geometries in just seconds and require very low hands-on subject matter expert (SME) input. This method is suitable for rapid-turnaround design studies employing hundreds of thousands to millions of solutions.

The mid-fidelity is an inviscid SPARC solution coupled with a momentum energy integral technique (MEIT) boundary layer correction. Solutions are achievable on complex geometries in just minutes, but unlike the MNA method, significant SME input is required to generate a computational mesh for the Euler solution. However, by decoupling the inviscid solution from the viscous terms, this approach avoids the gridding requirements of the viscous boundary layer in the non-linear solver, thereby reducing both cost and fidelity. This method is suitable for medium-turnaround engineering-fidelity analyses employing hundreds to thousands of solutions.

At the highest fidelity is the full RANS capability of SPARC, which models the majority of flow physics observed in real flight. Effects such as shock-boundary layer interaction, separated flow, and fin-fin interactions (critical for maneuvering systems) are naturally captured in RANS solvers. Solutions are achievable on complex geometries in hours. This fidelity comes with the highest computational cost and the highest burden of SME time to generate a computational mesh and slows down the time to solution considerably. This method is suitable for high-fidelity analyses employing up to hundreds of analysis points.

SPARC's MFTK is designed to quickly compute accurate performance with data from all levels of fidelity using a Hierarchical Kriging Interpolation algorithm. Generally, the methods utilize the trend information at lower-fidelity levels and anchor these trends to high-fidelity data. Thus, a full aerodynamics model for a vehicle can be achieved through a small number – typically fewer than one hundred – of high-fidelity points, anchoring thousands or more of low- and mid-fidelity points. This can cut the time to solution of such a model from months to just a couple of weeks.

Future plans for SPARC include further development of the MFTK to incorporate fast-running thermal response tools, including the 1D SPARC ablation capability, for an end-to-end one-way coupled aero/thermal analysis capability. This coupled aero/thermal analysis capability will replace existing legacy aero-thermal tools that have been employed at Sandia for flight vehicle analysis for over 40 years. The new capability will modernize/parallelize/automate the simulation process and generate a common interface for each of the software tools. ●

The Kokkos EcoSystem

TEAM *Christian Trott*

Contributing Writer:
Johann Snyder

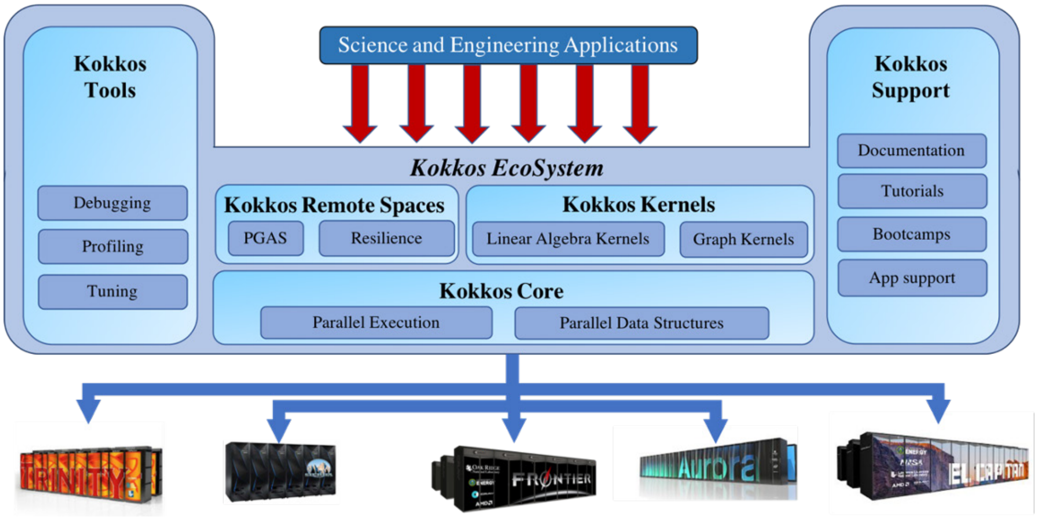
In 2016/2017, the field of HPC entered a new era driven by fundamental physics challenges to produce more energy and cost-efficient processors. Since the convergence on the Message-Passing Interface (MPI) standard in the mid-1990s, application developers enjoyed a seemingly static view of the underlying machine – that of a distributed collection of homogeneous nodes executing in collaboration. However, after almost two decades of dominance, the sole use of MPI to derive parallelism acted as a limiter to improved future performance. While MPI is widely expected to continue to function as the basic mechanism for communication between compute nodes for the immediate future, additional parallelism is required on the computing node itself if high performance and efficiency goals are to be realized.

When reviewing the architectures of the top HPC systems today, the change in paradigm is clear: the compute nodes of the leading machines in the world are either powered by many-core chips with a few dozen cores each, or use heterogeneous designs, where traditional central processing units marshal work to massively parallel compute accelerators which have as many as 200,000 processing threads in flight simultaneously. Complicating matters further for application developers, each processor vendor has its own preferred way of writing code for their architecture.

The Kokkos EcoSystem was released by Sandia in 2017 to address this new era in HPC system design by providing a vendor independent performance portable programming system for scientific, engineering, and mathematical software applications written in the C++ programming language. Using Kokkos, application developers can be more productive because they will not have to create and maintain separate versions of their software for each architecture, nor will they have to be experts in each architecture's peculiar requirements. Instead, they will have a single method of programming for the diverse set of modern HPC architectures.

While Kokkos started in 2011 as a programming model only, it soon became clear that complex applications needed more. It is also critical to have portable mathematical functions and developers need tools to debug their applications, gain insight into the performance characteristics of their codes and tune algorithm performance parameters through automated processes. The Kokkos EcoSystem addresses those needs through its three main components: the Kokkos Core programming model, the Kokkos Kernels math library, and the Kokkos Tools project.

Scan using SNLSimMagicApp



Kokkos Core is a programming model for parallel shared memory architectures. The model enables most application-written code to be performance portable across architectures. The programming model includes abstractions for frequently used parallel computing patterns, policies that provide details for how those computing patterns are to be applied, and execution spaces that denote on which compute resources the parallel computation is performed. The programming model also includes patterns for common data structures, policies that provide details for how those data structures are laid out in memory, and memory spaces that denote in which memory the data will reside.

The Kokkos Core programming model works by requiring that application development teams implement their algorithms in terms of Kokkos' patterns, policies, and spaces. Kokkos Core is then free to map these algorithms and data structures onto each target architecture according to architecture-specific rules necessary to achieve the best performance. While other programming models support execution patterns, execution policies, execution spaces, and memory spaces, only Kokkos supports memory layouts and memory traits, which are necessary for performance portability.

Kokkos Kernels is a software library of linear algebra and graph algorithms used across many HPC applications to achieve the best performance on every architecture. The baseline version of this library is written using the Kokkos Core programming model for portability and good performance. The library has architecture-specific optimizations or can utilize calls to vendor-specific versions of these mathematical algorithms where needed. This further reduces the amount of architecture-specific software that an application team will need to develop, thus further reducing their modification cost to achieve "best in class" performance.

Kokkos Tools is an innovative "plug-in" software interface and a growing set of tools that understand the Kokkos programming model and runtime. Providing debugging, profiling and tuning tools the project helps application developers during the entire life-cycle of a code. Debugging and correctness tools help identify complex software bugs and corner cases that often even evade manual inspection. Development teams can use the performance profiling tools to determine how well they designed and implemented their algorithms and to identify portions of their software that need improvement. Most recently, auto-tuning tools were added which allow applications to adapt to new hardware automatically, reducing the need for developers to fine tune the code for every new HPC platform their users want to leverage. Furthermore, the "plug-in" interface for tools allows third-party tool providers to hook into Kokkos codes the same way, enabling widely used profiling tools such as Tau and HPCToolkit to understand Kokkos.

Today, the Kokkos EcoSystem allows an ever-larger number of application teams to achieve portability and improve performance on advanced computing architectures. Not just a Sandia product anymore, the core Kokkos team now consists of developers distributed over five DOE national laboratories who work on maintaining and improving the EcoSystem as well as support Kokkos users at their institutions. New efforts such as Kokkos Remote Spaces cover a wider range of future HPC application needs and a dedicated support effort helps to train and educate software engineers and computational scientists. Kokkos is now used by hundreds of HPC developers around the world. Within DOE's Exascale Computing Project, it serves as the underlying portability layer for almost two dozen projects. Kokkos is also a basis for DOE laboratories to propose improvements to the ISO/C++ language standard such that, eventually, Kokkos capabilities will become native to the language standard. But until then: Performance Portability is Kokkos. ●

Advancing the Field of Reduced-order Modeling

TEAM

Irina Tezaur
Eric Parish
John Tencer
Patrick Blonigan
Francesco Rizzi

Despite improved algorithms and the availability of massively parallel computing resources, “high-fidelity” models are, in practice, often too computationally expensive for use in a design or analysis setting. The continuing push to incorporate the quantification of uncertainties, critical to many science and engineering applications, into modeling efforts presents an intractable computational burden for most real-world systems.

As a strategy for reducing the computational cost of such models while preserving high levels of fidelity in important components of the system, researchers have pursued projection-based reduced-order modeling—a technique that integrates ideas from data science, modeling, and simulation. Reduced-order modeling is a powerful tool that can enable real-time analysis and alleviate the computational burden posed by high-dimensional uncertainty quantification (UQ) problems.

The first step in building a reduced-order model (ROM) is to execute a set of analyses (such as running a full order model or FOM) during an offline 'training' stage. These analyses generate data that are used to extract important physical features, such as low-dimensional solution manifolds and interpolation points for approximating nonlinear functions. Next, the dimensionality and computational complexity of the high-fidelity model is reduced by projecting the governing equations onto a low-dimensional manifold and introducing other approximations where necessary. The resulting ROM can then be rapidly evaluated during an online 'deployed' stage, thereby enabling multi-query analyses such as UQ and optimization, as well as on-the-spot decision making and control.

Sandia's continued investment in reduced-order modeling research has contributed to the Labs' emergence as a leader in this growing field. It has also led to the development of cutting-edge algorithms and the creation of agile and performance-portable software libraries, which have made possible the deployment of ROM technologies on mission-critical problems and applications.

While advances have been made in reduced-order modeling during the past 15-20 years, there remain significant challenges in applying ROMs to complex mission-critical problems relevant to Sandia and the DOE. Sandia researchers are addressing these challenges through the development of novel model reduction methodologies that improve on existing methods and possess desirable mathematical properties such as stability and convergence. These methodologies borrow and adapt concepts from not only traditional numerical analysis/methods, but also from the up-and-coming fields of machine learning (ML) and artificial intelligence (AI).

Sandia's continued investment in reduced-order modeling research has contributed to the Labs' emergence as a leader in this growing field.

Development of novel reduced-order modeling methodologies

By leveraging ideas from computational fluid dynamics (CFD) and numerical discretizations such as closure modeling and energy-based formulations, Sandia researchers have developed techniques that give rise to ROMs with improved stability properties, as well as ROM formulations that preserve the intrinsic structure of the underlying physical system (e.g., Lagrangian or Hamiltonian structure, mass/momentum/energy conservation). Space-time and windowed least-squares model reduction approaches originating at Sandia are based on well-established concepts in time-discretization and optimization, and have addressed several key deficiencies in existing model reduction techniques, including their dependence on time discretization, their potential for exhibiting exponential error growth in time, and their parallel efficiency limitations.

Recently, significant progress has been made towards overcoming a well-known limitation of predictive ROMs: if a solution feature was not observed during the training stage of the ROM, the ROM will be incapable of capturing the feature when deployed online and may therefore lack robustness. This problem can be remedied through an online basis adaptation algorithm known as "adaptive h-refinement for ROMs," an approach developed at Sandia that is similar in flavor to adaptive mesh refinement (AMR).

In addition to improving ROM accuracy and robustness, Sandia's research in ROM has led to the construction of ROMs that are more efficient for large-scale multi-scale/multi-physics systems through, for instance, the development of a domain-decomposition-based "divide and conquer" model reduction methodology for decomposable systems that allows subdomain/component ROMs to be assembled into a full-system ROM in arbitrary ways.

The rise of ML/AI has further motivated Sandia researchers to explore ways to integrate concepts from these fields into the field of model reduction. This resulted in the creation of a machine-learning-based framework for quantifying and modeling ROM errors, and, more recently, the introduction of a novel nonlinear deep convolutional autoencoder architecture within a classical projection-based ROM.

Advancing ROM through ML and AI

Nearly all model-reduction techniques employ a projection of the governing equations onto a linear subspace of the original state space. Unfortunately, restricting the state to evolve in a linear subspace imposes a fundamental limitation on the accuracy of the resulting ROM. In particular, linear-subspace ROMs are often inadequate when applied to problems exhibiting a slowly decaying Kolmogorov n-width, e.g., advection-dominated problems such as the high-speed, turbulent flow over an airfoil or re-entry vehicle.

To address this, Sandia researchers developed a novel framework that involves projecting dynamical systems onto nonlinear manifolds (see Figure 1). These manifolds are generated in a computationally tractable manner by employing advanced ML techniques, namely convolutional autoencoders. The accuracy of the resulting ROMs can be improved further through the introduction of structure preserving ROM formulations, originally developed for ROMs with linear subspaces. This novel ML application resulted in Sandia researchers developing ML techniques to meet the unique challenges of this space such as the new convolutional autoencoder architectures that are compatible with the unstructured spatial discretization schemes common in computational physics. The use of nonlinear manifold projection (autoencoders) has resulted in a significant improvement in solution accuracy over more well-established linear methods, such as Least-Squares Petrov-Galerkin (LSPG) projection (see Figure 2). ●

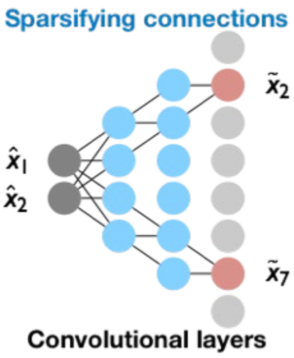


Figure 1

A sparsified Convolutional Neural Network (CNN)-autoencoder-based ROM that can be trained on simultation results on a hyper-reduced mesh.

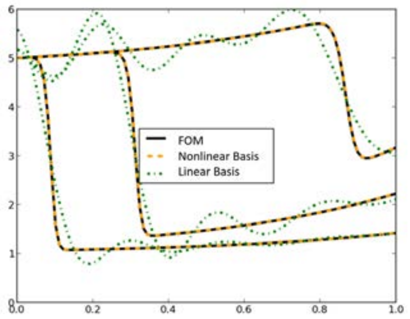


Figure 2

The use of a nonlinear (autoencoder) basis results in an improvement in the solution accuracy for LSPG ROMs for the solution of the 1D Burgers equation.

Scan using SNLSimMagicApp

Pressio: an HPC Library to Enable Reduced-order Modeling

TEAM

Francesco Rizzi
Patrick Blonigan
Eric Parish
John Tencer
Irina Tezaur

Efficiently implementing proven model reduction methods in large-scale simulation codes is, perhaps, the single largest barrier to their widespread adoption in industrial applications. Naïve implementations of model reduction methodologies require the modification of low-level operations and solvers for each simulation code of interest. Such an approach is not sustainable for institutions employing dozens of rapidly evolving simulation codes for different types of analysis, and commercial codes, which typically do not expose the required low-level operators and solvers.

As a potential solution to such an artisanal, one-off approach, researchers at Sandia have developed a software framework known as Pressio to mitigate implementation burdens without compromising performance. Pressio is an open-source project providing state-of-the-art model-reduction methods for any dynamical system expressible as a system of parameterized ordinary differential equations (ODEs). This simple, expressive mathematical framework is leveraged as a pivotal design choice to enable a minimal application programming interface (API) that is natural to dynamical systems, as illustrated in Figure 1. Pressio relies on modern C++ and generic programming to support applications with arbitrary data types and programming models. The library is also complemented with Python bindings to expose these C++ functionalities to Python users with negligible overhead and no user-required binding code.

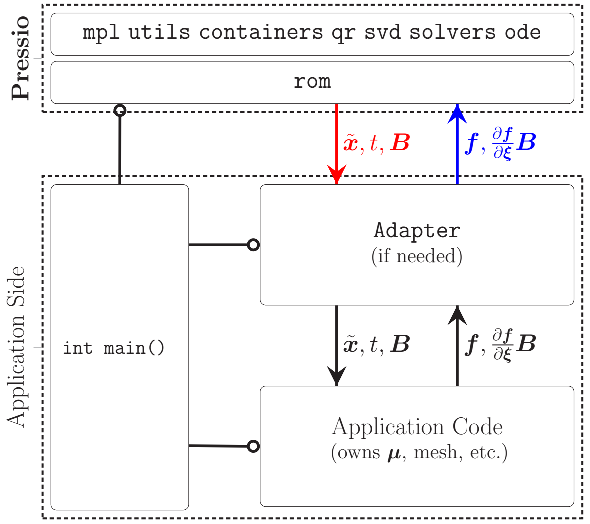

Pressio
mpl utils containers qr svd solvers ode
rom
$\tilde{\boldsymbol{x}}, t, \boldsymbol{B}$
$\boldsymbol{f}, \frac{\partial \boldsymbol{f}}{\partial \xi}\boldsymbol{B}$
Adapter
(if needed)
Application Side
int main()
$\tilde{\boldsymbol{x}}, t, \boldsymbol{B}$
$\boldsymbol{f}, \frac{\partial \boldsymbol{f}}{\partial \xi}\boldsymbol{B}$
Application Code
(owns $\boldsymbol{\mu}$, mesh, etc.)

Figure 1

Schematic of the interaction between Pressio and the application, highlighting elements belonging to Pressio and those belonging to the application. The red arrow is used to indicate information/data computed by and exiting Pressio, while blue is used to denote information/data computed by the application and entering Pressio.

Pressio is unique in that it enables extensive collaborations both across Sandia and with academia to develop, prototype, and test novel model reduction methodologies. Because Pressio has built-in HPC capabilities, it is able to uniquely target next-generation platforms and codes originally developed solely for high-fidelity modeling. Using Pressio, researchers can take a new method from academic partners and immediately test the method on large-scale mission critical problems.

Scan using SNLSimMagicApp

Deployment of ROM technologies on mission-critical problems and advanced HPC architectures

ROM methodologies are being deployed in a number of Sandia codes including SPARC, Sierra/ARIA, and Albany using Pressio, which has simplified the task of integrating model reduction into Sandia's engineering simulators. A number of applications have been targeted, including the design/qualification of the captive-carry environment, as well as analyses involving hypersonic aerodynamic simulations relevant to the study of re-entry vehicles and missiles. The primary HPC resources employed for ROM R&D at Sandia include the CEE (Common Engineering Environment), HPC clusters, and the Skybridge supercomputer. Additionally, the Synapse machine, a GPU platform for deep learning applications, has been used for ROM research involving nonlinear manifold projection, which integrates ideas from machine learning. Future platform targets include Sandia's GPU-based clusters, including Weaver and Vortex. ●

ROM for Hypersonic Aerodynamic Simulations

Hypersonic aerodynamics plays a crucial role in a range of aerospace engineering applications including the design and analysis of missiles and re-entry vehicles. The expense and difficulty of flight tests and experiments for hypersonic applications has resulted in greater reliance on computational models for design and analysis than in other flight regimes. This dependence drives a need for uncertainty quantification (UQ) to enable practitioners to study and characterize the sources and propagation of error and uncertainties in these computational frameworks.

As part of a Laboratory Directed Research and Development (LDRD) project funded by the Autonomy for Hypersonics (A4H) mission campaign, Sandia researchers have been developing and deploying a model reduction tool chain to accelerate path planning, design, and control of hypersonic vehicles. Pressio, together with the Sandia Parallel Aerodynamics and Reentry Code, SPARC, has been used to create ROMs for hypersonic problems relevant to this mission campaign. This ROM implementation has been evaluated on a 3D computational fluid dynamics simulation of the HIFiRE-1 hypersonic flight vehicle flying at Mach 7.1 (7.1 times faster than the speed of sound in the air surrounding the vehicle), with impressive results. The Pressio ROMs were able to simulate the environment under different air speeds and densities than the training data from which these models were generated. Errors of roughly 1% or less for a variety of engineering quantities of interest were noted (see Figure 3). Importantly, the ROMs achieved impressive speedups of up to 1000x compared to their corresponding full order models. These speed-ups were made possible in part by solving the governing equations on a “sample mesh” (see Figure 2), defined on a small subset of grid-points comprising the mesh on which a high-fidelity simulation is typically performed. Further improvements in the robustness, accuracy, and efficiency of these ROMs are expected, as cutting-edge ROM methodologies developed at Sandia and beyond are incorporated into the Pressio framework. ●

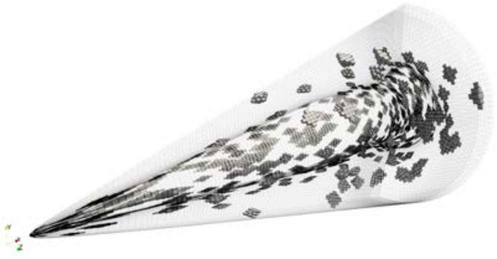


Figure 2

This picture shows the small subset of 813 randomly selected mesh cells (dark grey squares) used by the ROM. The cells comprising this so-called “sample mesh” make up roughly 1% of the entire mesh.

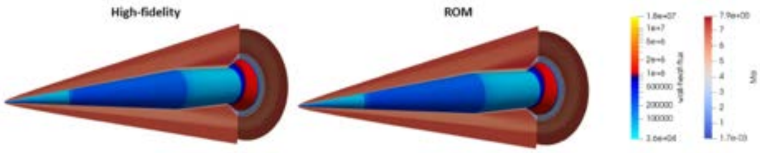


Figure 3

Example ROM solution for the HIFiRE-1 flight vehicle. The flow field around the vehicle is colored by Mach number, and the vehicle surface is colored by wall heat flux (an important quantity we use to determine how fast the vehicle heats up). The ROM solution (right) is indistinguishable from the analogous high-fidelity model solution (left) and achieves an error of 1% relative to the high-fidelity solution, all while requiring ~1000x less CPU-time.

Scan using SNLSimMagicApp

HPC RESPONDS IN REAL TIME

HOW RESEARCHERS HELPED INFORM THE COVID-19 RESPONSE

Unless otherwise indicated, this information has been authored by an employee or employees of National Technology and Engineering Solutions of Sandia, operator of Sandia National Laboratories with the U.S. Department of Energy. The U.S. Government has rights to use, reproduce, and distribute this information. The public may copy and use this information without charge, provided that this Notice and any statement of authorship are reproduced on all copies.

While every effort has been made to produce valid data, by using this data, User acknowledges that neither the Government nor operating contractors of the above national laboratories makes any warranty, express or implied, of either the accuracy or completeness of this information or assumes any liability or responsibility for the use of this information. Additionally, this information is provided solely for research purposes and is not provided for purposes of offering medical advice. Accordingly, the U.S. Government and operating contractors of the above national laboratories are not to be liable to any user for any loss or damage, whether in contract, tort (including negligence), breach of statutory duty, or otherwise, even if foreseeable, arising under or in connection with use of or reliance on the content displayed in this report.

Data-driven Epidemiological Inference and Forecasting

When COVID-19 began its spread across the United States in early 2020, it was important to understand the hitherto unknown rate of infection; to understand the rate of infection, one also needed to know the incubation period distribution of the disease. This was not an easy task because very little was known about the virus and how it spread. For Sandia scientists, the problem of forecasting the spread of an unknown virus sounded all too familiar.

In the early 1980s another unknown virus was killing individuals in San Francisco at an alarming rate. A prominent statistician, Prof. R Brookmeyer, of Johns Hopkins University (currently University of California, Los Angeles), wanted to understand the infection rate of this new virus. (DOI: 10.1080 /01621459.1988.10478599). Brookmeyer thought that by knowing how many people were *currently* exhibiting symptoms, he could work backwards and establish the rate at which they had been infected in the past. This was a difficult task because the incubation period – the time between infection and exhibition of symptoms – varies from person to person and can only be characterized via statistical summaries (i.e., it is a *random* variable characterized by a *distribution*). He created a modeling method to provide short-term forecasts of the outbreak.

TEAM

Jaideep Ray
Cosmin Safta

Contributing Writer:
Whitney Lacy

Sandia researchers were familiar with the modeling used by Brookmeyer from other disease modeling work at Sandia's Livermore campus. They thought a similar approach could be used to study the rate of spread of the new coronavirus. Its incubation period is much shorter than HIV's incubation period, and its infection rate far larger, which considerably simplified the modeling task. The goal was to produce short-term forecasts of the epidemic in a purely data-driven manner, free of any modeling assumptions, and thereby help the community understand how the virus would spread. By using a similar method to infer the latent infection rate curve from known data, Sandia researchers could predict the number of cases presenting symptoms over time. Infected cases observed on a given day are a consequence of people infected at various times in the past, coming out of incubation and presenting symptoms.

During the early days of the COVID-19 outbreak, *The New York Times* and Johns Hopkins University were collecting and publishing data on COVID-19 detections across the country, mostly from hospitals as the patients sought care. Using this raw data, Sandia researchers began a COVID-19 LDRD (Laboratory Directed R&D) project to model the outbreak in California and Italy, and specifically included Bernalillo County in New Mexico where Sandia has its largest number of employees.

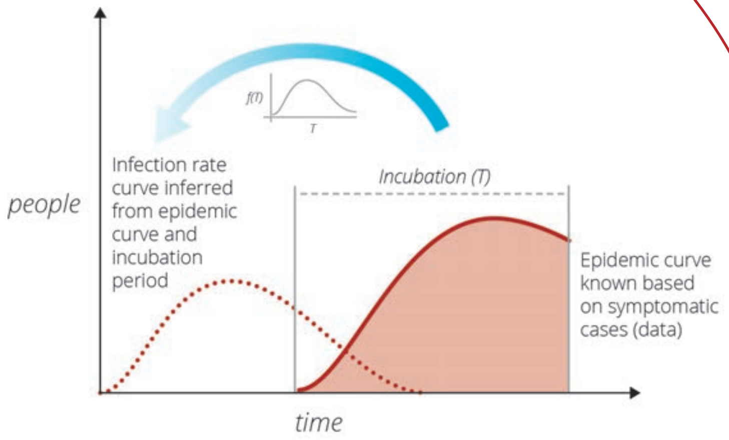
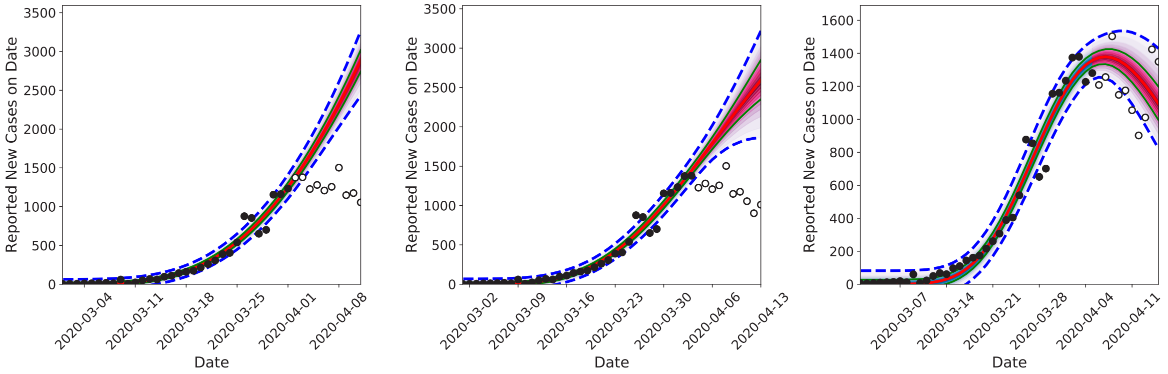

f(T)
T
Infection rate curve inferred from epidemic curve and incubation period
Incubation (T)
people
Epidemic curve known based on symptomatic cases (data)
time

Figure 1

Infection maps to the epidemic via incubation period

The modeling results surprised the researchers. In late March 2020, much of the United States entered lockdown. By the first week of April, the inferred infection curve showed that infections in California were flattening out, not going up. As more data became available and consistent inferences were drawn, the researchers concluded that the lockdown rules in place were slowing the spread of the virus.

Figure 2



Forecasts performed on April 1st, 3rd, and 5th for California. The black dots are the "new cases" data used to infer the latent infection rate. The red line is the mean forecast and the dashed blue lines bracket the 95% credibility interval on the forecasts. The red region is the inter-quartile range. The white circles are the data collected in the week after the forecast to check the quality of the forecasts. On the left, the forecasts on April 1st showed an increasing number of new cases, though in the succeeding week the new cases showed a dramatic decline. By April 3rd, the forecasts had begun to turn, as the inference procedure detected signals of an infection-curve flattening due to social distancing. By April 5th, the forecasts showed a downturn in new cases as the signal from the flattened (latent) infection rate is now evident in the forecast.

Scan using SNLSimMagicApp

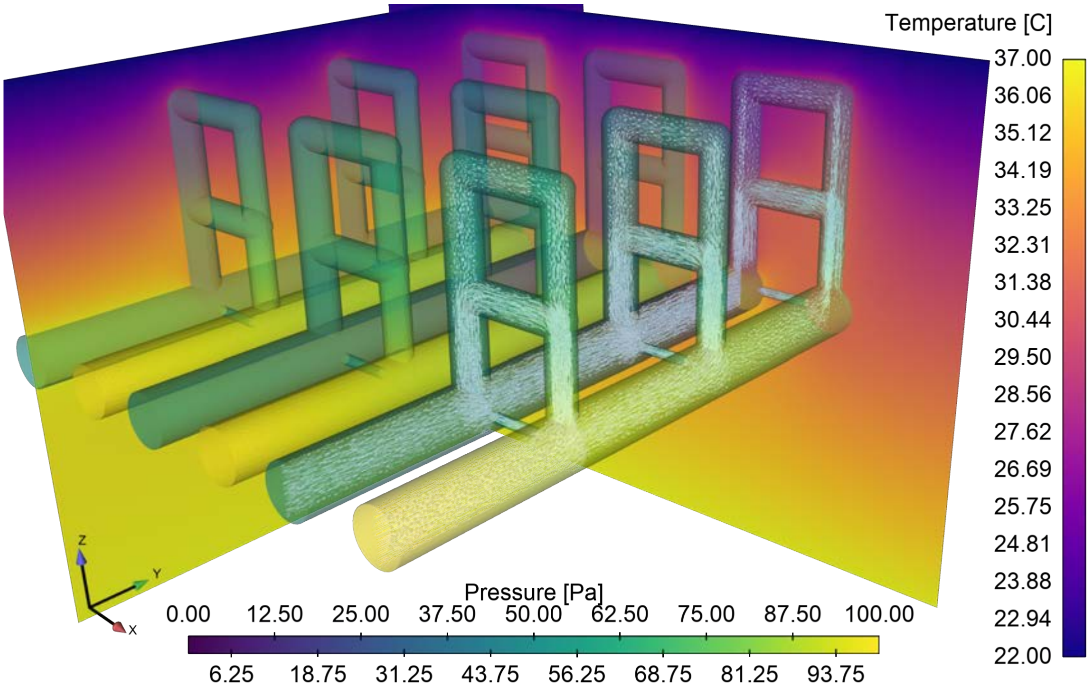
The team's accurate forecast caught the attention of other Sandia researchers studying the medical resource needs across the country (see *Sandia COVID-19 Medical Resource Modeling* story, page 50). and the two Sandia researchers in Livermore were asked to provide their daily infection rate predictions for all of the United States. Initially the team used individual personal computers (PC), but they quickly realized that running their model for every state would require harnessing the power of Sandia's HPC resources.

The researchers recruited a larger group of computational scientists with expertise in statistical inference, software engineering, and parallel computing. The prototypical software and algorithms were restructured for optimal performance on multi-core computers and scaled up for inference-and-forecasting procedures for all 50 states. Calculations that would run 10 hours overnight on a PC could now be performed in 30 minutes with HPC, and additional forecasting was added for a few foreign countries and regions, such as amalgamations of counties in New Mexico. Calculations were performed on the Common Engineering Environment Advanced Compute Servers every night, and forecasts were archived. These calculations revealed differences in flattening the infection rate curve where social distancing restrictions were in place and where they were not (e.g., in Northwest NM [primarily McKinley county]).

Less than three weeks after starting their LDRD research into understanding the spread of the new coronavirus, the team of computational researchers at Sandia were able to predict future infection rates based on data on new cases, via the process of inferring the latent past infection rate. Their results also highlighted that social distancing was a major contributor to slowing the spread of the disease. ●

Multiphysics Modeling Informs Early Fever Detection Sensor Development

Sandia HPC resources are being used to support a COVID-19 project focused on developing microneedle-based in-situ temperature monitoring technologies. Embedding temperature sensors 1 to 2 millimeters into the skin using Sandia-developed microneedles may allow for early detection of elevated body temperatures prior to it being detectable by external temperature monitors. If successful, this technology will be continually wearable and allow for continuous monitoring and early detection of deviations from the norm, possibly indicating the onset of an infection. Detailed representation of the upper layers of the skin, to include multiple layers of blood vessels and capillary loops, are being computationally modeled using Sandia's Sierra finite element modeling code suite to model heat transport within the layers, providing a detailed reference to the body conditions that the microneedle-based sensors will be observing. The computational model, pictured in the attached figure, combines blood flow through the complex vessel/capillary networks and the associated heat diffusion and convection. By parameterizing the geometry, the model can quantify the impact of dilation/constriction of blood vessels on heat transport to the skin surface through capillary loops, a key mechanism by which the body controls body temperature. ●



Parameterized geometric representation of the upper layers of skin, including blood vessels and capillary loops. The nearest vessel shows vectors representing blood flow rates through the vessel. Vessel walls are colored by blood pressure. Slices through the entire domain indicate the local temperature, showing that capillary loops provide the primary heat transport to the skin surface.

Written by Scott Roberts

Sandia COVID-19 Medical Resource Modeling

TEAM

Laura Swiler
Teresa Portone
Walter Beyeler

Contributing Writer:
Whitney Lacy

When the Ebola outbreak of 2014 ravaged West Africa, blood samples from ailing people would often be sent to a laboratory for testing, but the closest lab was hundreds of miles away. In more urban areas, blood samples were sent to the closest labs, but those labs were often already overwhelmed by the sheer volume of samples to test. Staff had no way of knowing that another lab, a little further away, had plenty of capacity. Sandia recognized this problem and quickly stepped in to help. Using the power of HPC, Sandia created transportation models to quickly determine possible locations for mobile diagnostic laboratories that could better support regions most affected by the outbreak.

In late 2019, a new coronavirus outbreak was quickly becoming a global concern, resulting in COVID-19, a deadly disease caused by the coronavirus with no vaccine in sight. In early 2020, federal and state officials worried that what was happening in Italy—overwhelmed hospitals with too few medical resources—would soon be happening in the United States. How could officials know, in over 3,000 counties across the country, which hospitals had enough of the right medical resources on hand to protect frontline workers and treat infected patients? Which states had excess capacity and might be able to lend resources to other states? In order to provide decision makers with the best answers to these critical questions, Sandia volunteered the power of their HPCs.

The power of HPC allowed quick delivery of results that can inform decision makers across the country.

Leveraging their experience in Africa, Sandia researchers began their effort by developing discrete event mathematical models to track patient progress through a hospital treatment system. As a patient enters the medical system, many factors affect possible treatment paths, such as a patient's underlying health conditions or a hospital's medical resources on hand. Researchers had to incorporate this treatment path uncertainty and the ranges of resource use per patient to provide risk indicators. For patients arriving at hospitals with COVID-19 symptoms, how many will need advanced care? Will the hospitals have enough consumable resources on hand (masks, gowns, gloves, face shields, sedatives) to protect the frontline staff? Will hospitals have enough fixed resources (regular or ICU beds, ventilators) for patients? And importantly, will hospitals have enough personnel resources (physicians, ICU nurses, respiratory therapists) to handle an influx of critically ill patients?

Using data from an epidemiological ("epi") model, which determines the spread of a virus, Sandia researchers created a "resource demand" model to predict medical resource needs at any geographic scale of information available. While a resource model can take any epi model data as input, for these studies Sandia used county-level patient streams generated from Los Alamos National Laboratory's existing EpiGrid model.

The resource model predicts the medical resources (consumable, fixed, and personnel) needed based on the epi model's patient arrival stream predictions.

Using two of Sandia's institutional program HPC clusters—Ghost and Uno—the generated patient streams were run through the resource model for each of 3,145 counties in the United States, where each county-level run involved 100 samples per scenario to perform uncertainty analysis. Three different social distancing scenarios were investigated. This resulted in approximately 900,000 individual runs of the medical resource model, requiring 15 node hours (540 processor hours), on the HPCs. The results included mean estimates per resource per county, as well as uncertainty in those estimates (e.g., variance, 5th and 95th quantile, and exceedance probabilities).

Within a few weeks of starting this study, Sandia was able to determine the maximum number of resource needs accounting for parameter uncertainties, such as the probability that a patient goes into the ICU, needs a ventilator, the length of stay, etc. Resource needs over time (i.e., the number of ICU beds needed over time) were calculated with uncertainty bounds. Lastly, Sandia calculated state and county risk indicators, such as the percentage of ICU beds available depending on capacity needed.

Sandia's resource modeling proved valuable in this very critical time in U.S. history. The power of HPC allowed quick delivery of results that can inform decision makers across the country. As updated patient stream projections become available from the latest epidemiology models, the analysis can be re-run quickly to provide resource projections in rapidly changing environments. ●

Evaluate. Innovate. Repeat.

Computing Teams at Sandia Chart New Course for Next-generation Architecture

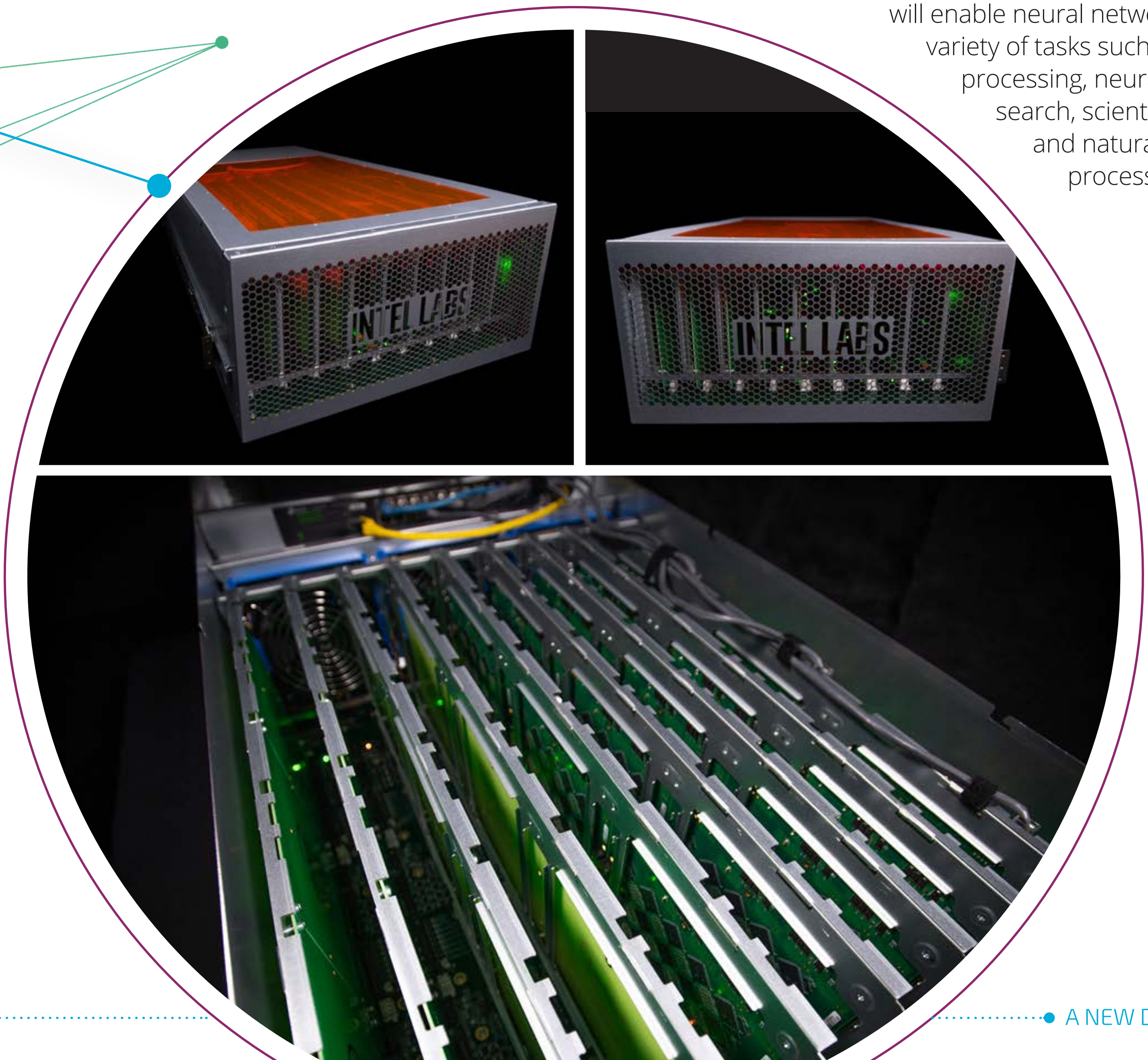
Traditional HPC has evolved into a broad array of architectures to meet a growing range of computing needs and workload. New technologies continue to deliver unprecedented performance in areas such as machine learning, biologically inspired computing, and quantum processing. Early investigation of how predictive simulation computational designs need to evolve to leverage the benefits of these architectures will help drive technology decisions at the national labs. By exploring these diverse computational methodologies to harness increased simulation fidelity, scalability and performance, Sandia is revolutionizing the national security mission space.

Sandia continues to be at the forefront of technology investigation through its acquisitions of new testbed systems. The following are some of the systems procured and managed in a partnership between Sandia's Extreme Scale Computing and IT Infrastructure Services groups.

Sandia’s new Fujitsu system is the first in DOE, and one of the first systems in the world, with A64FX processors. This new system couples ARM processors, wide vector units using Scalable Vector Extensions, and on-package High Bandwidth Memory (HBM) which provides more than double the memory bandwidth of traditional technologies. This system could benefit algorithms that do not perform well on GPUs.

Sandia acquired a large-scale spiking neuromorphic testbed to explore novel neural-inspired approaches to computation. As the first testbed in a multi-year partnership with Intel, the 50 million neuron coupled with a 50 billion synapse Loihi system is one of the five largest spiking neuromorphic platforms in the world. This capability will allow researchers to investigate how scale can enable a variety of applications and simulations. This novel approach to computation in systems like Loihi employs event—driven computation, thereby offering greater energy efficiency.

Purpose built for machine learning, Sandia’s Graphcore system is an artificial intelligence targeted system with a unique chip called an Intelligence Processing Unit (IPU). The IPU employs a much larger core count and access to more memory than central processing units or GPUs. This capability will enable neural networks for a variety of tasks such as image processing, neural architecture search, scientific computing, and natural language processing.



By exploring these diverse computational methodologies to harness increased simulation fidelity, scalability and performance, Sandia is revolutionizing the national security mission space.

The recent upgrade of the Cray compass software collaboration platform to the new HPE/Cray Shasta architecture incorporates a next-generation high-speed network interconnect with features that can mitigate message contention and congestion. This result was made possible by DOE investments in its Exascale Path Forward program. In addition, this upgrade will include new processors from AMD. AMD was recently selected as the technology provider for many leadership class platforms in the DOE complex.

Demand for leadership class computing cycles continues to grow. NNSA's largest current system, Sierra, deployed at LLNL, offers access to both high-performance CPU and GPU components, with the links to the most appropriate hardware devices for each type of physics or algorithm being used. In response to the increased demand and requirements for code porting, optimization and initial science runs, local testbed versions of Sierra were expanded by 33%. ●

ACKNOWLEDGEMENTS

Editor
Yasmin Dennig

Design
Stacey Long Reynolds

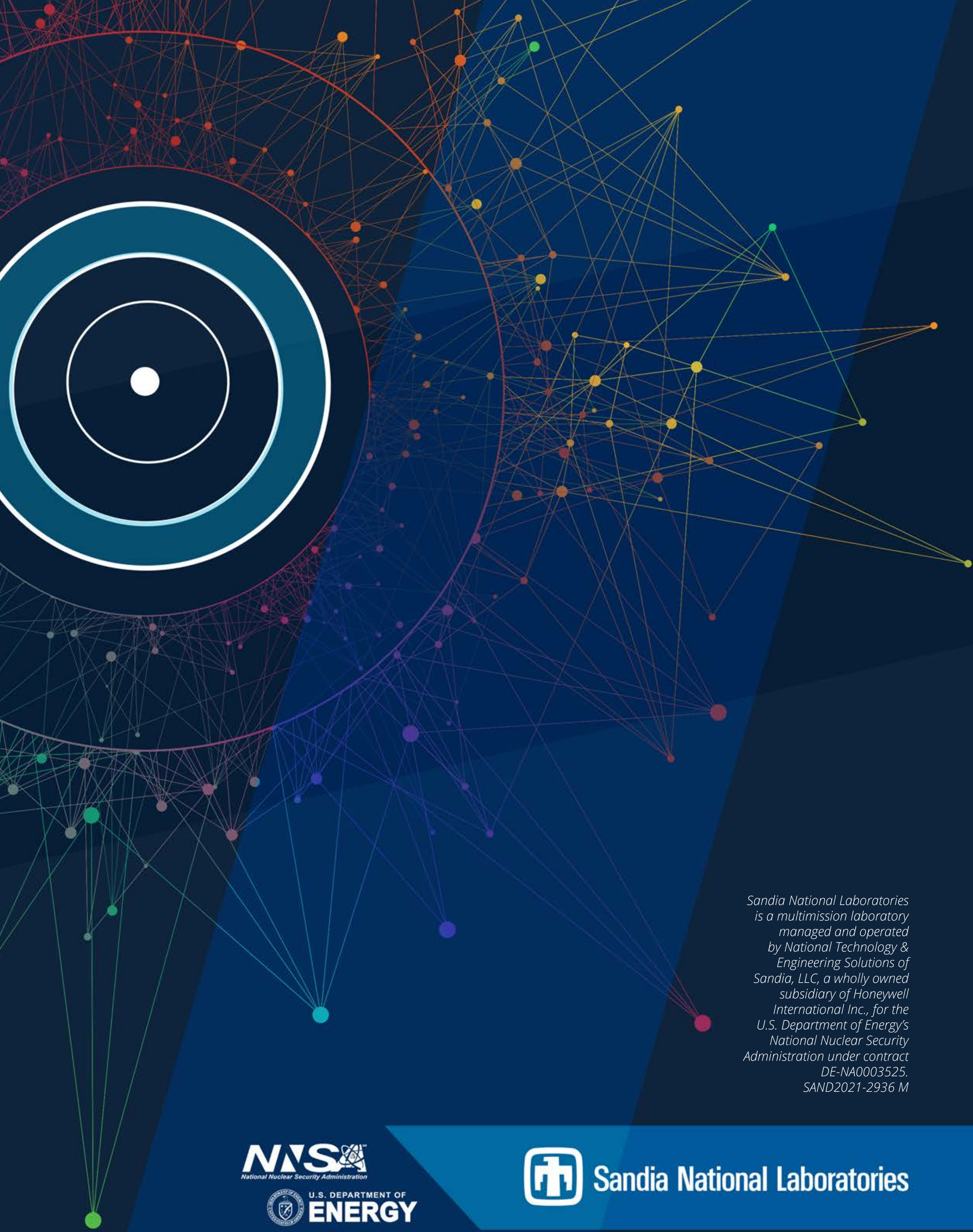
Video
Regina Valenzuela
Max Schwaber

SimMagic
Daniel Gomez

Proofreading
Patricia Crossno
Megan Davidson
Whitney Lacy

Production
Randy Bill

Management
Tom Klitsner
Scott Collis
Manoj Bhardwaj
Myles Copeland
David Littlewood
Kendall Pierson



Sandia National Laboratories
is a multimission laboratory
managed and operated
by National Technology &
Engineering Solutions of
Sandia, LLC, a wholly owned
subsidiary of Honeywell
International Inc., for the
U.S. Department of Energy's
National Nuclear Security
Administration under contract
DE-NA0003525.
SAND2021-2936 M

NNSA
National Nuclear Security Administration

U.S. DEPARTMENT OF
ENERGY

Sandia National Laboratories