HIGH PERFORMANCE COMPUTING

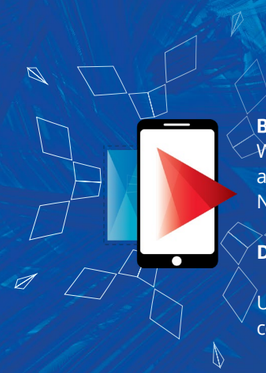
2024 ANNUAL REPORT

Sandia National Laboratories

TACKLING THE NATION'S TOUGHEST CHALLENGES

HIGH PERFORMANCE COMPUTING IN 2024

Sandia National Laboratories

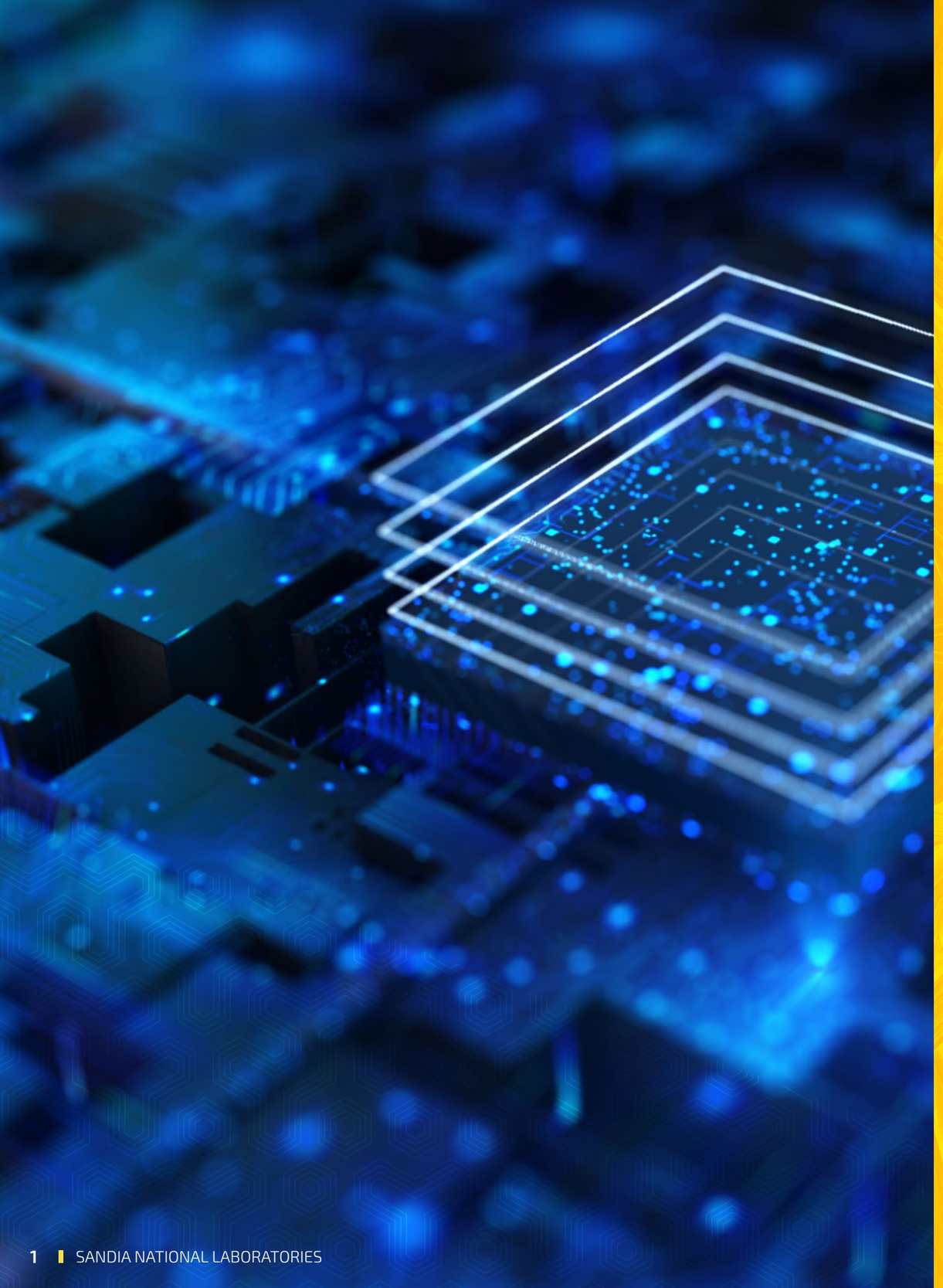


BUT WAIT, THERE'S MORE
Watch innovation come to life using **SNLSimMagic**, an augmented reality application developed at Sandia National Laboratories.

Download SNLSimMagic on your iOS device.

Use your iPhone to scan figures with this icon and watch content come to life.

Table of Contents



Director's message

I am pleased to present the 2024 HPC Annual Report, a culmination of our collective efforts in pushing the boundaries of high-performance computing (HPC) at Sandia National Laboratories. This report is a kaleidoscope that showcases the diverse and interconnected facets of our work, all converging on the common goal of advancing national security.

You will find an array of topics, each representing a piece of the puzzle that drives our mission. From cutting-edge simulations to data analytics and machine learning, our HPC work spans a wide spectrum of disciplines. With expertise and innovation we tackle the complex challenges that lie at the intersection of science, technology and national security.

As you explore the report, you will witness the synergy that happens when brilliant minds collaborate. Our researchers come together to explore new frontiers, using the power of HPC to unlock insights, propel discoveries and develop solutions that safeguard our nation.

This report highlights our achievements and underscores the importance of HPC in shaping the future. It shows the transformative impact of computing in such areas as signal mapping, energy systems, materials science and more. Each topic is a reflection in the kaleidoscope, revealing the intricate patterns that emerge when we combine our expertise, diversity and dedication.

I invite you to explore the 2024 HPC Report, to immerse yourself in the broad range of ideas and be inspired by the remarkable work being done within our Laboratories. We will continue to push the boundaries of what is possible as we ensure the safety and security of our great nation.

DR. JAMES PEERY
Laboratories Director

FIGURE 1

AWAKEN measurements will be used to validate wind plant models, which in turn will provide a better understanding of how wakes influence power production, wake steering strategies and optimal plant layout. (Drone footage in the video, above, by Josh Bauer and Bryan Bechtold, NREL 84025; photo, left, by Bryan Bechtold, NREL 84175)

AWAKEN the wind

One of the more challenging difficulties currently faced by the wind industry is understanding the effects of wakes from individual wind turbines and wind plants. The U.S. Department of Energy Wind Energy Technologies Office (WETO) has been investing in exascale codes to help improve understanding of the physics involved in wind turbines and wind plants. To help increase reliability and reduce uncertainty and risk, these models need to be validated and improved.

The American WAKE experimeNt, or AWAKEN, is a project funded by WETO and developed to address these challenges. A field campaign, involving partnerships with several national laboratories, industry and academia, covers five wind plants in north-central Oklahoma and has been collecting data from meteorological stations at 13 field sites, including scanning lidars, dual-Doppler radars, and three wind turbines fitted with over 50 unique instruments. Nacelle mounted lidars were also placed on five turbines. The measurements will be used to validate wind plant models, which in turn will provide a better understanding of phenomena like how wakes influence power production, wake steering strategies and optimal plant layout. The video in Figure 1 shows an overview of the project.

AUTHORS/TEAM

Tommy Herges, Myra Blaylock, Ken Brown, Lawrence Cheung, Nate deVelder, Dan Houck, Alan Hsieh, David Maniaci, Phil Sakievich, Gopal Yalla

CONTRIBUTING WRITER

Laura Sowko

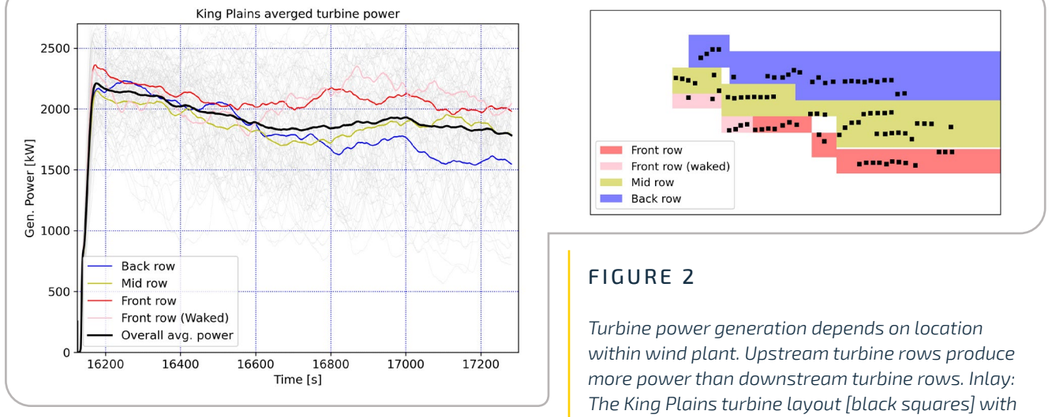


FIGURE 2

Turbine power generation depends on location within wind plant. Upstream turbine rows produce more power than downstream turbine rows. Inlay: The King Plains turbine layout [black squares] with colors indicating which row the turbine belongs to.

Computational fluid dynamics calculations are playing two important roles in this project. First, simulations were run before the instruments were deployed to help with instrument layout and scan strategies. And now that data has been collected for over a year, instances are being flagged for benchmark challenges, where multiple models will be compared in a blind study. These benchmark cases will evaluate models for a number of physical phenomena, such as plant wakes, wake controls and blockage effects. The effects of wakes from other turbines and placement of a turbine within the plant on power outputs can be seen in Figure 2.

Sandia has been relying on HPC resources to participate in these tasks and has run large eddy simulations (LES) on large sections of two of the wind plants in the study. The team has been using two codes from the Exa-Wind Suite, Nalu-Wind and AMR-Wind, which both have developers from Sandia. These codes solve the flow of the atmospheric boundary layer using incompressible Navier-Stokes equations with buoyancy and Coriolis forcing terms. Wind turbines are represented by coupled OpenFAST actuator disk models based on the actual turbines in the field. Figure 3 shows an example of the wake from a single turbine. A typical simulation of the wind plant will have a domain size representing 10x10x1 km using around 250 million elements and 30-40 turbines and would be run on over 1,500 processors.

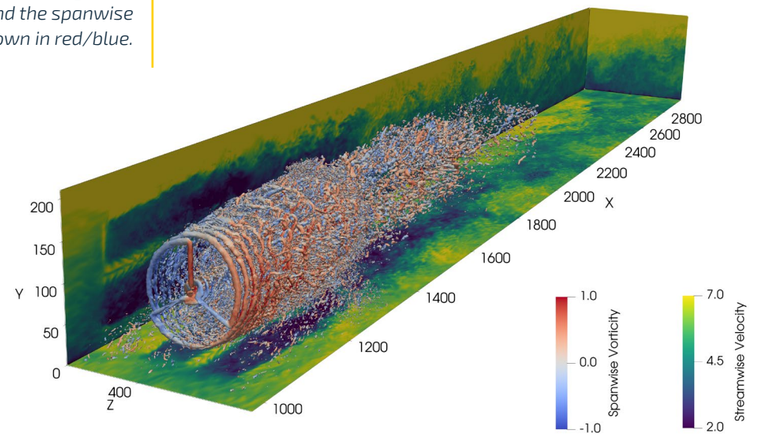


FIGURE 3

Wake visualization for a single turbine modeled with an actuator line. Streamwise velocity is shown in the blue/green/yellow planes and the spanwise vorticity is shown in red/blue.

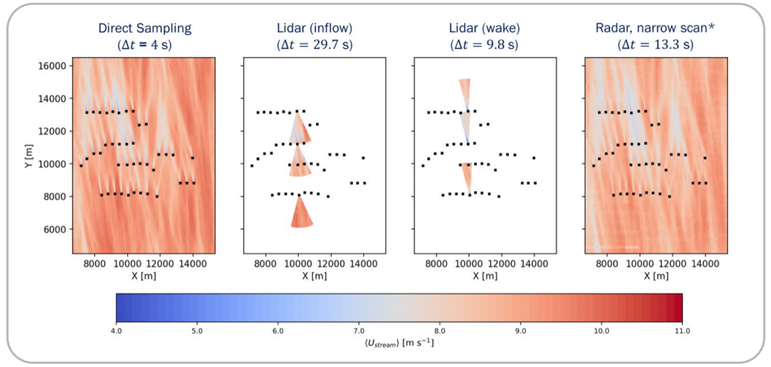


FIGURE 4

Comparison of a horizontal sample plane to forward and backward facing lidar and the synthetic dual radar image. Black dots indicate the wind turbine locations.

A unique development for the simulations done at Sandia is the implementation of "synthetic instrumentation." Instead of looking at simulation data along planes or lines, the sampling patterns of the actual instrumentation data collection are used to save velocity data, allowing for a more one-to-one comparison of simulation data to instrument data. For example, the lidars use line-of-sight Doppler data from lasers scattered off particles in the air. The pointing of the laser has a particular pattern that will map out the velocities in a specified area. The synthetic lidar records the velocities in the simulation using the same locations and timing patterns as the lidar in the field. The team is using this method for lidars, dual X-band radars, tethered balloons and meteorological stations. A comparison of a plane of data with the synthetic lidar and radar images is seen in Figure 4.

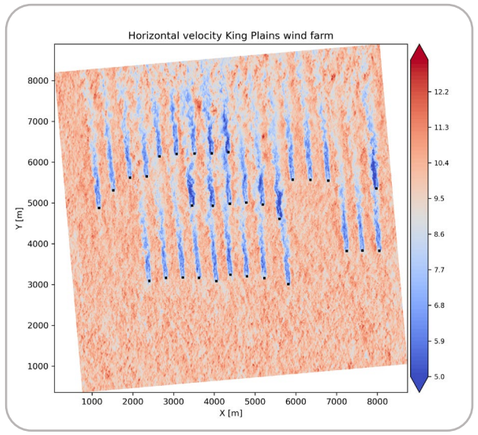
Because the benchmark cases are simulations of actual atmospheric conditions seen for a period at the site in Oklahoma, significant care must go into matching the details of the flow in the atmospheric boundary layer as well as making sure the turbine models are calibrated correctly. A "precursor" simulation is run without any turbine models to capture the wind velocity, turbulence intensity, shear strength and the temperature profile. Once the flow has been established and run for a long enough period of time to ensure the correct statistics of the flowfield, the simulation is run again with the same inflow conditions, but this time with the wind turbines included as either actuator line or disk models by coupling the boundary layer flow with the OpenFAST turbine models placed in the same locations as those in the actual wind plant.

FIGURE 5

Hub-height plane view of simulation of King Plains wind plant during stable atmospheric conditions.

Due to the large size of the domain and the fine grids needed for the LES calculations, it is impractical to save the full data set at every time step of the calculation. To ensure that enough data is captured to answer the scientific questions, there is substantial planning for what sub-sets of the data are important enough to be stored. In addition to the synthetic instrumentation mentioned before, more traditional planes are also captured. These include planes that capture the inflow and outflow to the plant, the sides and top to capture momentum transport and planes directly in front and in the wakes of individual turbines. There is also a plane at the turbine hub-height across the entire plant to capture general trends in the wakes (see Figure 5).

With this project, the team can determine how well wind plant models work and what needs to be done to improve them. This will help find ways to plan a more effective wind plant layout, which will help provide more efficient, reliable renewable energy for everyone.



(Photo by Bryan Bechtold, NREL 84186)

Following the sun over hill and over dale

Sun tracking algorithms have made utility-scale solar installations more profitable. What happens when solar is installed on uneven terrain?

Over the last 10 years, technology allowing rows of panels to rotate in place has become ubiquitous at large solar installations. These rotating installations, known as single-axis tracking (SAT) systems, are designed to allow panels mounted on a torque tube (the "single axis") to track the position of the sun over the course of the day. By allowing the panels to directly face the sun for the whole day – known in the industry as minimizing the angle of incidence – SAT systems boast a higher power output per row than a traditional "fixed tilt" system.

AUTHORS/TEAM

Emma Cooper, Kevin Anderson, Dan Riley, Madhuri Kumari

CONTRIBUTING WRITERS

Rory McClannahan, Monica Bigney

Single-Axis Tracking on a Horizontal Axis

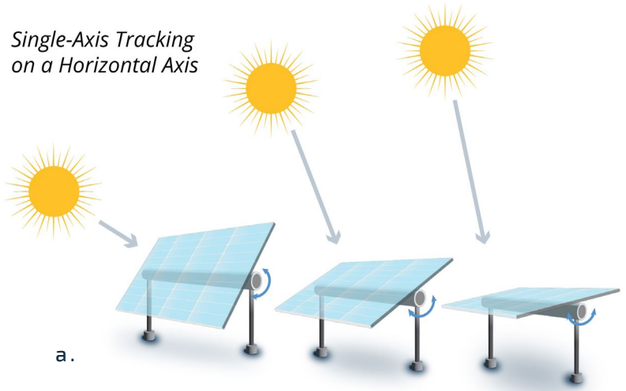


FIGURE 1

(a) The most popular design for single-axis tracking systems consists of a torque tube oriented in the north-south direction. Graphic adapted from Sinovoltaics, a solar developer.

(b) A single-axis tracking system deployed in the field. Image from Valsa, a solar supplier.

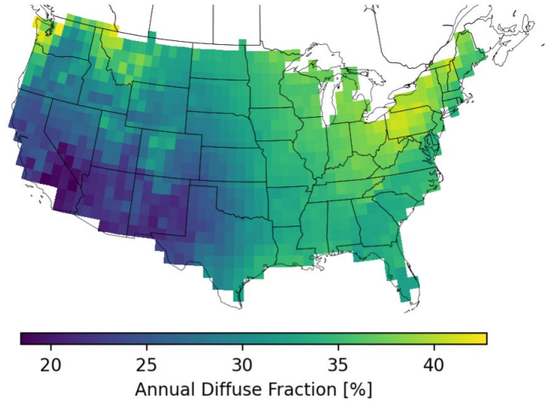


FIGURE 2

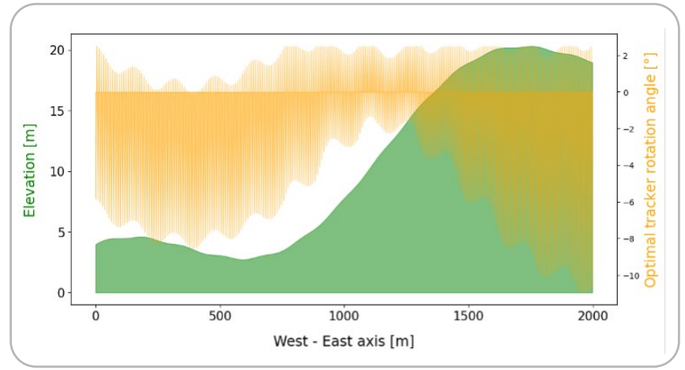
Diffuse fraction is one of the metrics used in the solar community to quantify the proportion of days that are overcast. A higher diffuse fraction indicates that the region experiences overcast weather more frequently. This graphic was created using HPDA resources and data from the NSRDB.

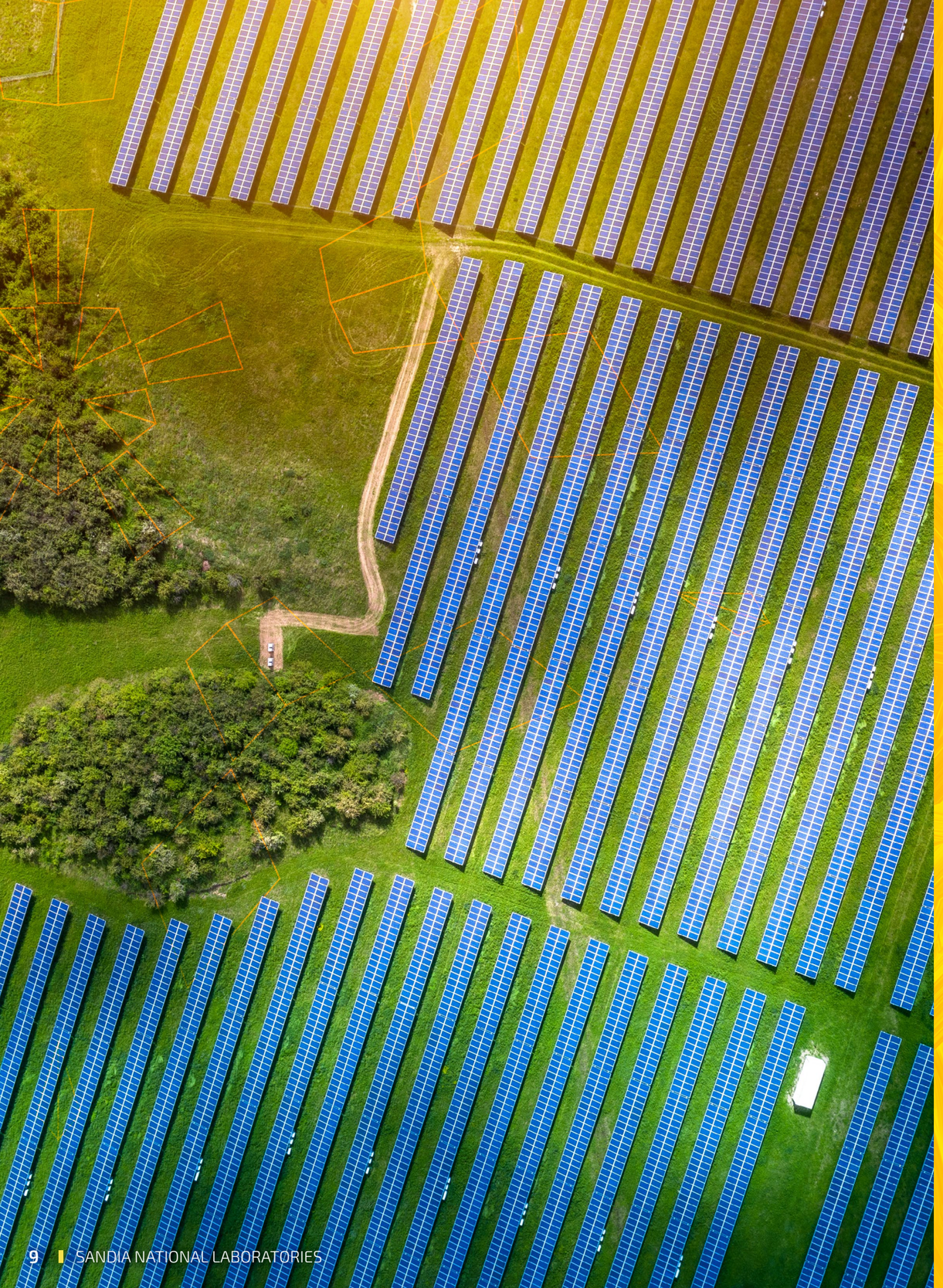
Local weather patterns have an impact on how much more power per row is realized. In Albuquerque, where conditions are usually clear, SAT systems will show large gains over fixed-tilt systems. But under the "socked in" cloudy conditions that are frequent in the northwest, solar panels can often collect more irradiance overall by lying flat than by tracking the position of the sun.

In this project, the team quantified how weather influences SAT system output across the United States by using local weather data provided by the National Solar Radiation Database (NSRDB) to simulate the power produced by four different sun-tracking algorithms in over 800 locations across the country. To parallelize these simulations, the team leveraged Kahuna, a high-performance data analytics (HPDA) research cluster hosted by Sandia California. Kahuna's implementation of a queue management system allowed the team to run batch simulations for hundreds of locations and monitor the outputs. By allocating calculations to multiple different nodes on Kahuna, the team was able to complete the calculations within a couple of days—as opposed to the several months it likely would have taken on a regular computer.

FIGURE 3

Tracker rotation angles calculated using HPDA resources with a global optimization algorithm. The example tracking system used in this study comprised 445 rows running north-south on synthetic terrain, which varies in the east-west direction. Here, the optimal rotations identified by the algorithm alternate between rows facing towards the rising sun (negative rotation angles) and rows facing away from the sun to avoid shading their neighbors (positive rotation angles).





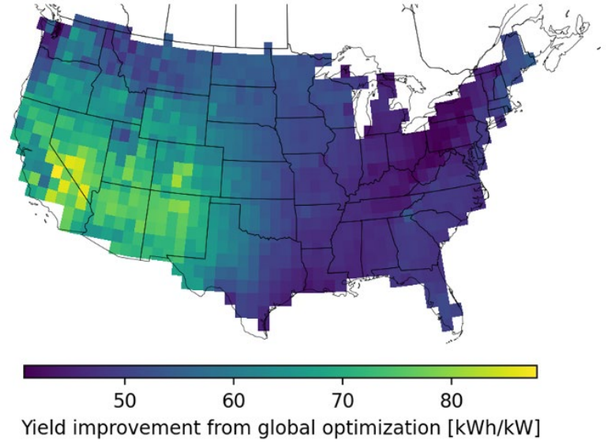


FIGURE 4

Annual yield improvement was strongly correlated with annual diffuse fraction. In other words, using the better sun-tracking algorithm led to greater increases in production in sunnier areas than in cloudy regions. However, even the cloudy regions saw an annual energy increase of 2% relative to the baseline when using the global optimization algorithm.

The study focused specifically on four sun-tracking algorithms that adapt their controls based on the surrounding terrain. Some of the algorithms work by finding a single optimal angle of rotation for all the rows in a system, while more computing-intensive algorithms optimize rotations row-by-row. For these latter algorithms, Kahuna's computing speed, enabled by its data processing unit (DPU) chips, came in handy. For our 400+ row SAT system, the DPU chips that make up Kahuna's hardware allowed for quick calculations of tracker rotation angles for the entire system in mere minutes.

Kahuna enabled the team to compare energy outputs from the four different sun-tracking algorithms and to identify regions of the country that experienced the highest and lowest gains. This demonstrated that an algorithm using a global optimization approach results in the largest energy gains, with increases in annual yield between 40 to 90 kWh/kW relative to the baseline algorithm the results were compared against. The Southwest region experienced the highest annual yield from using the global optimization approach, while the Midwest saw the lowest gains from using the algorithm.

The research results make a promising case for the use of adaptive tracking algorithms across the United States, but the energy gains found in this study are highly dependent on choice of terrain. Utility-scale solar sites are almost always graded, and there's not a lot of data available on the potential cost-savings or additional construction costs of building a system on an ungraded surface. Mechanical trackers that can follow adaptive algorithms also are more expensive, and it's unclear whether the additional expenses can be negated by grading-related cost savings.

This study is a first step towards a comprehensive value proposition for the use of uneven terrain over traditional graded terrain, but more research on cost is needed to make a compelling case to solar developers.

Forging a path for molecular-level simulations of turbulence in high-speed flows

AUTHORS/TEAM
Michael Gallis, Ryan McMullen, Stan Moore, Tim Koehler

CONTRIBUTING WRITER
Rebecca Cox

Turbulence is both a fundamental research problem and one of the most difficult engineering challenges of high importance to the nuclear deterrence program because it affects the vibrational loading and heating of reentry vehicles (RVs). The primary source of vibrational loading and heating during reentry is the turbulent boundary layer created by gas flowing over the vehicle's surface. Thermal-protection-system (TPS) materials that are used to coat RVs are often fibrous and may ablate due to extreme heating rates, creating a rough and permeable surface. This surface has the potential to significantly alter the turbulent boundary layer and thus alter the heating rate and vibrational loading. This in turn could affect vehicle performance. Consequently, there is great interest in the ability to accurately predict high-speed turbulent flows over RVs. The ability to make such predictions, however, is limited by the extreme complexity posed by the multi-scale nature of turbulence.

Simulating turbulence at the molecular level

Numerical simulations play a vital role in quantitatively understanding and predicting the effects of high-speed turbulent flows because of the difficulty associated with performing experimental measurements. Turbulent flows are almost universally simulated using the continuum Navier-Stokes (NS) equations, and their ability to accurately describe all length and time scales larger than molecular scales is rarely questioned. Typical NS simulations, however, do not include highly non-equilibrium effects or simulate molecular-level processes and neglect their impact on high-speed turbulence.

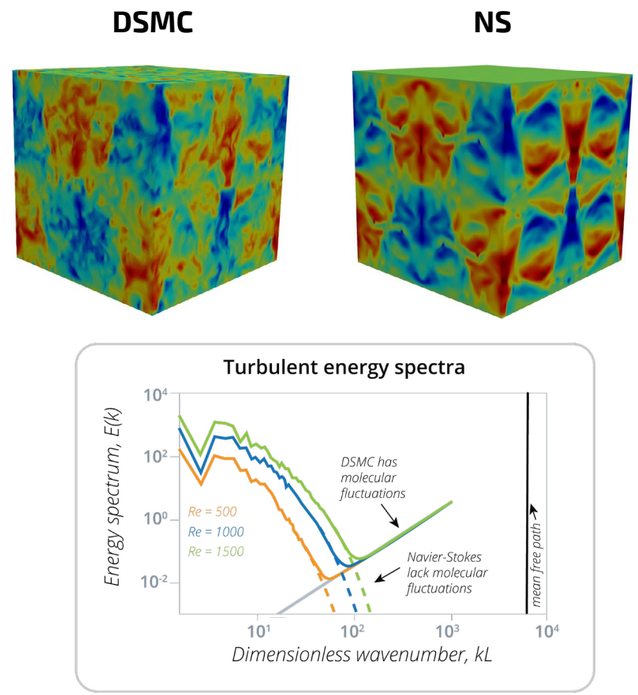
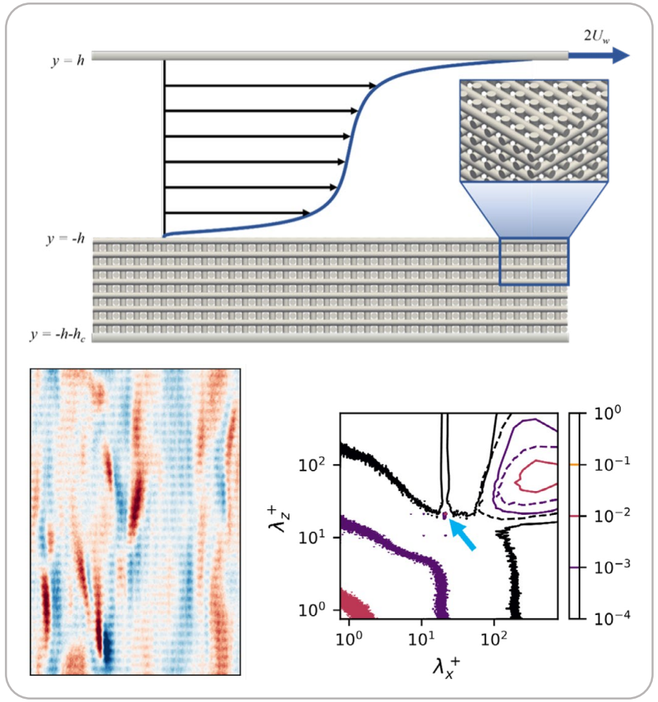


FIGURE 1

(Top) Comparison of velocity fields from DSMC and NS simulations of decaying turbulence. The large-scale features agree, but the small-scale features do not because of molecular fluctuations, which are not included in the NS equations. (Bottom) The distribution of kinetic energy across wavenumbers (inverse wavelength) in DSMC (solid curves) and NS (dashed curves) simulations of decaying turbulence. The gray line shows the spectrum for molecular fluctuations obtained from theory. DSMC simulations give the first direct evidence that the NS equations are not accurate in the dissipation range because they neglect molecular fluctuations.

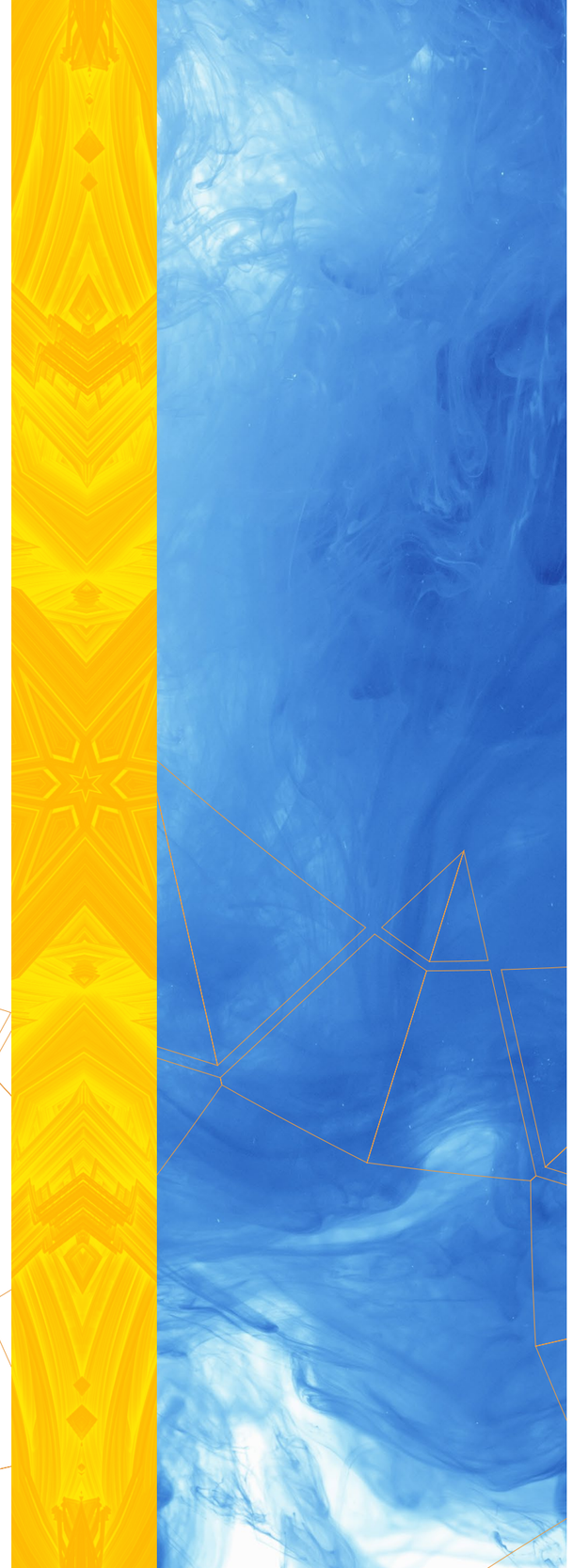
FIGURE 2

(Top) Schematic diagram of turbulent flow over a TPS-inspired surface comprised of an array of cylinders. The inset shows a close-up view of the cylinder structure. (Bottom) Visualization of velocity fluctuations (left) and their spectrum (right, solid lines) show the influence of the surface geometry on turbulent flow over the TPS-inspired surface. The wavelength corresponding to the pore size is indicated by the cyan arrow. The spectrum from a reference smooth-wall simulation (right, dashed lines) demonstrates how the permeable surface enhances the turbulent fluctuations. Continuum simulations of flow through the surface under-predict its effective permeability since they do not account for rarefaction effects.



In contrast, the direct simulation Monte Carlo (DSMC) method provides a high-fidelity representation of non-equilibrium gas behavior by directly simulating the motion of computational molecules. A computational "molecule" in DSMC may represent thousands of real molecules, leading to a more tractable simulation. DSMC has been recently shown to provide a computationally viable and physically superior way to investigate the relationship between molecular-scale processes and hydrodynamic-scale turbulence in a gas. Molecular methods, including DSMC, are computationally intense. However, petascale computational resources enable DSMC to accurately simulate non-equilibrium transitional and turbulent flows that contain length and time scales ranging from microscopic (molecular) to macroscopic (hydrodynamic).

For the last decade, Sandia has been developing SPARTA, a state-of-the-art massively parallel DSMC code. SPARTA is the world's premier open-source DSMC code, with hundreds of users worldwide. It is used extensively by academia, industry and space agencies. SPARTA has been developed to ensure portable performance and efficient use of Central Processing Unit (CPU) and Graphics Processing Unit (GPU) machines, enabling extreme-scale science on current and future supercomputer platforms. Turbulence simulations in this work typically utilized one third of National Nuclear Security Administration's (NNSA's) Advanced Technology System (ATS)-2 Sierra supercomputer at Lawrence Livermore National Laboratory, and several simulations were also performed using the entirety of the ATS-1 Trinity supercomputer at Los Alamos National Laboratory. Turbulence simulations currently run on ATS-3 Crossroads. SPARTA is also being ported to ATS-4 El Capitan and is being used as a benchmark for the future NNSA ATS-5 supercomputer because of its exceptional parallel performance and large usage on current ATS supercomputers.



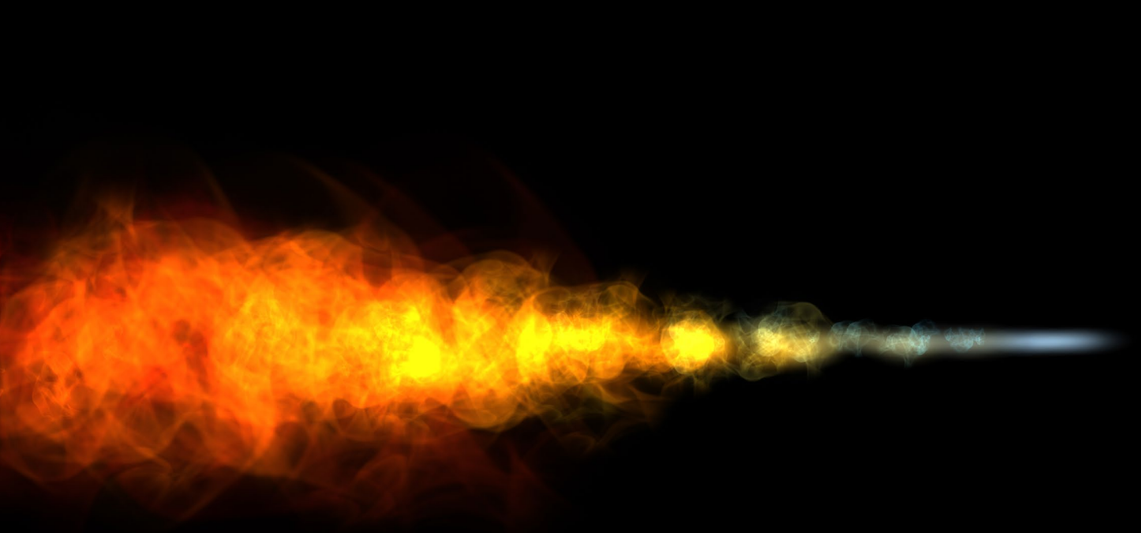
An important question that DSMC is well-suited to tackle is: What length scales in turbulent flows are affected by molecular-level phenomena? To address this question, decaying turbulence was simulated using SPARTA, and the results were compared with NS simulations (see Figure 1). The simulations revealed that the NS equations do not accurately represent scales in the so-called dissipation range in turbulent gas flows because they neglect molecular fluctuations. It was found that the DSMC turbulent kinetic energy spectra grow quadratically with wavenumber in the dissipation range due to molecular fluctuations (see Figure 1, bottom), which is in agreement with theoretical predictions, whereas the NS spectra decay exponentially. The length scale for which molecular fluctuations begin to dominate is notably much larger than the gas-molecular mean free path, specifically, in a regime that the NS equations are widely believed to describe.

Turbulent boundary layers over rough and permeable surfaces, like those found on RVs covered in TPS materials, are another example of turbulent flow where molecular-level phenomena play a significant role. The length scales associated with the roughness and permeability of typical TPS materials can notably be comparable to the gas-molecular mean free path, which itself may not be small compared to other characteristic flow length scales.

Recent DSMC simulations of turbulent flow over an idealized permeable fibrous substrate representative of TPS materials (see Figure 2, top) demonstrated that the near-wall turbulence is modulated by the permeable substrate with a wavelength equal to the pore spacing (see Figure 2, bottom). The flow within the substrate also shows significant rarefaction effects, resulting in an effective permeability that is significantly larger than the intrinsic permeability. This higher effective permeability means that gas can move more easily through the surface than the NS equations predict, which could have significant implications for thermal transport and surface chemistry in TPS materials.

Future work on high-speed turbulent flows

These investigations demonstrate the ability of DSMC, in particular of SPARTA, to simulate turbulent flows at the molecular level and highlight previously unappreciated limitations of the NS equations. This work opens a new path of research where molecular-level simulations can be used to quantitatively understand phenomena in high-speed turbulent flows that are currently inaccessible to more conventional methods. Because of SPARTA's excellent parallel performance, the advent of exascale supercomputers such as Oak Ridge National Laboratory Frontier, Argonne National Laboratory Aurora, and the future NNSA El Capitan will allow DSMC to push the state of the art in turbulence research even further.



Using decision boundaries to aid engineers in thermal safety applications

AUTHORS/TEAM

Jake Desmond, Marco Arienti, Maher Salloum, Ryan Keedy, Andrew Kurzawski, Timothy Walsh

CONTRIBUTING WRITER

Monica Bigney

Modern aerospace design blends a multidisciplinary set of engineering and physics tools with computational simulation to achieve the design requirements for a given vehicle. One such requirement is thermal safety, in which it is paramount to anticipate critical weakness thresholds for various machine parts to ensure they are manufactured to satisfy operational requirements. A crucial component of this design process is analyzing scenarios in which the design body becomes engulfed by a fire and intense amounts of heat are radiated throughout the system. To ensure the device does not malfunction, units called links are installed; these links typically come in pairs, one stronglink and one weaklink. Embedded within the engineering design is the assumption that the weaklink will reach its critical failure temperature before the stronglink (see Figure 1), thereby preventing the design from malfunctioning.

For a given thermal load scenario, the competing occurrences of the links' failure modes constitute what is known as a "thermal race." The success or failure of each potential thermal race traces out a decision boundary in the design parameter space. One can think of decision boundaries as dividing regions in space; on one side of the boundary, the designs are acceptable, whereas the opposite side results in a failed system. However, due to the chaotic nature of fires, it is impossible to simulate every possible combination of directions and heat intensities that may be encountered. These types of simulations are computationally expensive due to the high-dimensional nature introduced by material, geometric and environmental parameters. Any brute-force algorithm which attempts to find this decision boundary by discretizing the design space is therefore intractable and necessitates the use of adaptive learning algorithms to find these failure thresholds.

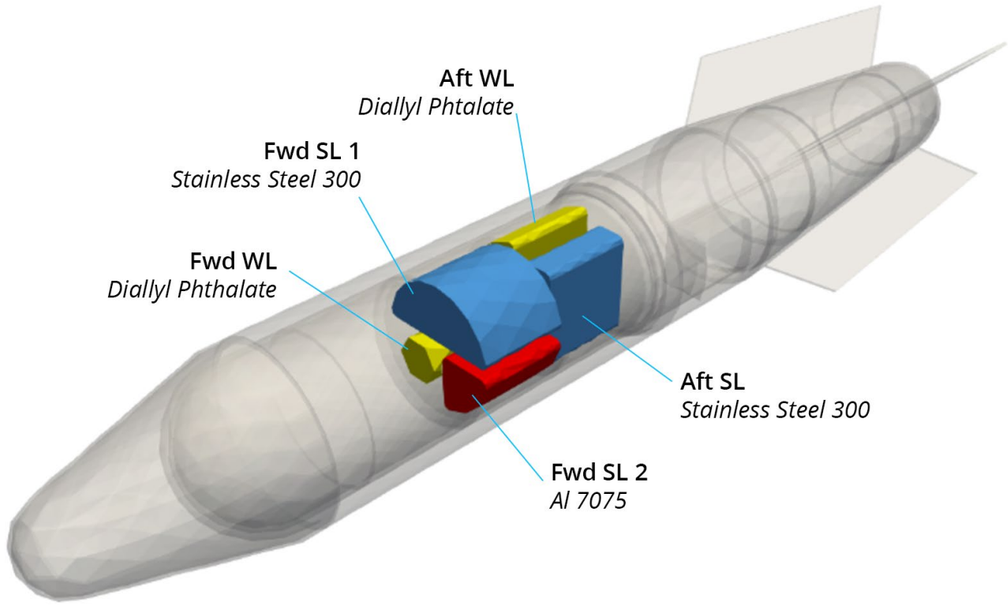


FIGURE 1

The locations of the weaklink and stronglink groups in a hypothetical design.

The Fusion of Simulation, Experiments and Data (FuSED) team has developed a decision boundary tool that complements the HPC environments by leveraging a machine-learning classification algorithm known as support vector machines (SVMs). This algorithm has two steps: (1) training a surrogate model using adaptive-learning SVMs that approximate the high-fidelity model of interest, and (2) using a post-processing step that calculates the probability that a certain design will exceed the designated threshold.

Training the surrogate is done by iteratively selecting points in the parameter space, which will yield information about the location of the decision boundary. Once a convergence criterion has been reached, the training step concludes and engineers can then query the surrogate as a proxy for the high-fidelity model and gain a computationally inexpensive estimate of whether parameter values will lead to an acceptable design. In general, there are two types of parameters that need to be considered: deterministic and uncertain. The deterministic parameters are concrete design options in the model that engineers have the capability of controlling. In contrast, the uncertain variables account for randomness in the system.

Sandia's novel approach trains the surrogate model over both deterministic and uncertain variables. The advantage is that once the surrogate is trained, a simple post-processing step can calculate the probability that a particular combination of design parameters will exceed the specified threshold. In this sense, the result is condensed down to the dimensionality of just the design space parameters while simultaneously adding parametric uncertainty into the model. One added difficulty of training surrogates over this combined space is that the dimensionality of the inputs increases. In general, many algorithms suffer from the curse of dimensionality but by leveraging SVMs, the team can ameliorate the issue and avoid the exponential growth in training data that is typically required. Details regarding this approach can be found in a recently published article in Structural and Multidisciplinary Optimization entitled "Assessing decision boundaries under uncertainty."

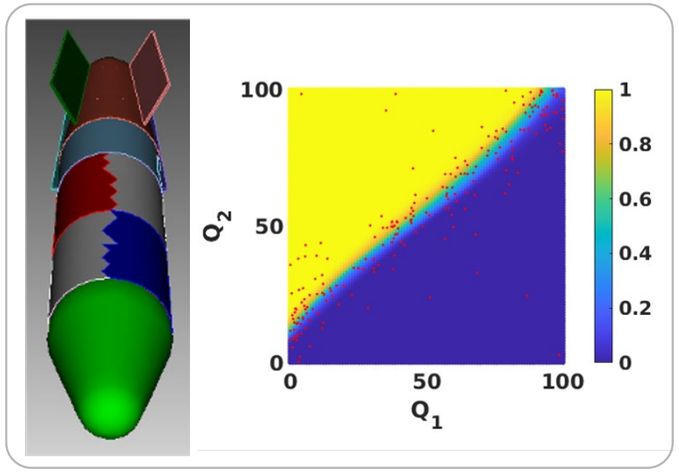


FIGURE 2

(Left) UBOMB geometry of fictitious reentry vehicle. The red and blue panels are prescribed a heat flux Q. (Right) the decision boundary obtained using SVM, where the red dots indicate the simulations required to obtain the interpolated solution. Colors indicate failure probability.

One core feature that was added to the algorithm is a concept referred to as "iteration concurrency." Due to the iterative nature of training the surrogate model, each step of training depends on the training points prior to that step. This typically precludes one from parallelizing the iterations since one must know the results of the current step to proceed forward. Iteration concurrency is a way to perform multiple training steps simultaneously by calculating hypothetical results. Since there are only a finite number of possible results from each step, the algorithm can suggest which points to train on, given every possible result. By calculating these simulations in parallel, two iterations of training can be executed concurrently. Once the simulations have finished, the model is trained on the actual results and the extraneous information is discarded. In this sense, the amount of time one must wait for the entire training phase to occur is cut in half. Depending on the available computational resources, the iteration concurrency can be performed on as many steps as desired, yielding further time reductions.

By using the geometry-based classification of the SVMs, researchers can provide the user with probability of failure estimates over the deterministic parameter space while minimizing the number of forward solves needed. In a brute-force approach known as the "lawnmower search," the determination of the decision boundary in an N-dimensional parameter space would require on the order of p^N forward evaluations, where p increases with the parameter resolution. This number is drastically reduced as the adaptive algorithm intelligently selects points to train on, based on the current estimate of the decision boundary in parameter space. Figure 2 illustrates an early result from this project where the team reduced the number of forward evaluations required by a factor of 10.4 during the search for the decision boundary, while preserving the accuracy of the brute-force approach. While this example only has two deterministic variables, this effect is expected to further amplify as dimensionality is increased. This allows an engineer to assess the reliability of a design and integrate this information into the design needs of a specific program by using high-fidelity simulations while simultaneously assessing parametric uncertainty in a timely manner.

Anemone taps into the power of HPC to advance science of sensors

AUTHORS/TEAM
David B. Karelitz, Todd A. Pitts, Aislinn J. Handley, Bridget K. Ford

CONTRIBUTING WRITERS
Rory McClannahan, Monica Bigney

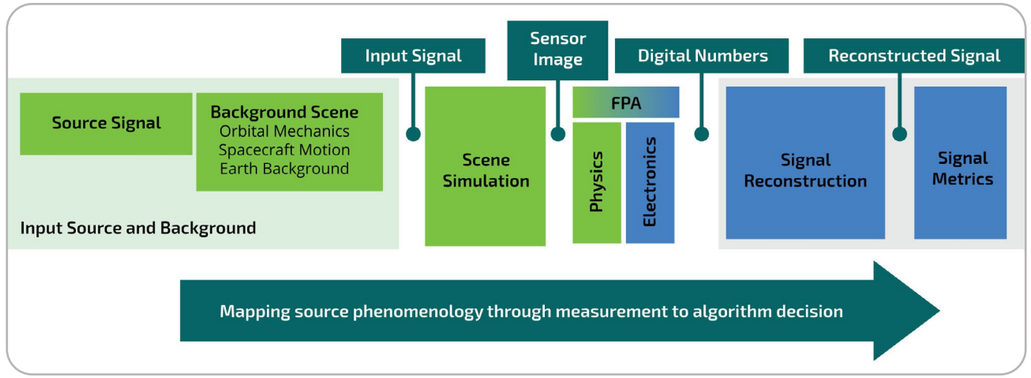
Modeling software tests numerous scenarios with an eye on assuring nuclear detonations are detected

Since the dawn of the nuclear age, it has been important to know where and when nuclear devices are used. This has led to the development and constant improvement of nuclear detonation detection sensors, work that continues today at Sandia using HPC resources to evaluate sensor performance across a wide variety of scenarios.

Sandia is developing the next-generation optical nuclear detonation detection sensor (SIGHTS). SIGHTS will be hosted on the Global Positioning System satellite constellation and will monitor Earth for nuclear detonations. HPC resources were used both to design the algorithms that operate onboard the sensor and to verify sensor performance as part of two critical program reviews: consent to ship (CTS) and system verification review (SVR).

Most advancements in science and technology come about due to the need to overcome challenges. One challenge for SIGHTS is that laboratory testing of mission performance is limited by the variety and fidelity of optical input signatures and backgrounds that can be generated in a lab setting, as well as the time available to test a range of representative scenarios. HPC simulations can generate much higher fidelity representations of both the input signal and Earth's background. In addition, the significant computing resources available on the HPC clusters allow a much wider variety of scene conditions to be tested. Mission performance predictions are complicated by SIGHT's extensive and complex onboard processing, which results in a highly non-linear response. By design, non-linear systems show large changes in output response for a small change in the input stimulus. Examples of such non-linearity include only sending data in response to an event of interest or when a signal exceeds a threshold; this makes it difficult to infer the sensor's response in cases that aren't explicitly tested or simulated.

An accurate sensor simulation is essential for overcoming these challenges and predicting sensor performance over a complex mission space. This need was identified early in the SIGHTS program and has led the SIGHTS Modeling and Algorithms (M&A) team to develop a high-fidelity system simulation utilizing HPC resources.



The SIGHTS M&A team includes subject matter experts in the areas of optical system modeling and design, signal and image processing, mathematics, electrical engineering, systems engineering and computer science. Utilizing models developed and implemented by these subject matter experts in combination with software developers, the multi-disciplinary team designed the sensor simulation in tandem with the onboard processing.

The result was Anemone, a multi-physics simulation code that models the sensor's view of the Earth and its response to that scene. For the scene simulation, Anemone includes models of Earth's background, incorporating its natural aspects such as the distribution of materials over the planet's surface, including clouds and the complex reflection of the sun off water. This is done while keeping other factors in mind, such as host vehicle position and attitude in relation to the Earth and sun, as well as the spectral and temporal optical characteristics of the signature of interest.

Anemone also includes models of the optical system, focal plane assembly and onboard digital signal processing, enabling the accurate prediction of sensor response to the expected on-orbit viewing conditions and signals. The development of Anemone utilized sustainable software development practices, including a robust automated testing suite, continuous integration and continuous deployment (CI/CD) and static and dynamic analysis of source code.

An important and unique feature of the automated testing suite for this code is that it includes a set of tests designed to ensure correlation between the simulation and actual hardware. These practices allowed for rapid development, including direct contributions from subject matter experts and rapid deployment to Sandia's production HPC environments. The ability to rapidly develop and deploy the simulation code to HPC resources enabled fast, iterative development of the sensor, its algorithms and their operating parameters.

As part of the CTS and SVR reviews, tens of thousands of Anemone simulations utilizing more than 1.1 million node hours across the Stout and Pecos HPC clusters were conducted and evaluated to assess sensor performance against a wide variety of scenarios. The volume of data generated also required HPC-based data post-processing and aggregation to evaluate, summarize and report sensor performance.

The completion of the CTS and SVR reviews would not have been possible without the availability and usability of Sandia's HPC resources and the support provided by the HPC Operations team.

FIGURE 1

Anemone simulation structure with the sensor's view of the scene on the left side (green boxes) through the sensor response on the right side (blue boxes). The green boxes represent the physical world and hence, generate a tremendous amount of data. This is true of both the physical sensor and its numerical simulation. Algorithms at the green/blue boundary perform a significant amount of data reduction. Further data reduction and decision making happens in the on-board electronics, represented by the blue boxes, to effectively utilize the communication bandwidth of the satellite link to the ground.

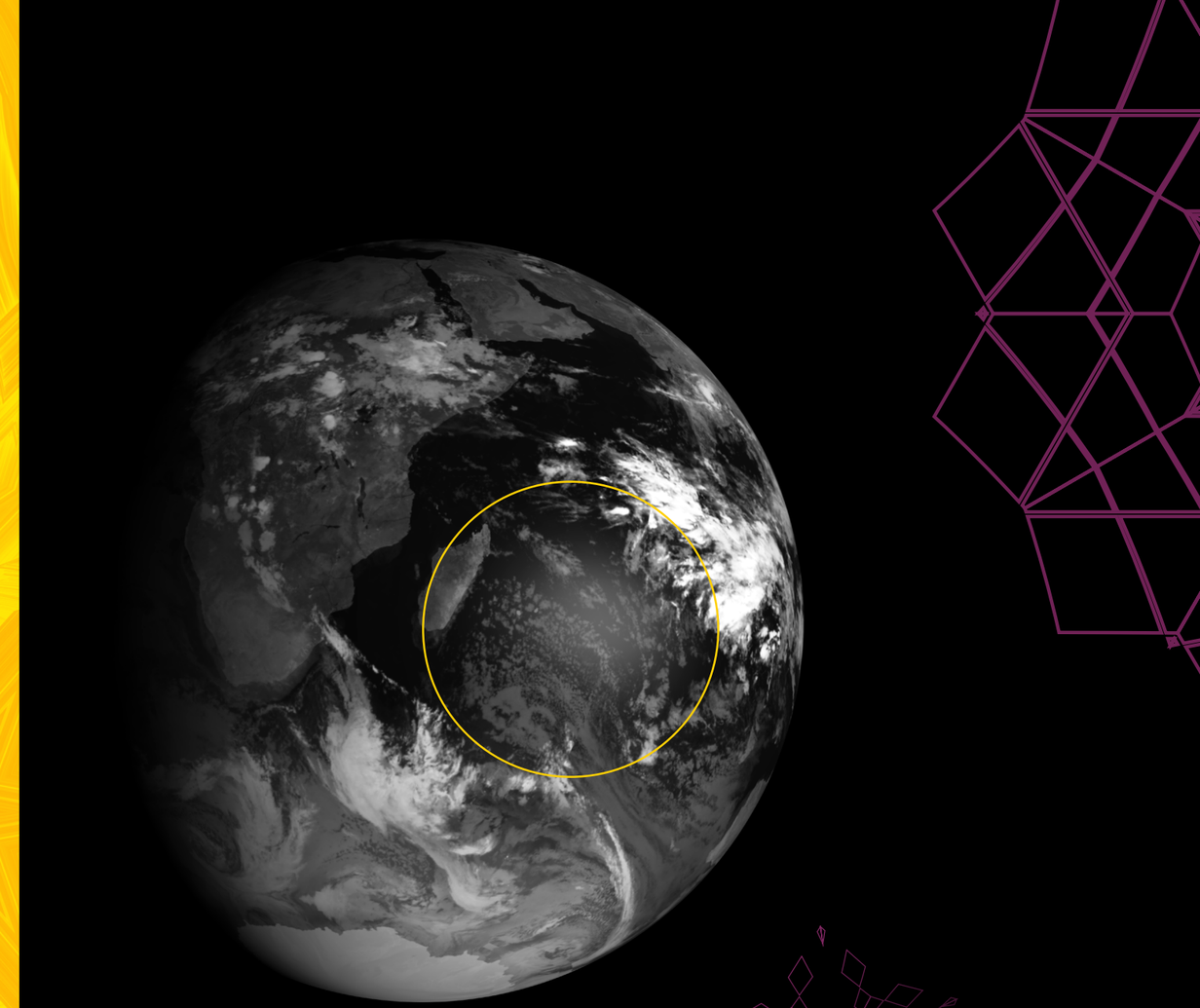


FIGURE 2

Anemone Earth background simulation, including models of the distribution of materials over the surface of the Earth, the complex reflection of the sun off water, clouds, and a model of the SIGHTS optical system. In addition, the simulation includes models of the vehicle orbit and attitude. Note the bright spot about halfway up and halfway from the center to the right edge, showing the reflection of the sun off the ocean.

MALA: accelerating scaled quantum chemical accuracy

AUTHORS/TEAM
Dayton J. Vogel, Siva Rajamanickam, J. Adam Stephens, Laura P. Swiler, Normand A. Modine, Aidan P. Thompson

CORE LDRD PROJECT TEAM MEMBERS/ COLLABORATORS OUTSIDE OF SANDIA
Lenz Fiedler, Atilla Cangi, Steve Schmerler

CONTRIBUTING WRITER
Alex Longo

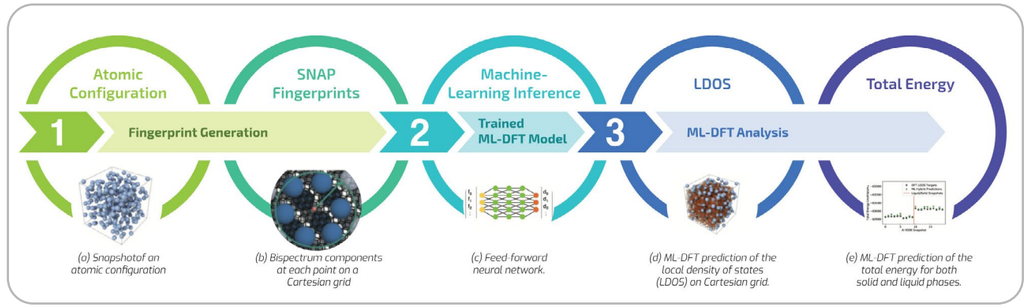


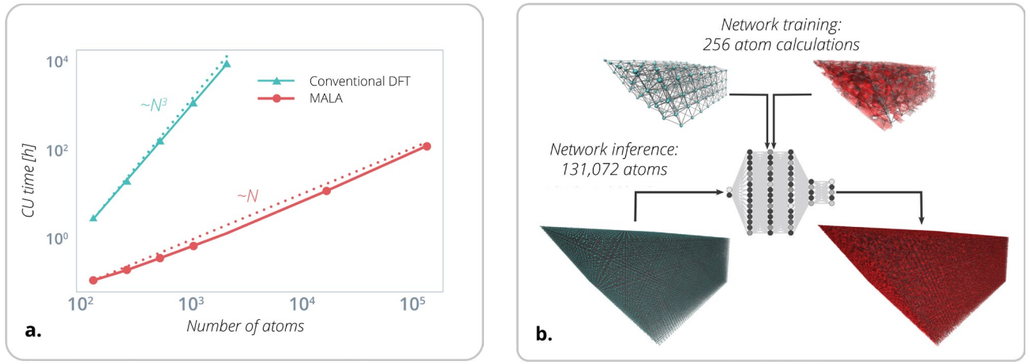
FIGURE 1

Sample workflow of MALA showing how physics-informed bispectrum components are generated from an atomic configuration (step 1) which is used as input to the ML model (step 2). The output of the ML model is used to compute quantities of interest (step 3).[5]

Discovering new materials and understanding the properties of existing materials are critical when addressing national security needs. To do this, Sandia studies molecular and material properties at the electronic structure level. Such calculations are expensive due to system size (number of atoms or length scale), simulation time (timescale) and accuracy needed. Materials Learning Algorithms, known as MALA,[1] is a newly developed software framework that uses machine learning (ML) to accurately predict the electronic structure of materials for previously unattainable lengths and timescales (Figure 1). MALA offers unprecedented accuracy and scale, surpassing any existing method and computing resource. In 2023, MALA was globally recognized as an R&D100 award winner for significant technological advancements in software.[2,3]

Scalable electronic structure calculations are critical for a materials modeling approach that can bridge several time and length scales. However, predictive atomistic materials modeling is hampered by the costly and redundant generation of first-principles data. Large-scale molecular dynamics (MD) simulations (LAMMPS) of material behavior require accurate interatomic potentials (IAPs), such as the spectral neighbor analysis potential (SNAP), that must be fit to higher fidelity datasets. These datasets are commonly generated using density functional theory (DFT), which is complex, expensive, limited to small scales (nanometers and femtoseconds) and exhibits N^3 scaling in a system size where N is the number of atoms (see Figure 2a). The amount of data needed to construct an IAP increases exponentially with the number of chemical elements, thermodynamic states, phases and interfaces. Despite its limitations, DFT remains the most heavily used approach for electronic structure calculations and has increased in popularity since its creator won a Nobel prize in 1998.

MALA addresses the issue of scaling by using a combination of ML and novel physics-based methods to enable research that was previously unfeasible. By predicting local quantities of interest with ML and computing global quantities of interest with physics-based approaches, MALA balances the trade-off between accuracy and speed. Unlike conventional methods, large MALA calculations can be distributed over multiple processors in a straightforward way, so that the time to solution is independent of the system (number of atoms) and grid sizes. One can learn the physics of the systems with just a few atoms and compute the electronic structure for much larger systems without the loss of chemical accuracy (see Figure 2b). Moreover, this approach is unique in providing both electronic structure information and total system energy. These properties uniquely position MALA to study and predict the behavior of material defects, among others, which provides fundamental physics information on the reliability of materials.



How MALA works

MALA aims to address a few key questions, namely, "Can machine learning help accelerate these first-principle electronic structure calculations?" and, "Can this be more scalable with respect to system size?" The methods developed in the MALA framework answer both questions in the affirmative. MALA employs both ML – specifically, deep neural networks (DNNs) – and physics-based approaches for predicting the electronic structure of materials. ML has become a fundamental technology for problems such as image classification, language modeling, caption generation and text or voice recognition. However, its application in the scientific domain is still in relatively early stages. While there are examples of physics-informed neural networks solving partial differential equations, applying ML to first principle calculations poses significant challenges as these calculations are the "gold standard" on which a multi-scale modeling framework is built. Surrogate models for such calculations have to be extremely accurate in order to be useful.

Instead of relying on a fully data-driven approach to understand every aspect of physics, MALA uses a hybrid approach alongside the DNN to predict local quantities of interest with very high accuracy, and a physics framework around it to compute the global quantities of interest. MALA's grid-based approach, in which the DNN model predicts local quantities at each grid point, achieves very high accuracy while being invariant of the system size. The physics-based framework should also be able to predict system properties from the predictions of the DNN models in a scalable way. MALA's approach to go from local density of states to quantities of interest, such as total energies, meets this requirement. These properties result in a technology that has the potential to transform multi-scale modeling methods over the next decade by surpassing the scale of traditional DFT calculations while remaining more accurate than other scalable methods. Figure 3 shows an example where stacking faults in a 131,072 atom Beryllium simulation cell were accurately predicted after training on models with 256 atoms.

MALA's framework is a natural candidate for HPC use cases, and the DFT-based training is the classic use case for training resources at HPC facilities. Furthermore, as the system size increases, the grid size increases as well. While this results in more work (still linear in system size), the grid points can be parallelized by decomposing the grid on a supercomputer and doing independent inference at different grid points. This allows MALA's framework to use HPC resources, especially graphics processing units (GPUs), to accelerate electronic structure calculations effectively.

FIGURE 2

(a) Direct comparison of computational time in CPU hours needed by MALA and conventional DFT as the number of atoms is increased. DFT is unfeasible beyond 2000 atoms. MALA scales to 100,000 atoms at a fraction of the cost. (b) MALA scalability for an ML model trained on DFT data for 256 atoms performs inference on a 131,072 atom system.[4]

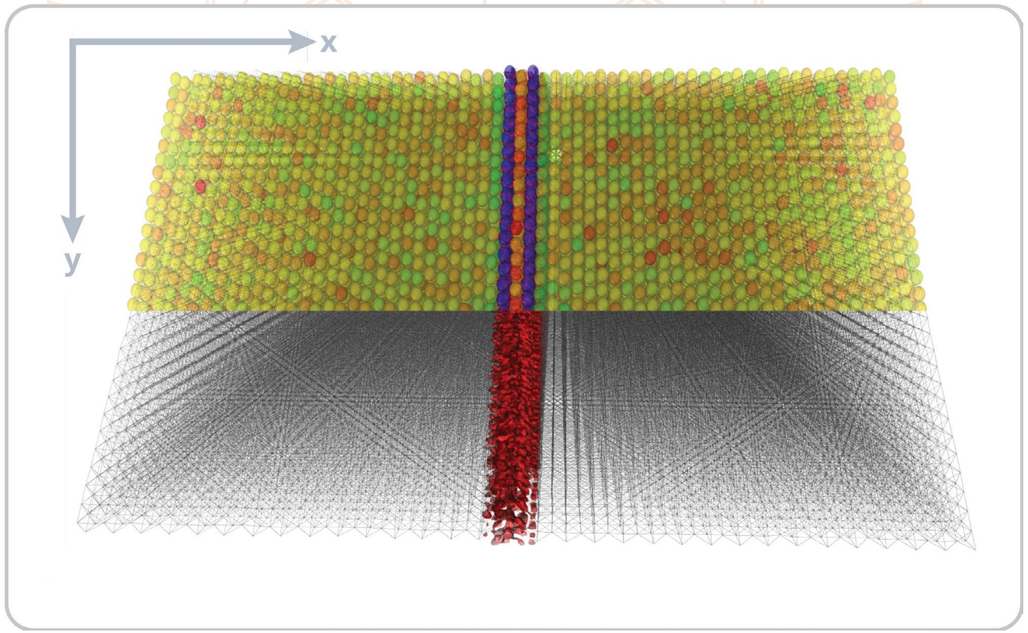


FIGURE 3

Illustrating size transferability of the MALA ML model. Beryllium simulation cell of 131,072 atoms with a stacking fault, generated by shifting three layers along the y-axis, creating a local face-centered cubic (FCC) geometry as opposed to the hexagonal close-packed (hcp) crystal structure of Beryllium. The colors in the upper half correspond to the centrosymmetric parameter calculated by the OVITO visualization tool, where blue corresponds to FCC and red-to-light-green represents hcp local geometries. The lower half of the image, generated with VMD software, shows the difference in electronic density for 131,072 Beryllium atoms with and without a stacking fault.

References

1. Ellis, J. A., Fiedler, L., Popoola, G. A., Modine, N. A., Stephens, J. A., Thompson, A. P., Cangi, A., & Rajamanickam, S. (2021). Accelerating finite-temperature kohn-sham density functional theory with deep neural networks. Physical Review B, 104(3). https://doi.org/10.1103/physrevb.104.035120
2. Fiedler, L., Modine, N. A., Schmerler, S., Vogel, D. J., Popoola, G. A., Thompson, A. P., Rajamanickam, S., & Cangi, A. (2023). Predicting electronic structures at any length scale with machine learning. Npj Computational Materials, 9(1). https://doi.org/10.1038/s41524-023-01070-z
3. Mala-Project. (n.d.). Mala-project/mala: Materials learning algorithms. A framework for machine learning materials properties from first principles data. GitHub. https://github.com/mala-project/mala
4. Mala. Research & Development World. (2023, August 12). https://www.rdworldonline.com/rd-100-2023-winner/mala/
5. Sandia National Laboratories. (2023, August 24). R&D 100 winner 2023: Mala: Materials Learning algorithms. YouTube. https://www.youtube.com/watch?v=ylhy5L-4Eg0. SAND2023-04048V

Design and certification of a plutonium air transportation package

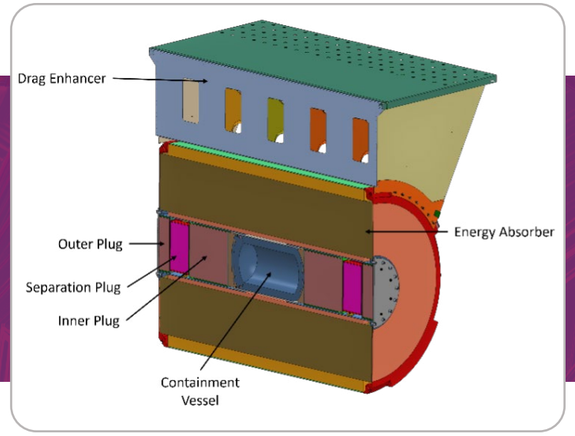
AUTHORS/TEAM

John Bignell, Doug Ammerman, Mike Starr, Sal Rodriguez, Gregg Flores, Lindsay Gilkey, Bob Kalan

CONTRIBUTING WRITER

Alex Longo

Sandia researchers have developed a large plutonium air transportation (LPAT) package that utilizes HPC resources and test facilities. This package is capable of transporting hazardous plutonium materials over-the-road or by air and was designed to satisfy the plutonium air transport safety criteria specified by U.S. regulations. These regulations stipulate that a hazardous materials transportation package must maintain containment of its hazardous contents, provide sufficient shielding against radiation leakage from the package and prevent the package contents from going critical during an accident. The regulations define a series of hypothetical accident condition (HAC) tests for which the package must be designed. These tests are intended to bound realistic accident scenarios.



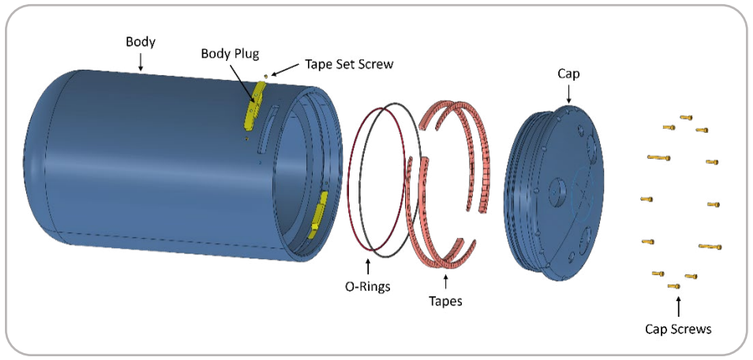


FIGURE 1

Large Plutonium Air Transport (LPAT) package and components.

Of the eight defined tests, the impact and fire tests are by far the most challenging to withstand. Six tests are to be applied sequentially:

1. Impact: the impact of the package at a velocity of at least 129 m/s into a flat, essentially unyielding surface
2. Static Load: the application of a 312 kN static compressive load on the package
3. Puncture: the drop of the package from a height of 3 m onto a steel puncture probe
4. Slash: two sequential impacts of a steel angle dropped from a height of 46 m onto the package
5. Fire: the exposure of the package to a fully engulfing jet-fuel pool-fire for 60 minutes
6. Immersion: the immersion of the package in water to a depth of at least 0.9 m

Two are stand-alone tests:

7. An individual terminal velocity impact test, which is only required if the package's terminal velocity is greater than 129 m/s
8. A deep submersion test where an undamaged package is submerged in water and subjected to an external water pressure of 4 MPa

The LPAT package (Figure 1) is cylindrical in shape with a diameter of 1.5 m and a length of 2.0 m. It has a mass of about 4400 kg. The package consists of three primary safety critical components: a Containment Vessel (CV), an Energy Absorber (EA) assembly and a Drag Enhance (DE). The high strength stainless steel CV prevents the release of radioactive material contents and provides shielding to limit radiation exposure to those around the package. The CV, which is about 380 mm in diameter and 700 mm long, is enclosed within the EA assembly. The EA assembly, which includes a body and six plugs, protects the CV from mechanical and thermal loads generated during the specified regulatory tests. The EA assembly components are comprised mostly of a unique composite material developed at Sandia that is made from successive layers of perforated aluminum sheets and Kevlar® fabric. This composite material provides excellent energy absorption, as well as exceptional structural and thermal protection. The DE is a lightweight aluminum structure that limits the free-fall terminal velocity of the package to less than 129 m/s. The DE was designed using Sandia's HPC resources and fluid dynamics analysis code (Figure 2).

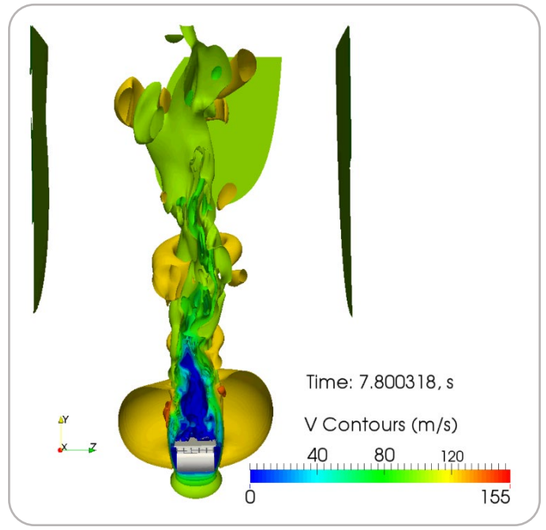


FIGURE 2

Drag enhancer fluid dynamics simulation.

An initial design for the LPAT package was developed from an earlier and smaller preliminary PAT package developed at Sandia. An engineering unit of the initial design was built and subjected to testing that intended to prove-out the feasibility of such a large PAT package. In 2013, the engineering unit was impacted at 133.8 m/s into a large concrete target at Sandia's 600 m rocket sled track facility. It was subsequently burned for 60 minutes in a jet-fuel pool fire at Sandia's Lurance Canyon Burn Site. Results of the tests demonstrated that a plutonium air transport package of the size of the LPAT package was feasible.

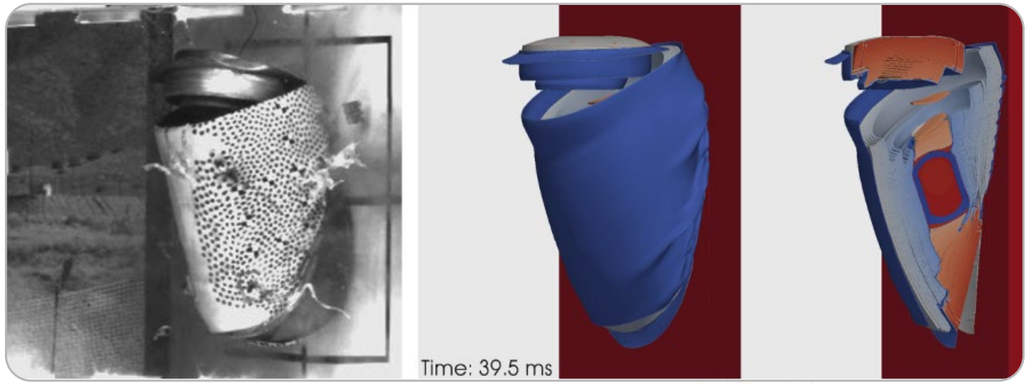


FIGURE 3

Comparison of the FEM response with the outcome of the engineering unit impact test.

Data from the tests were used to construct a detailed finite element model (FEM) of the package (Figure 3) for use with Sandia's SIERRA Solid Mechanics (SIERRA/SM) finite element analysis code. The half-symmetry model, consisting of 2.5 million elements, was used to refine the initial package design. Hundreds of simulations were completed using Sandia's HPC resources for each new design iteration to assess its response to the accident condition scenarios, culminating in a final package design in 2016. The half-symmetry final design model comprises approximately 4.2 million elements and has been used extensively in certification evaluations of the LPAT package.

A Safety Analysis Report for Packaging (SARP) is a document that describes the design and safety-critical features of a package and provides the technical evidence demonstrating that the package meets the regulatory requirements applicable to the package. In December 2019, a 1600+ page LPAT package SARP was submitted to regulators. Thousands of normal and accident condition simulations, completed over several years on Sandia's HPC machines, provided most of the data used to establish the performance of the package against regulatory requirements. Results from a series of

FIGURE 4

Corner certification impact test (real time).

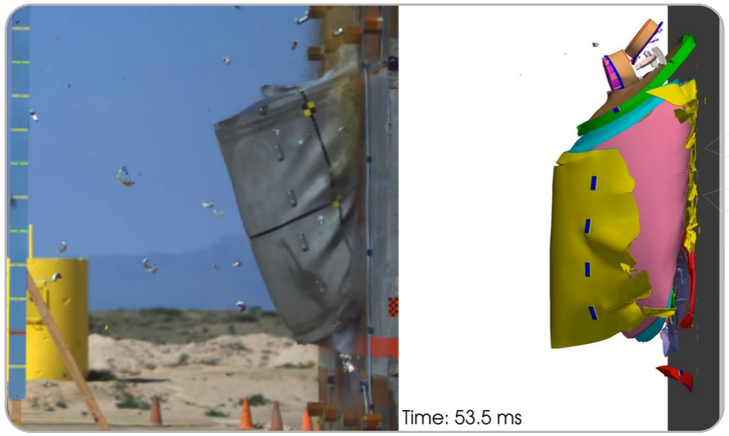


FIGURE 5

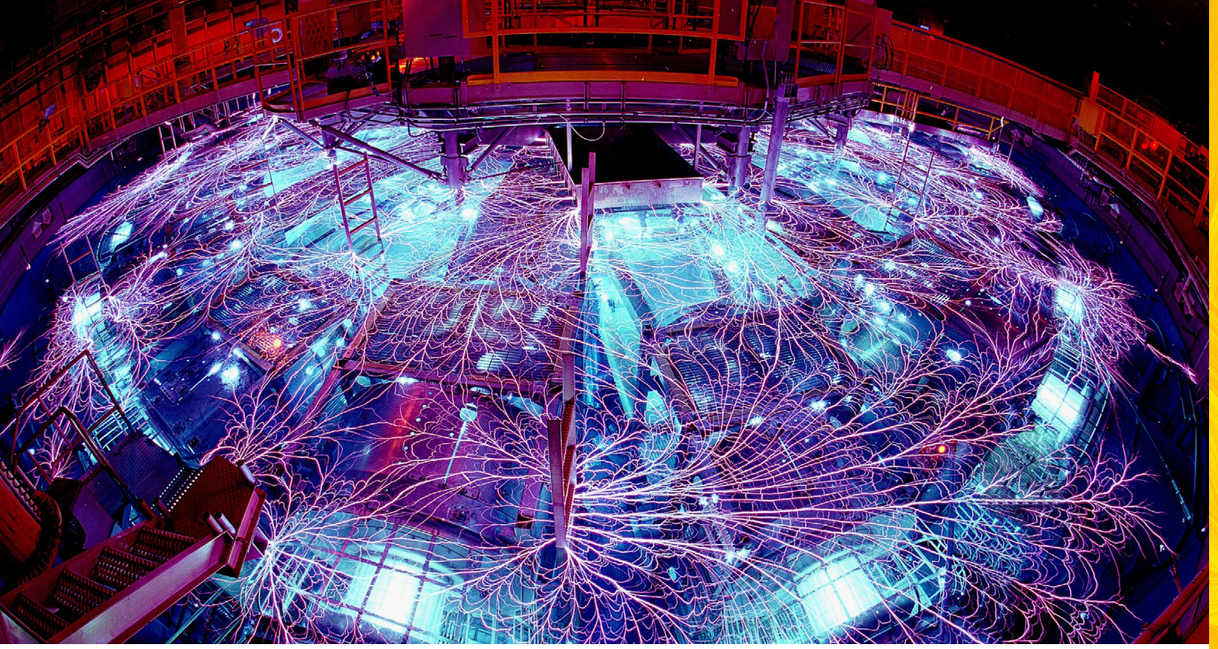
Comparison of the FEM response with the outcome of the corner certification impact test.

FIGURE 6

Certification fire test.

plutonium air transport certification tests are also described in the SARP. Data derived from the extensive set of simulations completed for the LPAT package provided a wealth of information to the design team, allowing them to identify the critical test scenarios that eventually comprised the certification test series. In 2018, three packages, in three impact orientations, were tested at Holloman Air Force Base: corner (see Figures 4 and Figure 5), side and end. The side orientation impact velocity was 131.2 m/s, corner orientation was 137.1 m/s and the end orientation was 139.2 m/s. Following the impact tests, the corner impact package was subjected to the remaining sequence of regulatory tests, including a fire test (Figure 6). Results from the tests demonstrated the exceptional performance of the LPAT package under the plutonium air transport HAC test scenarios.

With the SARP's submission in 2019, a decade-long process of developing the LPAT package came to a head at Sandia, with the Labs' HPC resources playing an instrumental role in the realization of this monumental achievement.



Unraveling material dynamics in intense radiation environments

AUTHORS/TEAM

Nathan Moore, Mikhail Mesh, Jason Sanchez, Kyle Cochrane, Shawn Pautz

CONTRIBUTING WRITER

Antonia (TJ) Cardella

Experiments at Sandia's Z Machine—the largest pulsed-power device in the world today—generate some of the most extreme laboratory conditions on our planet, providing glimpses into the behavior of planet cores, asteroid impacts and accretion around black holes; however, the enormous pressure, temperature and radiation intensity necessary to study these phenomena are often destructive to diagnostics and other equipment. The success of future fusion energy facilities will rely on wall materials and devices designed to operate within such harsh conditions. Potential insults include radiation, shock, heating, debris, magnetic fields and even vibration akin to small earthquakes—necessitating computer models of experimental designs that incorporate many different types of physical effects.

Recent advances at Sandia have enabled accurate, high-fidelity, large-scale simulations of many of these processes to be conducted for the first time. Central to these developments is ALEGRA (Arbitrary Lagrangian Eulerian General Research Application), a code with new algorithms that expand Sandia's capabilities. Unlike other codes customized for one application, ALEGRA was envisioned to allow simultaneous simulation of many different, interwoven processes. For example, ALEGRA can simulate the motion of individual x-ray photons as they scatter through a material, leaving tracks of energy that alter its flow—a process called radiation-hydrodynamics. As the flow is altered, the location and magnitude of the energy deposition is also altered. These processes can occur quickly. The cumulative effect of many photons striking a surface can induce temperatures up to 100,000K, ejecting plumes of hot gas at hypersonic speeds and altering the flow within nanoseconds.

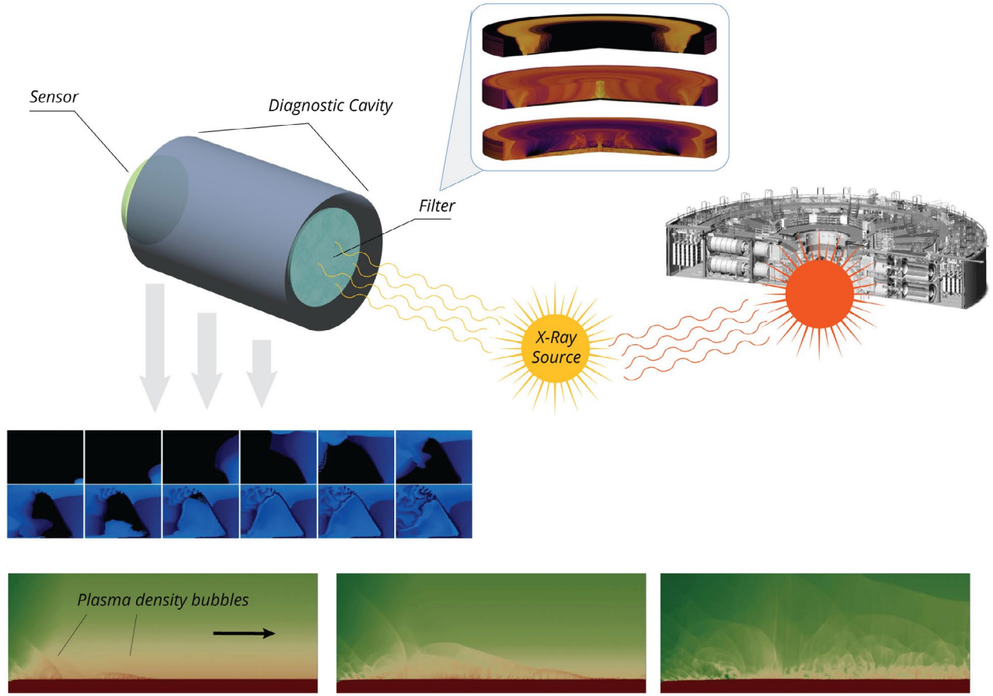


FIGURE 1

Despite their simplistic appearance, diagnostic cavities at the Z-Machine involve complex, dynamic material processes that often require multiphysics computer simulation to properly design.

A key advance is the ability to perform stochastic radiation-hydrodynamic calculations on axisymmetric geometries to allow rapid assessment of experimental cavities. Figure 1 shows an external filter that protects and controls illumination of an internal experiment. Radiation from the center of the Z Machine heats the x-ray filter enough to form an ionized plasma, which changes its ability to transmit x-rays. Some of the x-rays scatter into the nearby cavity walls where they generate a highly energetic plasma that travels at speeds up to 80 km/s.

This plasma overtakes the filter, leading to its complete destruction within a few millionths of a second (see Figure 2). Understanding the dynamics of these events is crucial for ensuring that the filter will maintain the desired illumination and protect the experiment inside the cavity.

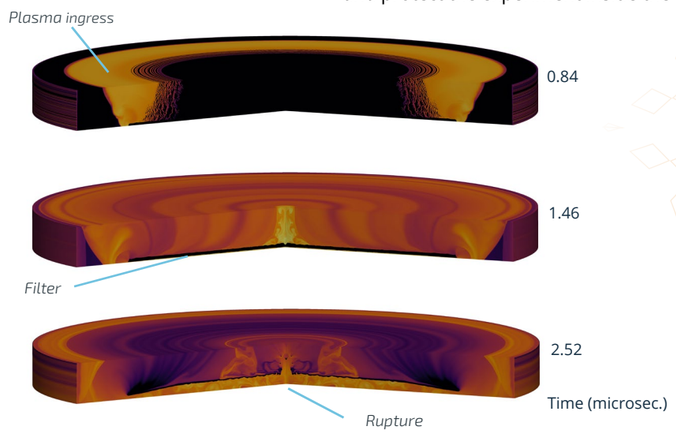
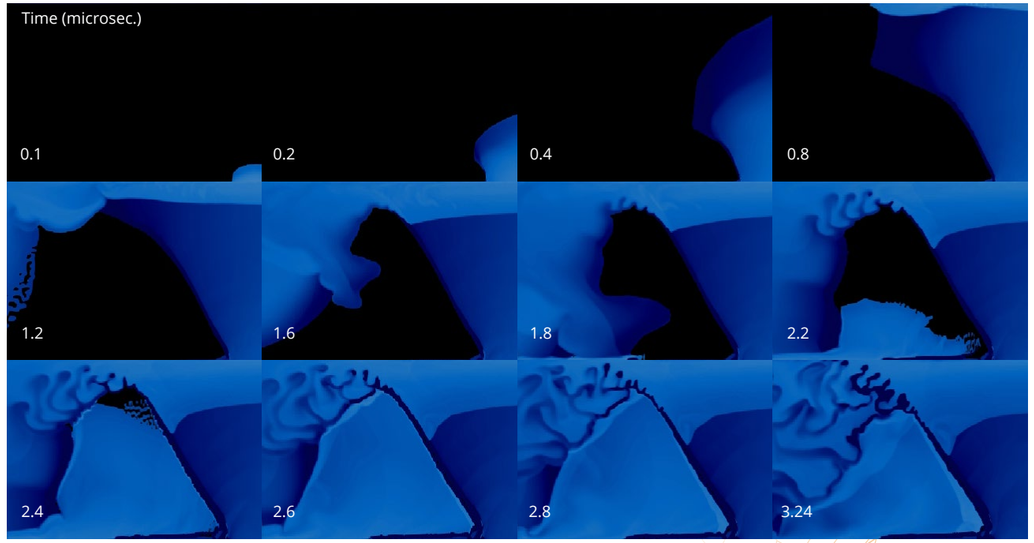


FIGURE 2

Snapshots of fully coupled radiation-hydrodynamics simulation using ALEGRA to predict the time of destruction of an x-ray filter used in an experimental cavity at the Z Machine.



The new capabilities have also been used to predict complex plasma flows inside experiments. As shown in Figure 3, hot plasma from an irradiated corner of the material begins to fill a void in the nearby vacuum. At first, the plasma is unimpeded and expands quickly, rebounding off the fixture walls. But the corner, which was superheated by the radiation, continues to eject a plume of hypersonic gas. The collision of plasma fronts restricts the expansion to a triangular region, while also inducing the fanciful vorticity seen around the edges. These plasma dynamics, which can only be understood through multiphysics simulation, influence the strength of optical diagnostic beams that must transmit through the cavity. ALEGRA simulations have enabled the design of such novel experiments.

FIGURE 3

Snapshots of plasma evolution within an experimental cavity.

Radiation-hydrodynamic simulations have also been used to improve our understanding of surface erosion from intense radiation bursts, such as those that would be needed to deflect an asteroid on a collision-course with Earth, or to predict the reliability of a containment wall in a fusion reactor. Certain Z Machine experiments can be used to emulate these processes. Figure 4 shows a simulation of an experiment heating the edge of a metallic disc with an intense x-ray pulse. Although the entire surface vaporizes, perturbations near the outer edge cause a series of plasma density bubbles to race toward the center of the cylindrical panel. These density fluctuations cause different rates of erosion to occur across the disc. Understanding the interactions of these plasma jets with material surfaces may be particularly important for deflecting asteroids, many of which have rough surfaces and uneven topology.

Sandia's broad mission space requires flexible, multiphysics codes that can tackle a wide range of problems. Toward that end, radiation-hydrodynamic capabilities have been further extended by coupling ALEGRA to SCEPTRE under the RAMSES code suite. This coupling was the result of a multi-year collaboration between the code teams, combining over two decades of research and development. Like ALEGRA, SCEPTRE tracks the movement of photons in a material, but it also calculates the trajectories of electrons jettisoned from atoms that are struck by x-rays. Other collisions will simply

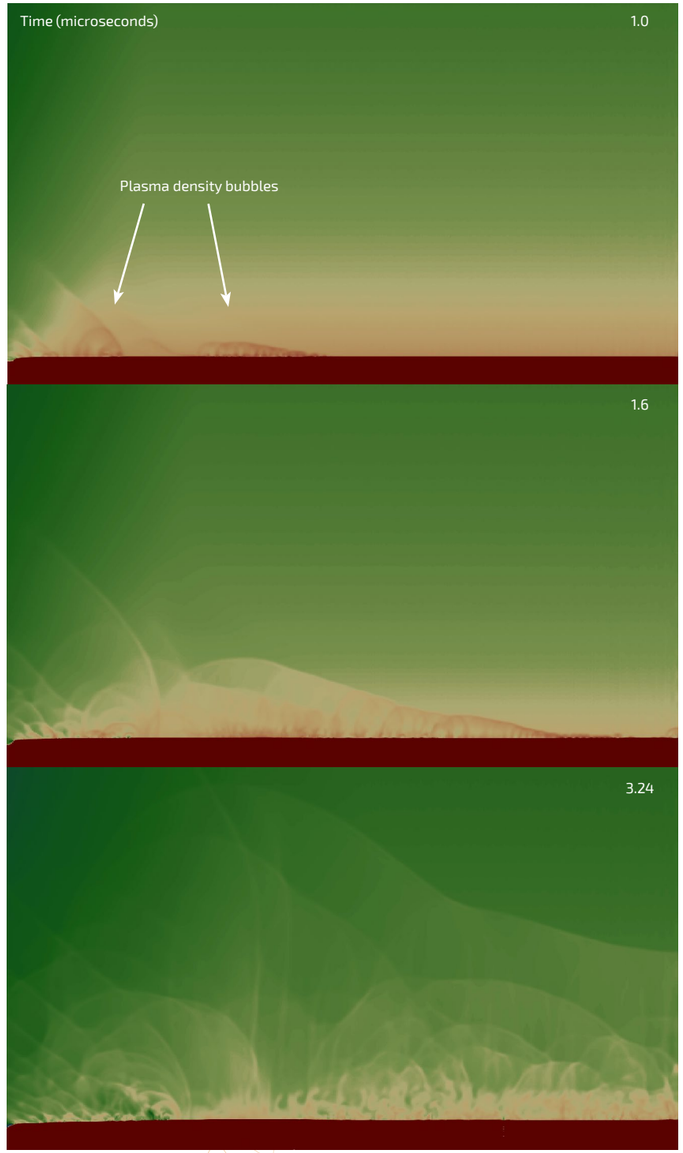


FIGURE 4

Snapshots of plasma density bubbles following x-ray irradiation of a metal surface. Disturbances in plasma density race toward the center of a cylindrical disc, leading to differences in surface erosion rates.

change the direction of a photon, but some of the energy is lost. SCEPTRE contains advanced algorithms that recalculate the resulting change in the energy spectrum of the x-ray field as it propagates through a material. The result is increased accuracy in calculating where energy is deposited, which affects erosion of walls and coatings. SCEPTRE can also simulate intense streams of electrons impinging on materials during particle accelerator experiments.

Results from SCEPTRE are communicated back to ALEGRA in real time using a multiple program, multiple data (MPMD) algorithm, which works like an exchange broker: each code contributes new information about the material state, which is updated in real-time as the material evolves. Besides increasing accuracy, this strategy allows developers to work on each code independently, streamlining future improvements. A recent upgrade to the x-ray absorption library in ALEGRA allows calculations to be performed for mixtures of any elements in the periodic table, enabling agile design of materials and experiments. Another critical advance is the ability to apply SCEPTRE to selected regions of a material, which has led to a 16x reduction in computational cost and enabled larger scale, multiphysics solutions.

Future developments will further improve the capability, usability and flexibility of ALEGRA. Work is underway to more seamlessly couple ALEGRA to codes that use different numerical schemes to represent the same geometry, and to allow more seamless transfer of data between different models. These advances will expand the number of ways that models for different physical processes can be combined to solve complex, multiphysics problems.

Radiation-aware Xyce simulations of memory circuits

Integrated circuit simulation for design

AUTHORS/TEAM

Eric Keiter, Heidi Thornquist, Ting Mei; Pranita Kerber, Biliana Paskaleva, John Teifel, Spencer Nelson, Clark Dohrmann, Ichitaro Yamazaki

CONTRIBUTING WRITER

Antonia (TJ) Cardella

The development of reliable, radiation-hardened microelectronics is an important aspect of Sandia's core mission. For many decades, Sandia has had a prominent role in advancing the state-of-the-art in microsystems, R&D radiation effects, reliability physics and failure analysis. The process of radiation-aware integrated circuit design involves many steps including architectural design, logic design, physical design, physical verification and final sign-off. Each of these steps involves different types of computational tools, which comprise a "tool flow." Such tool flows are a standard practice in the microelectronics community and form the basis for the electronic design automation industry.

One type of computational analysis is circuit simulation[1] that involves a detailed, transistor-level description of the circuit to generate a system of network-coupled differential algebraic equations. Originally made popular by the Berkeley SPICE program,[2] transistor-level simulation becomes impractical for large-scale circuits due to reliance on sparse direct linear solvers. As a result, while SPICE-style simulation is standard practice for analog circuit design, it is less commonly used in design of larger digital circuits. Digital designs are subject to constraints not present in analog designs, and these constraints can often be exploited to expedite computational analysis. Hence, the traditional tool flow for digital integrated circuit (IC) design only applies transistor-level simulation codes in a very limited manner: to characterize standard cells. Analysis of entire ICs—which may comprise billions of transistors—is subsequently performed using specialized, lower-fidelity tools. These specialized tools include applications such as timing simulators, and use standard cell models as input.

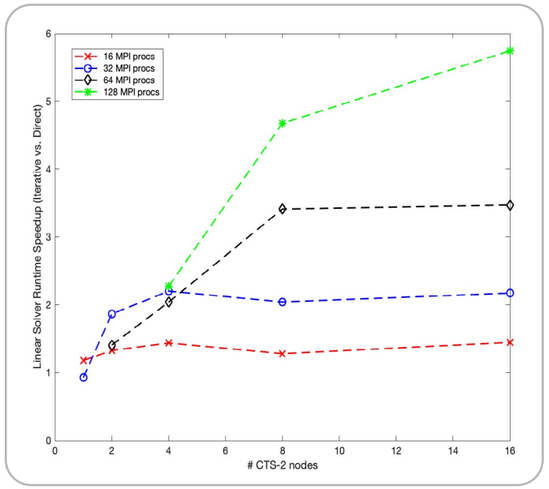


FIGURE 1

Runtime speedup of block subdomain preconditioned method (overlap = 1, Intel MKL Pardiso, 4 threads) compared to direct method (Intel MKL Pardiso, 16 threads) in simulating SRAM circuit on CTS-2. The MPI processors are varied from 16, 32, 64, and 128, while the number of CTS-2 nodes are varied from 1,2,4,8, and 16.

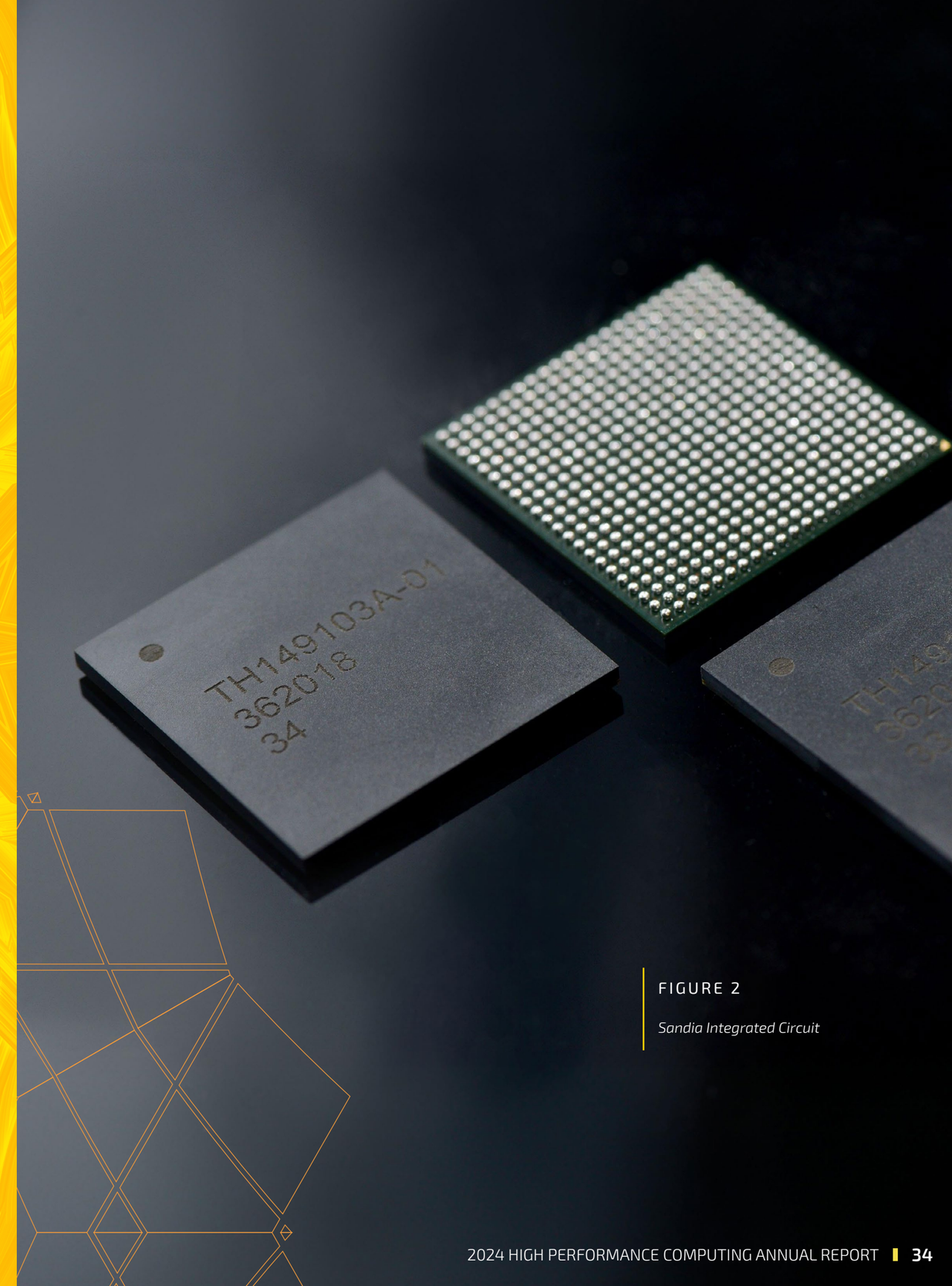


FIGURE 2

Sandia Integrated Circuit

In the context of radiation-hardened design, it can still be beneficial to perform transistor-level circuit simulation on an entire digital IC to provide high-fidelity information of the circuit response. This is because low-fidelity digital design tools are not intended to account for certain types of radiation effects, and their presence can violate some of the assumed constraints, potentially affecting simulation accuracy. Hence, to efficiently perform SPICE-style simulations on large ICs, novel computational tools, like Xyce, are necessary.

Xyce is an open source, SPICE-compatible, high-performance transistor-level circuit simulator developed at Sandia. It is capable of solving a broad range of circuit problems, from small-scale circuits on desktop computers to extremely large circuits on large-scale, high-performance computing platforms. Actively developed since 1999, Xyce provides the capability to investigate general network systems and has been integral to simulating radiation effects on Sandia-designed circuits as well as biological/neural networks and power grids. Diverse requirements in capacity, application and analysis have necessitated R&D of unique algorithms and techniques that facilitate simulation in both the time and frequency domains. Network simulation tools like Xyce are generally part of a larger analysis tool flow that has motivated recent work in improving capabilities for workflow integration.

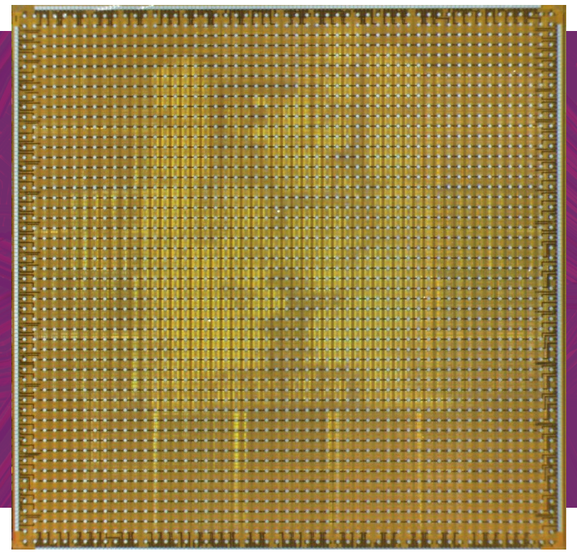


FIGURE 3

Sandia Integrated Circuit.

Xyce is used to analyze ionizing radiation effects in mission-relevant, application-specific integrated circuits (ASICs). This work is supported by the Advanced Simulation and Computing program's Accelerated Digital Engineering initiative and facilitates the use of radiation-aware digital engineering techniques within the standard ASIC design flow.

A recent performance study illustrates the scalability that Xyce can achieve on the new CTS-2 HPC Production Cluster, called Amber. This computing resource has 2.0 GHz Intel Sapphire Rapids processors with dual sockets and 56 cores, for a total of 112 cores, and 256 GB RAM per node. There are 1496 nodes available, but the results from this study are limited to 16 nodes.

The circuit of interest for this study is a large, static random-access memory (SRAM) circuit designed for a 180-nm complementary metal oxide semiconductor technology. The SRAM circuit comprises 1.6 million transistors, leading to a coupled system of equations with 8.8 million total unknowns. The computational time required to perform transistor-level simulation on a circuit of this magnitude is dominated by repeatedly solving a linear system of equations. The assembly of the linear system, which evaluates device models to obtain contributions to the matrix and right-hand side vector, is scalable. However, the matrices generated in circuit simulation are typically sparse, have heterogeneous non-symmetric structure and are often ill-conditioned. Traditional circuit simulators rely on direct methods to solve this challenging linear system because they are dependable and easy to use, but such methods scale poorly to this size of problem and dictate the total simulation time. Iterative methods have greater potential for scalability, but their performance is predicated on finding an adequate preconditioner.

The performance study compared the observed scaling on CTS-2 when using a direct method versus a block subdomain preconditioned iterative method to perform transistor-level simulation. The direct method and block subdomain preconditioner used the same solver, Intel MKL Pardiso, with 16 threads and 4 threads, respectively, to either perform the global or local solve. This strong scaling study illustrated that the linear solver speedup can be substantial when iterative methods are used, up to 5.7x faster on 16 CTS-2 nodes with 128 MPI processors. The optimal number of CTS-2 nodes that can accelerate the simulation is dependent upon the number of MPI processors.

In the end, the result of this study showed that Xyce's capability allows detailed, radiation-aware circuit performance predictions to be produced in a reasonable amount of time, which in turn can improve understanding of circuit margins, allowing more time for informed decisions to be made.

References

1. Najm, F. N. Circuit Simulation. John Wiley & Sons, Inc., 2010.
2. Nagel, L. W. SPICE2, a computer program to simulate semiconductor circuits. University of California, Berkeley, Technical Report, Memorandum ERL-M250, 1975.

CaaS infrastructure to support Accelerated Digital Engineering

The CaaS journey at Sandia

AUTHORS/TEAM

Angel Beltre, Kevin Pedretti, Chris Garasi, Sylvain Bernard, Andrew Younge

CONTRIBUTING WRITER

Rebecca Cox

The Accelerated Digital Engineering (ADE) initiative at Sandia is pioneering the creation of radically streamlined and highly interactive user interfaces to advance modeling and simulation capabilities. The goal is to improve end-user productivity and increase the impact on early design and surveillance activities. Underpinning ADE is a Computing-as-a-Service (CaaS) layer that integrates cloud and HPC technologies to support the deployment of interactive, user-facing services and the orchestration of back-end computing jobs running across Sandia's computing resources. Realizing this vision required an interdisciplinary team of simulation, user experience and computing infrastructure experts working across organizational boundaries to push computing at Sandia forward.

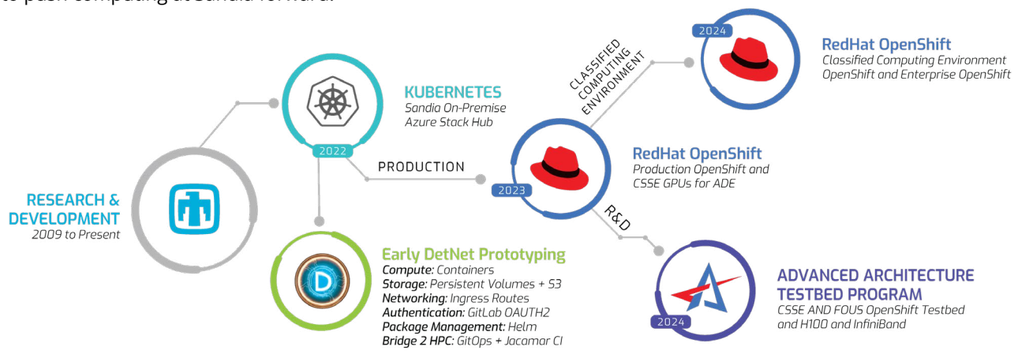


FIGURE 1

Container orchestration as a service has reached critical mass at Sandia

Figure 1 provides a high-level view of how CaaS infrastructure has evolved over the past several years. Early research explored cloud technologies such as virtualization and containers. In 2022, an on-premises deployment of Microsoft's Azure platform was leveraged to explore the use of Kubernetes (K8s) a cloud operating system for containerized applications and services, used to deploy ADE exemplar applications.

The team collaborated with the ADE DetNet team to explore how to provision computing, storage and networking resources on the K8s platform and link to Sandia's existing HPC infrastructure. A novel scheme leveraging GitLab Jacamar runners was used to enable front-end applications running on K8s to link to and manage jobs running on HPC systems, which is a technology developed by the Exascale Computing Project (see Figure 1). Throughout 2022 and 2023, K8s usage migrated to production instances of the RedHat OpenShift Container Platform (OCP), a commercial version of K8s, running on Sandia's restricted and classified computing environments. These production systems were augmented with additional computing resources, including servers and graphics processing units (GPUs) to support ADE CaaS workloads.

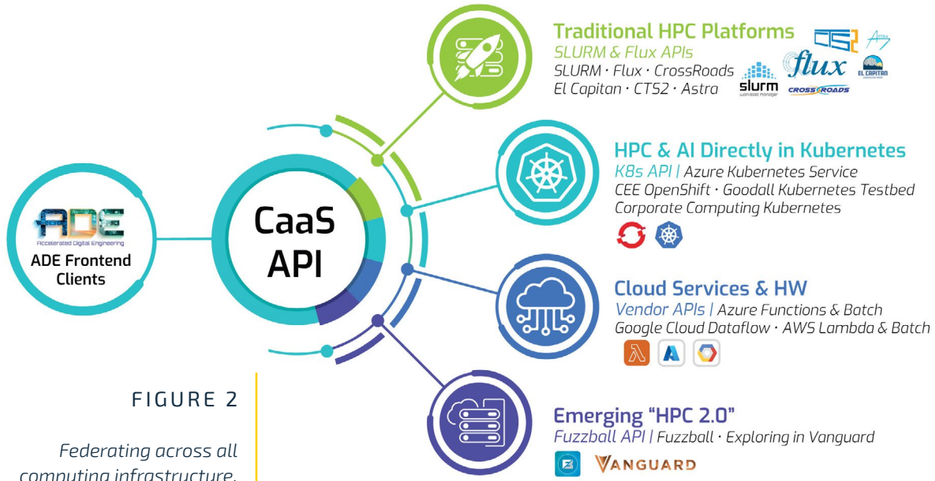


FIGURE 2

Federating across all computing infrastructure.

The role of R&D on MPI, InfiniBand, CaaS API and multi-cluster federation

Recent R&D efforts at Sandia have led to the deployment of an OpenShift testbed designed to support exploratory activities that are challenging to execute in production environments. This initiative is part of a broader commitment to enhancing computing capabilities on cloud-based platforms. The Sandia team is adding Message Passing Interface (MPI) and InfiniBand (IB), to K8s to run HPC jobs, while investigating Kubernetes-like Control Planes (KCP) to enable routing jobs seamlessly to both K8s and HPC clusters. These endeavors are crucial for investigating the integration of cloud computing with HPC, aiming to create a seamless convergence that can eventually be applied to production environments. The techniques and technologies developed within the OpenShift R&D testbed are intended to be transitioned to production once successful, benefiting ADE workloads and end-users. Current approaches involve leveraging K8s operators to facilitate some of these advanced capabilities. These operators, developed by the Cloud Native Foundation or major industry vendors, provide a foundational framework for Sandia's in-house research, testing and development processes.

The CaaS team has successfully executed multi-node MPI jobs within the elevated security architecture of the OpenShift Cluster (OC) thereby enabling scalable ADE workflows on the platform. In addition, an effort to enable IB support aims to enhance node and core affinity for application deployment, utilizing OC as their backend.

Figure 2 illustrates the CaaS Application Programming Interface (API) as a federation layer designed to communicate with the underlying computing infrastructure at Sandia. Additionally, the CaaS team has initiated R&D efforts to enable use of KCP as a federation layer. The KCP initiative is developing a unified back-end control plane that effortlessly merges traditional and modern HPC environments, cloud-based infrastructures like K8s and OpenShift services from major cloud vendors, and emerging workflow managers such as Fuzzball. This integration is facilitated through an API-based approach, aiming to improve the way developers interact with HPC systems and how they utilize them in their workflows.

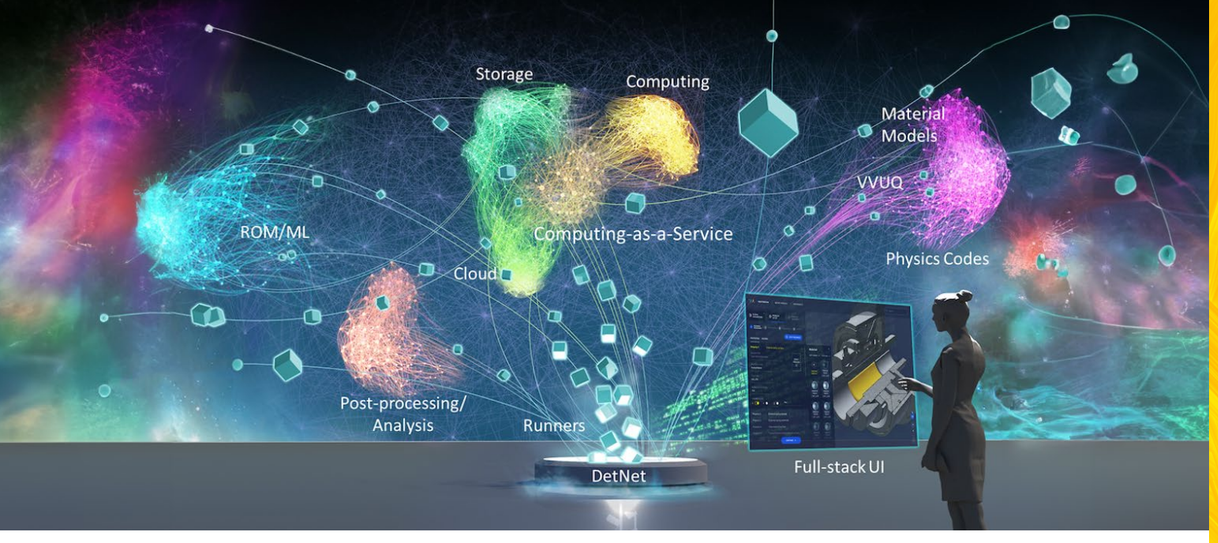


FIGURE 3

Graphic artist representation of a complex digital ecosystem, coordinated by CaaS elements, enabling a digital engineering workflow for a user. (Credit: Todd Hebenstreit & Christopher Garasi - 2020)

Exploring K8s for AI services

The OCP infrastructure at Sandia has proven to be an asset for deploying Artificial Intelligence (AI) services efficiently. By leveraging OCP, Sandia has been able to harness the power of containerization and orchestration to streamline the deployment, scaling and management of AI and Machine Learning (ML) workloads. This approach not only capitalizes on OCP's inherent strengths in performance, reliability and scalability, but also benefits from its elasticity and flexibility, making it a well-suited platform for the dynamic requirements of AI services. Today, Sandia is positioned to deliver AI services and continues to move ADE forward to fulfill mission demands more efficiently.

Sandia's future posture for adopting cloud technologies

Sandia's CaaS team is working to integrate new cloud technologies (containers, K8s, Fuzzball, etc.) with our traditional HPC environment to modernize and improve these technologies. Our goal is also to leverage the improved computing environment to explore delivering ASC modeling and simulation capabilities as turnkey services to end-users, making them more accessible and easier to use.

Figure 3 shows a graphic artist interpretation of a complex digital ecosystem created approximately four years ago. At that point, DetNet and other ADE applications like it were simply a vision. Four years later, Sandia's CaaS team has realized this vision through collaboration with the integrated codes teams, web application and user experience experts and CaaS infrastructure.

Beyond 2025, Sandia aims to maintain its forward-looking posture by actively participating in holistic co-design efforts with cloud technologies. These efforts aim to provide critical insights to enhance the understanding of HPC requirements and increase Sandia's impact in the larger computing community.

References

1. Advanced Simulation and Computing (ASC). Vanguard. n.d. 08 06 2024. <https://vanguard.sandia.gov>.
2. Sandia National Laboratories. Advanced Simulation and Computing. n.d. 26 05 2024. <www.sandia.gov/asc/>.
3. Younge, Andrew J. Sandia "Researchers Collaborate with Red Hat on Container Technology." 2020. 24 05 2024. <https://www.sandia.gov/ccr/news/sandia-researchers-collaborate-with-red-hat-on-container-technology/>.

Acknowledgments

EDITORS
Sophia Corwell
Ariana Stern
Sheina MacCormic

DESIGNER
Johanna Pearman

VIDEOGRAPHER
Vince Gasparich

SIMMAGIC
Aaron Moreno

COPY EDITOR
Whitney Lacy

MANAGEMENT
Tom Klitsner
Sophia Corwell
David Littlewood
Kendall Pierson

Sandia National Laboratories is a multi-mission 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. SAND2024-134970

Sandia National Laboratories

U.S. Department of ENERGY

National Nuclear Security Administration