Scientific Discovery on Exascale Systems

June 15, 2015

Submitted by:

Bruce Hendrickson
Director, Center for Computing Research
Sandia National Laboratories

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000

Contents

Title/Name of Your Application: The Accelerate Climate Model for Energy (ACME)

Name, Institutional affiliation, and e-mail address: Dr. Mark Taylor, ACME Chief Computational Scientist, mataylo@sandia.gov

Overview Description and Impact: Today's climate models have demonstrated the ability to simulate the earth's climate of the last century with remarkable skill for continental spatial scales over multiple decades. Despite this success, adaption and mitigation strategies require accurate probabilistic simulation of changing climate statistics over spatial scales as small as watersheds and over ten-year periods. Based on our long history of steadily improving the skill of climate models, we know this will require a combination of increased resolution, improved parameterizations and additional complexity. Three aspects of climate and Earth system simulation require the power of future exascale platforms.

Spatial Resolution: Coupled Earth System Models will have increased spatial resolution. For example, ACME's first model will be at the hydrostatic limit in the atmosphere and eddy-resolving in the ocean. Today's 10-30 petaflop systems make it possible to run long climate simulations at resolutions relatively high resolutions (25km) but that are still well within the hydrostatic limit (10km). The also enable very short exploratory simulations with global cloud resolving (1km) resolutions. Assuming an increase of 50x -100 x in performance over today's systems multi-decade climate simulations will push into non-hydrostatic weather resolving resolutions (5km) and it will become more practical to run short exploratory global cloud resolving simulations. With the land grid defined as watersheds, we have the opportunity to implement fully 3-D surface/subsurface implicit solves at high resolution (~30 m) within each watershed-gridcell.

Multi-scale Simulation: Exascale computing systems will enable the replacement of sub-grid parameterizations with more explicit simulation sub-models, such as cloud-resolving "super parameterization" recently developed for atmospheric models. Scientifically, this approach has shown some promising early results. These types of models are ideal for Exascale computing systems and would make climate length simulations possible with many cloud-resolving features.

Large Ensembles: Because of the non-deterministic nature of the coupled Earth system, probabilistic prediction of the changes in climate statistics that result from the small, but persistent long term changes in the Earth's energy budget necessitates executing large ensembles of simulations. This is a valuable but straightforward use of exascale computing.

System Requirements: Like many mixed PDE-ODE simulation problems, climate simulation requires scalable algorithms and software to utilize Exascale hardware. A first step was taken by ACME with every component in the v1 model running on fully unstructured grids, removing the largest scalability bottleneck in Earth system models. This increased scalability has taken us well into the petascale regime, but further increases in scalability are required to move beyond this regime. Currently, we have several alternatives to increasing parallelism. Based on current trends, we anticipate Exascale systems will have similar or fewer nodes then today's systems, which each node relying heavily on accelerators. Thus new levels of parallelism will need to be exploited through increased use of on node parallelism such as threading and vectorization.

Code and Tools: The ACME project is currently investing heavily in openACC (for GPU accelerated systems) and openMP for Intel Phi based systems. This support comes from a variety of programs in DOE's Office of Science, including BER, ASCR, and LCF early readiness

programs. It remains to be seen if these approaches can coexist in a single code base, or if multiple implementations will be required to support both approaches. If the later, than a key tool will be suitable programming model which can provide both performance and performance portability.

Models and Algorithms: Time integration remains a bottleneck even with perfect weak scaling. New algorithms and approaches are needed, such as multiscale techniques where fast processes can be treated locally on accelerators, and more efficient time stepping methods which can make better use Exascale machines, such as multi-rate, implicit and parallel in time approaches. Improvements in model coupling will also be needed, to allow for both increased concurrency and better accuracy as the models resolve finer scales.

End-to-End Requirements: Exascale climate models will generate 10^2-10^3 PetaBytes of simulation output that are valuable to the full climate community. ACME has an explicit component for server side analysis and visualization using data engines close to the data source. The solution is scalable to the Exascale, but requires investments in infrastructure close to the data sources, presumably near the LCFs.

Related Research: In the era of simulation software with $10^6 – 10^7$ lines of code, advancing climate simulation and prediction on Exascale architectures requires BOTH computational science research and software engineering, the "development" part of R&D. Computational science libraries and tools can have a positive and sustained impact on many code projects. Libraries amortize verification/maturation/performance-tuning expenses when adopted by multiple codes requiring the same or similar algorithms. Furthermore, unified deployment of libraries decreases barriers to interdisciplinary research. Software Quality Tools allow expert software engineers to be shared among multiple projects common tools improves agility of staff between projects

10-Year Problem Target: We envision a fully coupled atmosphere-ocean-land-land ice-sea ice model that would run at 5 simulated years per day at fully-resolved deep convective scales in the atmosphere (approximate grid spacing of I km horizontally), 100 m grid spacing on the land and land ice, and nominal 10 km grid spacing in the ocean and sea-ice to capture the dynamics of mesoscale ocean eddies. The application would execute an ensemble of 10 simulations currently in an ensemble, resulting in a century of simulation in a month of wall-clock time. This will allow the scientific community to address the following questions:

- How will the sea level, sea-ice coverage and ocean circulation change as the climate changes?
- How will the distribution and cycling of water, ice, and clouds change over the coming decades?
- How will climate and extreme weather change on the local and regional scales?

Direct Numerical Simulation of Turbulent Combustion, S3D and AMR

Jackie Chen, Sandia National Laboratories, jhchen@sandia.gov

Co-leads: John Bell, Lawrence Berkeley Laboratory, jbbell@lbnl.gov
Ray Grout, National Renewable Energy Laboratory, rwgrout@nrel.gov

Overview Description and Impact: Considerations of energy and environmental security and sustainability, as well as economic competitiveness, demand accelerated development of advanced industrial and transportation combustion technologies that combine high efficiency, low emissions, and the ability to reliably operate on an increasingly diverse range of fuels, including bio-derived and synthesis fuels, as well as an evolving feed of fossil fuels. Development of these technologies is significantly hampered by the lack of robust, predictive computational design tools for advanced combustion systems, particularly in new mixed-mode combustion regimes where stringent efficiency and emissions legislation are driving future technologies.

In stationary gas turbines for power generation natural gas is currently the dominant commercial fuel. On the other hand hydrogen-rich gaseous fuels derived from coal and biomass gasification processes are attractive alternatives to natural gas and may offer lower carbon emissions. Furthermore, with tightening emissions regulations industry is embracing novel combustion concepts such as lean-premixed combustion wherein the fuel is mixed with an excess of oxidizer prior to entering the combustion zone. While clearly attractive from an emissions and efficiency perspective, lean premixed combustion poses serious design challenges. Using direct numerical simulations of canonical turbulent flame configurations we propose to study four primary design challenges faced by gas turbine manufacturers: 1) flame stabilization and NO_x emissions; 2) autoignition in reheat burners; 3) thermoacoustic instabilities; and 4) flashback.

In internal combustion engines used in the transportation sector innovations have been hampered by the lack of scientific understanding of how variations in fuel composition affect engine performance, and hence by the inability to predict impacts when conventional petroleum-based fuels are modified or replaced. Even for conventional fuels, existing models are often unsatisfactory, particularly near the limits of stable operation of advanced engines. At the ragged edge, strong sensitivities to subtle differences in chemical fuel properties are amplified by the stochastic nature of turbulence, which may lead to undesirable effects such as misfire or knock. Therefore, fundamental research in combustion science is essential to understand and predict the behavior of a diverse range of fuels in aero-thermochemical environments representative of emerging low-temperature premixed combustion engines. We propose to perform DNS in engine-relevant configurations to understand fundamental coupling between isentropic compression, autoignition kinetics, turbulent mixing from multiple fuel injections, and flame and soot chemistry. These fundamental studies will enable us to develop predictive models for the co-design of fuels and future advanced engines that avoid engine knock, stabilize a lifted diesel jet flame at aero-thermo-chemical conditions that minimize soot production, modulate combustion phasing or heat-release rate to provide maximum fuel efficiency and near-zero emissions, or provide mixing conditions needed to reduce emissions. Many of these turbulence-chemistry interactions are governed by mixed-modes of combustion including flame propagation into a pre-igniting mixture, lifted flame stabilization by cool flame chemistry, and effects of thermal, composition, and reactivity stratification on compression ignition. Exascale DNS capabilities are required to achieve the high Reynolds number, high pressure, and complex chemical kinetics required to tackle these gas turbine and IC engine science problems. Larger domains and longer simulations are needed to provide statistical convergence, and a larger dynamic range of scales are required to resolve high Reynolds number flame/turbulence phenomena at pressure.

System Requirements: The evolving architectures towards exascale compound the problem of programmer productivity. As machines integrate heterogeneous processors such as GPUs and deep memory hierarchies, the programmer becomes responsible for deciding how to effectively map applications onto a target

architecture. Existing programming models force the programmer to implement these mapping decisions directly in the source code. This entangles the functionality of the code with its mapping in a way that inhibits either aspect of the code from being modified without a complete understanding of the other. Even worse, achieving performance portability across different architectures requires implementing two or more different mappings in the same source code, which is commonly done with a preponderance of conditional compilation directives and/or complete forks of the source tree. Therefore, we need a raised level of programming abstraction using a task-based programming model, i.e. high productivity computing for combustion via the construction of performance-portable programs enabled by data-centric asynchronous dynamic task-based programming environments/runtimes, autotuning and optimization of mappers encapsulating domain specified policies, and embedded domain specific language compilers for stencil-based PDE's, thermo-chemistry&transport, for analytics, UQ and viz. We also need hardware support for vectorized division and fast exponentials, increased memory size and bandwidth, large registers, and NVRAM/burst buffers to perform in situ analytics, UQ, multi-level checkpointing, and for resiliency.

Codes and Tools:

S3D is a massively parallel direct numerical solver to simulate turbulent combustion in canonical configurations and thereby gain fundamental insights into the physical and chemical interactions in turbulent reacting flows. S3D is a complex piece of software that has evolved over the course of thirty years with support from DOE BES and ASCR. The original version of S3D consists of approximately 200K lines of Fortran and uses MPI for communication between threads. A second version of S3D targets heterogeneous systems by combining MPI with OpenACC directives in a hybrid implementation. By leveraging both CPUs and GPUs, the MPI/OpenACC version of S3D is roughly two times faster than the Fortran/MPI version. A third version of S3D was written in Legion, a novel task-based programming system. Legion automates details of scheduling tasks and data movement, and separates the specification of tasks and data from the mapping onto a machine. The result is a robust, easily modified large (*i.e.* 100+ transported species) chemical mechanism implementation that improves time to solution by nearly 9X over the S3D MPI code and obtains over 80% of the achievable performance on the target systems, 13,000+ nodes on Titan at OLCF and 5000+ nodes on PizDaint at CSCS.

Models and Algorithms: We need to develop new communication-avoiding asynchronous algorithms for solving both the compressible and low-Mach reacting Navier-Stokes equations including high-order adaptive mesh refinement, spectral deferred correction methods for AMR, and asynchronous finite-difference stencils. Alternatively, to overcome the challenges related to both thin spatial layers and temporal stiffness we need to develop a wavelet adaptive multilevel representation (WAMR) in space and an adaptive model reduction method (G-Scheme) in time. The G-Scheme is an explicit solver developed for stiff problems that is built upon a local decomposition of the dynamics in three subspaces involving slow, active and fast time scales.

End-to-End Requirements: We need to develop data management infrastructure (i.e. middleware, runtimes, algorithms, and hardware support) to support in situ, in-transit computation of adjoint sensitivity UQ, combustion and turbulence analytics, multi-level I/O and visualization.

Related research: There are companion combustion experiments for each of the target problems described above that provide complementary information to the DNS. The exascale computations are needed to provide sufficient ensemble sizes for converged statistics and put the DNS in the same parameter regimes for comparison with experiments. A combustion gateway or portal is needed to share the DNS and experimental data and tools with the global modeling community (academia, industry, labs).

10 year target problem: One exascale target problem is performing DNS of dual fuel reactivity controlled compression ignition with a gasoline blend primary reference fuel (PRF) which has the potential for 60% efficiency and near-zero emissions. To perform an engine relevant DNS at a turbulent Reynolds number of 4300, pressures between 30-60 atm, temperatures between 750-2000K would require 1 PB of state, 3 PB high water memory use, O(1 million time steps), wall clock time (20 days at billion-way concurrency), and generate 1 exabyte of data. Including in situ adjoint sensitivity UQ would increase this cost by a factor of three.

Development of Predictive Combustion Models through Coupled Application of Simulations and Experiments using Exascale Systems

Joseph C. Oefelein (oefelei@sandia.gov), Jonathan H. Frank (jhfrank@sandia.gov)
Sandia National Laboratories, Combustion Research Facility, Livermore CA
Ramanan Sankaran (sankaranr@ornl.gov)
Oak Ridge National Laboratory, Oak Ridge Leadership Computing Facility, Oak Ridge, TN

Overview Description and Impact: Aggressive national goals for reducing petroleum use by 25 percent by 2020 and greenhouse gas emissions by 80 percent by 2050 will require major improvements in all aspects of our nation's energy use. At the same time, the U.S. transportation and energy sectors are under tremendous pressure from international competitors and challenging economic conditions. Achieving reduced fuel usage and emission goals will require significantly shortened product development cycles for cleaner, more efficient engine technologies. Concurrently, fuels will also be evolving, creating additional complexity and further underscoring the need for efficient product development cycles. Under the current cut-and-try approach, design cycles simply take too long. These challenges present a unique opportunity for U.S. leadership in supercomputing to develop predictive simulation tools for combustion. Simulations that reliably predict engine efficiency and pollutant emissions using both conventional and new fuels require higher fidelity modeling than is possible with current computing resources. Exascale platforms will facilitate model development and validation, while providing the necessary speedups in the time to solution and fidelity of numerical data required to make revolutionary advances in the design of combustion systems.

While substantial progress has been made to understand basic principles of turbulent combustion through advances in experimental and computational capabilities to-date, this understanding is still limited due to the inherent highly nonlinear multiscale/multiphysics nature of the phenomena. Direct Numerical Simulation (DNS) can only be applied for moderate turbulence levels and small combustion volumes. Thus, some form of modeling will always be necessary to access device-relevant conditions and geometry, even with exascale computing. Recognizing the limitations for application of DNS, Large-Eddy-Simulation (LES) approaches have been advanced for both science and engineering and have significant potential for design. Specific research priorities for development of predictive models have been defined in recent workshops such as the SC-BES sponsored "Workshop on Clean and Efficient Combustion of 21st Century Transportation Fuels," and the jointly sponsored SC-BES and EERE-VT "Workshop to Identify Research Needs and Impacts in Predictive Simulations for Internal Combustion Engines." Needs are centered on addressing the basic-science questions that limit our ability to simulate turbulent combustion, and the critical questions related to model development and validation. These (and related) reports consistently outline the need for a synergistic combination of LES and experiments. Because LES depends on the use of models that approximate physical and chemical reality, high-fidelity experimental data over a range of space and time scales is needed to properly evaluate and calibrate these models. With the recent development of high-speed (10 – 100 kHz) imaging of chemical species and flow velocities in two and even three dimensions, the computational requirements for fully evaluating this data and for directly comparing this data with LES simulations is immense.

System Requirements: Optimizing the workflow associated with development of predictive models will enable an entirely new generation of high-fidelity multiscale/multiphysics

simulations that identically match key engine operating conditions and geometries (e.g., gas turbine and reciprocating internal combustion engines) with direct coupling to complementary experiments. This is necessary to understand how to control and optimize advanced engines with significantly improved efficiencies and minimal emissions using a variety of fuels.

- **Code and Tools:** The main solver to be used is the RAPTOR code framework developed by Oefelein. This is a Computational Fluid Dynamics (CFD) solver designed for application of LES to a wide variety of turbulent combustion problems. The theoretical framework solves the fully coupled conservation equations of mass, momentum, total-energy, and species for a chemically reacting flow. It is designed to handle high-Reynolds-number, high-pressure, real-gas and/or liquid conditions over a wide Mach operating range, including liquid fuel injection and sprays. It accounts for detailed thermodynamics and transport processes at the molecular level, and is sophisticated in its ability to handle a generalized sub-filter model framework in both the Eulerian and Lagrangian frames. RAPTOR has been ported to the full hierarchy of DOE computer systems and has been part of a variety of INCITE and ALCC grants over the past decade. It was recently selected as one of 13 partnership projects for the Center for Accelerated Application Readiness (CAAR) in collaboration with the Oak Ridge Leadership Computing Facility. Application of RAPTOR spans across both the SC-BES and EERE-VT programs with the objective of enabling predictive simulations of combustion for transportation, propulsion, and power systems.
- **Models and Algorithms:** The algorithmic framework in RAPTOR has been modernized to migrate it from a pure MPI model, to a hybrid MPI + OpenMP + OpenACC model. Refactoring the code to many-core systems with accelerators involves optimizing FLOP intensive kernels while simultaneously modifying baseline data structures in the code so that they perform optimally on heterogeneous nodes. Compute intensive kernels are being ported using modern programming approaches such as Kokkos, a portable C++ programming model.
- **End-to-End Requirements:** The dramatic increase in the amount of data generated in future simulations and experiments will necessitate a relative decrease in the amount of I/O possible. Currently, the size of respective datasets from DNS and LES is doubling every year. Similarly, recent advances in high-speed multi-dimensional imaging capabilities are transforming experimental studies of the complex dynamics and three-dimensional structure of turbulence-flame interactions using techniques that generate on the order of 100GB of data every second. Data from both simulations and experiments can now take weeks to process on current architectures. This volume of data will necessitate a large movement toward in-situ data processing and visualization at runtime. The role of scientific data management middleware will be key in providing the software infrastructure that exploits exascale architectures ability to overlap data analysis with I/O.

Related Research: The need to develop predictive models for combustion is a well recognized need across many industries and government agencies (e.g., DOE, DOD, NASA). It is well aligned with DOE missions in both basic and applied research, with broad impact in the general development and application of CFD algorithms, I/O, and data reduction and analysis tools.

10-Year Problem Target: A progression of sub-targets over a 10 year span will lead to end-to-end simulations of both gas turbine and reciprocating internal combustion engines with measured uncertainty quantification and two order of magnitude decrease in the time to solution.

Title/Name of Your Application: Reactive Molecular Dynamics Simulations of Initiation and Deflagration-to-Detonation Transition in Shocked Energetic Materials

Name, Institutional affiliation, and e-mail address: Aidan Thompson, Multiscale Science, Org. 1444, Sandia National Laboratories, athomps@sandia.gov

Overview Description and Impact:

Understanding the physics and chemistry of detonation in energetic materials is important to Department of Energy (DOE) and DOE's National Nuclear Security Administration's missions. Nevertheless, much of the knowledge about shock-induced initiation of energetic materials is empirical in nature, based on macroscopic observations. The spatial and temporal scales (nanometers and femto- to nanoseconds) are such that it has been difficult or impossible to resolve fundamental phenomena in the buildup from incident shock pressures to steady-state detonation. Modeling and simulation have played a critical role in improving our understanding of these phenomena. On the very low end of the scale, atomic/molecular heterogeneities can be studied using quantum mechanical (QM) methods that are parameter free and can be truly predictive, but in a very limited fashion. Continuum models for energetic materials design, largely parameterized empirically with best-guess assumptions about what physical processes must be explicitly captured or potentially ignored, are well suited for treating device-scale phenomena, but attempting to describe grain-scale phenomena is precluded by computational mesh issues and also by the breakdown of the continuum assumption itself.

Large-scale reactive molecular dynamics (MD) using the LAMMPS code bridges these two scales, enabling reactive MD simulations of shock-induced initiation with a realistic description of chemistry validated against QM methods. Thus far, these ReaxFF simulations have been limited to several million atoms. Thompson and co-workers have improved the speed, functionality, robustness, real-time diagnostics, and post-process analysis of LAMMPS reactive MD simulations, so that validated models with up to 10 million atoms can be run on petascale computers, such as the Sequoia and Chama ASC platforms. This has provided unprecedented insight into the effect of small isolated heterogeneities on the shock initiation threshold (Figure 1)

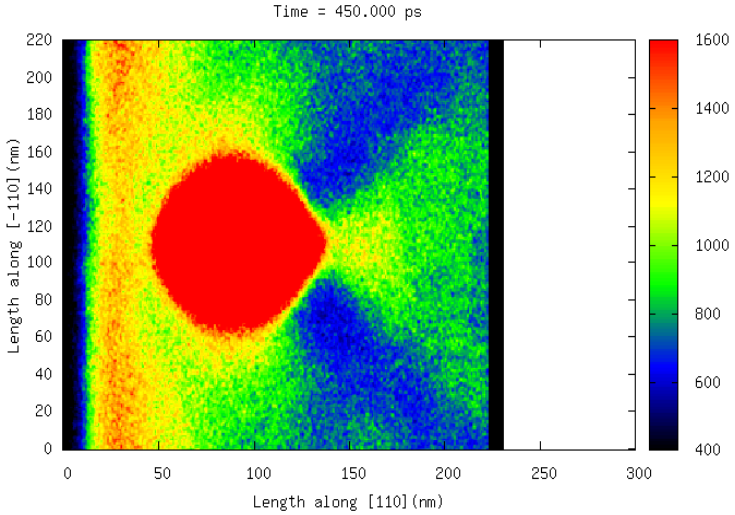

Figure 1 Temperature field around a collapsed void in PETN from a 10 million atom reactive MD simulation.

With 100x increase in computer resources, we will be able to apply the same methodology to proportionately larger systems, with atom counts in the range of 10^8-10^9. This will allow us to simulate energetic material samples containing 100's of heterogeneities, instead of just one. This will enable the direct observation of emergent collective behavior on the grain-scale, including initiation and the deflagration-to-detonation transition. By growing the atom count in proportion to the computer resources, we will achieve roughly the same throughput as today. However, the physically relevant timescales will also increase with system size, requiring a larger number of timesteps and a longer time to solution. This can be offset to some extent by further increasing computer resources.

System Requirements:

- **Code and Tools**: Our reactive MD simulations use the ReaxFF interatomic potential, as implemented in LAMMPS. Depending on density, chemistry, and other factors, the raw cost of this potential is about 100,000 Flops/atom/timestep. Due to memory access, non-vectorization, MPI communication, and other losses, actual performance is about 10^-3 core-secs/atom/timestep on current platforms e.g. 10 million atoms on 16k nodes of Sequoia. We recently switched to an MPI+OpenMP implementation on Sequoia, which significantly reduced memory requirements and provided modest improvements in speed.
- **Models and Algorithms:** Making full use of advanced processors such as Intel phi will require extensive changes to memory access patterns. This can be achieved in a performance-portable manner using a templated library such as Kokkos.
- **End-to-End Requirements:** Post-processing and off-line visualization is currently an essential and time-consuming step for extracting knowledge from these simulations. With 100x larger datasets, post-processing will no longer be viable. Visualization and post-processing will have to be performed in-line during the simulation.

Related Research: Looking beyond energetic materials, there are a wide range of materials science applications that require atomistic simulations of coupled chemistry and mechanical processes. Examples include: stress corrosion, thin-film deposition, glass-metal interfaces, ceramic sintering, and combustion. The availability of reactive MD simulations on exascale resources would have a major impact on these fields.

10-Year Problem Target: By 2025, we expect that billion-atom reactive MD simulations will be routine. Coarse-grid continuum fields of temperature and pressure will be generated on the fly, allowing systematic investigation of key parameters, in a manner that is currently only possible using phenomenological continuum models.

Other Considerations/Issues: A key element of this work will be direct comparison of outputs with experiment and continuum models.

Modeling Turbulent, Reacting, Hypersonic Flow on Realistic Flight Vehicles

Ross Wagnild, Sandia National Laboratories, rmwagni@sandia.gov.
Michail Gallis, Sandia National Laboratories, magalli@sandia.gov.
Steve Plimpton, Sandia National Laboratories, sjplimp@sandia.gov.

Overview Description and Impact:

Accurately simulating hypersonic flow poses a difficult, but very relevant problem for several different organizations across government, academia, and industry. The high temperatures and high fluid velocities stretch transport and internal energy distribution models that describe the gas beyond what is currently capable for laboratory measurement. Many of the models used to describe the transport phenomena and internal energy distribution were empirically derived 60 years ago, many of them based on low temperature combustion data. Further improvement on such models has been limited due the computation scale required to simulate the physical phenomena with sufficient fidelity. Many questions and uncertainty remain for the effect of non-equilibrium on gas-phase chemical reactions, turbulence-chemistry interaction, and gas-phase relaxation.

In order to better understand the gas processes in hypersonic flow, a molecular –level approach is necessary. While it is still not feasible to perform a molecular dynamics simulation at the necessary length scales, a direct simulation Monte-Carlo (DSMC) method has the ability to model the non-equilibrium effects of the gas. The use of an exascale system would allow DSMC to simulate the fully turbulent flow at scales seen in hypersonic flight with sufficient fidelity to capture the necessary physical and chemical phenomena. The data extracted from these simulations would be compared to the latest relevant experimental data and then applied to derive continuum-level models for direct numerical simulations (DNS) of realistic flight geometry in a realistic hypersonic flow. This effort would begin to address the following fundamental science questions about:

- Energy selectivity in chemical reactions and energy specificity in reaction products
- Effects of turbulent mixing on the progress of chemical reactions in hypersonic flows
- Effects of gas temperature and pressure and solid material on the rate at which the surface ablates in a non-equilibrium state.

The answers to such questions would have a large impact on the scientific community as there are currently several groups around the nation analyzing small aspects of each of these questions, including NASA, DOD, NNSA, and several universities.

System Requirements:

There several limitations of existing HPC systems that prevent the full realization of the goals of this proposal. The first is computational power. Using current petaflop HPC computers, DSMC has begun to simulate fluids in the turbulent regime. However, to achieve the level of simulation needed to model realistic, hypersonic, turbulent flow would need an excess of 1 exaflop. A similar limitation exists for the DNS capability. Reacting turbulent flow simulations require many chemical species and minute grid scales, requiring tens of equations being solved on trillions of grid cells to achieve the necessary accuracy.

Once a simulation of the necessary size is possible, the data stream from such a simulation would be enormous. Currently, SPARTA generated output on Sequoia contain multiple terabytes of data and take 4 days to transfer from Sequoia to a local processing computer. Exascale simulations are expected to generate several orders of magnitude more data. Thus, future simulations would need to intelligently reduce the data on the fly using data

transforms such as dynamic modal decomposition and machine learning to comprehend the enormous volumes of data and distill important aspects for the code operators to understand.

- **Code and Tools**:
  - The DSMC code intended for use on this project is called SPARTA. SPARTA can simulate gas flows with a very high degree of accuracy, scaling extremely well on the current HPC architecture. SPARTA can currently utilize the entire Sequoia computer, simulating a maximum of one trillion cells and 3-5 trillion particles. A weak scaling study indicates that advancing from one to 1.57 million cores 10% of peak performance is lost. The code has in-situ visualization that can overcome the limitation of storing the projected petabyte file sizes required for the simulation data. Necessary updates for the code to run on future architectures include reimplementation of kernels for GPU, Phi, and other nodal hardware an exaflop system may utilize. The code is supported and developed by the ASC program.
  - The DNS code intended for use on this project is called SPARC. The code is currently in development stages. SPARC is based on the finite volume approach to solve the Navier-Stokes equations that describe continuum fluid motion. DNS type flux evaluation methods and reacting flow models are in the development plan. The code currently runs on hundreds of processor cores, however, no scale testing has been performed to determine what the parallel computation limitations are. The code is supported and developed by the ASC program.
- **Models and Algorithms:**
  - In order to process the data generated by exascale simulations, machine learning strategies will be employed. The algorithms employed in machine learning strategies have shown promise in processing large amounts of data to determine underlying pattern and generate models to describe the data. Models will be derived to describe the DSMC-generated data for use in the DNS code.
- **End-to-End Requirements:**
  - Methods for generating, storing, and reading the surface and gas-phase grids on realistic flight vehicles will need to be developed.

Related Research:

Development of an exascale DSMC and DNS capability will inform lower-fidelity, lower-cost simulation codes and allow them to better capture the relevant physics seen in hypersonic flow. Currently, the general practice for continuum codes is the use turbulence models that have known inaccuracies and are unable to capture turbulent heat flux and shear forces to a vehicle wall as well as boundary layer separation.

Additionally, the chemical reaction models used in continuum codes are based on equilibrium data at low for hypersonic entry temperatures. Models derived from exascale computations would allow for a more accurate calculation of non-equilibrium chemical reactions and internal energy relaxation and redefine the field of gas-phase and gas-surface chemical interactions.

10-Year Problem Target:

Exascale DSMC simulations are expected to contain several trillion cells and model many trillions of particles on relevant problem sets to determine continuum models for DNS. The Exascale DNS simulations are expected to contain trillions of cells.

Other Considerations/Issues: None

Title: Extreme Scale Infrasound Inversion and Prediction for Weather Characterization and Acute Event Detection

Name, Institutional affiliation, and email address: B. van Bloemen Waanders (bartv@sandia.gov) and C. Ober (ccober@sandia.gov) (Sandia National Laboratories)[1]

Overview Description and Impact: Accurate and timely weather predictions are critical to many aspects of society with a profound impact on our economy, general well-being, and national security. In particular, our ability to forecast severe weather systems is necessary to avoid injuries and fatalities, but also important to minimize infrastructure damage and maximize mitigation strategies. The weather community has developed a range of sophisticated numerical models that are executed at various spatial and temporal scales in an attempt to issue global, regional, and local forecasts in pseudo real time. The accuracy however depends on the time period of the forecast, the nonlinearities of the dynamics, and the target spatial resolution. Significant uncertainties plague these predictions including errors in initial conditions, material properties, data, and model approximations. To address these shortcomings, a continuous data collection occurs at an effort level that is even larger than the modeling process. It has been demonstrated that the accuracy of the predictions depends on the quality of the data and is independent to a certain extent on the sophistication of the numerical models. Data assimilation has become one of the more critical steps in the overall weather prediction business and consequently substantial improvements in the quality of the data would have transformational benefits. This paper describes the use of infrasound inversion technology, enabled through exascale computing, that could potentially achieve orders of magnitude improvement in data quality and therefore transform weather predictions with significant impact on many aspects of our society.

Traditional data acquisition consists of spatially and temporally limited measurements from a range of sources, such as satellites, balloons, ground sampling, and flight sensors. This is an extensive process resulting in voluminous data but yet the overall information content is limited because the sampling is spatial and temporally sparse in addition to being heterogenous. For instance, satellites provide averaged quantity of interests over some spatial area at a certain temporal rate. Ground samples and balloon measurements provide highly accurate data at specific point locations but at a slower and mostly inconsistent temporal rate. In the context of the complex dynamics, these data acquisitions provide sparse information and cannot provide continuity that would eliminate the spatial and temporal uncertainty associated with atmospheric material properties. If somehow atmospheric material properties, such as density, temperature, humidity, and wind velocities, could be measured in a continuous spatial and temporal manner, the quality of the calibrated models realized would potentially revolutionize weather forecasting.

For the last decade, geophysical full waveform inversion (FWI) technology has matured to provide quality reconstructions of subsurface material properties by acquiring data from systematic explosion-based experiments and solving a large scale optimization problem in which the difference between data and numerical predictions are minimized. The numerical predictions are generated by sophisticated simulations of the wave propagation physics. This full waveform inversion provides the foundational technology to reconstruct material properties in the atmosphere through similar mechanics. The instantiation of a source generates acoustic waves which transmit and reflect as a result of changes in properties in the atmosphere. By recording the reflected waves through sensors, a large scale inverse problem can be solved that reconciles the differences between observations and numerical predictions. The study of sound in the atmosphere is known as infrasound and typically operates in a low frequency range below 20 Hz. Sources occur through natural events such as volcano eruptions, earthquakes, acute weather systems and through man-made events, such as explosions, rocket launches, aircraft induced

[1]Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000; SAND2015-4756 R

shock waves. The culmination of FWI technology, the study of sound wave in the atmosphere known as infrasound, and the prospect of exascale computers can provide a paradigm shift in data quality for weather models.

System Requirements: We estimate that a single deterministic inversion of atmospheric material properties will require approximately 24 hours of simulation on an exascale computer. This estimate is based on a three dimensional mesh (1E12 cells) to resolve wave propagation around the globe at a frequency of 0.3 Hertz. Since the instantiation of a source will generate traveling waves in both direction, the wave front only need to be simulated half-way around the globe. The resource estimate also assume 10 percent processor efficiency, 100 optimization iterations and 10 forward solutions to calculate adjoint, gradient, globalization metrics per optimization iteration. A global inversion with these specifications would potentially be sufficient to resolve 1 km variations in the atmosphere. More detailed resolution would be potentially beneficial but dictates corresponding increases in computational resources. Furthermore, managing uncertainties in the infrasound inversion problem would also improve the material property reconstruction, but the computational requirements would far exceed even the exascale capability.

Code and Tools: Sandia has invested considerable effort into the development of large scale inversion of engineering and geophysical applications. Existing large scale inversion capabilities for wave propagation can be applied to the infrasound problem today for simplistic atmospheric conditions at coarse resolutions. Our software leverages the Trilinos framework which is a component-based design, offering software flexibility and extensibility. In particular, the software is uniquely positioned to address exascale computing issues.

Models and Algorithms: To realize a complete and accurate capability suitable to address general weather conditions, a moving media capability would need to be incorporated into the wave propagation dynamics and corresponding adjoint calculations. Furthermore, target inversion parameters such as temperature, humidity, and wind velocities, would require development of algorithms to connect to existing inversion parameters (density and wave speed).

End-to-End Requirement: Interfaces will be required to automatically communicate updates of atmospheric properties and quality metrics between infrasound and weather models.

Related Research: In addition to the impact on weather predictions, infrasound information is currently used to help detect nuclear detonations in support of the comprehensive nuclear test-ban treaty. These approaches however use simplistic algorithms and although capable of determining locations, the quality of the source reconstruction is not sufficient to conclusively determine the type of explosion. FWI technologies applied to both seismic and infrasound measurements will drastically improve the quality of the reconstruction by simultaneously inverting for material properties (in the subsurface and atmosphere) and source terms. Although the infrasound inversion in this case can now be limited to a smaller region, the frequency of the source will encompass a much larger range and therefore require finer mesh resolutions, which consequently will increase the computational requirements. In combination with seismic inversion, we estimate a similar order of magnitude computational requirement would be required in comparison to the global infrasound inversion. The FWI-based infrasound technology can also provide accurate characterization of initial conditions for chemical, biological, and nuclear plant disasters where a malfunction causes a catastrophic failure and nuclear, chemical, or biological fallout need to be tracked in real time. Not only could infrasound reconstruct the initial conditions to provide more accurate forecasts (using a transport model) but subsequent infrasound inversions could determine atmospheric density changes and help track the plume, the quality of which is critical for evacuation and mitigation purposes.

10-year Problem Target: The 10 year target would be to achieve a 10 fold improvement in weather predictions spatially and another 10 fold improvement in temporal improvement, solely due to infrasound data acquisition and FWI inversion. Additionally the target would be to significantly reduce overall damage, injuries, and cost from acute events.

Title/Name of Your Application: Next Generation Codes for Plasma Physics: Multi-Scale, Multi-Physics for a Broad Range of Applications

Name, Institutional affiliation, and e-mail address: Thomas A. Gardiner, Sandia National Laboratories, tagardi@sandia.gov

Overview Description and Impact: Computational physics traditionally approaches a problem by considering the characteristic (length, time, energy)-scales, constructing a system of equations at that scale, and discretizing that system for solution on a computer. This means that solving problems that span multiple scales, e.g. kinetic to continuum, are very complex and often require simplifications to be made such as breaking the problem into different spatial / temporal domains and using the solution from one code to drive the solution in another, or introducing "knobs" that modify the solution in some regions to obtain the right behavior for the wrong reason. While this has been a necessary evil for many years, exascale computing could break the shackles of this approach by enabling a consistent, simultaneous solution across multiple scales. Such a breakthrough would enable simulations of key phenomena in high-energy-density (HED) physics, including inertial confinement fusion (ICF), laboratory astrophysics, and properties of materials under HED conditions. These research areas are of fundamental importance to the Stockpile Stewardship Mission of NNSA as well as the broader science mission of DOE/OFES. Concretely, the new simulation capability will not only accelerate experimental progress on Z, NIF, and Omega, but may well prove to be a necessary step towards ignition in the laboratory, a most important goal.

While there are many physical problems where a multi-scale solution approach would be valuable, the problems we focus on here are magneto-inertial fusion (MIF), dynamic materials, and radiation effects experiments which are three core research programs on the Z-machine at Sandia National Laboratories (SNL). A basic element of all experiments performed on Z is the power flow down a series of magnetically insulated transmission lines (MITLs) to a convolute where multiple parallel transmission lines are combined. The physical description of the electromagnetic waves and plasma in the MITLs and convolute are most appropriately described using a kinetic model such as a particle in cell (PIC) code. Beyond the convolute a single transmission line carries the power flow to the load for which an accurate physical description requires a continuum approximation such as magnetohydrodynamics (MHD) using highly accurate equation of state and electrical conductivity tables generated via quantum molecular dynamics. A challenge for predictive simulations of all experiments on Z lies in spanning the range of scales from the kinetic power flow regions to the continuum MHD regions. MIF experiments that generate conditions sufficient to drive thermonuclear reactions introduce an additional complication that the particles in the high energy tail of the distribution function, responsible for thermonuclear reactions, have mean-free-paths comparable to the size of the fuel implying that a fully coupled MHD / kinetic description is required to fully understand the evolution. The application of a fully coupled MHD / PIC code to this research has wide ranging implications for stockpile stewardship, energy independence, and national security.

System Requirements: Two significant hurdles must be overcome in order to realize fully coupled MHD / kinetic calculations. These are (1) coping with the shear size of the kinetic calculation and (2) developing algorithms that can consistently and seamlessly integrate a coupled MHD / kinetic description of a plasma.

Code and Tools: The relative simplicity, and high degree of parallelism associated with PIC algorithms make them nearly ideal for current and future computer architectures. For example, recent publications on porting electromagnetic PIC codes to GPU architectures achieve an acceleration of a factor of 30 to 100 relative to current CPU implementations. Furthermore, with a judicious choice on particle placement, an explicit integration approach could be utilized, thereby limiting communication to passing particles and fields across domain boundaries with very little to no global communication.

Models and Algorithms: Recent publications have begun to address the question of coupling kinetic (i.e. single particle distribution functions) and two-fluid (electron, ion) 5-moment approximations in the collisionless plasma limit. While our applications require a solution algorithm that spans the range of scales from collisionless to strongly collisional, these studies are quite relevant. These publications stress the importance of a strong coupling of the integration algorithms for the kinetic and two-fluid approximations. They also indicate that the computational cost of the two-fluid solver is negligible compared to the kinetic solver. Incorporating collisions into the kinetic solver is likely to further exacerbate this difference. Hence, constructing a multi-scale MHD / kinetic solver that will run efficiently on future exascale computing architectures depends on accelerating the solution of the kinetic solver and developing algorithms that tightly couple it to the MHD solution with a consistent physical treatment for collisional transport.

End-to-End Requirements: Multi-scale MHD / kinetic simulations running on exascale systems will readily generate 10 – 100 petabytes of data during the course of a calculation. At this scale, incorporating some degree of fault tolerance during data write is likely to become imperative in order to prevent minor errors from corrupting the data. Data analysis will also likely benefit from a hierarchical approach with successively improved analysis results in order to allow the user to analyze the data in real time.

Related Research: Exascale computing holds the promise to deliver fully coupled simulations that span the scales from continuum to kinetic regimes. In the MFE community, a multi-scale MHD / PIC application would assist in studies of the diverter region. In laser driven ICF and MAGLIF (an SNL MIF concept) studies such an application would greatly improve our understanding of laser / plasma interactions and non-local heat transport. In astrophysical sciences this application would improve our understanding of accretion processes around black holes, shed light on long standing questions of the generation of large-scale magnetic fields from small-scale physical processes, as well as enable studies of cosmic ran generation and the feedback on interstellar medium. In the area of solar physics, a multi-scale MHD / PIC code could address the role of kinetic-scale reconnection processes on the large-scale magnetic fields around the sun and Earth's magnetosphere, and with it bring an improved understanding of these processes and solar weather.

10-Year Problem Target: Consider a simulation that spans 100 nanoseconds in time with a zone size of 10 microns and time steps governed by the speed of light. This requires 3 million time steps to complete and each time step must finish in 1/10 second for this calculation to complete in about 3e5 seconds (about 3.6 days). Recent publications on porting electromagnetic PIC codes to GPUs report that the time required in a PIC calculation is about 2-9 ns for each particle push. Using a value of 5ns per particle push, we can expect to push about 20 million particles in 1/10 second on each GPU. Utilizing 1e5 GPUs we can push 2e12 particles which amounts to 100 particles per zone on a 20 billion zone calculation resulting in a well-resolved, 3D multi-scale MHD / PIC simulation.

The next frontier in materials mechanics: science-based, application-relevant, multi-scale materials design with embedded uncertainty quantification

Corbett C. Battaile, Sandia National Laboratories, ccbatta@sandia.gov

Hojun Lim, Sandia National Laboratories, hnlim@sandia.gov

Overview Descriptions and Impact: The properties and performance of most materials originate at length and time scales well below those accessible by any conventional continuum method. The inhomogeneities that exist at these scales play a dominant role in determining properties, and can lead to substantial variability in performance. The details associated with materials microstructures, defects, and chemistry all drive properties, but are homogenized into continuum constitutive treatments when we analyze mechanical behavior using today's technologies. These sub-continuum physics in materials are the main drivers of stochastic behaviors and are the key attributes in uncertainty quantification (UQ). To enable meaningful UQ analysis using models and methods that explicitly treat sub-continuing details like microstructure and chemistry, larger length scale models require concurrent coupling to lower length scale methods, such as density functional theory (DFT), molecular dynamics (MD), and dislocation dynamics (DD). Today's scientific simulation codes are unable to perform studies with the pertinent details to accomplish truly predictive systems performance analysis. Preliminary work has been done to couple mechanics to lower level physics calculations, and UQ has garnered significant interest from a number of scientific fields of study. Nonetheless, a science-based predictive capability for UQ studies with statistically significant populations of multi-scale simulations is not yet within our grasp using today's computational resources. Limitations in today's computer technologies leave us little choice but to analyze materials properties in a phenomenological, empirical, and deterministic fashion. However, next generation exascale platforms provide promise to the scientific community to answer fundamental questions like, *"How can we generally and robustly predict the detailed and complex connections between materials processing, chemistry, structure, and properties? How can we better predict materials' stochastic performance to legitimately enable high-throughput, virtual materials design, fabrication, testing, and qualification?"* This vision for a future capability, based on fundamental science and the statistics that underlie materials behavior, will benefit society by shifting the paradigm from expensive and time-consuming trial-and-error approaches, to a fully predictive and science-based philosophy. Furthermore, these fundamentally new capabilities in computational UQ will accelerate materials innovation, materials design, production, application, qualification, and safety in nearly every product imaginable.

System Requirements: Crossing this horizon in materials mechanics will require codes, methods, and platforms that drastically overshadow today's technology. Today's state of the art is to use direct numerical simulations (DNS) incorporating microstructure on simplified geometries, and this approach has proven heroic in magnitude and far too expensive for day-to-day, production purposes. For example, metal plasticity models informed from atomic-scale data or dislocation dynamics to simulate components on relevant scales ($\sim m^{-3}$) require that we account for the behaviors of $\sim 10^{10}$ crystallographic grains and $\sim 10^{12}$ finite elements (FE) in a continuum solid mechanics. This is well beyond the capability of even the most advanced current computing architectures. In fact, the largest FE simulations of this type reported to date considered only $\sim 10^{8}$ elements to resolve only $\sim 10^{-6}$ m^3 of material. Clearly, the chasm between today's capabilities, and those required to enable the vision of science-based materials mechanics predictions at relevant scales, is enormous and can only by crossed through the promise of exascale computing.

- **Code and Tools**: Proposed work requires next generation software platforms (like Albany, a multi-scale, multi-physics mechanics code funded by Office of Science and currently deployed to a wide range of architectures including Titan and Mira), which are able to take advantage of new emerging and paradigms like CUDA and other GPU/many-core architectures
- **Models and Algorithms**: Some of the foundational tools for this purpose already exist, e.g. crystal plasticity methods for microstructure-explicit solid mechanics, and lower-scale methods for physics calculations using DFT, MD and DD. However, there is to date no robust and general infrastructure for coupling these methods, and past successes, while impressive and promising, have primarily been on-off demonstrations using specific codes and methods. The proposed work will require advanced solvers capable of multi-physics-based multi-scale/multi-grid DNS of billions of grains, in order to achieve true component-scale analysis with explicit treatment of sub-continuum effects.
- **End to End Requirements**: The data requirements underlying the proposed scenario would require next generation architectures for data storage and access; and the communication requirements involved in obtaining constitutive behaviors via concurrent calls to physics simulations will demand interconnect technology that does not exist today, and is not in the same class as the sorts of capabilities we envision to simply perform today's calculations "bigger, better, faster." Defining, representing, and storing all the acquired information about structure, chemistry, other necessary state data will require advances in informatics and "big data" technology. All of these end-to-end requirements are compounded by the need to sample different possible realizations and internal states in order to harness the potential of structure- and chemistry-aware paradigms for UQ studies.

Related Research: The proposed effort has direct relevance to various initiatives at the federal level, including Integrated Computational Materials Engineering (ICME)[1] and the Materials Genome Initiative (MGI)[2]. These initiatives are drivers behind a wide range of activities across corporate, laboratory, and academic research. In particular, an approach like this one would be particularly impactful in applications like welding, joining, and additive manufacturing, where the complex and spatially inhomogeneous microstructures and chemistries create a breakdown of scale separation between sub-continuum structures and observable properties. The proposed work will be supported by verification and validation (V&V) via multi-level experiments.

10 year Problem Target: In 10 years, we anticipate that the realization of exascale computing will shift the philosophy of the nation's scientific community about how we predict and analyze materials behavior and properties, away from today's model where application-relevant scales are addressed using only phenomenological, empirical, deterministic methods, and toward a paradigm that explicitly incorporates sub-continuum details and physics with quantification of the inherent uncertainty operating at the relevant length and time scales.

Other considerations/issues: This work requires multi-disciplinary/ multi-scale approaches among the fields of materials, solid mechanics, physics, computer science, and statistics.

[1] Integrated Computational Materials Engineering: A Transformational Discipline for Improved Competitiveness and National Security," a report by the National Research Council, 2008.
[2] "Materials Genome Initiative for Global Competitiveness," a report by the US Office of Science and Technology Policy, 2011.

Title/Name of Your Application: Planetary defense and asteroid deflection technology

Name, Institutional affiliation, and e-mail address: Mark Boslough, Sandia National Laboratories, mbboslo@sandia.gov

Overview Description and Impact: Asteroid deflection is a key element of both planetary defense and NASA's Asteroid Redirect Mission (ARM) to move an asteroid into a stable orbit around the Earth. Both have clear national security implications, although that is not the stated purpose of ARM. Exascale computing can contribute in two different ways: (1) shock physics simulations and (2) optimization.

Planetary defense refers to mitigation of the impact risk from near-earth objects (NEOs). Mitigation includes active schemes to prevent an impact, such as deflection or disruption and dispersion, which have the potential to substantially lower the risk and potential consequences. The DOE labs have ongoing planetary defense programs, and work is accelerating because of the recent signing of an interagency agreement between NNSA and NASA. One component of the planetary defense research has been focused on quantitative assessment of risk, which is characterized as a low-probability, high-consequence risk. The main thrust of current projects is the modeling active mitigation scenarios which involve the hypervelocity collision by a kinetic impactor (KI), or the explosive shock loading due to prompt neutron flux or direct coupling from a nuclear explosive device (NED).

We expect another area of research to emerge as the rate of asteroid discovery continues to accelerate. There are currently about 10,000 earth-crossing asteroids in the catalogue, but with new capabilities coming online this could increase to millions, or even tens of millions, in the next decade. Each of these objects represents a possible impact threat, but also an exploitable resource if it can be nudged into a useful and accessible orbit. There are also national security implications associated with this

We anticipate that a 100-fold increase in computational performance to the exascale range could begin to address the following relevant questions that are far beyond current computing resources:

1) What is the precise mechanism by which an asteroid breaks up, explodes in the atmosphere, and couples energy to the ground? This question requires solving the governing equations over many orders of magnitude, from the 10-100 km of flight-path length through the atmosphere to the 10-100 m scale of the asteroid itself, to the mm and small grain size and fragment size of the exploding object.
2) How can momentum be transferred in a predictable way to an asteroid from a KI or NED? Momentum transfer requires a detailed knowledge of the structure of an asteroid and the ability to model shock and radiation interactions across many spatial scales.
3) In a large (>million) population of NEOs, can a globally optimal solution be found for which can be captured or placed in a desired orbit (e.g. captured into Earth orbit for a hardened instrument platform, or transfer flyby orbit to Mars for radiation shielding of astronauts)? This would require the forward integration millions of orbits with potentially trillions of different perturbations into chaotic paths that result in the desired placement.

System Requirements: Current limitations for continuum modeling in support of risk assessment and active mitigation schemes are the inability to run with sufficient resolution over the full relevant domain of the problem. For example, simulating the first ten seconds of the 2013 asteroid entry over Chelyabinsk, Russia, requires a multi-day run on thousands of processors, and is still under-resolved. Increasing resolution from meters down to millimeters in three dimensions would allow physics at all

relevant scales to be modeled, but would require 12 orders of magnitude increase in performance for a naïve decomposition scheme. With modeling, this performance requirement could be relaxed, but this problem will still be resolution-limited up to and beyond exascale.

Chaotic orbital optimization is an almost entirely unexplored research field. Current efforts have primarily relied on trial-and-error, coupled with human intuition, to define low energy orbits to pre-selected asteroid targets. Evolutionary search methods would likely be the best way to automate the process, but scaling up to millions of asteroids in a multi-dimensional parameter space is severely limited by computational performance.

- **Code and Tools**: The dynamic continuum problems (airbursts and deflection technology) would make continued use of the Eulerian shock physics code, CTH, but at higher spatial resolution. Multiscale and multi-physics issues can be addressed using specialized capabilities of ALEGRA, an arbitrary Lagrangian-Eulerian code that provides flexibility, accuracy and reduced numerical dissipation over pure Eulerian code treatments, as well as high-deformation solid dynamics codes (e.g., Sierra/SM). ALEGRA is especially well suited for large distortions, strong shock propagation, and high energy deposition environments associated with nuclear or kinetic impact deflection methods. For the early phase of atmospheric entry, non-continuum codes using DSMC methods would be used. For the orbital optimization, the DAKOTA suite of tools would be applied, as well as open-source and Sandia-developed evolutionary algorithms.
- **Models and Algorithms:** Because of the multiscale and multiphysics nature of the problem, which spans a range of states from extreme high pressure and high density to near vacuum, specialized equations of state would need to be developed. New multiscale fracture and fragmentation algorithms may need to be developed to accurately model the details of ejecta formation, which dominates momentum coupling.
- **End-to-End Requirements:** The orbital optimization problem will require the capability to download orbital elements for newly discovered asteroids at a rate of thousands to tens of thousands per day.

Related Research: There are many other shock and impact physics problems that could exploit resolution increases of many orders of magnitude. These problems include explosive initiation, penetration mechanics, armor/anti-armor, space debris shielding, fragmentation mechanics, nuclear weapons effects, and impact cratering. Potential limitations on proposed research include material models (equations of state, strength, etc.) over the range of states of interest.

10-Year Problem Target: For continuum modeling: the ability to achieve tens of seconds per day on a problem in a tens of km domain down to mm scales over some part of the domain, using adaptive mesh refinement. For orbital optimization, the ability to exhaustively search the perturbation space for and identify low-energy chaotic capture or transfer orbits for up to ten thousand orbits per day.

Title/Name of Your Application: Fracture dynamics in polymeric adhesives

Name, Institutional affiliation, and e-mail address: Mark Stevens, Org. 1814, Sandia National Laboratories, msteve@sandia.gov

Overview Description and Impact:

Failure of polymeric adhesives is a long standing issue in materials science important to both Office of Science interests and the National Nuclear Security Administration's missions. While we have a lot of knowledge based on experience, fundamental knowledge especially of the fracture dynamics near an interface is severely lacking. We do not have a criterion for when fracture will initiate in a polymeric system. Moreover, connecting the molecular structure and interactions to the fracture behavior simply cannot be done presently. This greatly limits the design of materials and the prediction of failure. Experimentally probing the interfacial region is difficult and has been a barrier for progress. Simulations can treat the interfacial region and provide the connection among molecular structure, interactions and the deformations that yield fracture, but treating the large length and time scales are a major challenge. With exascale computing we can advance the range of systems simulated and recent work at Sandia suggests there is the opportunity to make ground breaking progress into this problem. Recent simulations at Sandia have found that the failure strain decreasing with system size for highly crosslinked polymer adhesives (i.e. common structural adhesive such as epoxies). This decrease cannot continue beyond systems 2-4 times larger as the failure strain will go negative. Thus, there must be a critical length scale on the order of 100 nm, where the change in behavior occurs. As most adhesives are thicker than this scale, the corresponding structure must significantly control the deformation and fracture behavior. In metals, we know that grains introduce such a length scale, but in amorphous polymers, there is no clear structure that has been identified. With faster computers, we could understand an important physical regime in polymer adhesives and provide practical information in their design.

There are multiple aspects that require substantial increase in computational resources. The primary source requiring faster computations is the time scale. For example, a simulation that does a strain of 100 nm in a time of 100 ns is moving the top surface at a speed of 1 m/s, but most experiments are of order 1 μ/s. Presently, in order to deal with this time scale limitation, polymer simulations use coarse-grained models that remove most of the chemical detail, but maintain the connectivity of the network structure. Even with these simplified models, pull velocities of only about 1 mm/s are possible. With present resources we are only able to do a handful of the largest simulations in a year and really need to many more simulations/year to obtain the variation in systems. Initial future work, would be to do such system parameter variations only lowering the pull rate enough to determine the trends and not attempt to reach the experimental regime. Longer term there would be a transition to the detailed atomistic models which are more expensive per time step, but

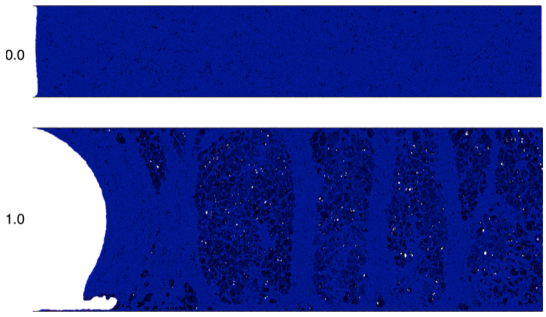

Figure 1 Images of system at 0 and 100% strain in tensile pull simulations. The strained system shows voids (dark regions) and crack initiation in lower left corner.

contain the ultimate information we seek.

System Requirements: Presently, we are using 1000s of nodes, but a single system takes about 30 days to simulate. Standard parallelization already implemented will enable treating larger systems, but we also need to reduce the total CPU time.

- **Code and Tools**: The molecular dynamics code is LAMMPS. There is some support for LAMMPS in the ASC program. There is an LDRD which ends this FY that is funding research in the polymeric adhesives. Presently the adhesive simulations are not using the new developments for the coprocessors (GPU or Intel phi), which are just being developed. The LAMMPS development is using the Kokkos library to accomplish this. Further development of LAMMPS usage of Kokkos will be required as well as development of KOKKOS for the new hardware. In addition, for more advanced polymer models, new features such as bond creation and breaking will be needed.
- **Models and Algorithms:** Model development includes going from the present simple coarse-grained models of the polymers to more advanced models and finally to atomistic models. There are various methods already developed to do this work, but they have not been applied to these systems.
- **End-to-End Requirements:** Most analysis to date has been handled in post-processing. At the present system size we are reaching the limits of post-processing on local workstations and will probably have to start using HPC resources to perform these calculations or to include the calculations within the LAMMPS simulations. The concern in the past has been that such calculations could slow down the main simulation. Recently LAMMPS has implemented the ability to process the configuration dumps, which bring the full parallelization to these calculations without affecting the main simulation. Visualization is an important part of analysis and the system sizes are also an issue, but visualization is best done at the desktop level. In this case we will have to implement means of visualizing parts of the system or coarse-graining the data.

Related Research: Fracture of materials is a fundamental problem and many aspects of this research on polymeric systems is similar and sometimes identical, to fracture in other materials. As such new understanding can have broad and profound effects. In addition, the problem inherently involves interfaces, which is another broad challenge, albeit not the focus for coarse-grained simulations. As we transition to treated atomistic models, the atomistic details of the interface naturally become essential to treat.

10-Year Problem Target: In a 10-year program, the main effort would be to extend the model from coarse-grained to atomistic. The extra CPU time per MD time step in the fully atomistic model is substantial. In addition, one would like to have newer force-fields that explicitly treat bond breaking, which are absent from polymer force-fields , although present in related molecules. In any case, such force-fields are even more expensive and thus require further algorithm and hardware developments.

Other Considerations/Issues: Ideally, we would like experimental data to compare to directly, but experimentally probing the interfacial regime is limited. Hopefully new simulation data would shed light on what needs to be measured and indicate a means to perform such experiments.

Title/Name of Your Application: Large-Scale Quantum-Accurate Atomistic Simulations of Materials using Exascale Computing Resources

Name, Institutional affiliation, and e-mail address: Aidan Thompson, Multiscale Science, Org. 1444, Sandia National Laboratories, athomps@sandia.gov

Overview Description and Impact:

Classical molecular dynamics simulation (MD) is a powerful approach for describing the mechanical, chemical, and thermodynamic behavior of solid and fluid materials in a rigorous manner. The material is modeled as a large collection of point masses (atoms) whose motion is tracked by integrating the classical equations of motion to obtain the positions and velocities of the atoms at a large number of timesteps. The forces on the atoms are specified by an interatomic potential that defines the potential energy of the system as a function of the atom positions. Typical interatomic potentials are computationally inexpensive and capture the basic physics of electron-mediated atomic interactions of important classes of materials, such as molecular liquids and crystalline metals. Efficient MD codes running on commodity compute resources are commonly used to simulate systems with N = 10^5 - 10^6 atoms, the scale at which many interesting physical and chemical phenomena emerge. Quantum molecular dynamics (QMD) is a much more computationally intensive method for solving a similar physics problem. Instead of assuming a fixed interatomic potential, the forces on atoms are obtained by explicitly solving the quantum electronic structure of the valence electrons at each timestep. Because MD potentials are short-ranged, the computational complexity of MD generally scales as $O(N)$, whereas QMD calculations require global self-consistent convergence of the electronic structure, whose computational cost is $O(N^k)$, where $2 < k < 3$ and N is now the number of valence electrons. For the same reasons, MD is amenable to spatial decomposition on parallel computers, while QMD calculations allow only limited parallelism. As a result, while high accuracy QMD simulations have supplanted MD in the range N = 10-100 atoms, QMD is still intractable for $N > 1000$, even using the largest computing resources. Conversely, typical MD potentials often exhibit behavior that is inconsistent with QMD simulations. This has led to great interest in the development of MD potentials that match the QMD results for small systems, but can still be scaled to the interesting regime N = 10^5 - 10^6 atoms. These quantum-accurate potentials require many more floating-point operations per atom compared to conventional potentials, but they are still short-ranged. So the computational cost remains $O(N)$, but with a larger algorithm pre-factor. It has already been demonstrated that prototypes of these quantum-accurate potentials can be run efficiently on petascale resources. However, many practical challenges must still be overcome, mostly related to the difficulty of achieving quantum-accurate forces for a wide range of local atomic environments.

The availability of near-exascale resources would enable several radically innovative approaches to overcome these challenges:

1. During the potential fitting process, large sets of quantum calculations must be performed targeting a wide range of local atomic environments. Using large computer resources, many independent quantum calculations can be performed simultaneously. These calculations need to be supervised in an automated and adaptive manner using ideas from machine learning and uncertainty quantification to systematically improve the quality of the potential.

2. Uncertainty in the MD simulation outputs due to uncertainty in the interatomic potential can be quantified by running ensembles of simulations with different

interatomic potentials. These are drawn from a distribution generated as part of the fitting process.

3. On-the-fly estimates of accuracy during MD simulations can be used to trigger additional improvements in the interatomic potential, essentially combining strategies from the first two items above. This requires pausing the MD simulation to perform new quantum calculations, refitting the potential, and then continuing the MD run.

4. A database of quantum calculations can be used to eliminate the interatomic potential entirely. MD forces can be interpolated directly from the existing database. As new local atomic environments emerge during the MD simulation (e.g. due to melting of a crystal), new quantum calculations can be added to the database.

All of these approaches require augmenting large-scale MD with other types of calculations, including quantum electronic structure calculations, optimization, machine learning, and database operations. Sophisticated algorithms and software will be required to orchestrate the interplay between these different components, ensuring that each is allocated adequate compute resources.

The ability to run quantum-accurate MD simulations of arbitrary materials for millions of timesteps with 10^5-10^6 atoms will give access to a huge body of science and technology knowledge that is currently beyond our reach. For example, the performance characteristics of optical and electronic devices are strongly affected by the presence of coordination defects and other structural motifs in thin amorphous oxide layers such as SiO2. Neither MD nor QMD is capable of simulating the growth and annealing processes that control these structures. Quantum-accurate MD would enable us to directly simulate these processes, allowing us to examine how process variables such as oxygen pressure and temperature can be used to achieve specific optical and electronic properties.

System Requirements:

- **Code and Tools**: The underlying codes for MD (LAMMPS) and QMD (NWChem, SeqQuest) are already well-developed for petascale platforms. The new approaches will require weak coupling of many sub-petascale computations.
- **Models and Algorithms:** Effective implementation of these approaches will require overcoming challenges related to the heterogeneous nature of the computations. Robust, adaptive algorithms will be required to manage many quantum calculations and use them to drive large-scale quantum-accurate MD simulations.
- **End-to-End Requirements:** A key aspect of these approaches will be the use of large dynamic databases.

Related Research: The impact of these approaches will extend into every area of material science, condensed-matter physics and chemistry where atomistic simulation is currently being used.

10-Year Problem Target: In this time frame, these approaches will generate petabyte-scale databases of quantum calculations for specific chemical systems, such as Si/O/H, enabling quantum-accurate MD simulations to be run using moderate computing resources e.g. 1 ns/day with 10^6 atoms on a petascale computer.

Other Considerations/Issues: N/A

Bayesian parameter and state estimation for regional-scale land models

J, Ray (POC; jairay@sandia.gov), L. Swiler and G. Hammond — Sandia National Laboratories
M. Huang (maoyi.huang@pnl.gov) and X. Chen, — Pacific Northwest National Laboratory
W. Riley (wjriley@lbl.gov) and G. Bisht, — Lawrence Berkeley National Laboratory

Overview Description and Impact: Land Surface Models (LSMs) such DOE-BER's ACME land model, are run at coarse resolutions (1^0x1^0) for climate change predictions at global scale. Variable resolution LSMs, such as the Community Land Model (CLM), can also be run at *fine resolutions* to assist decision-making at regional and site scales. However, in such a case, they require calibration to be predictive. Limited regional and site scale data, and CLM's structural errors lead to parameter estimates with large uncertainties; consequently, Bayesian calibration has been used to develop parameter estimates as probability density functions. Surrogate models ("curve-fits" of CLM) are used to alleviate the formidable computational cost of Bayesian inference, but they can be constructed only for a few geographic locations. Thus, the approach is not readily extensible to regional scale LSMs.

Exascale platforms hold two promises in this regard. First, their computational power may allow calibration of CLM (i.e., without recourse to surrogates) at regional scales. Further, coupling CLM to a 3D subsurface simulator, e.g., PFLOTRAN, can reduce its structural errors, and simplify calibration by removing non-physical artifacts. We propose to extend our toolset, currently used for parameter and state estimation using CLM and PFLOTRAN at the site level, to exascale platforms. The toolset consist of a scalable, multichain hybrid genetic-algorithm/Markov chain Monte Carlo (MCMC) sampler, called SAChES (used for Bayesian parameter estimation), and Ensemble Kalman Filters (EnKF) which are used in state estimation problems with PFLOTRAN. The toolset will be used to address regional-scale parameter and state estimation (henceforth *calibration*) problems using an integrated CLM+PFLOTRAN composite.

Impact: The effort will yield regional LSMs with enhanced predictive skill. They will improve integrated assessment studies, management of subsurface contaminants and decision-making in light of a changing climate. The enhancement will be achieved via calibration of LSMs to observational data, enabled by our scalable Bayesian calibration framework. The framework could also be reused to estimate land model parameters that affect hydrological, ecosystem, and biogeochemical dynamics. Bayesian calibration can also yield causal models of structural errors that can then be used to improve process representations in LSMs and PFLOTRAN.

System Requirements: The fundamental obstacle in calibrating a regional-scale CLM+PFLOTRAN composite lies in the computational cost of the process. It arises mainly from the large number of times the composite will be invoked in a Bayesian calibration setting. Prior work with CLM surrogates and single-chain MCMC has revealed that $O(10^4)$ *sequential* CLM invocations are required to solve inverse problems of modest dimensionality (e.g., three hydrological parameters). The computational cost (number of invocations) of an inverse/calibration problem increases combinatorially with the number of independent parameters being estimated; inclusion of PFLOTRAN increases the per-invocation cost dramatically. In addition, during calibration, the composite will be forced using reanalysis meteorology (from files), resulting in frequent short seeks and reads and stressing the file system. State estimation with EnKF and PFLOTRAN at the Hanford site has revealed a need for frequent restarts (from checkpointed data), which can also impose sever filesystem loads for large problems. Consequently, while the current codebase could be extended to perform multisite calibration on petascale platforms, a regional scale calibration will require resources beyond what is available today.

Our calibration approach imposes a few preferences for the proposed exascale platform. CLM and PFLOTRAN adopt different domain decomposition strategies and their coupling imposes much exchange of data at their interface. Being able to place CLM and PFLOTRAN instances explicitly to manage and optimize memory access and communication costs will be helpful. Further, the short reads of forcing data would benefit from a memory-based or a staged file system. However, the bulk of the effort will involve code restructuring and a reorganization of the CLM-PFLOTRAN coupler. The details are below.

- **Code and Tools**: The calibration tools and the individual simulators will require extension to be efficient at exascale. The use of the composite inside a calibration loop will require a 10x-100x speed-up in its execution. Exascale platforms allow such enhancements in performance, without fine-grained domain decomposition, by exploiting SIMD operations on the abundant cores. The codes will require extension to MPI+X, where X could be OpenMP, OpenACC or the task-based runtimes currently being explored by ASCR's co-design centers, to exploit heterogeneous cores and deep memory hierarchies expected on future platforms. Parameter estimation, posed as a statistical inverse problem, will be solved using SAChES. It is currently employed for single-site calibration of CLM to observational data. Funded by ASCR, SAChES is implemented using C++/MPI and scales to 128 chains (on Hopper at NERSC), with each chain driving a separate instance of a parallel simulator. In order to enable ensembles with $O(10^3)$ chains, SAChES will require redesign as an ensemble of chains, loosely coupled via MPI-2 remote memory access paradigm. Loose coupling of large chain ensembles also assists resilience to silent errors expected on exascale platforms; corrupted chains can be decoupled and killed, with little statistical impact on the inferred parameter distributions. EnKFs driving PFLOTRAN are currently used to assimilate spatiotemporal observations at the Hanford site and infer spatially distributed states ($O(10^5$-$10^6)$ correlated unknowns) with an $O(10^3)$-sized ensemble. At exascale, we will investigate OpenDA and DART as the basis for scalable state estimation with the CLM+PFLOTRAN composite.

- **Models and Algorithms:** The SAChES algorithm will require a "tempering" modification of its MCMC sampler to efficiently and accurately infer multimodal solutions to statistical inverse problems. The differing domain decompositions of CLM and PFLOTRAN, along with thread-like and task-based programming models being proposed for exascale platforms will likely require a re-design of the coupling mechanism to preserve computational efficiency. EnKF-based state estimation will focus on non-stationary fields, requiring advances in wavelet-based models and sparsity enforcement.

- **End-to-End Requirements:** Petascale capabilities currently available may suffice.

Related Research: The methods and software developed in this effort could be repurposed for other calibration problems where quantified uncertainty is helpful; such scenarios arise in material property and reaction rate estimation in energy storage (batteries, fuel cells etc.), as well as inference of permeability/porosity fields in CO_2 sequestration sites.

10-year Target Problem: We envision Bayesian estimation of $O(10^2)$ independent parameters, using forward models of computational complexity consistent with the 3D simulator PFLOTRAN and CLM composited together. We expect this estimation may be possible using $O(10^3)$ – $O(10^4)$ chains, each driving the forward model on $O(10^4)$ cores. In the state estimation context, the advances will be algorithmic rather than in size. We expect to infer non-stationary fields, using methods that combine sparsity enforcement and filtering, without imposing Gaussian or smoothness assumptions.

Title/Name of Your Application: Exascale Computing for Subsurface Science

Name, Institutional affiliation, and e-mail address: Dr. Susan J. Altman, Sandia National Laboratories, sjaltma@sandia.gov

Overview Description and Impact:
Subsurface energy resources provide more than 80 % of total US energy needs today. Radioactive waste disposal and carbon sequestration capitalize on isolation offered by subsurface media. Discovering and effectively harnessing subsurface resources while mitigating impacts of their development and use are critical pieces of the Nation's energy strategy. The Subsurface Technology and Engineering crosscutting initiative (SubTER) is focused on a fundamental objective – Adaptive Control of Subsurface Fractures and Fluid Flow – common to all subsurface applications. Central to SubTER's objectives are fundamental and necessary scientific advances in our understanding of reactive multi-phase transport in permeable geomaterials, the ability to image flow and heterogeneity within the permeability field, and to safely engineer systems that actively modify, monitor and utilize an evolving permeability field. Three challenges of subsurface modeling will be greatly aided by exascale computing:

Multi-Scale Physics: The multi-scale nature of the subsurface naturally suggests a multiscale algorithmic approach to theory and model development. The effectiveness of present-day models depends on their calibration to a particular site, via expensive laboratory and field tests. Exascale computing could answer the question, for example, of how the physics and chemistry at the molecular/pore scale couple with fine-scale geologic texture and affect measurements at the lab scale where the empirical data lie? From a top down perspective, ultra-fine gridding of continuum geophysical methods should agree with upscaled pore-scale multi-physics.

Coupled Processes: In both geological and engineered subsurface systems *in situ* processes will evolve subsurface properties and state through time. Therefore, multiscale multiphysics models are necessary to compute realistic probabilities of future states and outcomes. Advanced numerical methods and algorithms are needed to simulate fully coupled thermal, hydrological, mechanical, chemical and biological (THMCB) processes with disparate spatial and temporal scales.

Streaming Analysis of Observational Data: Real-time analysis and inversion of large-N sensor networks in electrical resistance tomography, is an example where high-performance computing facilitates improved permeability monitoring and active decision-making. Furthermore, modern full-waveform inversion of broadband seismic data already pushes the limits current petascale computing capabilities. A hybrid joint inversion approach (seismic, electric, electromagnetic, ground-deformation, hydrologic, etc.) at the scale needed by SubTER would require exascale computation and drive additional fundamental understanding of coupled, multi-scale flow processes from molecular dynamics to continental-scale processes.

System Requirements: Integrating multiscale multiphysics models involves scalability challenges for petascale and exascale computing. Progress is being made to utilize GPUs, MPI, and OpenMP tools when appropriate to take advantage of modern computing systems, and at the both ends of the physical modelling scale there is a need for increased computing power to handle larger and more detailed systems.

- **Code and Tools**: Massively parallel simulators like the LAMMPS molecular dynamics simulator and the PFLOTRAN reactive multiphase flow and transport code are making progress in taking advantage of the technologies that will allow exascale computing, but geomechanical models and coupled process models still need to be extended for these new

technologies, including handling realtime data acquisition for larger sensor networks. Nvidia claims that current codes in these areas are showing a six to fifteen times speedup with proper accelerators. Increased scalability will allow improvements to the resolution for field-scale models and increased system sizes and quantum bases at the molecular-scale, but the finite-element methods and quantum methods will need vectorization.

- **Models and Algorithms:** Algorithms and methods to effectively vectorize calculations using a combination of GPU, MPI and OpenMP accelerators will need to be developed and algorithms to distill information for efficient flow from one physical scale to its neighbors will need to be enhanced. New algorithms to handle the multiscale datasets in near realtime will need to be developed. Scalable numerical methods for solving sparse nonlinear systems of partial differential equations (e.g. multiphase flow, energy, geomechanics) implicitly in time at the petascale and beyond still need to be developed.
- **End-to-End Requirements:** As models become more detailed, either through increasing system size (molecular scale) or refining the mesh (field scale), increased data processing and storage requirements will be needed to handle both input and outputs.

Related Research: Exacscale computing can be used to simulate coupled THMCB processes over very long time scales and very large domains required for long-term assessment of nuclear waste and carbon sequestration security. Near-field chemical and thermal processes (i.e., at injection point or at waste canisters) control time steps of possibly very large (i.e., sedimentary basin) domains across simulations running millions of years.

Deep disposal boreholes for radioactive waste in crystalline rock (5 km) will demand coupled geophysical inversion of multiple surface geophysics (seismic, gravity, self-potential and electromagnetic), predicting thermohaline convection and reactive transport through discrete fracture networks at continential physical scales and geologic time scales.

10-Year Problem Target: Numerical modeling of physical systems is trending towards increasingly physically realistic and coupled processes. Large-scale systems are moving towards more direct upscaling of properties from smaller-scale physics-based models, rather than utilizing geostatistical parameterization to quantify uncertainty. The sources of uncertainty are then moved to either the upscaling process or the smaller-scale physics. Uncertainty analysis is also moving towards more direct Bayesian MCMC simulation to resolve non-linear uncertainty relationships, rather than relying on linearizing relationships derived from frequentist statistical approaches. Many millions of individual forward model realizations may be needed to quantify uncertainty, with each realization being a coupled multiscale multiphysics model. As model coupling becomes tighter and model complexity increases the previous linearization approaches used to understand uncertainty will be inadequate to capture real uncertainty inherent in advanced modeling.

For nuclear waste disposal, we envision modeling performance assessment for an complete nuclear waste repository with nested integration of complex process models and refined discretizations at fine time scale in the near-field (near the actual repository waste packages) and coarser discretization and reduced-order representations of processes at the far field (near receptors). Current petascale capabilities can model a subset of this problem; however, exascale computing with complementary advances in numerical methods will enable simulation of a full repository with increasing mechanistic process models embedded in the near field. These higher-fidelity models will better position DOE in the determination of a suitable repository for nuclear waste disposal within the US.

Bayesian calibration and machine learning of turbulence models

Team: S. Arunajatesan, L. Dechant, H. Kolla, S. Lefantzi, J. Ling and J. Ray, Sandia National Laboratories.
Point of contact: J. Ray, jairay@sandia.gov

Overview Description and Impact: Reynolds-Averaged Navier-Stokes (RANS) simulations are routinely used in energy-related engineering design applications. They rely on empirical submodels that require calibration to be predictive. Recent work has shown that Bayesian calibration of RANS parameters, using experimental data from a high Reynolds number jet in crossflow (JinC) interaction, can significantly improve prediction accuracy. The formidable cost of Bayesian calibration was alleviated by the use of statistical surrogates ("curve-fits") of the simulator. However, this approach faces two obstacles when generalizing to other flow regimes. First, surrogates can be constructed for only a few cases; this brittleness makes their use unsuitable in a design environment. Second, the parameters were considered to be constant over the entire flow field; in reality, they may be flow regime dependent and vary over the flowfield.

The 100-fold increase in computing power promised by exascale platforms can be used to overcome both the obstacles. First, it will become feasible to perform Bayesian calibration with the RANS simulator itself "inline", instead of relying on surrogates. It will yield parameters as probability density functions, thus capturing estimation uncertainty. Second, it will become possible to create large training sets of RANS simulations and use machine learning (ML) strategies to develop flow-regime dependent parameter calibrations. In addition, robust Bayesian calibrations could be accomplished by using topological flow features in formulating the likelihood ("objective function") rather than the point-wise measurements currently used. We propose to develop these scalable Bayesian calibration and ML techniques and demonstrate them on turbulence model calibration in high Reynolds number JinC interactions. It is an idealization of fuel-air mixing, film cooling of combustors and turbine blades, interaction of volcanic plumes with wind and a host of other engineering and natural flows.

Impact: Calibration to experimental data can improve the predictive skill of RANS models in the flow regime of interest, improve design and optimization studies and promote a simulation-driven design process. This will be enabled by our scalable Bayesian calibration and ML capability. Calibration is particularly important in high Reynolds number regimes where Direct Numerical Simulations cannot be used to cover the operational envelope. Further, our method segregates the model improvement process, conducted on exascale and big-data analytics platforms, to ease its incorporation into established design workflows.

System Requirements: The primary obstacles in accomplishing our scientific goals is the lack of sufficient computing power and a data platform to hold an ensemble of turbulence simulations, for interrogation by big data analytics to learn turbulence model enrichments. Our current (modest) Bayesian calibration effort (3 parameters), performed with surrogates, required $O(10^7)$ CPU-hours on Sequoia (LLNL); performed *without* surrogates, it would require $O(10^{10})$ CPU-hours. Since the cost of inverse problems rise combinatorially with dimensions (i.e., number of parameters estimated), a practical calibration effort on existing petascale platforms is simply infeasible.

Our proposed approach imposes a set of preferences on future exascale platforms. Our RANS simulator can significantly benefit from SIMD parallelism and the efficient use of accelerators to speed up critical code sections. Data staging capabilities, such as burst buffers, will be helpful for in-situ analysis, when generating training data for ML of turbulence models on the exascale machine. However, the bulk of the

effort will involve restructuring of our codes and algorithmic changes to better exploit exascale architectures. These are described below.

- **Code and Tools:** The tools for this effort will be drawn from ongoing projects. Our RANS simulator, SIGMA CFD (**S**andia **I**mplicit **G**eneralized **M**ult-Block **A**nalysis **C**ode for **F**luid **D**ynamics) is implemented in Fortran/MPI, and routinely used in ASC-supported, full-system simulations on $O(10^5)$ CPUs on Sequoia. It is also used to generate training data for surrogates. It will be replaced by SPARC, a hybrid multi-element structured/unstructured CFD code written explicitly for exascale applications. A suitable MPI+X programming model will be implemented to extract optimal performance on a heterogeneous architecture, where X could be OMP, OpenACC, Kokkos or a task-based runtime parallelization. Our multichain Bayesian inverse problem solver, SAChES, a hybrid genetic-algorithm and Markov chain Monte Carlo (MCMC) sampler, is currently used to calibrate climate models. It is supported by ASCR, is implemented in C++/MPI-1 and has scaled to 128 chains. Note that a chain can drive a parallel simulator. SAChES will be re-implemented in MPI-2, as a loosely coupled Markov chain ensemble, communicating via asynchronous remote memory access. Loose coupling is necessary for scalability beyond $O(10^3)$ chains. It also enables resilience since a chain corrupted by a silent error can be decoupled and "killed" without significantly impacting parameter estimates. Our ML analytics are implemented in Python, supported by internal (LDRD) funds and used in the identification of flow markers of RANS prediction errors. They will be scaled up using the Apache Spark platform. Its component, MLlib, contains Numpy-based scalable implementations of ML algorithms for distributed computing platforms. Our in-situ topological analysis tools are implemented in C++/MPI, used in turbulent combustion simulations on Titan (ORNL) and are developed within EXACT, an ASCR funded exascale co-design center. They will be integrated with SPARC to develop flow descriptors to be used in calibration and ML of turbulence models.

- **Models and Algorithms:** Algorithmically, the RANS simulator would require modification to incorporate higher-order (non-linear) turbulence models, upgrades in the time-integrator to accommodate the resultant increase in stiffness and coupling with the topological analysis tools. SAChES will require the inclusion of a tempered version of an MCMC algorithm to better accommodate multimodal (estimated) parameter distributions.

- **End-to-End Requirements:** Our use of Spark requires an HDFS file storage system. Consequently, an HDFS-based data-store shared with the exascale platform (to allow saving of turbulence simulations / training data for machine learning) and a Spark platform (to run the big-data analytics) would be welcome pieces of computing infrastructure.

Related Research: Our calibration and ML facility could also be repurposed to calibrate Large Eddy Simulations, climate models, and material models in high energy density environments (e.g., Z-pinch). These applications involve empiricisms and consequently, calibrations.

10-Year Problem Target: We expect that our framework would be able to calibrate turbulence models for high Reynolds numbers ($Re = 10^6$-10^8) applications, yielding predictions with quantified uncertainties. This would imply calibration of O(10) parameters, using $O(10^3)$ chains, each driving a realistic (laboratory-scale model) forward problem on $O(>10^5)$ cores. The entire process would be free from the limitations of current parameter calibrations and the vagaries of surrogate-model creation and consequently amenable to inclusion in a design workflow.