Sierra: Science on a Mission

GTC 2019

San Jose, CA

Presenter: Rob Neely
Weapons Simulation and Computing Program Coordinator for Computing Environments

March 19, 2019

LLNL-PRES-767803

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

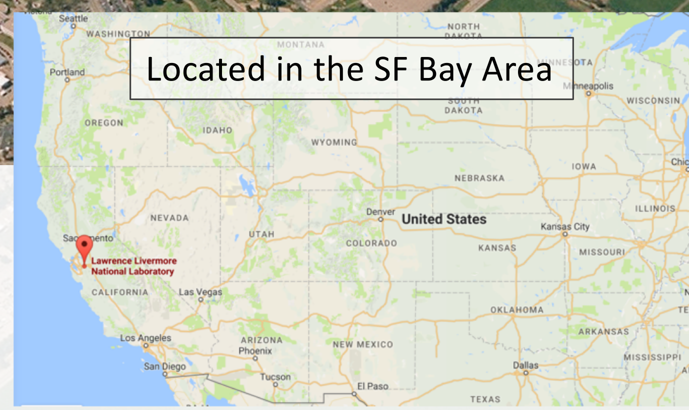
Lawrence Livermore National Laboratory

Outline

- LLNL's Mission for Stockpile Stewardship
- Sierra system overview
- Motivation for Sierra
- Application preparation with the "Sierra Center of Excellence"
- Exemplar unclassified NNSA application areas
- Open Science Period Results
  - Earthquake modeling with SW4
  - Biological modeling of Cancer
  - Neutron lifetime
  - Extreme-scale RANS modeling
- Future directions: AI/ML in scientific computing

LLNL's Mission Requirements for Extreme-scale Computing

Lawrence Livermore Nat'l Lab (LLNL) is a multidisciplinary national security laboratory



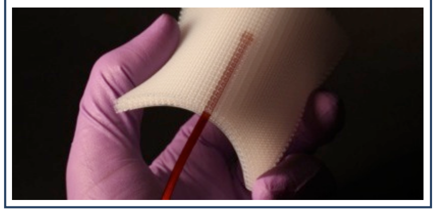
- Established in 1952
- Approximately 7,000 employees
- 1 square mile, 684 facilities

Science and Technology

Nuclear Security → Other National Security

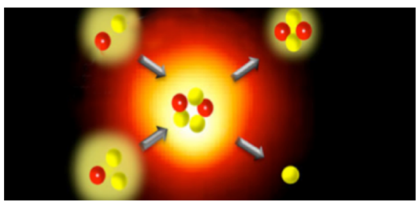
High-Energy-Density Science

Advanced Materials and Manufacturing

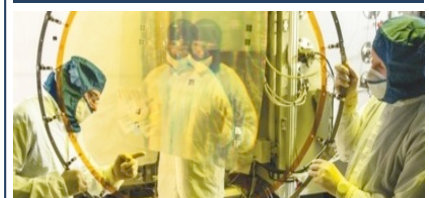


Bi0-Science and Bio-Engineering

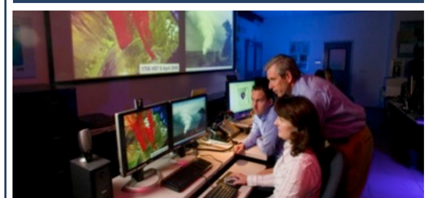
Nuclear, Chemical, and Isotopic S&T

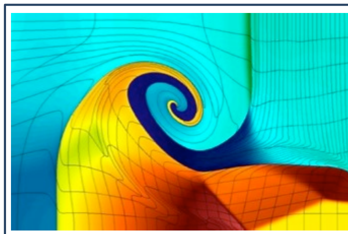


Lasers and Optical S&T

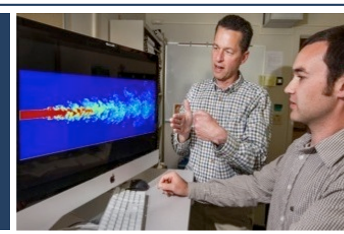


Earth and Atmospheric Science





High-Performance Computing, Simulation, and Data Science



Nuclear security and the Stockpile Stewardship Program

- Assess annually the safety, security, reliability and, effectiveness of the stockpile
- Extend life of stockpile warheads; adapting safety and security features to evolving requirements
- Strengthen underpinning science, technology, and engineering
- Nuclear nonproliferation and counterterrorism

The Stockpile Stewardship Program has successfully maintained the nuclear stockpile without explosive nuclear testing since 1992

Life extension programs can improve safety and security while providing no new military capabilities (consistent with policy)

Remanufacture approach

Legacy System

Remanufacture as close as possible to original

Improved safety and security

Improve safety and security approach

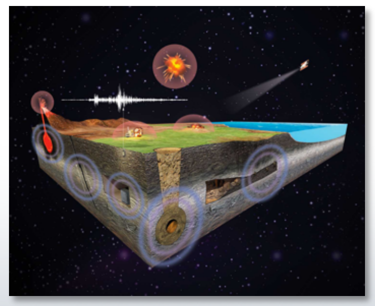
Still cannot take it on the track

Nuclear security is more than nuclear weapons

Non-proliferation

- Verification monitoring
- Sampling and analysis
- Safety and security of special nuclear material

Treaty monitoring

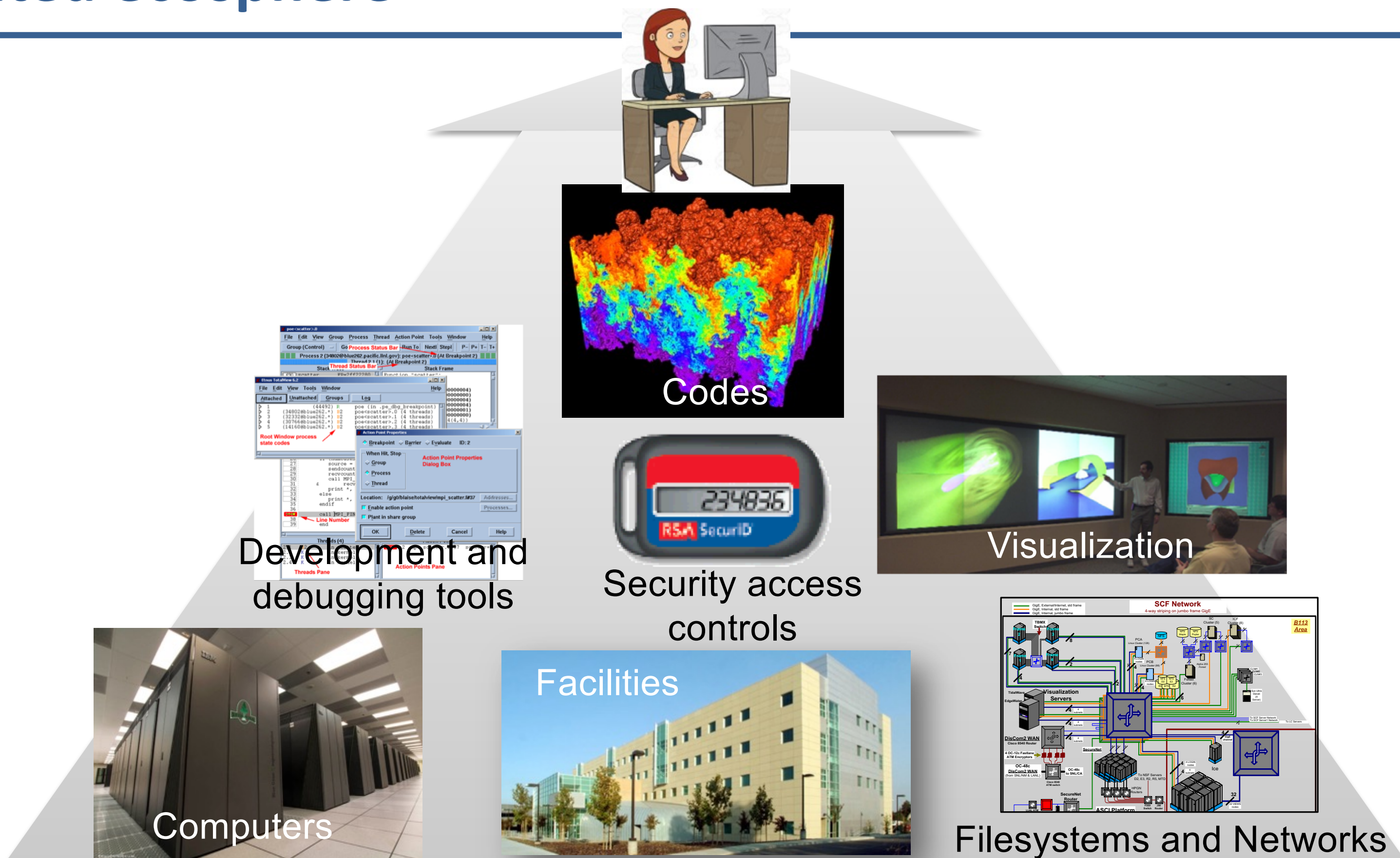


- Monitoring and detection for nuclear test ban treaty agreements
- Analysis and characterization of nuclear tests

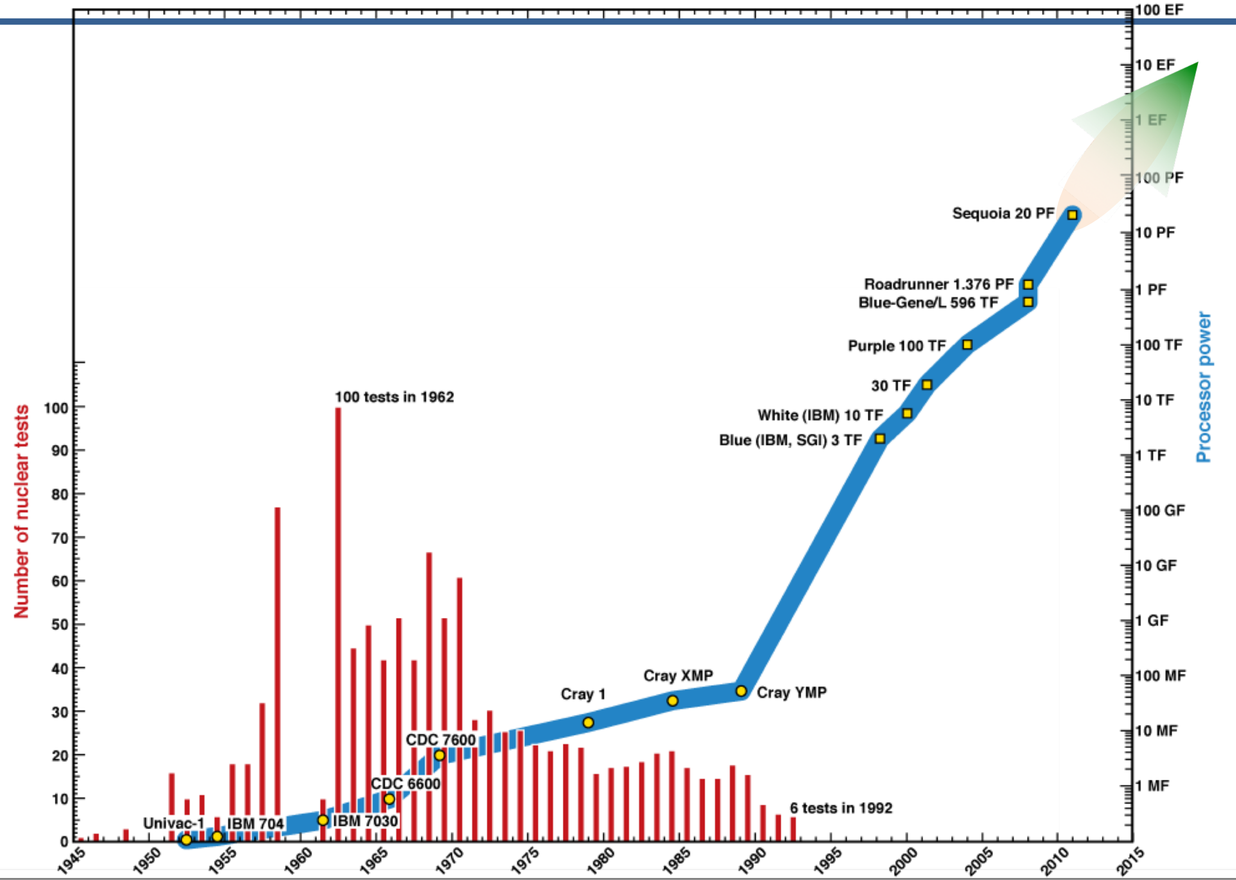
Counterterrorism

- Forensics
- Detection and diagnostics
- Incident response

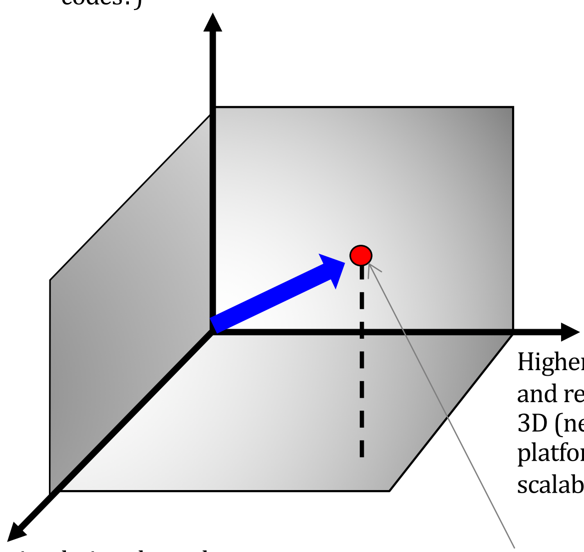
Providing simulation capabilities to the end user involves a large, integrated ecosphere

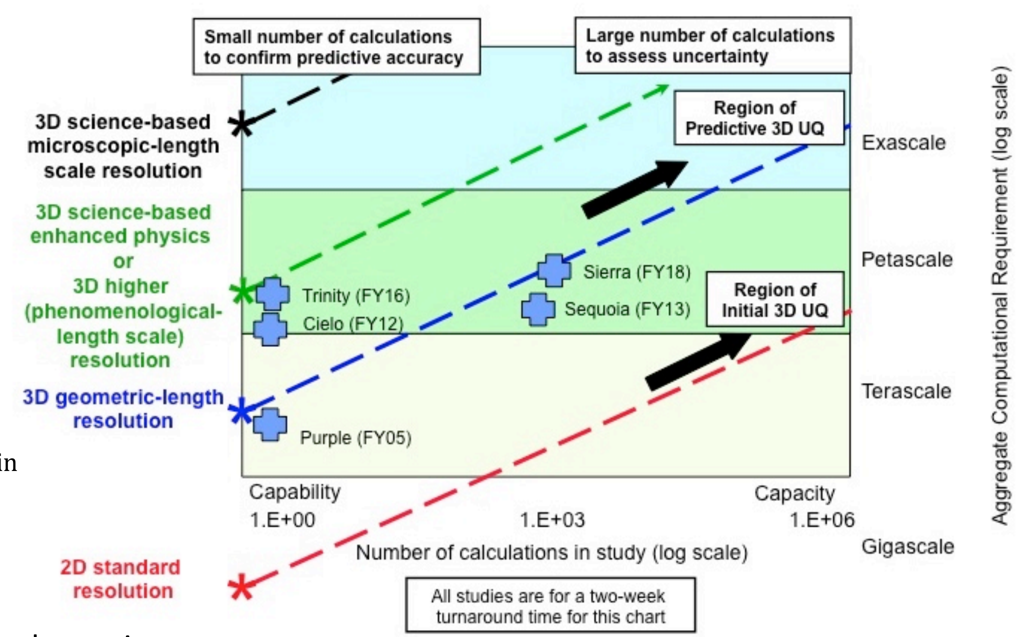


High-performance computing (HPC) is central to sustaining the nuclear stockpile in the absence of additional nuclear tests



What drives super-exascale requirements is 3D high resolution, enhanced physics calculations done in an uncertainty quantification regime





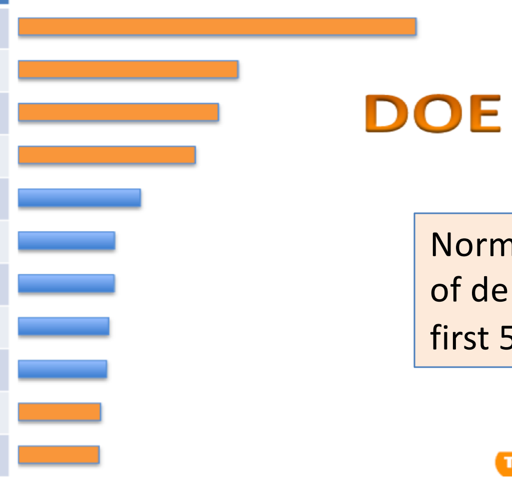
Our mission demands require progress along all three axes to advance our predictive capabilities

LLNL has a long history of fielding Top500 systems

Courtesy of Erich Strohmaier (LBL) SC’17 Invited talk
https://www.top500.org/static/media/uploads/presentations/top500_ppt_invited_201711.pdf

Dominant Sites

#	Site	Country	Norm-HPL
1	LLNL	USA	1,466
2	LANL	USA	810
3	ORNL	USA	737
4	SNL	USA	652
5	NSCC Guangzhou	China	449
6	RIKEN AICS	Japan	354
7	NASA/Ames	USA	352
8	FZ Jülich	Germany	333
9	JAMSTEC	Japan	325
10	NERSC	USA	303
11	ANL	USA	298



Normalized HPL performance of deployed systems over the first 50 Top500 lists (25 years)

What is not captured here is the derivative. Like a “moving average”, this will change over the next decade as *non*-US/DOE sites dominate the top 3 positions.

Sierra System Overview

Sierra is a 125 Petaflop peak system based in the Department of Energy’s Lawrence Livermore National Laboratory supporting national security mission advancing science in the public interest

System Overview

System performance

- Peak performance of 125 petaflops for modeling and simulation

Each node has

- 2 IBM POWER9 processors
- 4 NVIDIA Tesla V100 GPUs
- 320 GiB of fast memory
  - 256 GiB DDR4
  - 64 GiB HBM2
- 1.6 TB of NVMe memory

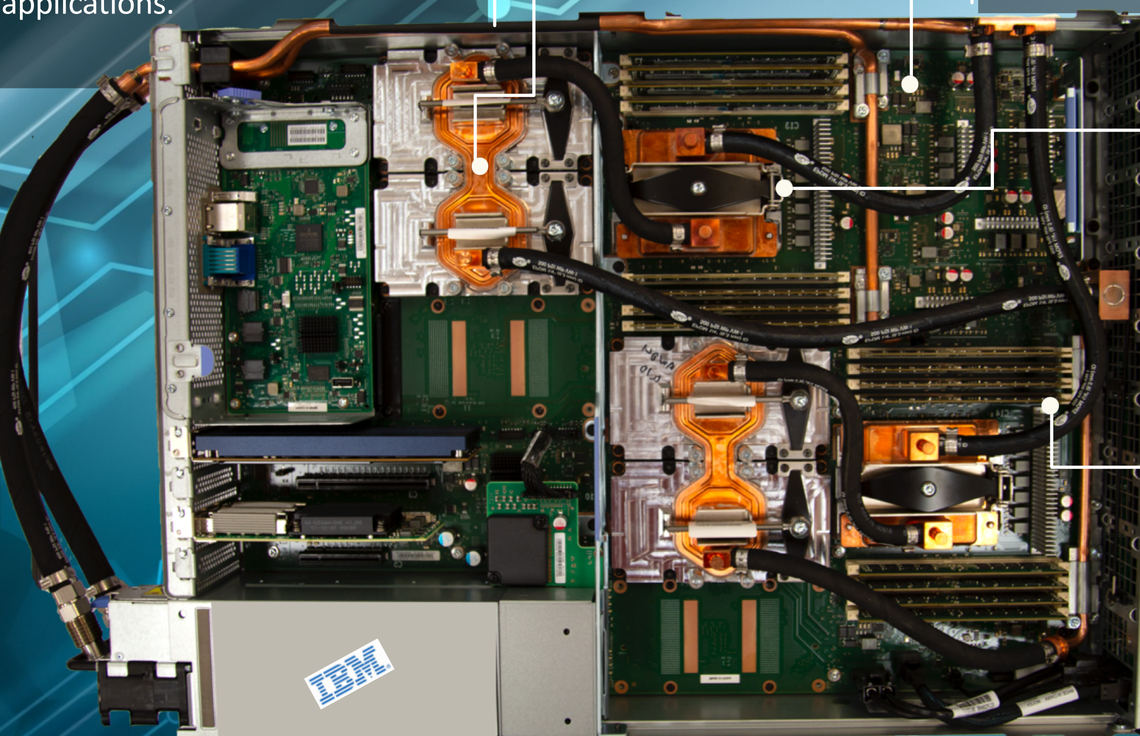
The system includes

- 4,320 nodes
- 2:1 tapered Mellanox EDR InfiniBand tree topology (50% global bandwidth) with dual-port HCA per node
- 154 PB IBM Spectrum Scale file system with 1.54 TB/s R/W bandwidth

What makes Sierra so effective?

GPUs: Sierra links more than **17,000** deep-learning optimized NVIDIA GPUs with the potential to deliver exascale-level performance (a billion-billion calculations per second) for AI applications.

High-Speed Data Movement: High-speed Mellanox interconnect and NVLink high-bandwidth technology built into all of Sierra's processors supply the next-generation information superhighways.



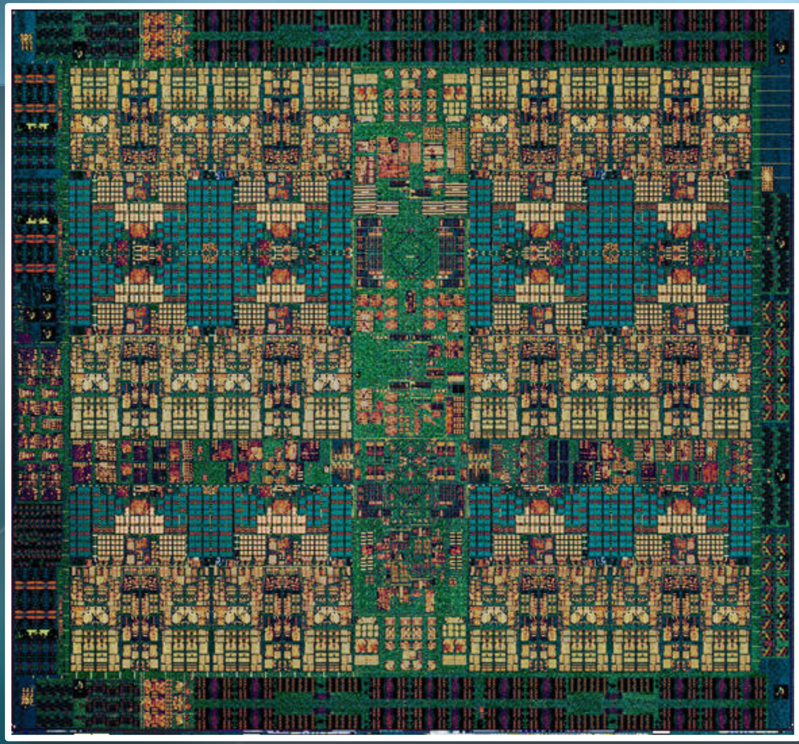
CPUs: IBM Power9 processors to rapidly execute serial code, run storage and I/O services, and manage data so the compute is done in the right place.

Memory: Sierra's sizable memory gives researchers a convenient launching point for data-intensive tasks, an asset that allows for greatly improved application performance and algorithmic accuracy as well as AI training.

Lawrence Livermore National Laboratory

IBM Power9 Processor

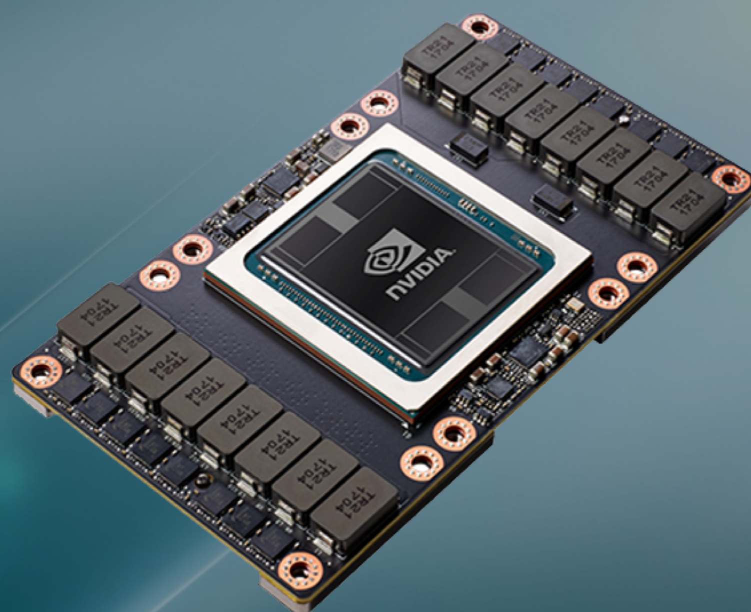
- Up to 24 cores
  - Sierra's P9s have 22 cores for yield optimization on first processors
- PCI-Express 4.0
  - Twice as fast as PCIe 3.0
- NVLink 2.0
  - Coherent, high-bandwidth links to GPUs
- 14nm FinFET SOI technology
  - 8 billion transistors
- Cache
  - L1I: 32 KiB per core, 8-way set associative
  - L1D: 32KiB per core, 8-way
  - L2: 256 KiB per core
  - L3: 120 MiB eDRAM, 20-way



Lawrence Livermore National Laboratory

NVIDIA Volta Details

	Tesla V100 for NVLink	Tesla V100 for PCIe
Performance with NVIDIA GPU Boost	DOUBLE-PRECISION 7.8 TeraFLOPS SINGLE-PRECISION 15.7 TeraFLOPS DEEP LEARNING 125 TeraFLOPS	DOUBLE-PRECISION 7 TeraFLOPS SINGLE-PRECISION 14 TeraFLOPS DEEP LEARNING 112 TeraFLOPS
Interconnect Bandwidth Bi-Directional	NVLINK 300GB/s	PCIE 32GB/s
Memory CoWoS Stacked HBM2	CAPACITY 16GB HBM2 BANDWIDTH 900GB/s	



Lawrence Livermore National Laboratory

Lassen

Lassen is an unclassified system similar to Sierra, but smaller in size.

Its peak performance is 20.9 petaflops vs. Sierra's 125 petaflops peak performance.

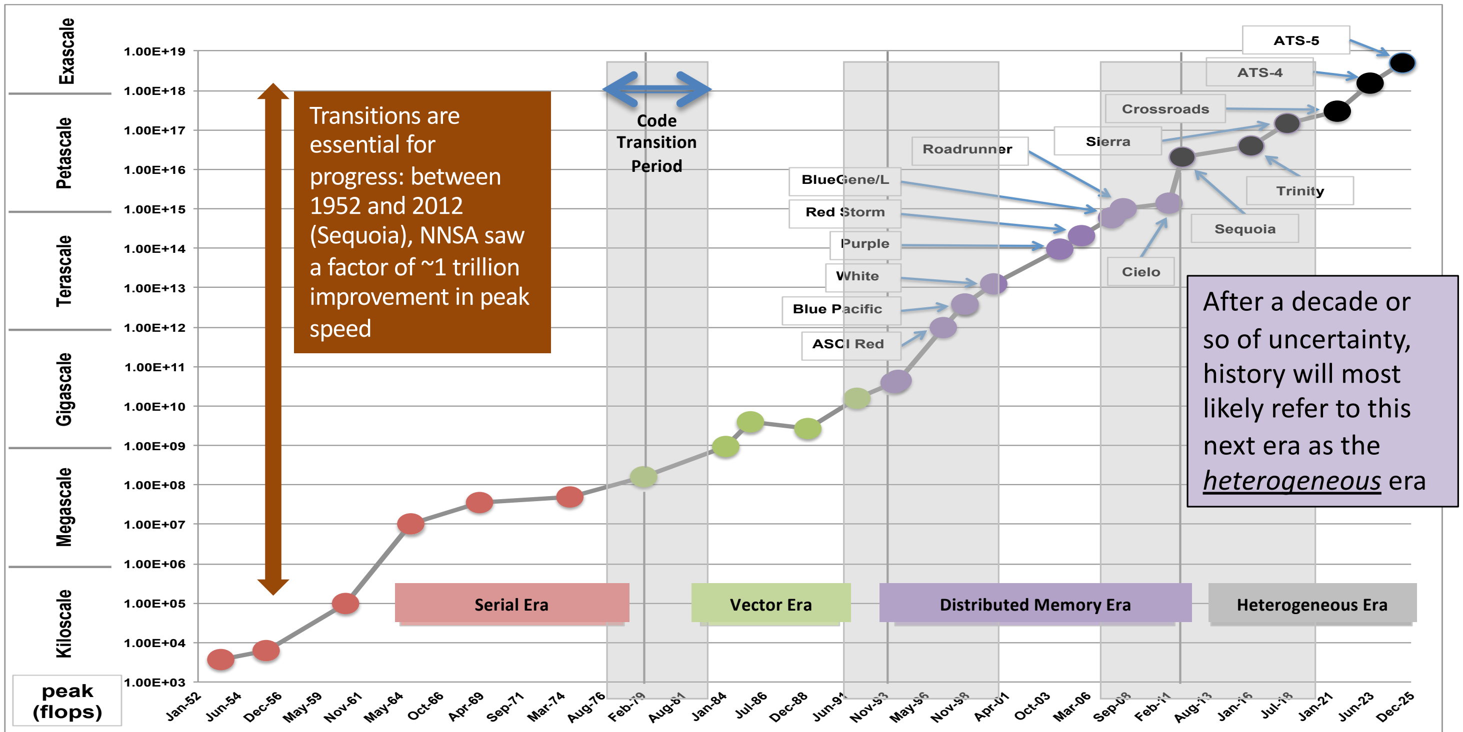
Multiple programs and institutional users will use Lassen for leading-edge science, such as predictive biology, computer-aided drug discovery, or exploring novel materials, on a world-class system.

System Details (subject to change on final install)	Lassen
Zone	CZ (Unclassified)
Nodes	720
POWER9 Processors per Node	2
GV100 (Volta) GPUs per Node	4
Node Peak (tFLOP/s)	29.1
System Peak (pFLOP/s)	20.9
Node Memory (GiB)	320
System Memory (PiB)	0.225
Interconnect	2x IB EDR
Off-Node Aggregate b/w (GB/s)	45.5
Compute Racks	40
Network and Infrastructure Racks	4
Storage Racks	4
Total Racks	48
Peak Power (MW)	~1.8

LLNL's Sierra System Installation

The Need to Pivot Toward Exascale

Advancements in HPC have happened over three major epochs, and we're settling on a fourth (heterogeneity)



Each of these eras defines a new common programming model and transitions are taking longer; DOE/NNSA is entering a fourth era.

The drive for more capable high-end computing is driving disruption in HPC application development

The trickle down…

External Constraints → Hardware Designs → Software Supports → Applications React

Assume: Continued increased computational power is desired

Power
- Unreasonable operating costs
- Today's tech = $100M/year – exceeds capital investments

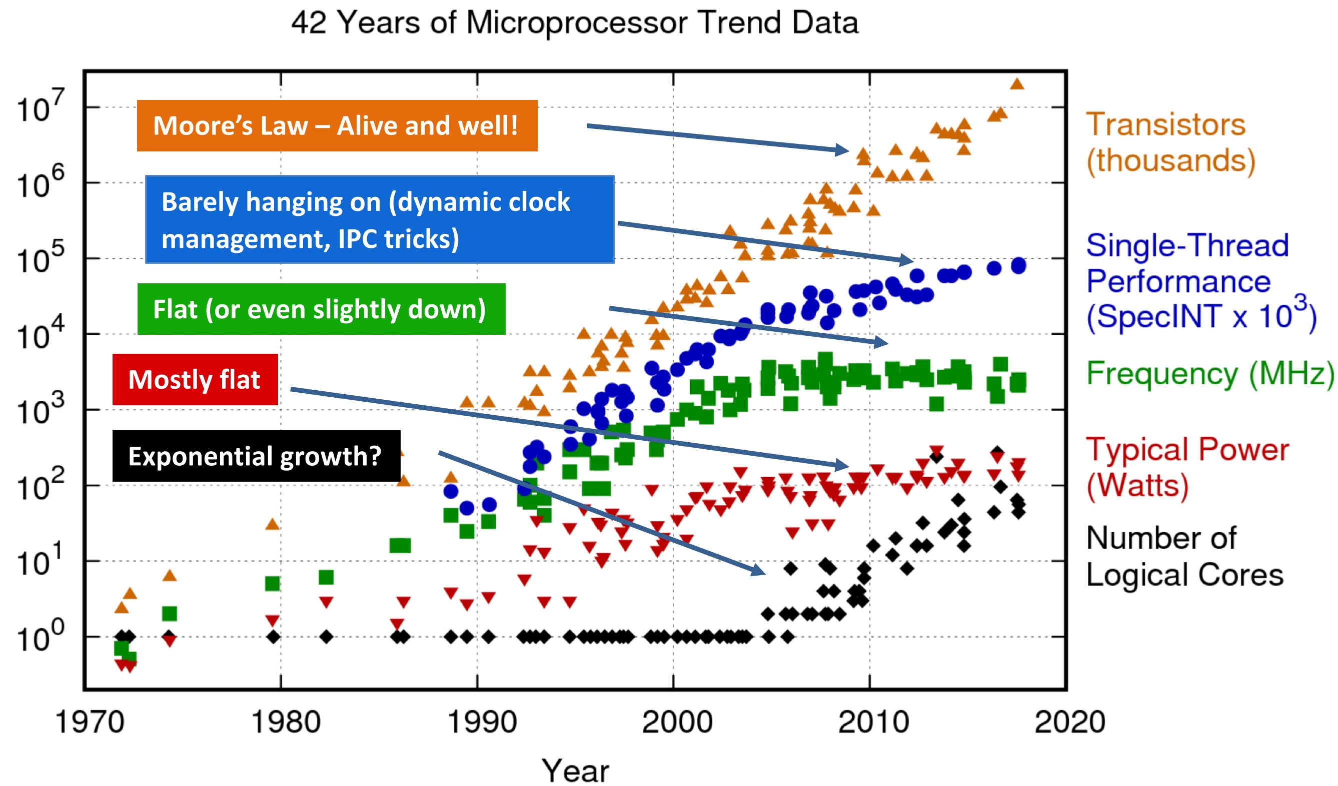
Chip / processor efficiency
- GPUs / accelerators
- Simpler cores
- Unreliability at near-threshold voltages – compounded by scale of systems

On-node "inscaling" challenges
- Complex hardware
- Massive on-node concurrency requirements
- New programming and memory models

Disruptive Changes
- Code evolution
- Algorithm rewrites
- Platform uncertainty

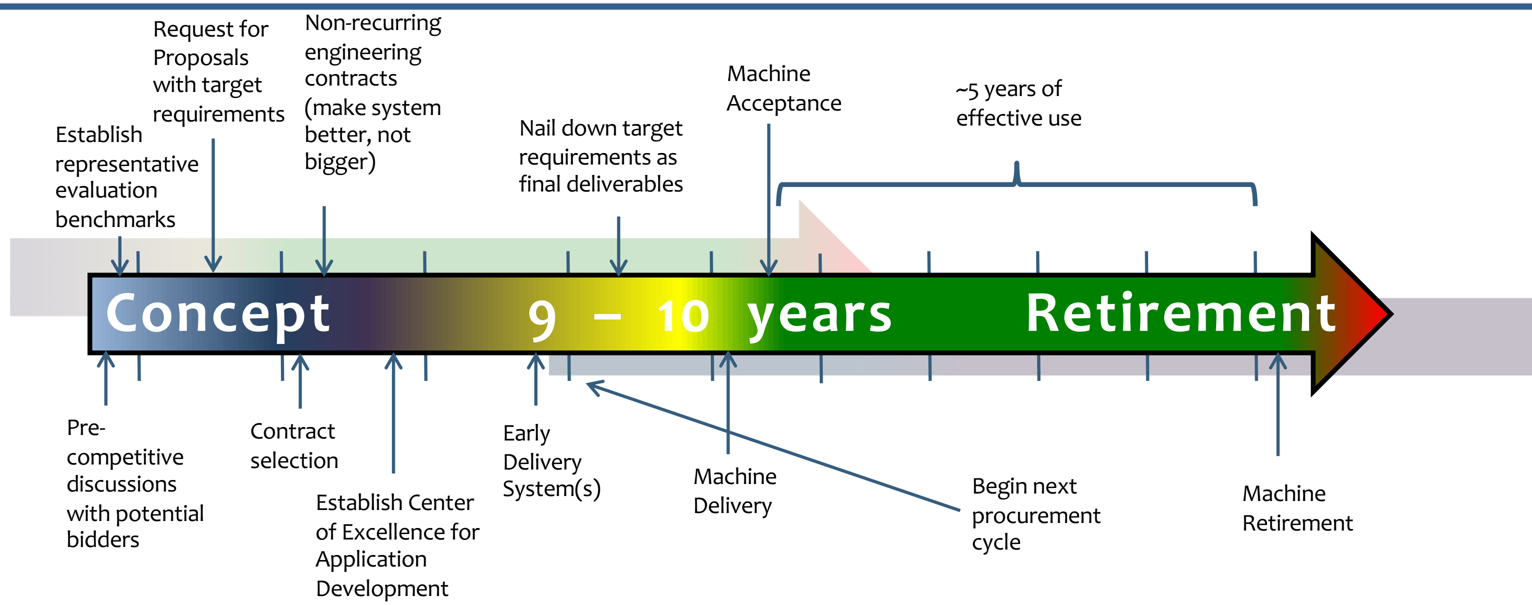
Meanwhile, “Big Data” and AI is driving industry trends and investments
- Social Networking, Machine/Deep Learning, Cloud Computing, …
- Software innovation in data-centric HPC is thriving
- Traditional HPC is facing strong headwinds if it can’t adapt

Processor trends tell the parallelism story….



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

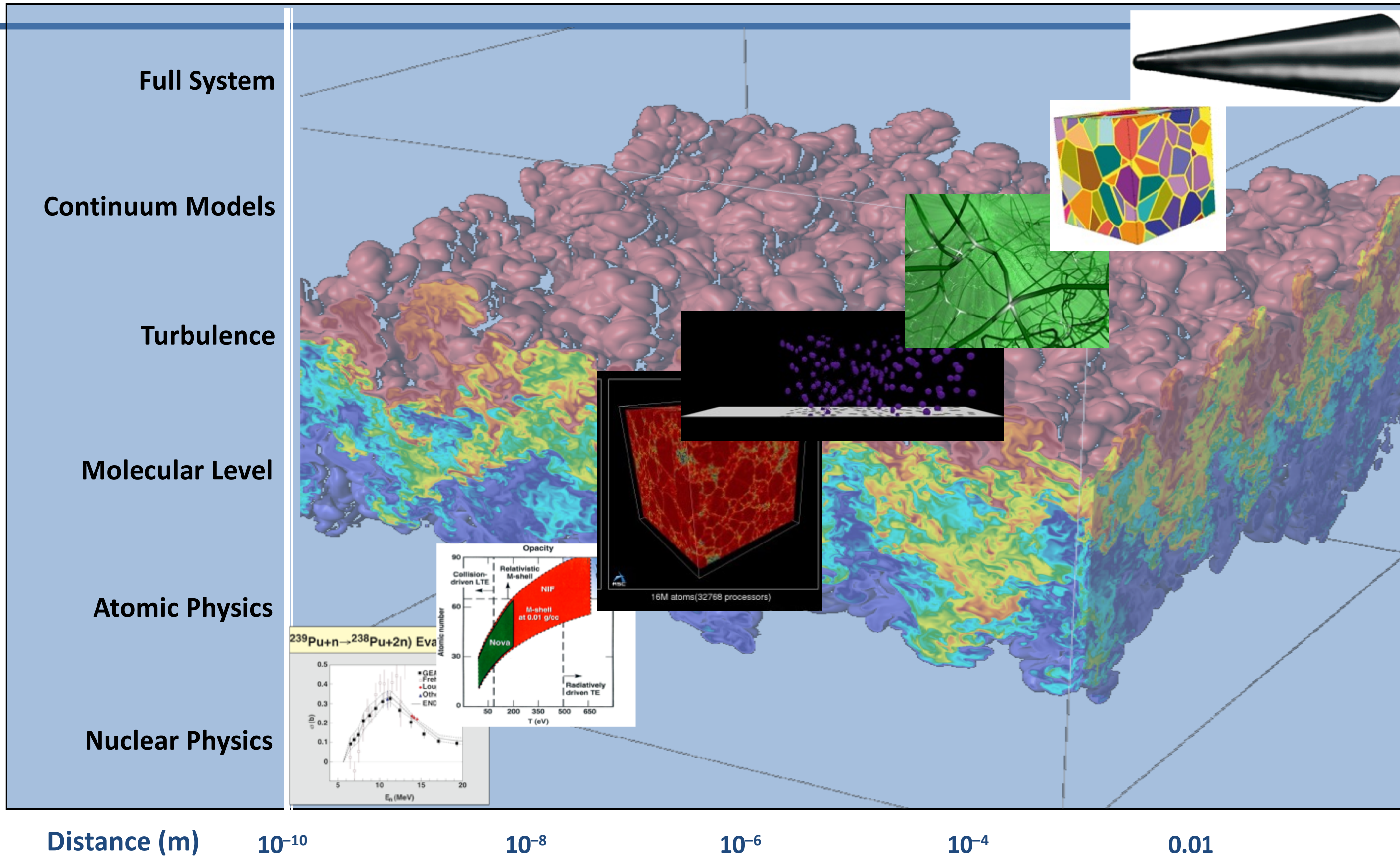
The LLNL/ASC procurement strategy relies on long-lived partnerships with HPC vendors



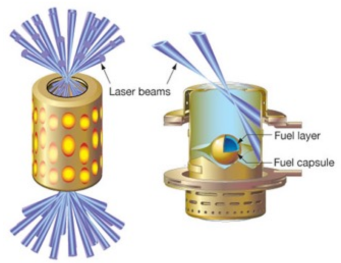
- Long lead times in procurements allow vendor roadmaps to be clear enough to bid, but still fungible
- Use of target requirements reduces risk to vendor, and thus cost to the Program
- Testbed architectures make up the unofficial 3rd tier of the ASC platform strategy: evaluation of competing and potential future technologies at smaller scale

Application Preparation and the Sierra Center of Excellence

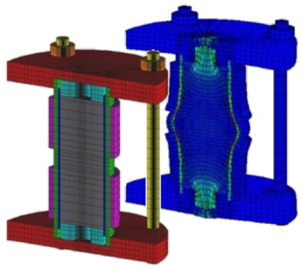
Predicting the life cycle behavior of nuclear warheads requires simulation at a wide range of spatial and temporal scale



Large, Integrated multi-physics Codes provide the basic simulation capabilities for broad range of application domains

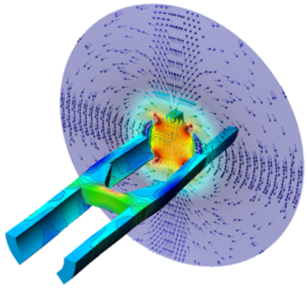


Inertial Confinement Fusion



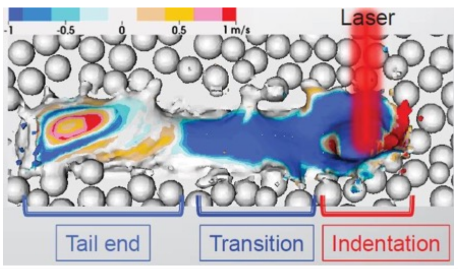
HE Cookoff

Blast on Structures

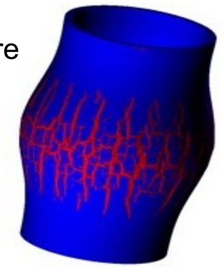


Navy Railguns

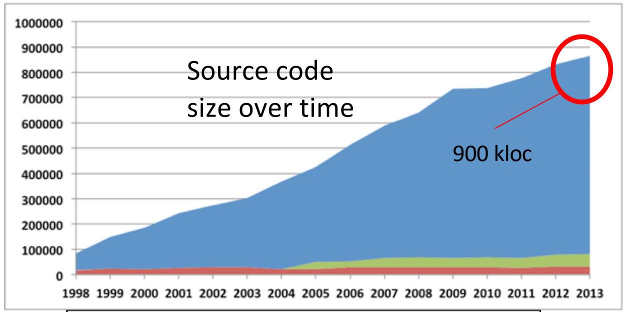
Additive Manufacturing



Fracture and Failure



- Long life-time projects
- 15+ years of development by large teams (10 – 20+ people, ~50% CS & 50% Physicists)
- Algorithms tuned for minimal turn-around time *based on today's hardware!*



Average of 60Kloc/year of additional code to support in a representative IC

Sierra Represents a Major Shift For Our Mission-Critical Applications on Advance Architectures

2004 - present

- IBM BlueGene, along with copious amounts of linux-based cluster technology dominate the HPC landscape at LLNL.
- Apps scale to unprecedented 1.6M MPI ranks, but are still largely MPI-everywhere on ”lean” cores

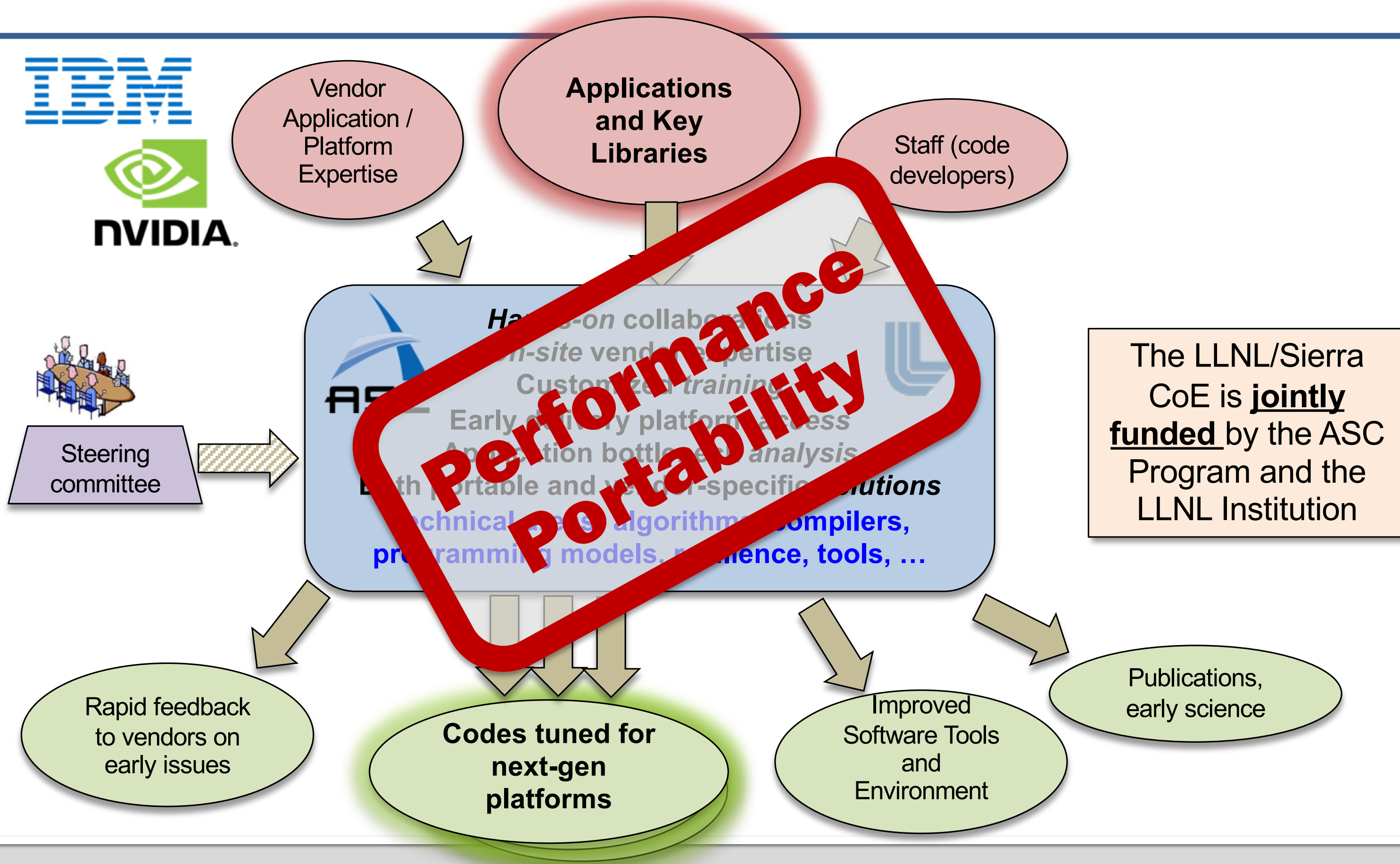
2014

- Heterogeneous GPGPU-based computing begins to look attractive with NVLINK and Unified Memory (P9/Volta)
- Apps must undertake a 180 turn from BlueGene toward heterogeneity and large-memory nodes

2015 – 20??

- Heterogeneous computing is here to stay, and *Sierra* gives us a solid leg-up in preparing for exascale and beyond

Application enablement through the Center of Excellence – the original (and enduring) concept



High level goals of the Center of Excellence

To train ASC and LLNL institutional application teams in the art and science of developing high performance software on the Sierra architecture

To have our flagship applications utilizing advanced features of Sierra as soon as the machine is generally available, and fully optimized (4-6x improvement over Sequoia) for the architecture within 1-2 years of deployment

To develop performance-portable solutions that allow application teams to focus on maintaining a single source code that will be effective on platforms deployed at our sister laboratories

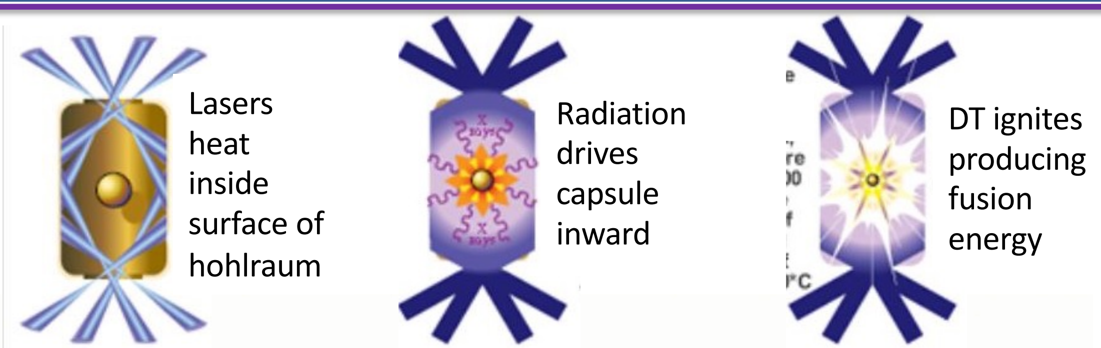
To build strong ties with our vendor partners as a key element of our co-design strategy – leveraging their deep knowledge of the hardware and algorithmic approaches specific to their platform, and informing them of our long-term application requirements as they will impact future machine designs

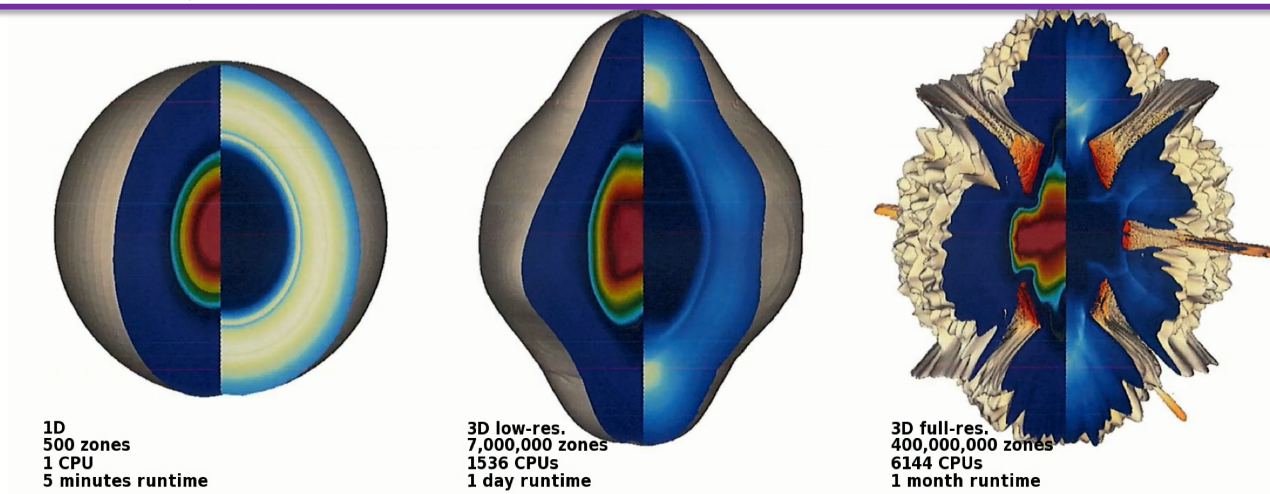
Specific strategies for each of these goals were captured in an "Execution Strategy" document and used to guide our implementation

Applications on Sierra

High resolution ICF simulations help us understand yield degradation mechanisms

Basics of Inertial Confinement Fusion



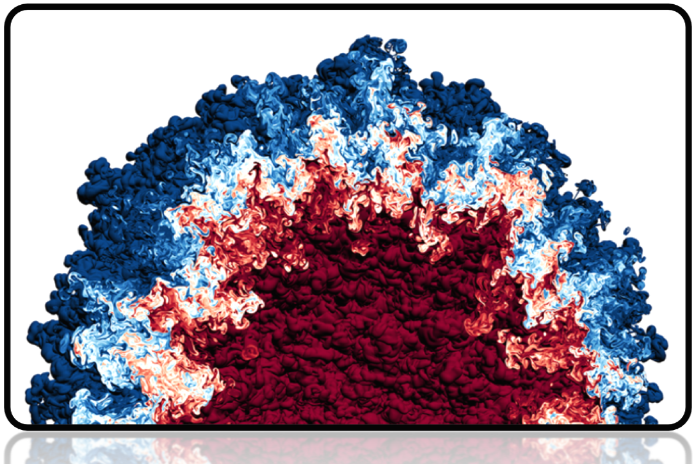


Taking Inertial Confinement Fusion (ICF) simulations to a higher level

Chris Clouse
LLNL

Until recently, 3D simulations were often run as a double check on routinely employed 2D approximations. These 3D simulations were computationally expensive and, as a result, were run as sparingly as possible. Design and analysis in LLNL's National Ignition Facility (NIF) Program until today relied primarily on 2D approximations because 3D simulations could not be turned around quickly enough to make them a useful routine design tool.

Sierra's architecture, which is expected to bring speed-ups on the order of 10X for many of our 3D applications, will be able to process these crucial simulations efficiently, changing the way users compute by making the use of 3D routine.



"Sierra provided an unprecedented (98 billion cell) simulation of two-fluid mixing in a spherical geometry. Understanding hydrodynamic instability and the transition to turbulence process is important in inertial confinement fusion and High Energy Density (HED) Physics."
—Chris Clouse

Developing the next generation of algorithms and codes

Rob Rieben
LLNL

To maximize the full potential of new computer architectures like Sierra's, we are developing new algorithms that play to the strengths of the machine—specifically the high intensity of compute operations available to data that remain local.

We are also developing next-generation simulation codes for inertial confinement fusion and nuclear weapons analysis that employ high-order, compute-intensive algorithms that maximize the amount of computing done for each piece of data retrieved from memory. These schemes are very robust and should significantly improve the overall analysis workflow for users. These advanced simulation tools enabled by Sierra will improve throughput along two axes: faster turnaround and less user intervention.

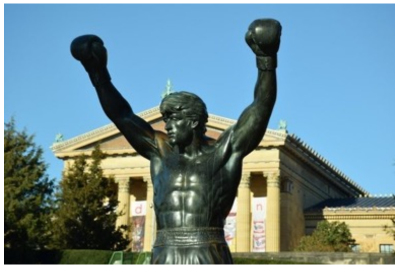
"The combination of our next generation, compute intensive, high-order algorithms with the processing power of Sierra will deliver an exciting new era in multi-physics simulations."
—Rob Rieben

Early Science Applications

Early Science - That time between when the system is first booted up (~3/1/18) and when it is transitioned to the classified network for Programmatic use (~1/20/19) when friendly users help "shake down" the system.

Characteristics: Buggy or incomplete software stack, random bad hardware, rebooted on a whim, nursing problems along through the night, etc…

Benefits… when it's up, you can sometimes get the whole system to yourself for as long as you can keep things running. Invaluable feedback to the facility and vendors, and a one-time shot for open science apps

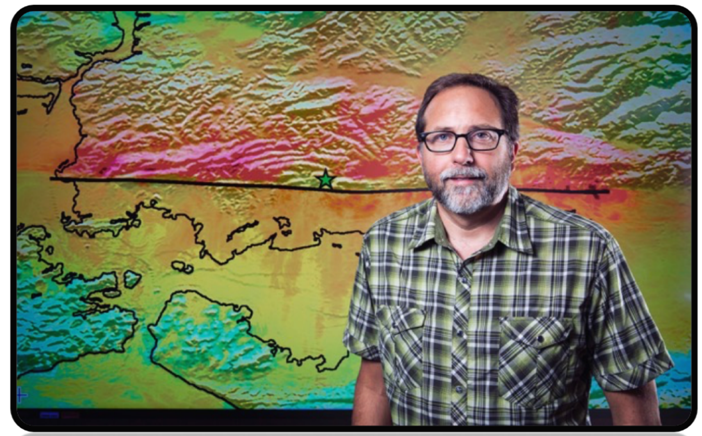


Sierra advances resolution of earthquake simulations

Arthur Rodgers
LLNL

The character of strong earthquake shaking near a fault is highly variable and poorly constrained by empirical data. Supercomputers enable simulation of earthquake motions to investigate the hazard and risk to buildings and infrastructure before damaging events occur. The large scale (~100 km) and fine detail (~10 m) of high frequency waves (>5 Hz) from damaging earthquakes require today's most powerful computers.

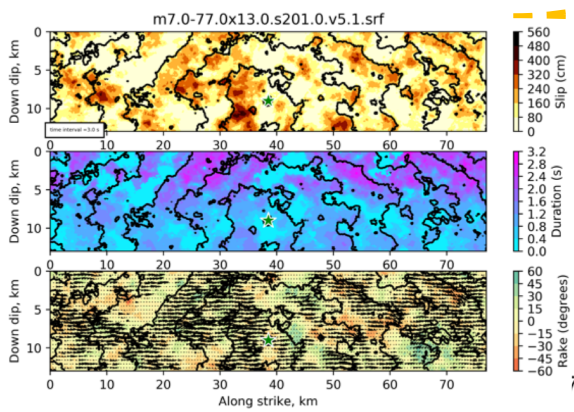
Using SW4-RAJA, a 3D seismic simulation code ported to GPU hardware, LLNL researchers are able to increase the resolution of earthquake simulations to span most frequencies of engineering interest on regional domains with rapid throughput to enable sampling of various rupture scenarios and sub-surface models.

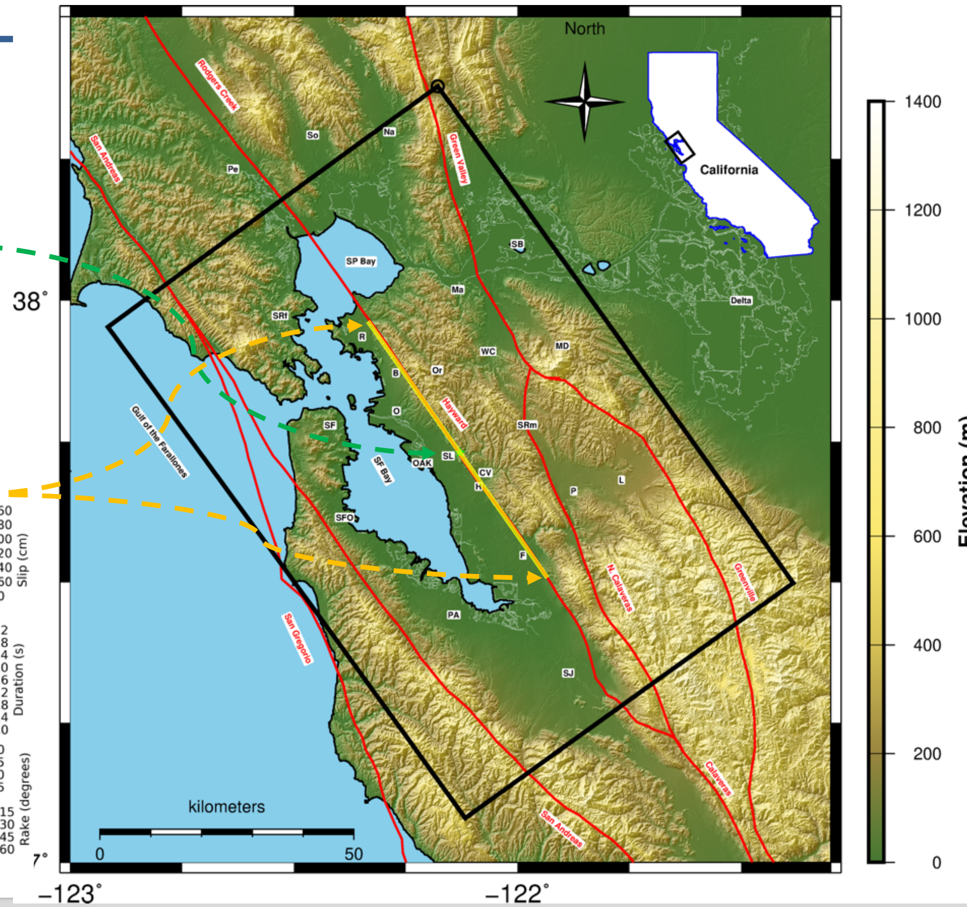


"Sierra's thousands of GPU-accelerated nodes allow SW4-RAJA to compute earthquake ground motions with 100's of billions of grid points in shorter run times so we can resolve high-frequency waves and investigate different rupture scenarios or earth models."
—Arthur Rodgers

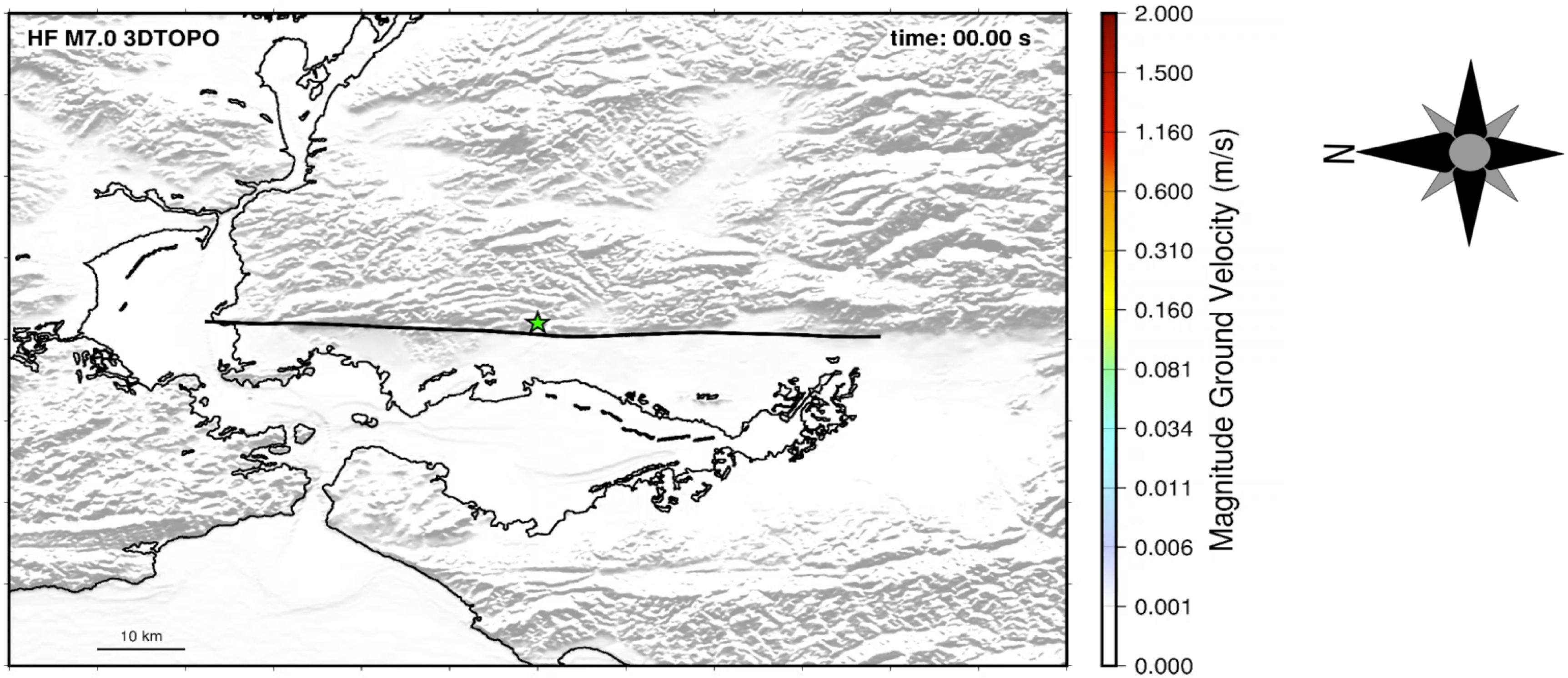
Computational Domain and Rupture

- 120 km x 80 km x 35 km
- Centered on:
  - San Leandro
  - hypocenter (green star)
- Rupture: M_W 7.0
  - Dimensions: 77 x 13 km
  - Draped onto 3D geometry of the Hayward Fault





Modeling a 7.0 quake on the hayward fault (SF Bay Area)



Courtesy: Artie Rogers, LLNL

National Lab HPC resources we run on:
Cori-II at LBNL/NERSC and Sierra at LC/LLNL

Cori Phase-II at LBNL/NERSC
#10 on Top 500
27 PetaFLOPS (peak)
68 Intel Xeon Phi (CPUs) per node

EXASCALE COMPUTING PROJECT

BERKELEY LAB

NERSC National Energy Research Scientific Computing Center

SIERRA

SW4 port to GPU platforms supported by 3-year LLNL internal project

Sierra at LLNL
#3 on Top 500
119 PetaFLOPS (peak)
2 IBM POWER (CPU) per node and
4 NVIDIA GPUs per node

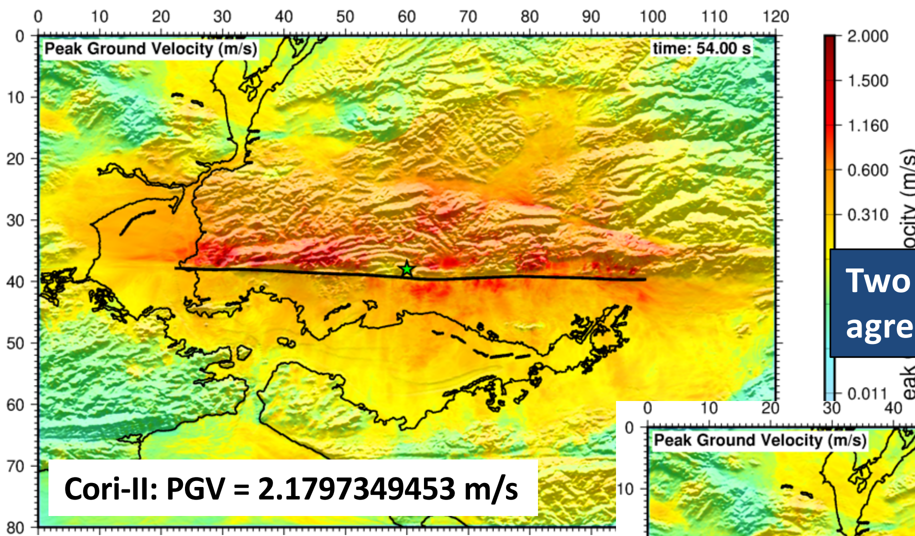
Lawrence Livermore National Laboratory

Future platforms will rely heavily on GPU's

Summer 2018
Hayward Fault M_W 7.0 runs

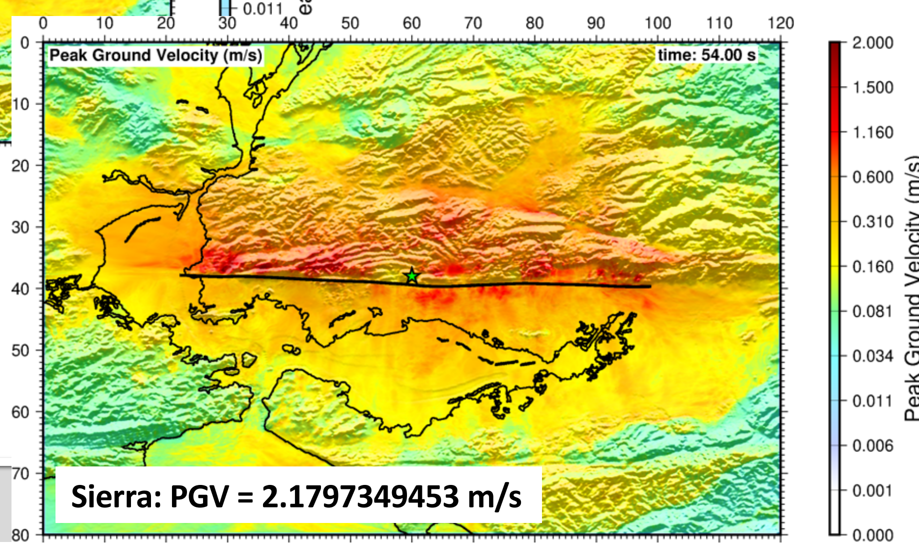
SW4 (CPU) f_{max} = 5.0 Hz on Cori-II	SW4-RAJA (GPU) f_{max} = 5.0 Hz on Sierra
v_{Smin} = 500 m/s	v_{Smin} = 500 m/s
PPW = 8	PPW = 8
T = 90 s	T = 90 s
h_{min} = 12.5 m	h_{min} = 12.5 m
N_{points} = 25.8 billion	N_{points} = 25.8 billion
$N_{time\text{-}steps}$ = 63,063	$N_{time\text{-}steps}$ = 63,063
Checkpointing every 4000 time-steps	No checkpointing
Ran on 8192 nodes (85%) of Cori-II (524,288 cores CPUs)	Ran on 256 nodes (6%) of Sierra (1024 GPUs)
Solver time 10.3 hours	Solver time 10.3 hours

Verification of SW4 and SW4-RAJA: Peak Ground Velocity maps

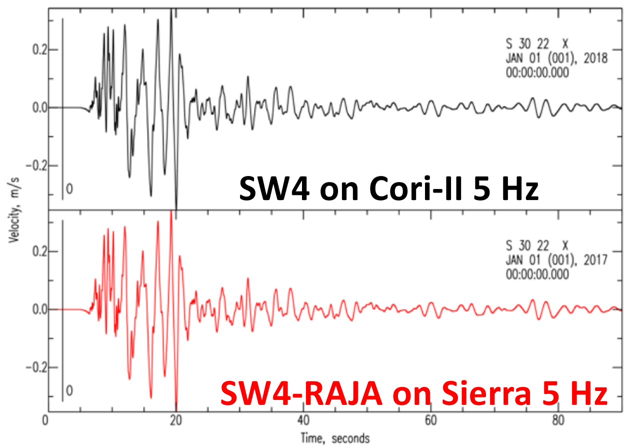


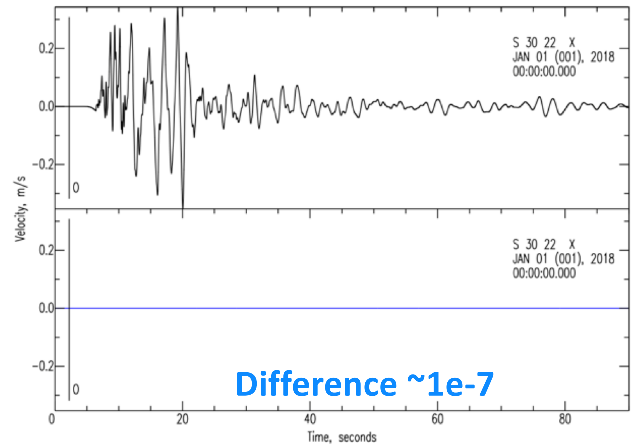
HF M_W7.0 3DTOPO
f_{max} = 5 Hz
h_{min} = 12.5 m

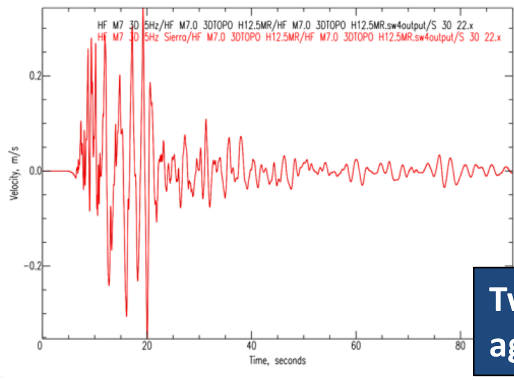
Two codes, different platforms PGV maps agree to machine precision!



Verification of SW4 and SW4-RAJA: Waveforms agree to single precision (SAC files)







SW4-RAJA on Sierra 5 Hz
SW4 on Cori-II 5 Hz
overplotted

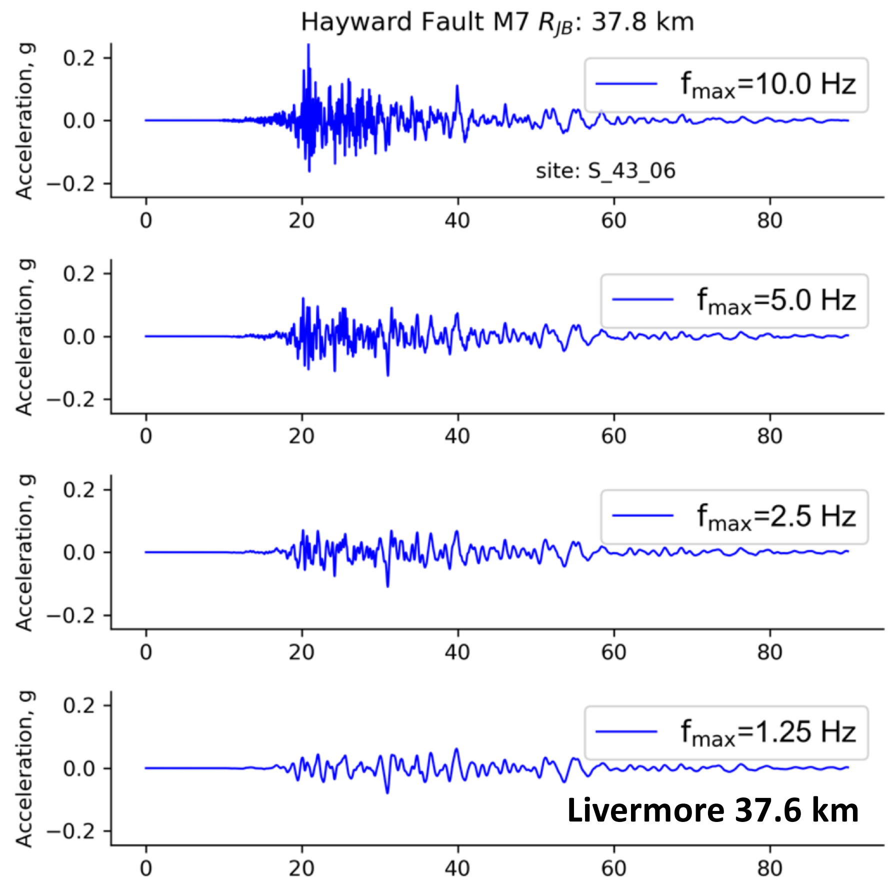
Two codes, different platforms waveforms agree to machine precision!

Comparison of near-fault accelerations for a range of resolutions (f_{max} = maximum frequency)

Sierra run Aug. 2018
f_{max} = 10.0 Hz

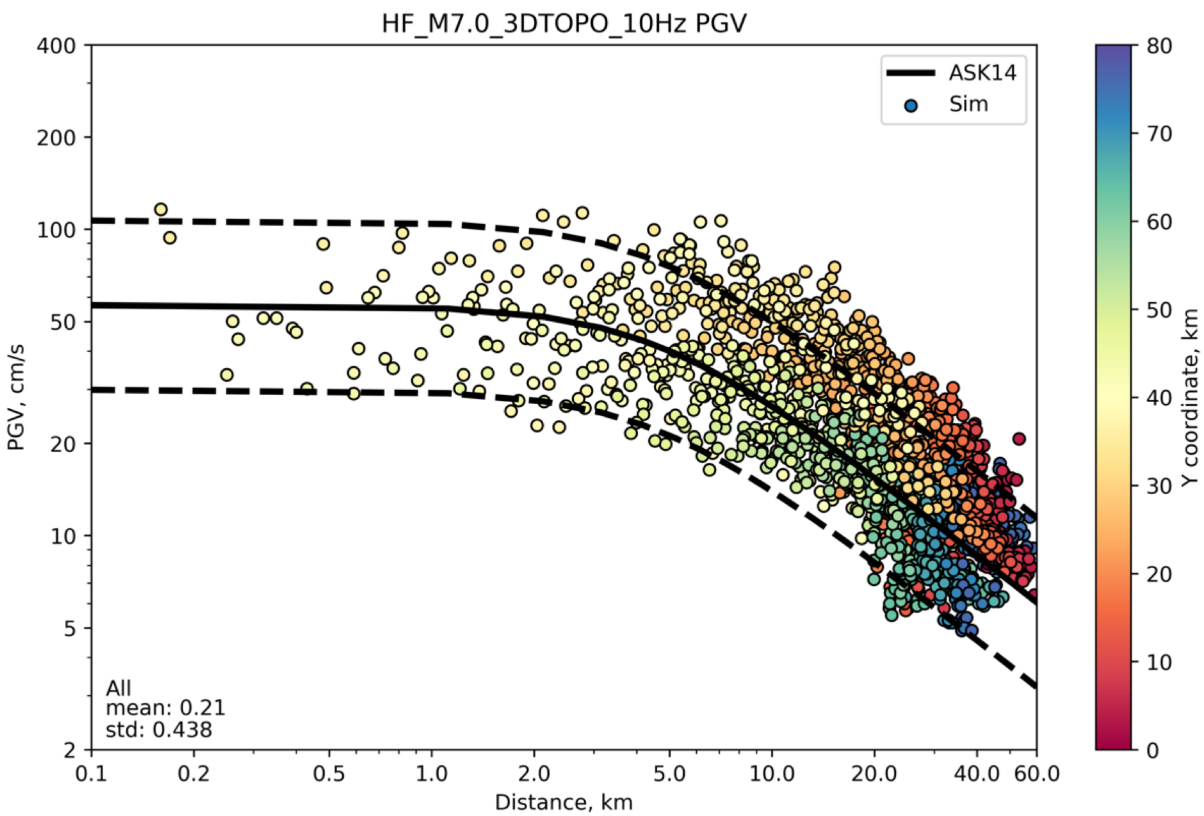
Doubling f_{max} requires 16x more computational effort

As frequency content increases, amplitudes increase and are more consistent with observations

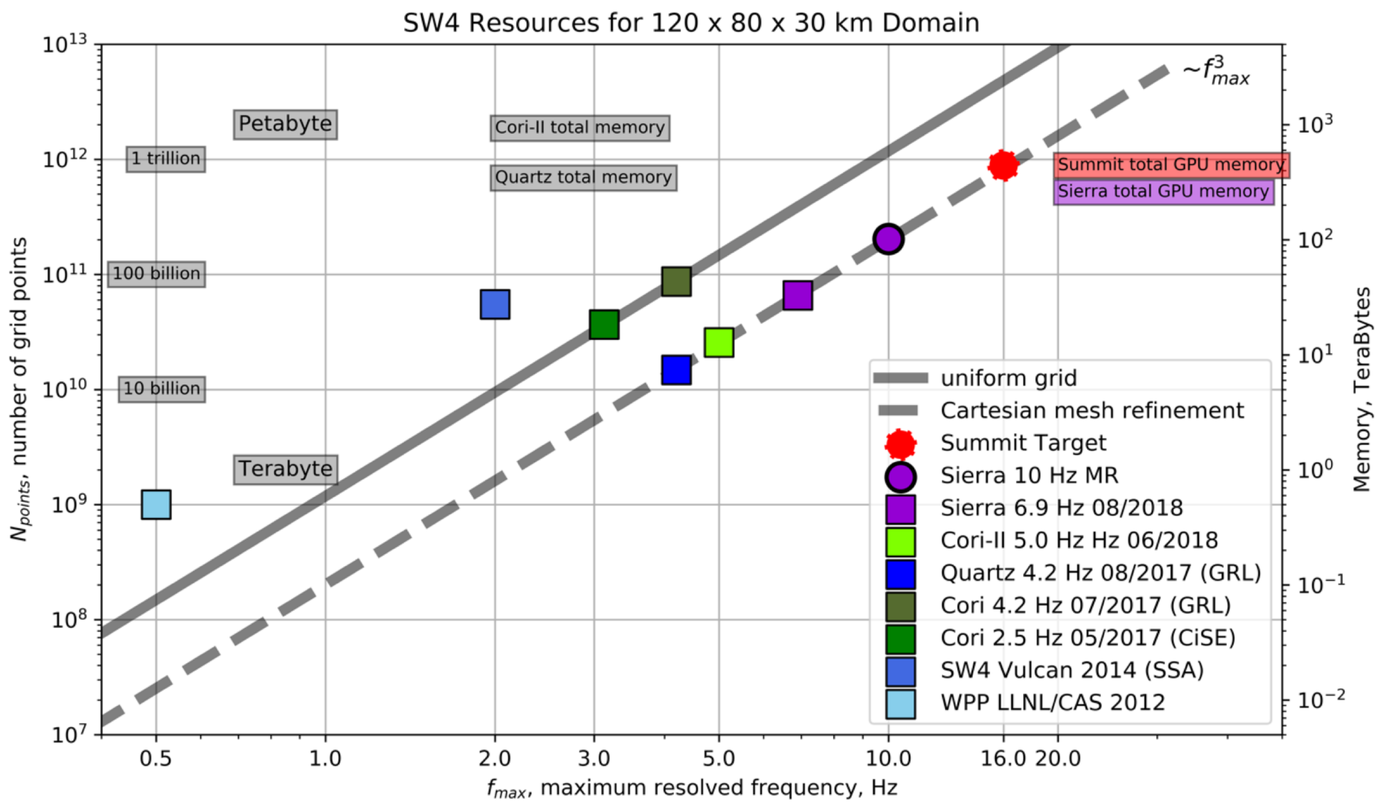


GMIM's versus distance with ASK (2014) GMM

HF M_W 7.0 10.0 Hz - PGV



Perspective on previous regional-scale Hayward Fault runs on various machines



Joint Design of Advanced Computing Solutions for Cancer

A collaboration between DOE and NCI

CANCER MOONSHOT

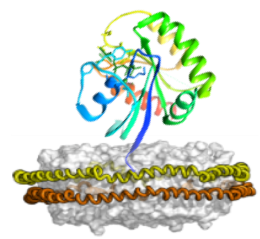
Pilot 1

Pre-clinical Model Development

Yvonne Evrard (FNLCR)
Rick Stevens (ANL)

Pilot 2

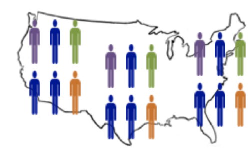
RAS Biology in Membranes



Dwight Nissley (FNLCR)
Fred Streitz (LLNL)

Pilot 3

Precision Oncology Surveillance



Lynn Penberthy (NCI)
Gina Tourassi (ORNL)

Courtesy: Fred Streitz, Jim Glosli, LLNL

Oncogenic RAS is responsible for many human cancers

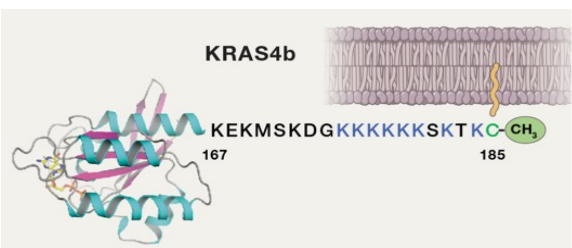
93% of all pancreatic

42% of all colorectal

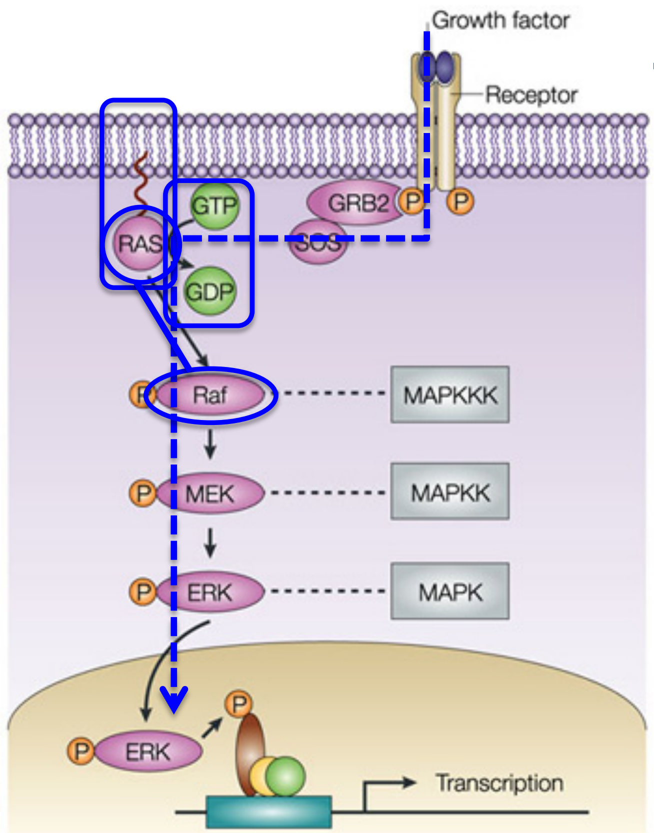
33% of all lung cancers

1 million deaths/year world-wide

No effective inhibitors



Simanshu,Cell 170, 2017



Nature Reviews | Molecular Cell Biology

Pathway transmits signals

RAS is a switch
oncogenic RAS is “on”

RAS localizes to the plasma membrane

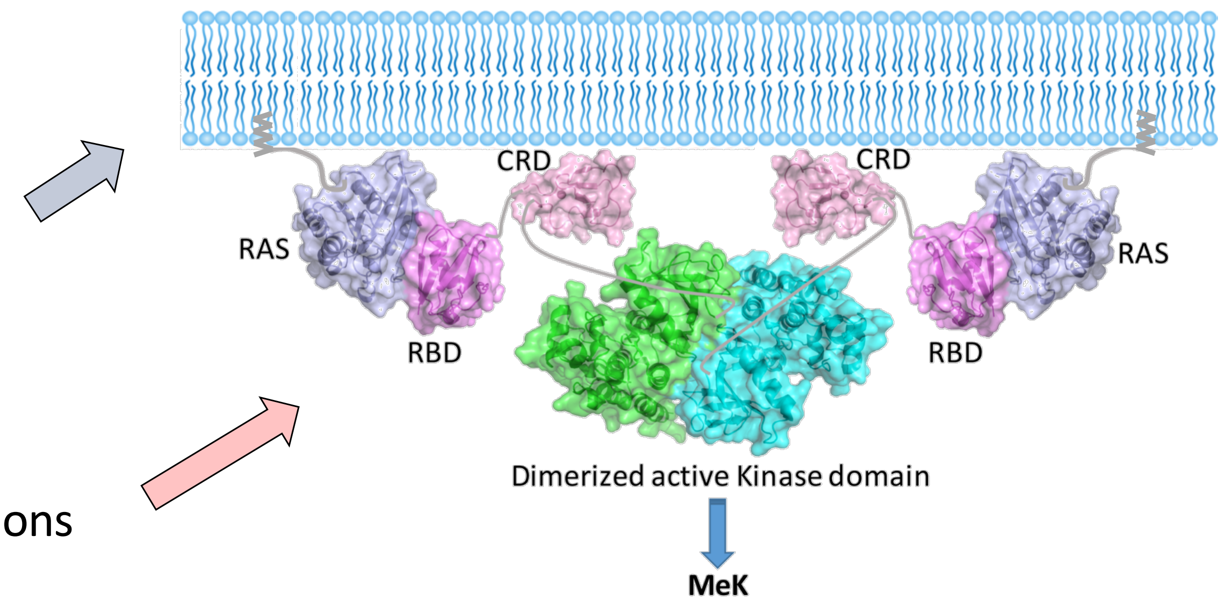
RAS binds effectors (RAF) to activate growth

Essential strategy: utilize appropriate scale methodology for each component

A problem of length and time scale:

Membrane evolves on ms time frame across μm…

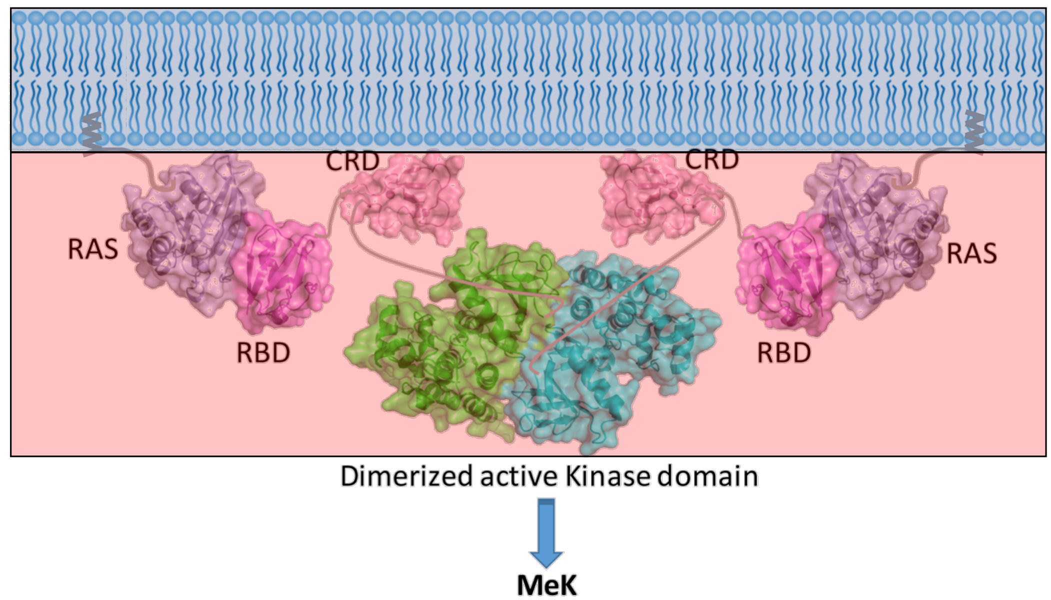
…while protein interactions involve μs across nm



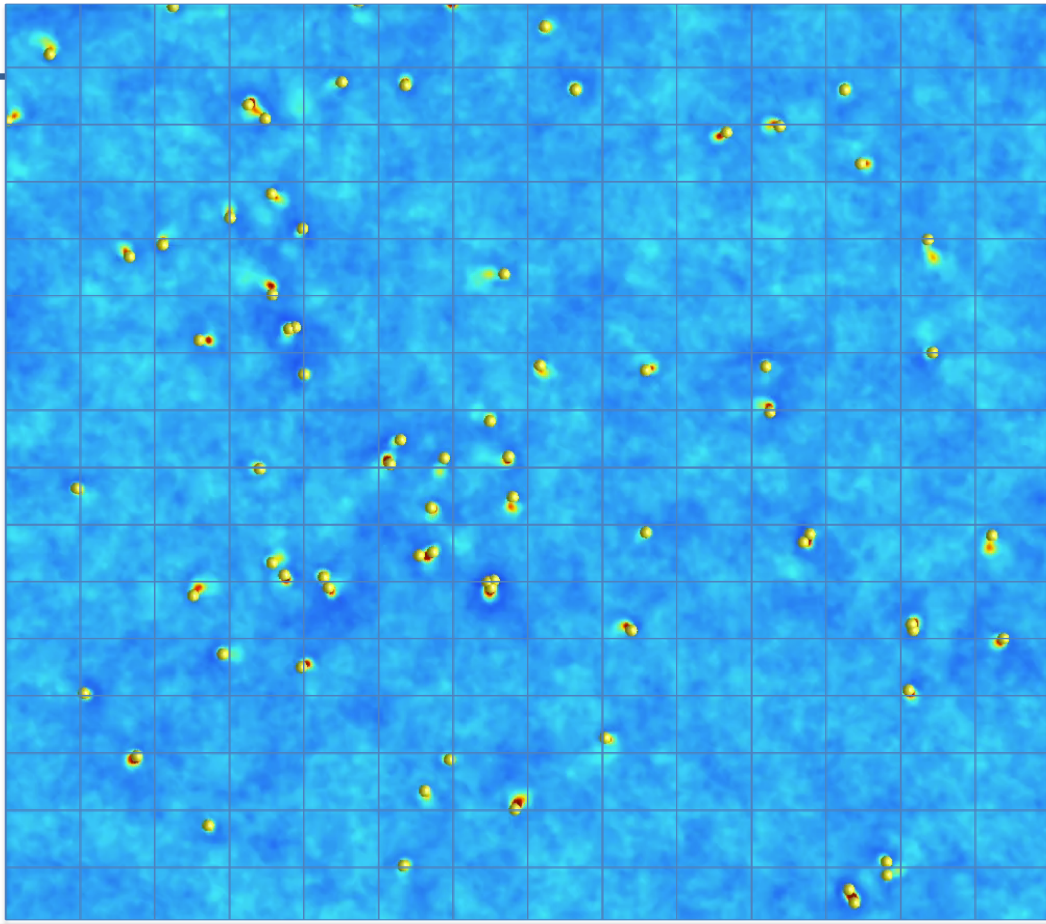
Essential strategy: utilize appropriate scale methodology for each component

Model membrane with RAS at micron (continuum) scale

Model protein behavior at molecular scale

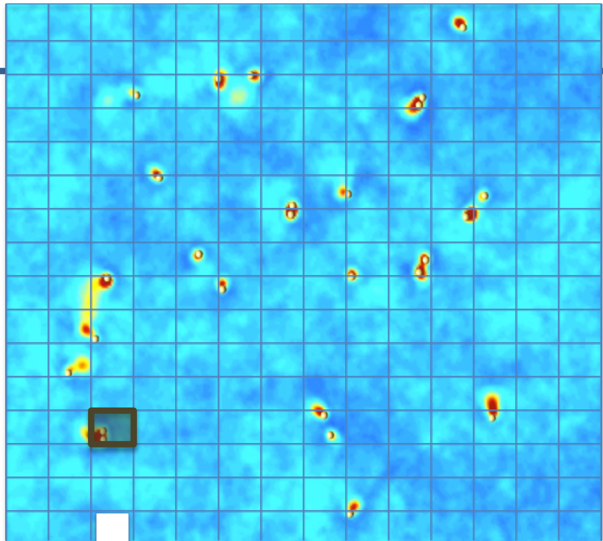


From continuum to molecular modeling

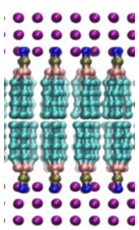


1. **Identify regions of interest using AI techniques**
   - Use variational autoencoder to learn reduced order model (ROM) of system
   - Map every region of simulation into this ROM
   - Rank every patch in the simulation by distance from others (”uniqueness”) in ROM
   - Identify the most unique patches as the most interesting

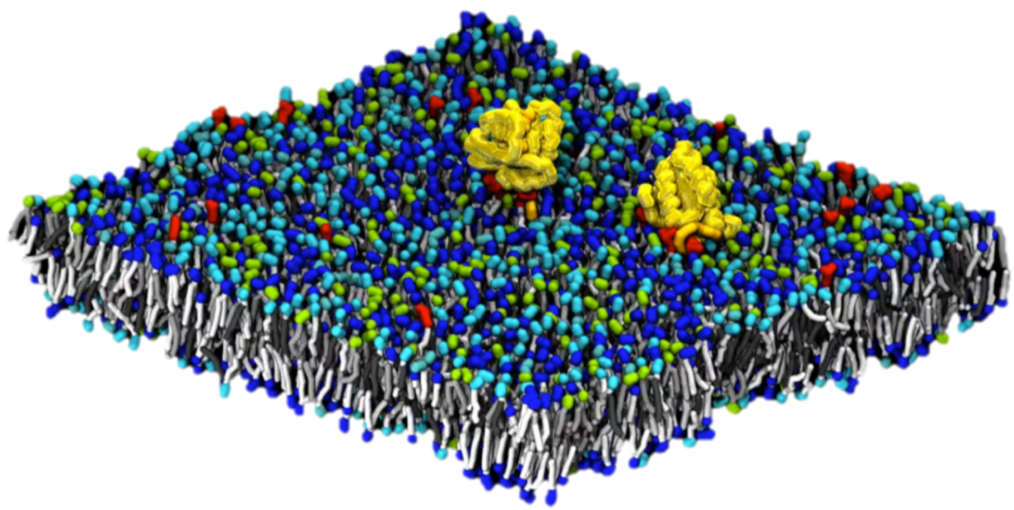
From continuum to molecular modeling



1. **Identify region**

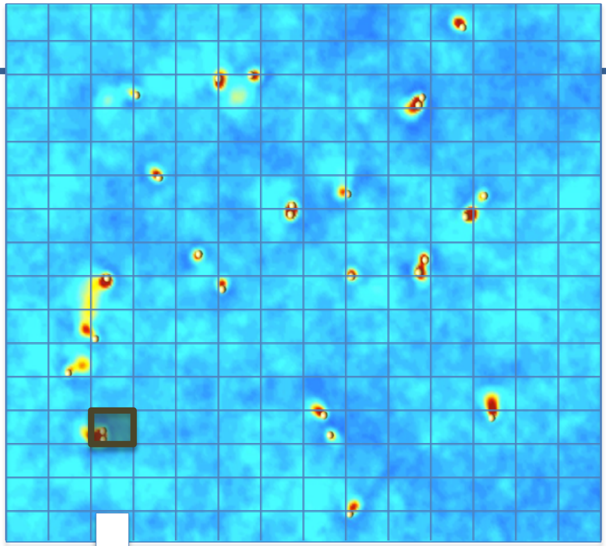


2. **Generate particle positions**

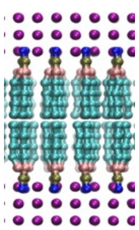


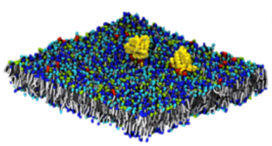
- Lipid positions are randomly generated consistent with composition
- Ras proteins, if present, are inserted with appropriate conformation
- System is equilibrated to “smooth edges”

From continuum to molecular modeling



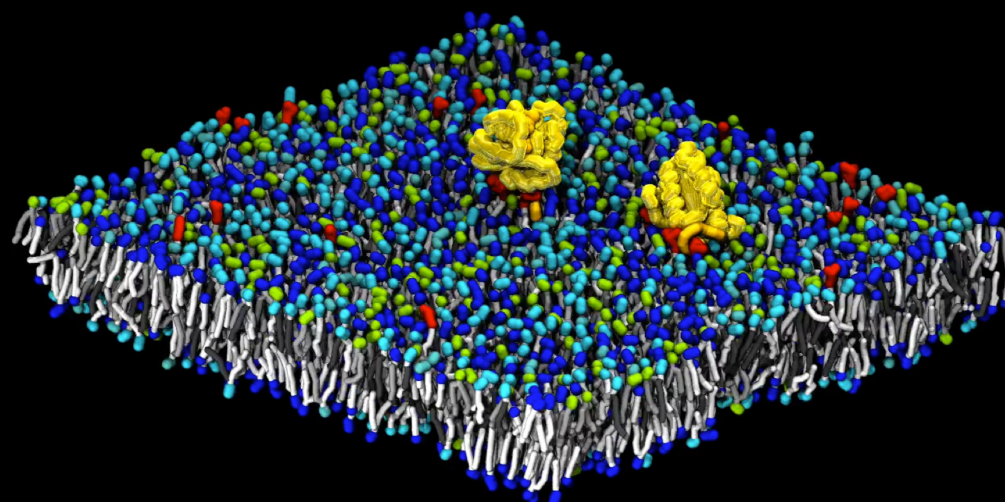
1. Identify region



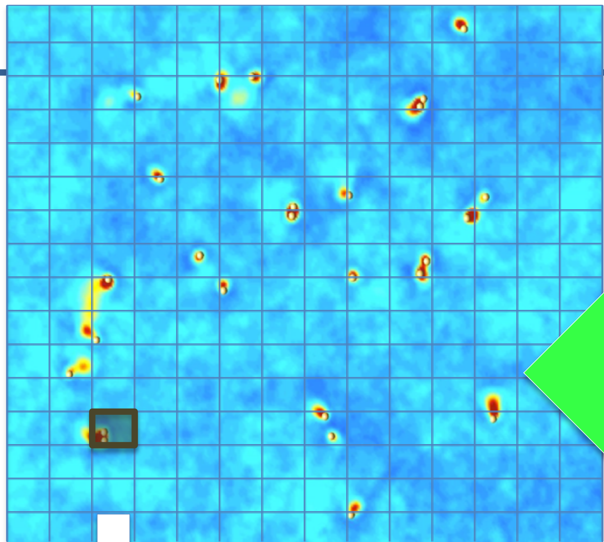


2. Generate particle positions

3. Simulate at molecular resolution

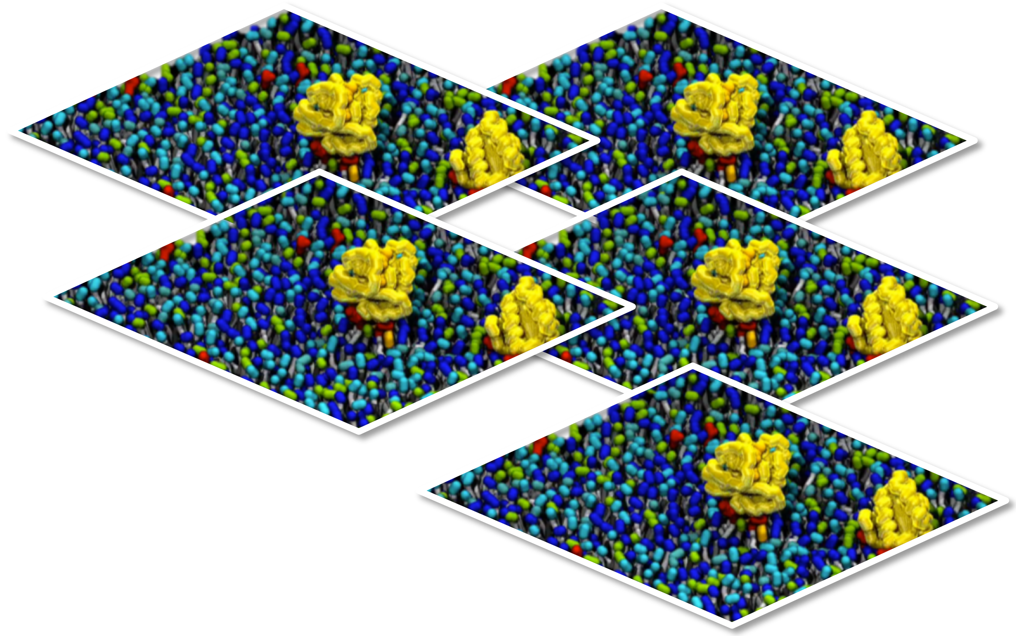


From continuum to molecular modeling

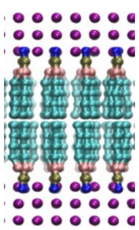


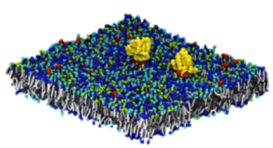
3. Simulate at molecular resolution
4. Feedback molecular information to continuum model

Feedback



1. Identify region

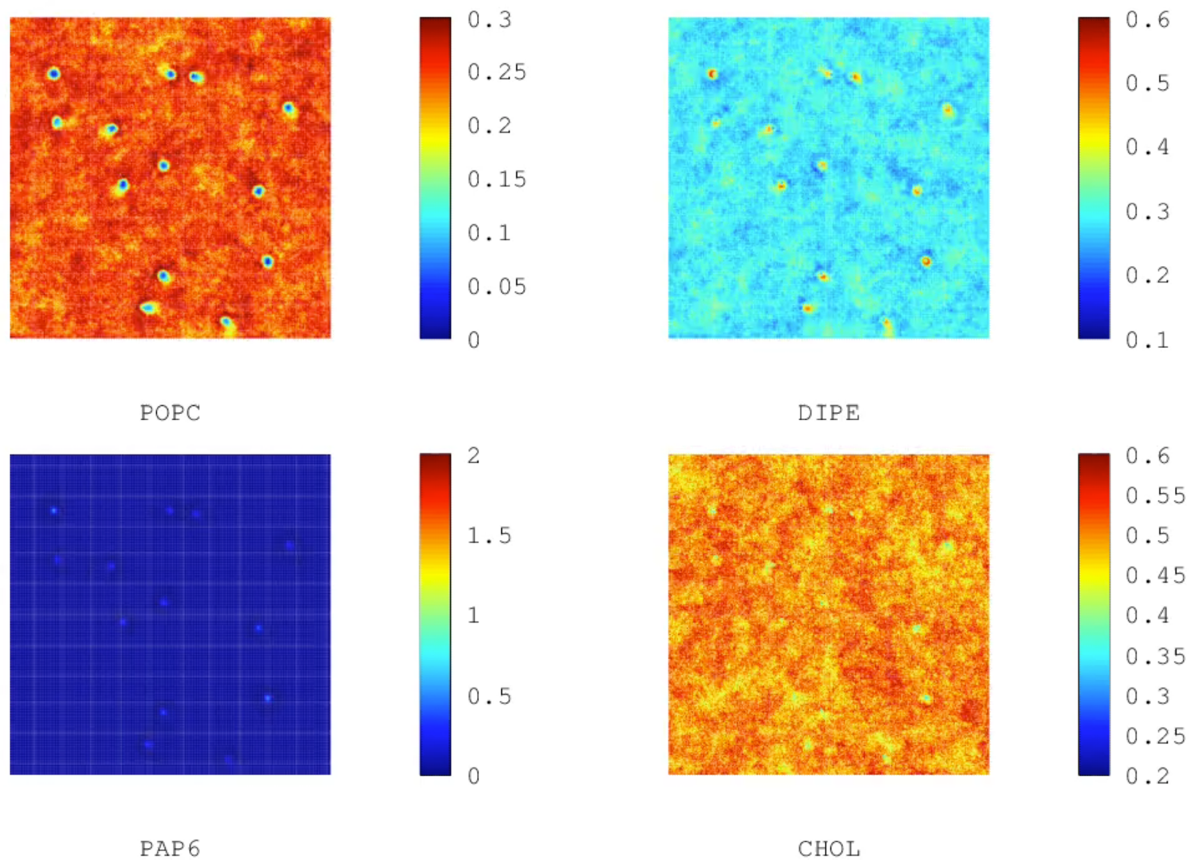




2. Generate particle positions

Preliminary Results

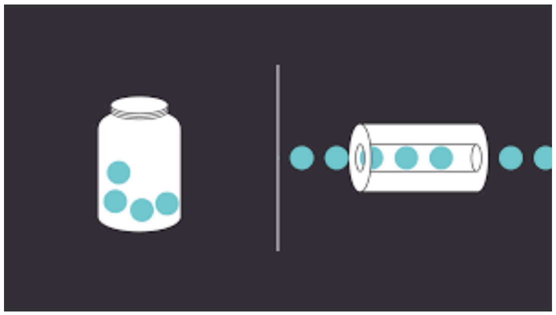
- 50nm X 50nm high-res study w/ 2 RAS proteins (40 µs)
- Investigate phenomena witnessed originally in µm X µm scale simulation
- Aggregation/repulsion of charged lipids (PAP6, PAPS, DIPE, CHOL) following “collision” of RAS
- Results demonstrate importance of time and length-scale for simulation



The Neutron Lifetime on Sierra Early Science

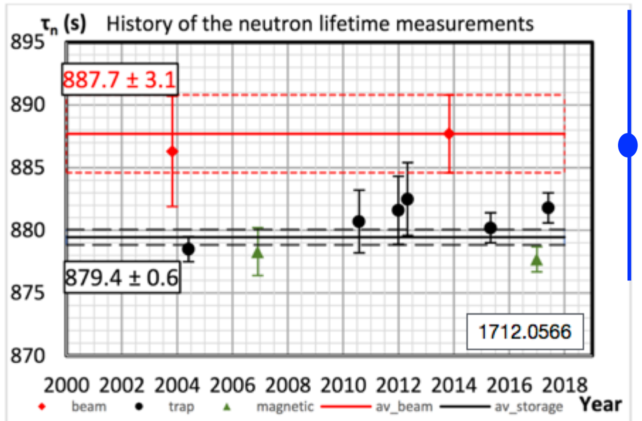
Pavlos Vranas LLNL, André Walker-Loud LBNL, et. al. (CalLat Collaboration)

- The lifetime of a neutron is ~15 min. (~881 sec), and its value has a profound effect on the mass composition of the universe.



bottle
τ = 879.4 ± 0.6s

beam
τ = 887.3 ± 3.1s



In the known theory of physics (QCD) neutrons decay to protons. Beam experiments count how many protons emerge from a beam of neutrons.

Bottle experiments trap neutrons in a "bottle" and measure how many are left after a period of time. If neutrons also decay to other new unknown particles (? dark matter ?) results will not agree with the known theory of physics (QCD).

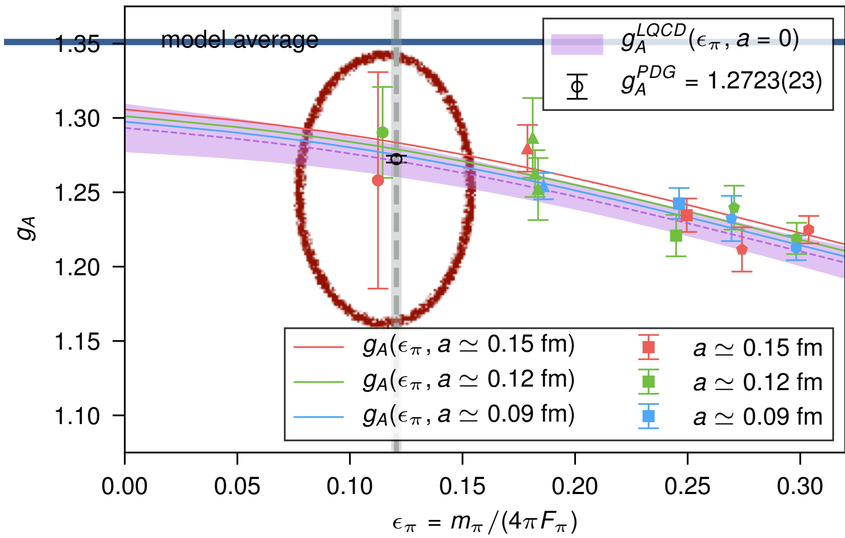
- These two methods result in measurements of the lifetime that have a 99% probability of being incompatible - why?
  - Is there an unaccounted for systematic?
  - Or more exciting, is there new physics causing them to disagree?
- Which method is consistent with the prediction from the Standard Model?
  - Answering this question requires access to leadership class supercomputers

- Using Sierra the team simulated the fundamental theory of Quantum Chromodynamics (QCD) and calculated the lifetime to an unprecedented theoretical accuracy (blue bar). This preliminary result has an **uncertainty that is 4 times smaller** than the next best result in the field (equivalent to a factor of 16 times more data).
- This early science time allowed us to form a concrete plan to further reduce the uncertainty and achieve a **discriminating** level of precision in the theoretical prediction.

The Neutron Lifetime on Sierra Early Science

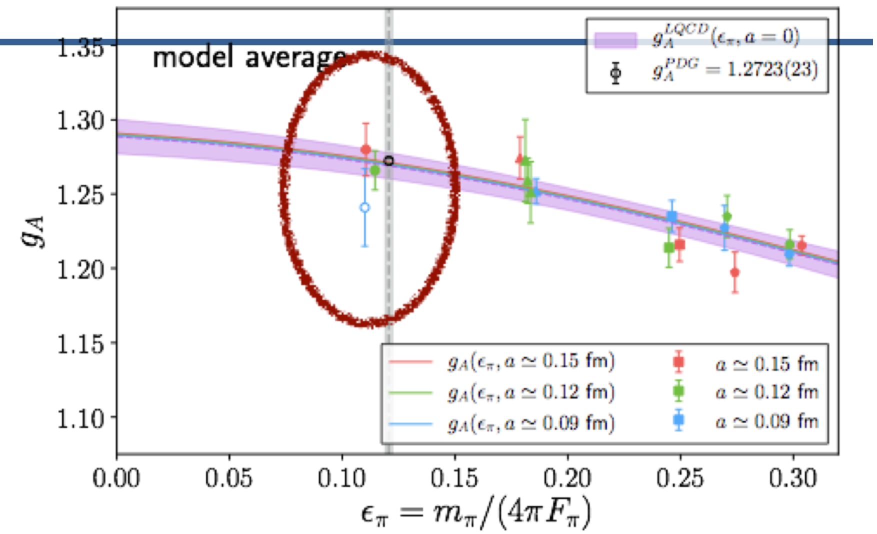
Pavlos Vranas LLNL, André Walker-Loud LBNL, et. al. (CalLat Collaboration)

California Lattice Collaboration



Nature 558 (2018) no. 7708, 91-94

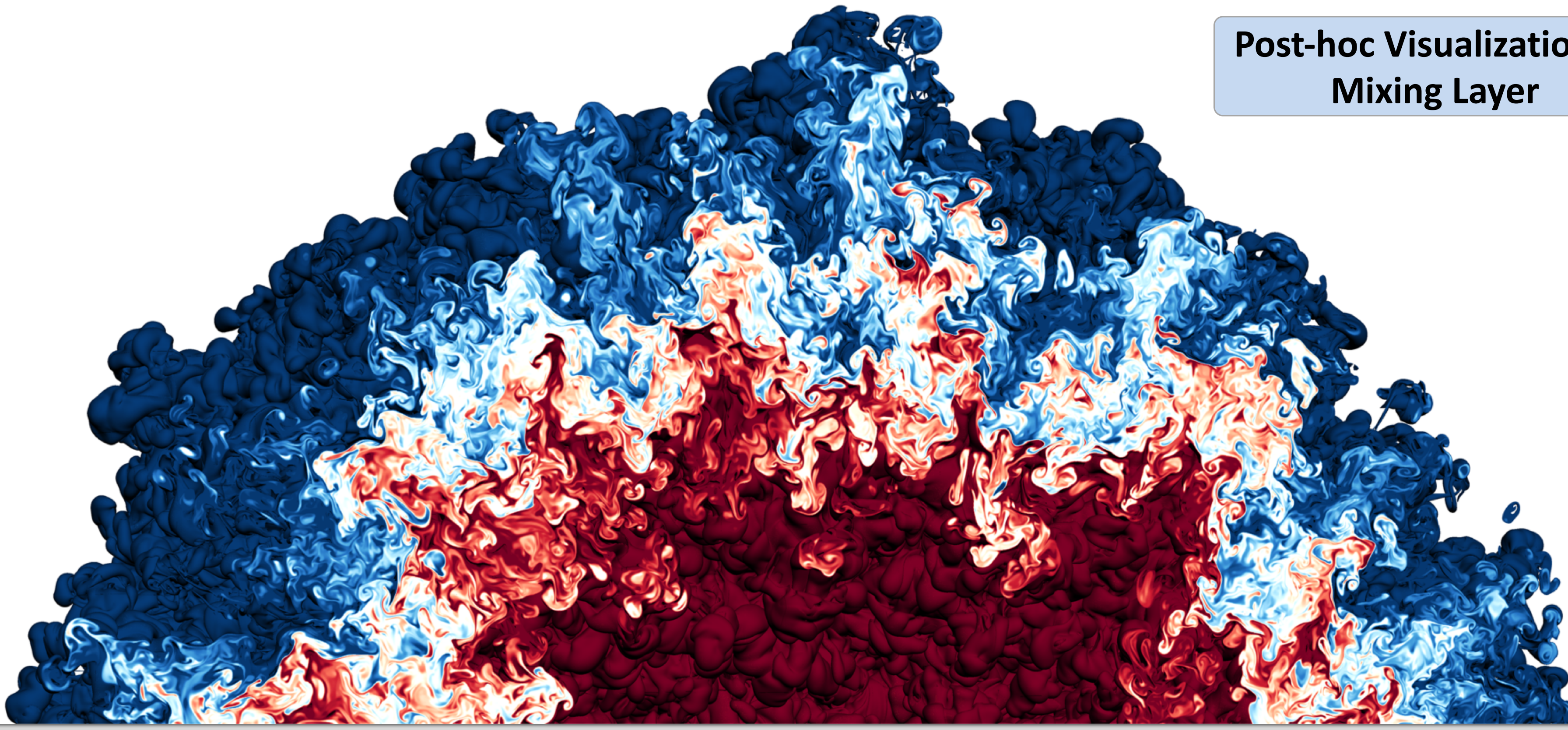
Sierra



Sierra Early Science (preliminary)

- The vertical gray band denotes the physical value of pion mass - these points are significantly more expensive than the rest to compute - but the most valuable for the final predictions
- The green point in our publication cost as much computing time as all the other points combined
- The green point from Sierra has **10x** more statistics than our publication
- The red point from our publication was not useful
- The red point from Sierra came from an entirely new calculation and is now very useful
- The blue point from Sierra was entirely unattainable from previous computers (it still needs more statistics to be useful)

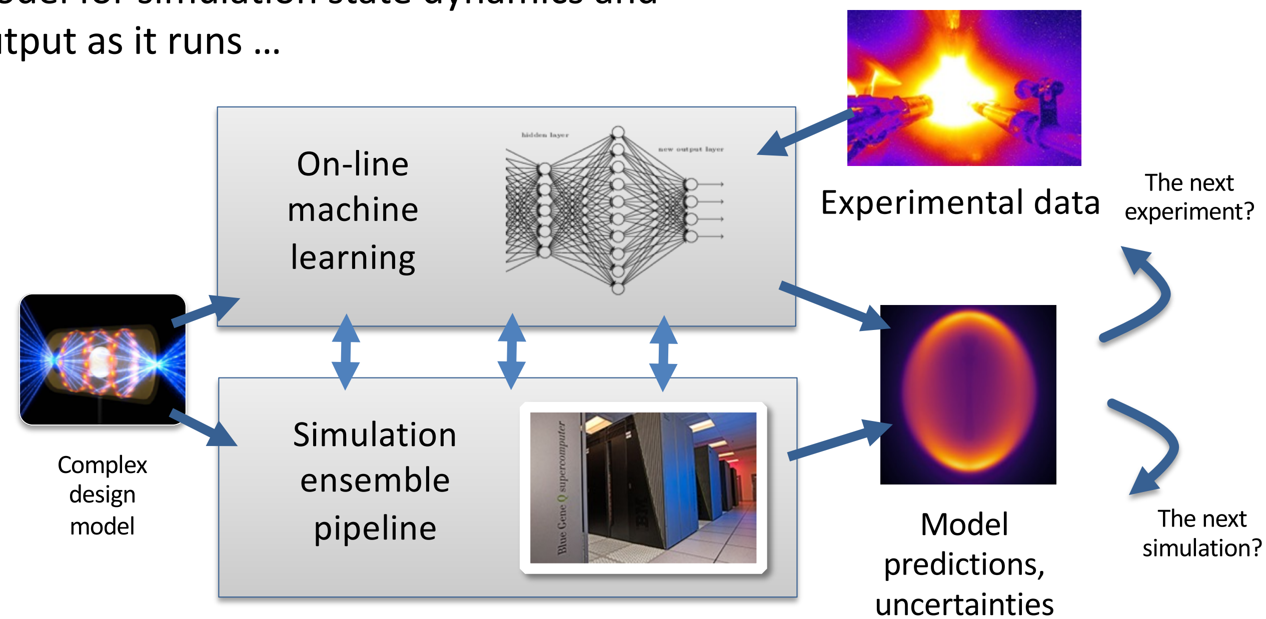
We ran a massive turbulent fluid mixing simulation on Sierra in October 2018



Post-hoc Visualization of Mixing Layer

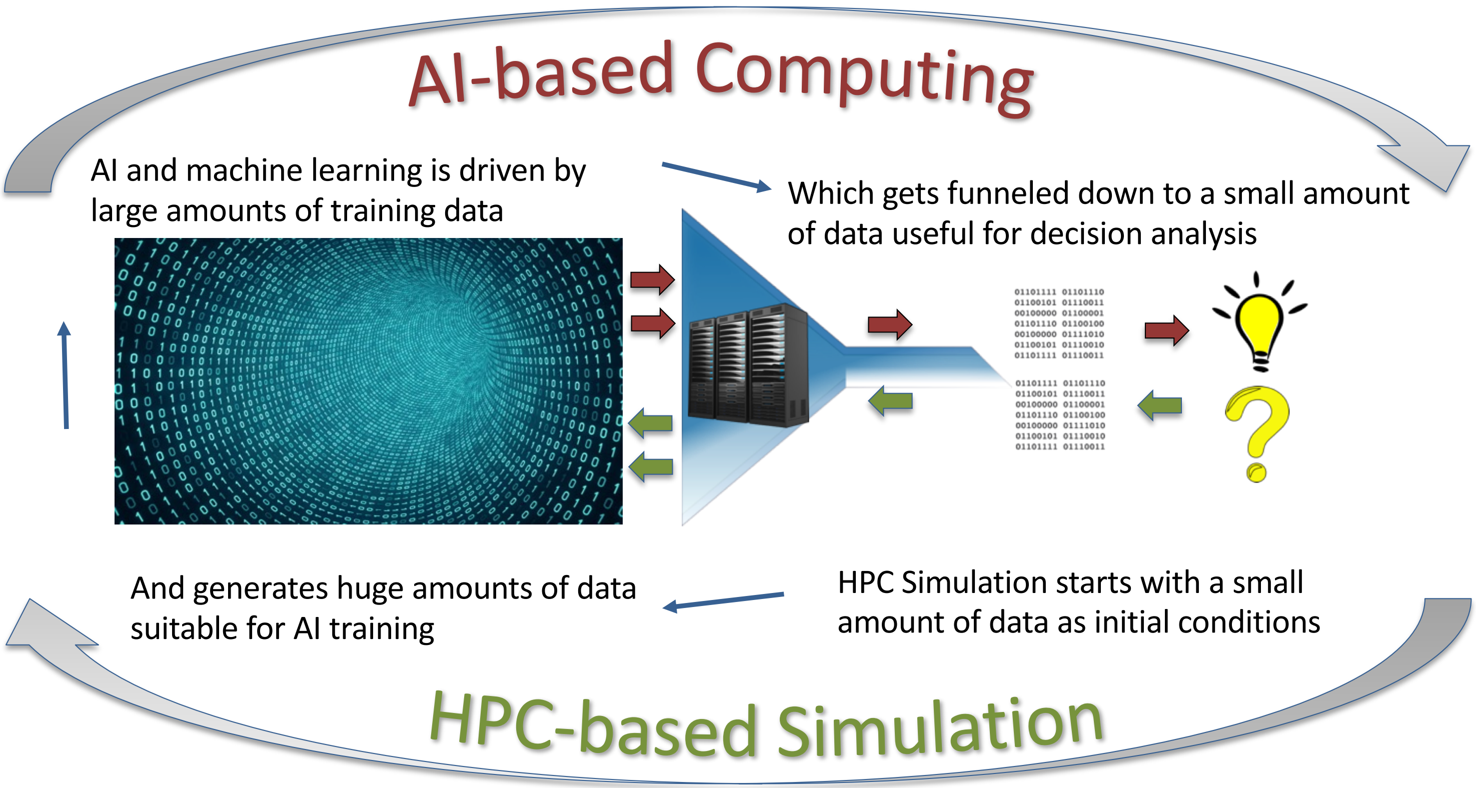
Future simulations for complex systems will be guided by machine learning

Cognitive simulation learns an approximate model for simulation state dynamics and output as it runs …



We're developing an integrated roadmap for NNSA and LLNL in cognitive simulation

The Virtuous Circle of HPC Simulation and Machine Learning



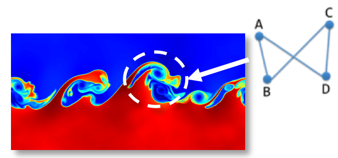
Cognitive simulation integrates machine learning with simulation at multiple levels

Learning in the simulation loop →

Molecular dynamics systems that decide when and where to switch resolution

→ **100X longer simulation time spans**

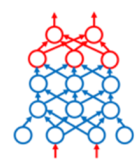
Learning on the simulation loop →

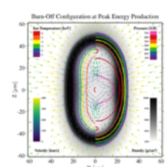


Simulation systems that can predict and correct mesh tangling

→ **Greatly reduced time on mesh tuning**

Steering the simulation loop →

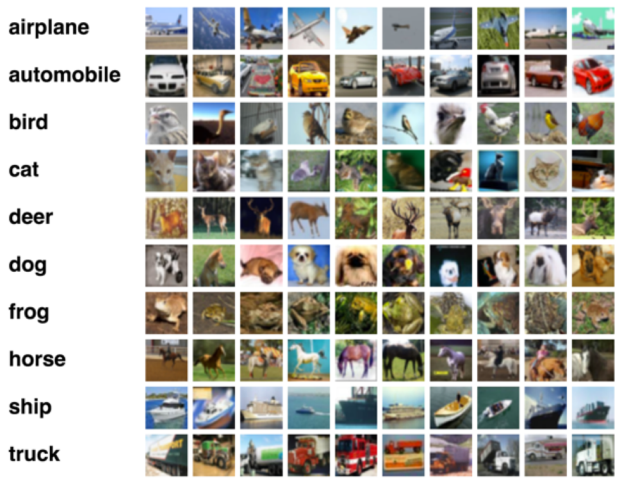




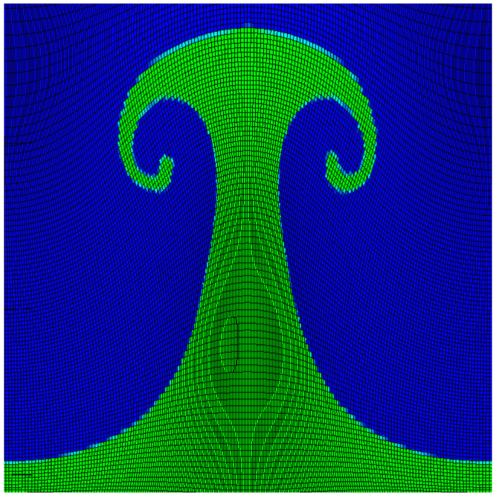
Learned models integrating simulation and experiment

→ **Improved prediction, Efficient UQ, New designs**

Scientific computing challenges differ from industry challenges



- Different type of scaling
- Data augmentation techniques
- Image classification defined by fine features
- Laws of physics



Goal is to leverage industry advancements and specialize when needed.
Validation needs to be a part of this!

Open Research Questions for Machine Learning in Science

Physics Constraints	• Results must honor laws conservation (e.g. energy, mass)
Sparse Data	• It is sometimes intractable to generate sufficient training data
Explainability	• Interpretability of results necessary for predictive science
Data Collection	• Lots of data with no standards for specification or sharing

These are all critical research areas to pursue if we are to demonstrate the value of AI techniques in the pursuit of predictive science

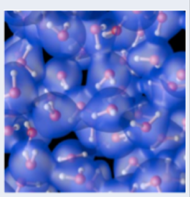
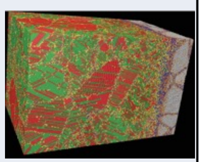
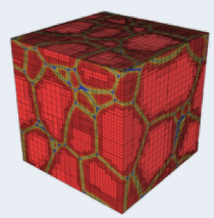
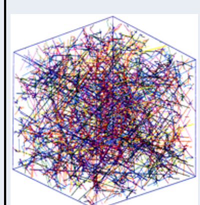
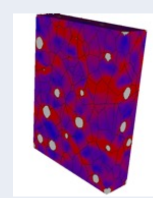
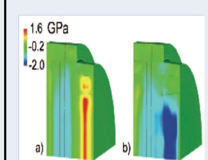
Intelligence built into HPC simulations will change how we tackle problems that have overwhelming amounts of data

Exascale Approach		AI-based Approach		Expected Benefits
Higher-fidelity subgrid models run inline with multi-physics integrated codes	→	AI mimics high fidelity models at a fraction of the computational cost	→	Applications run faster, we can increase number of simulations for UQ and validation
Analysis is performed by knowing where to look for interesting features in the data	→	Smart integrated AI identifies features of interest in throngs of data	→	Insights from exascale simulations aren't lost in a data deluge
Performance optimized in applications to be best for average case	→	On-the-fly recompilation of code with optimal settings for specific run	→	More effective use of complex exascale architectures

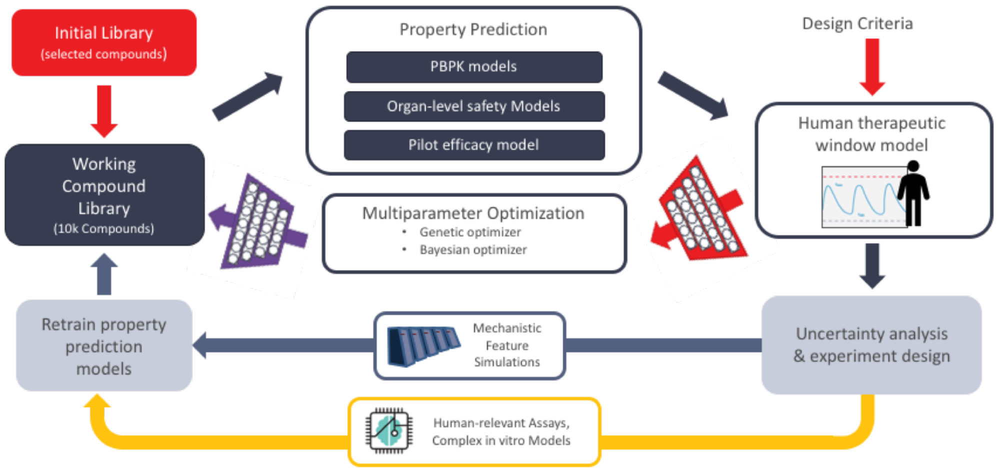
Intelligent Simulation complements our exascale investments.
Sierra will allow us to research and explore the possiblities

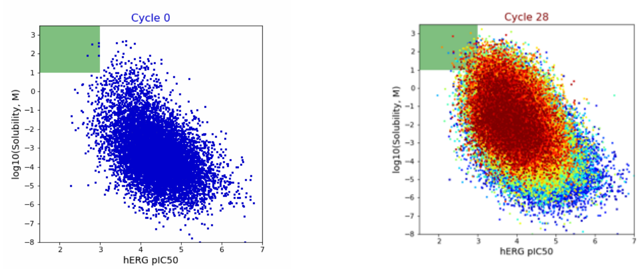
Surrogate models to increase simulation coupling for multi-scale problems

- Multi-scale problems are difficult to model
  - —Often solved using low-fidelity table lookups
  - —More accurate models are user-intensive and/or computationally expensive
- Surrogate models can be trained on high-fidelity physics models
  - —Expensive calculations may be reduced to function calls

Ab-initio	Atoms	Long-time	Microstructure	Dislocation	Crystal	Continuum
Inter-atomic forces, EOS, excited states	Defects and interfaces, nucleation	Defects and defect structures	Meso-scale multi-phase, multi-grain evolution	Meso-scale strength	Meso-scale material response	Macro-scale material response
						

The ATOM partnership is exploring new active learning approaches to accelerate drug design





Approach: An open public-private partnership

- -Lead with computation supported and validated by targeted experiments
- -Data-sharing to build models using everyone's data

Partners: LLNL, GSK, NCI, UCSF

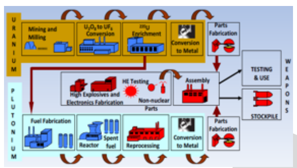
Product: An open-source framework of tools and capabilities

25 FTE's in shared Mission Bay, SF space

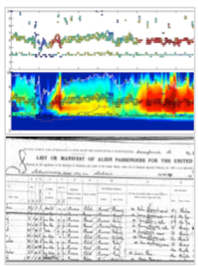
R&D started March 2018

ADAPD: Advanced Data Analytics for Proliferation Detection

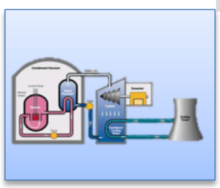
NNSA R&D Testbeds



Indirect Data

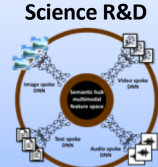


SME process and physics models



Focus on Hard Problems

Focused Data Science R&D



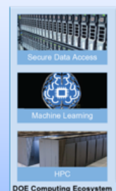
- Predictive Models
- MultiModal Detection
- Crossing Facilities and Scenarios

Leveraging Growing Industry Investments

NVIDIA IBM intel

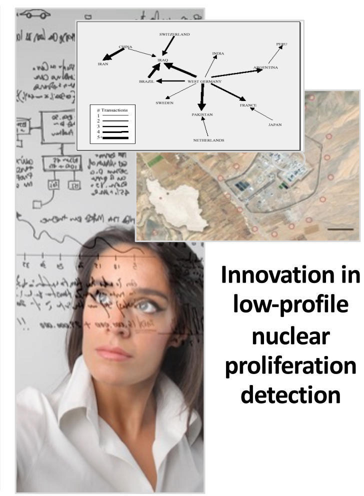
DOE/NNSA Partnerships

ASC



Student and University Programs

New and responsive analytics capabilities at the Labs

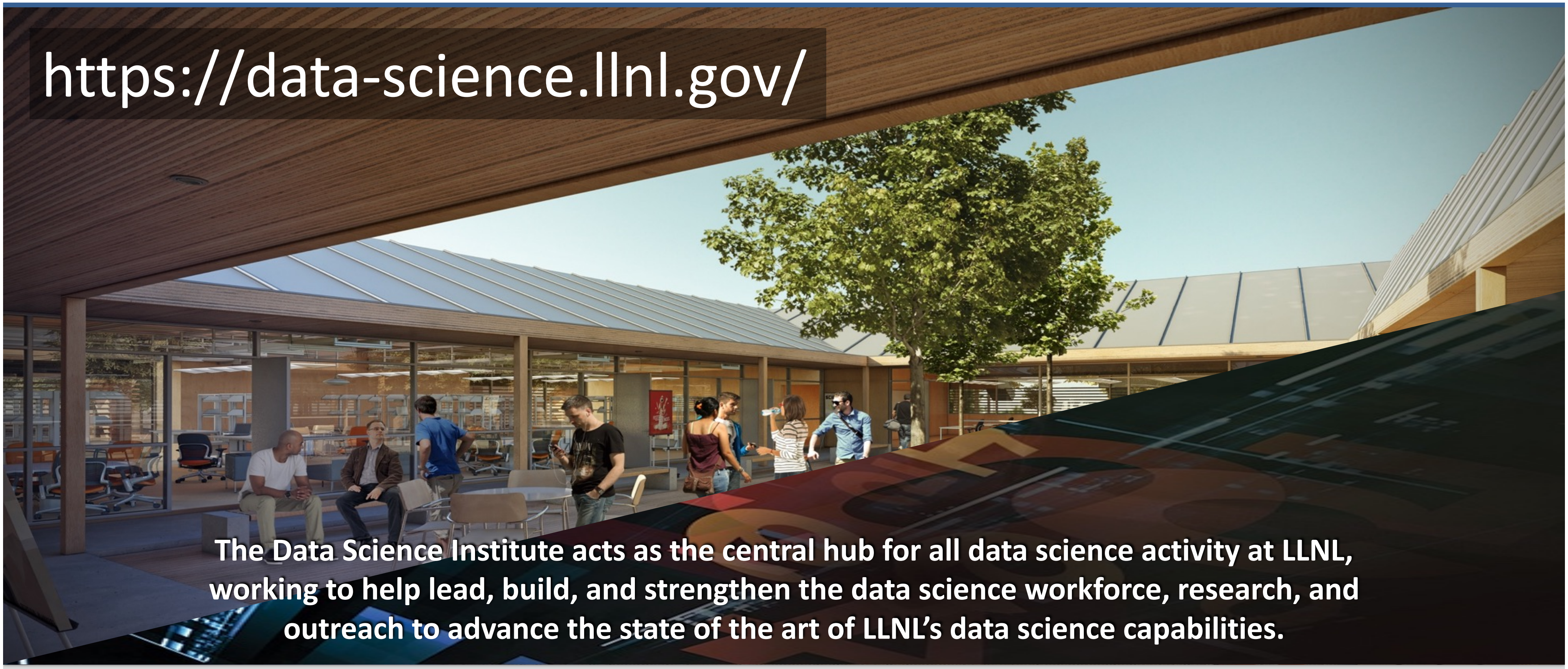


Innovation in low-profile nuclear proliferation detection

Establishing the next generation of proliferation analytics expertise

LLNL’s Data Science Institute (established 2017)

Big Machines. Big Data. Big Ideas.



Strengthening the LLNL Data Science Workforce

Fostering a sense of community

- **Strengthen and sustain LLNL's data science workforce**
- **Enhance internal coordination and communication of the data science workforce across disciplines and programs**
- **Promote and expand the visibility of data science work at LLNL**
- **Evolve strategic data science vision and guide S&T investments**

Data Science Council

- Inform S&T investments
- Communication

Workforce Development

- Reading groups
- Courses
- Summer Program

Web Presence

- External/Internal website
- Mailing list
- Slack

Outreach

- Academic engagement
- Info sessions
- Recruiting

Seminar Series/ Workshop

- Invited speakers
- UC/LLNL multi-day workshop

Data Science Summer Institute

The Focus is the Future

12-week flexible internship focused on data science

Students funded 50% to work on solutions to challenge problems, attend courses and seminars, and strengthen skills.

1000+ applicants in FY18

50 total students in FY17+FY18

- 26 students in FY18
- 24 students in FY17

Visiting Faculty

- James Flegal, UC Riverside
- Robert Gramacy, Virginia Tech

Challenge Problems

- Topology Optimization
- Machine Vision
- Multimodal Data Exploration
- Cyber

Conclusions

- The 125 PF Sierra system is currently being stood up for use in the Stockpile Stewardship mission of the NNSA
- Preparation for the pivot to heterogenous computing was difficult, but is demonstrating significant benefits in computational results
- The Open Science period for Sierra allowed open applications significant allocations to demonstrate cutting-edge science results
- LLNL is pursuing an application strategy that will encompass both traditional exascale computing needs, as well as the burgeoning AI/ML trends.

The following two talks will complete the story – stay tuned!

CASC

Center for Applied Scientific Computing

Lawrence Livermore National Laboratory