Experiences using Accelerators at ORNL
Application Readiness, Early Science, and Industry Impact

John A. Turner

Group Leader
Computational Engineering and Energy Sciences
Chief Computational Scientist
Consortium for Advanced Simulation of
Light-Water Reactors (CASL)

HPC User Forum
Sept 15-17, 2014, Seattle, WA

Overview

- U.S. DOE Leadership Computing Program
- OLCF-3: The Titan Project
- Application Readiness and Early Science on Titan
- The Consortium for Advanced Simulation of Light-Water Reactors (CASL) – a U.S. DOE Innovation Hub
  - connection to Titan Application Readiness and Early Science
- Industry Impact

What is the U.S. DOE Leadership Computing Program?

- Collaborative DOE Office of Science program at ORNL and ANL
- Mission: Provide the computational and data resources required to solve the most challenging problems.
- 2-centers / 2-architectures to address diverse and growing computational needs of the scientific community
- Highly competitive user allocation program (INCITE, ALCC).
- Projects receive 10x to 100x more resourc than at other generally available centers.
- LCF centers partner with users to enable science & engineering breakthroughs (Liaisons, Catalysts).

Oak Ridge National Laboratory

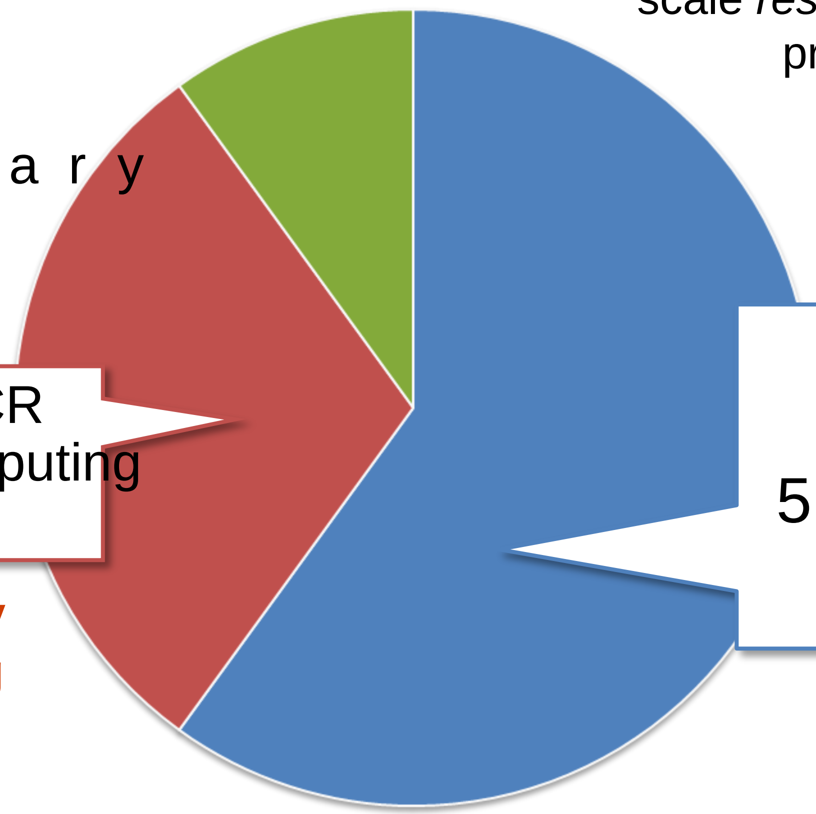
Argonne National Laboratory

Three primary ways for access to LCF

Distribution of allocable hours

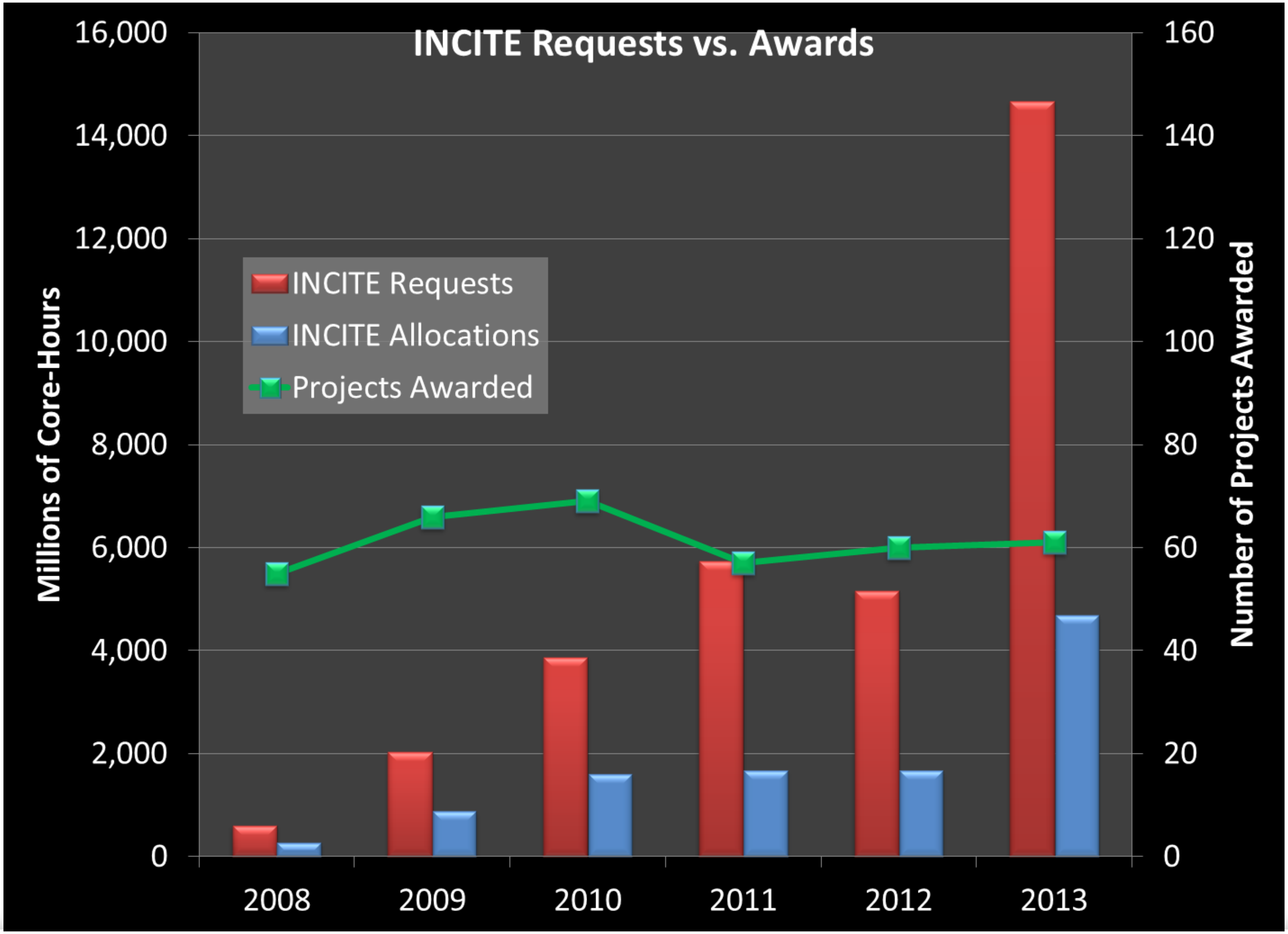
Leadership-class computing

INCITE seeks computationally intensive, large-scale *research* and/or *development* projects with the potential to significantly advance key areas in science and engineering.



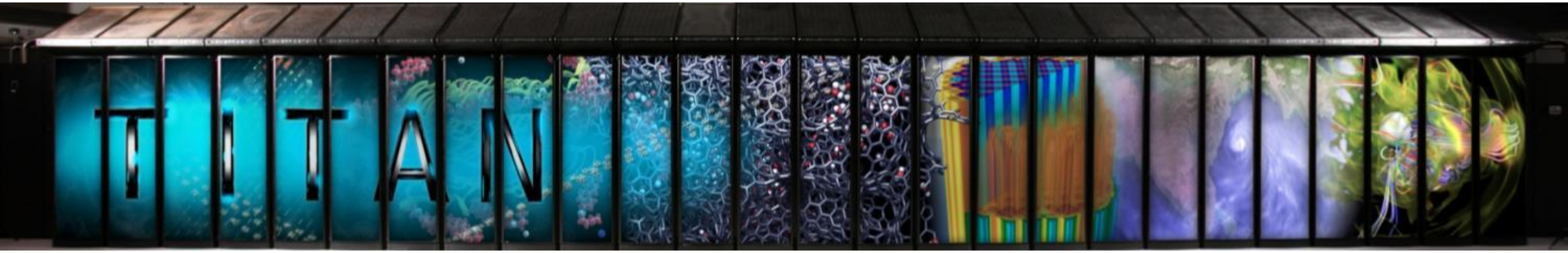
DOE/SC capability computing

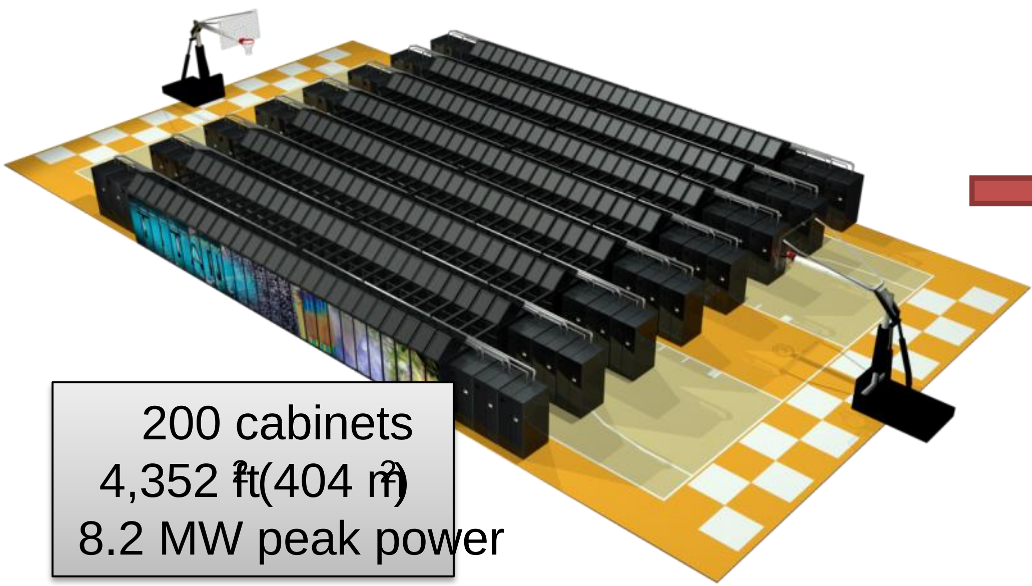
Demand for INCITE resources outstrips supply with 3x more time requested than available ðNumber of projects remains flat



ORNLõs òTitanó Hybr

Cray XK7 with AMD Opteron + NVIDIA Tesla processors





200 cabinets
4,352 ft² (404 m²)
8.2 MW peak power

SYSTEM SPECIFICATIONS:

Å27.1 PF/s peak performance

Å24.5 GPU + 2.6 CPU

Å17.59 PF/s sustained perf (LINPACK)

Å18,688 compute nodes, each with:

ï16-Core 2.2 GHz AMD Opteron 6200

ïNVIDIA Tesla K20x

ï32 GB DDR3 + 6 GB DDR5 memory

Å710 TB total system memory

Å32 PB parallel filesystem (Lustre)

ÅCray Gemini 3D Torus Interconnect

Å512 Service and I/O nodes

Throwing away 90% of availab
performance if not using GPUs

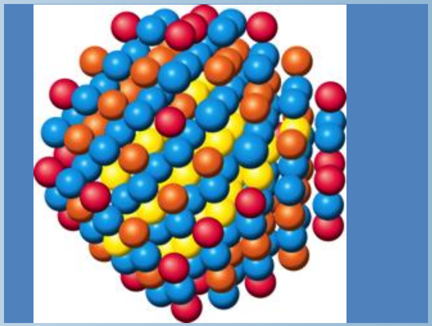
http://energy.ornl.gov/

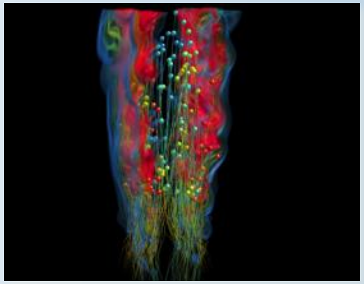
Center for Accelerated Application Readiness (CAAR)

- We created CAAR as part of the Titan project to help prepare applications for accelerated architectures
- Goals:
  - Work with code teams to develop and implement strategies for exposing hierarchical parallelism for our users applications
  - Maintain code portability across modern architectures
  - Learn from and share our results
- We selected six applications from across different scienc domains and algorithmic motifs

Early Science Challenges for Titan

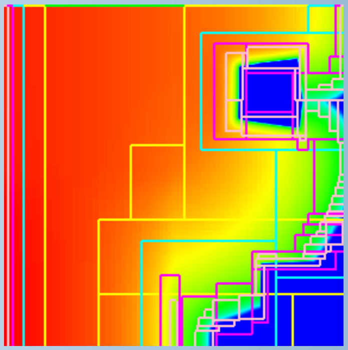
WL-LSMS
Illuminating the role of material disorder, statistics, and fluctuations in nanoscale materials and systems.

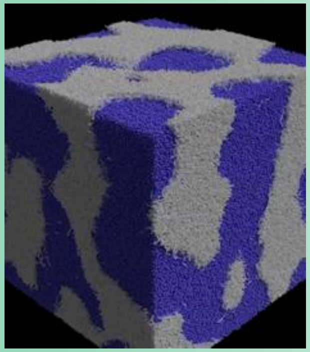




S3D
Understanding turbulent combustion through direct numerical simulation with complex chemistry.
.

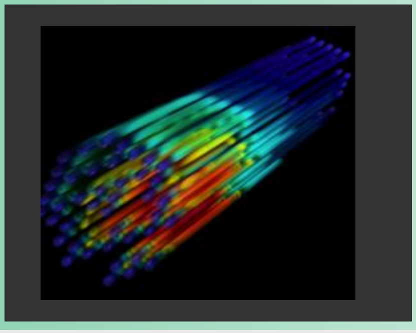
NRDF
Radiation transport ï important in astrophysics, laser fusion, combustion, atmospheric dynamics, and medical imaging ï computed on AMR grids.





LAMMPS
A molecular dynamics simulation of organic polymers for applications in organic photovoltaic heterojunctions , de-wetting phenomena and biosensor applications

CAM-SE
Answering questions about specific climate change adaptation and mitigation scenarios; realistically represent features like precipitation patterns / statistics and tropical storms.



Denovo
Discrete ordinates radiation transport calculations that can be used in a variety of nuclear energy and technology applications.

http://energy.ornl.gov/

CAAR Plan

- Comprehensive team assigned to each app
  - OLCF application lead
  - Cray engineer
  - NVIDIA developer
  - Other application, local tool/library developers, computational scientists
- Single early-science problem targeted for each app
  - Success on this problem is ultimate metric for success
- Particular plan-of-attack different for each app
  - WL-LSMS – dependent on accelerated ZGEMM
  - CAM-SE – pervasive and widespread custom acceleration required
- Multiple acceleration methods explored
  - WL-LSMS – CULA, MAGMA, custom ZGEMM
  - CAM-SE – CUDA, directives
  - Two-fold aim
    - Maximum acceleration for model problem
    - Determination of reproducible acceleration paths for others

Effectiveness of GPU Acceleration?

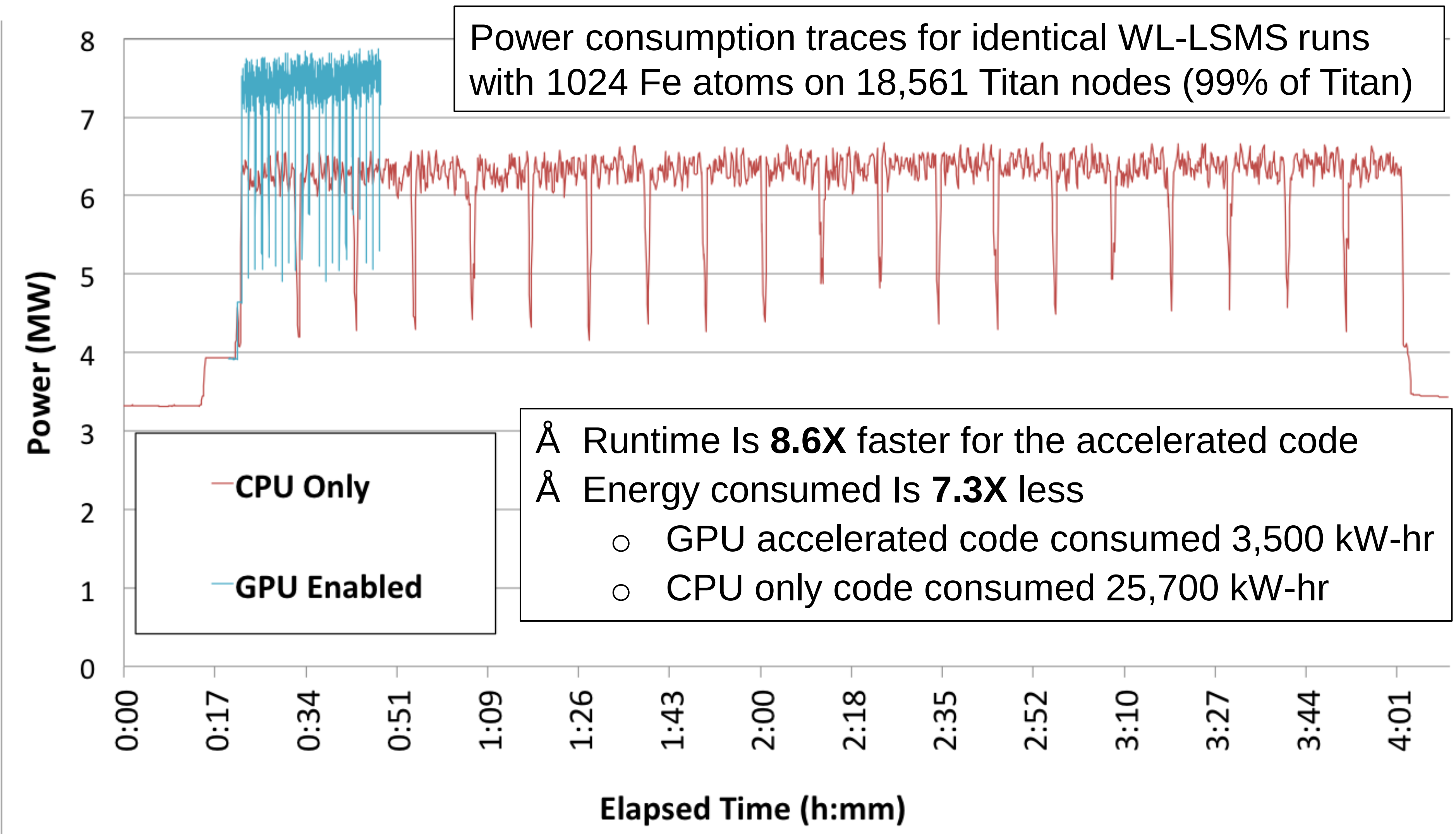
OLCF -3 Early Science Codes -- Performance on Titan XK7

Application	Cray XK7 vs. Cray XE6 Performance Ratio*
LAMMPS* Molecular dynamics	7.4
S3D Turbulent combustion	2.2
Denovo 3D neutron transport for nuclear reactors	3.8
WL-LSMS Statistical mechanics of magnetic materials	3.8

Titan: Cray XK7 (Kepler GPU plus AMD 16-core Opteron CPU)
Cray XE6: (2x AMD 16-core Opteron CPUs)
*Performance depends strongly on specific problem size chosen

Application Power Efficiency of the Cray XK7

WL-LSMS for CPU-only and Accelerated Computing



Power consumption traces for identical WL-LSMS runs with 1024 Fe atoms on 18,561 Titan nodes (99% of Titan)

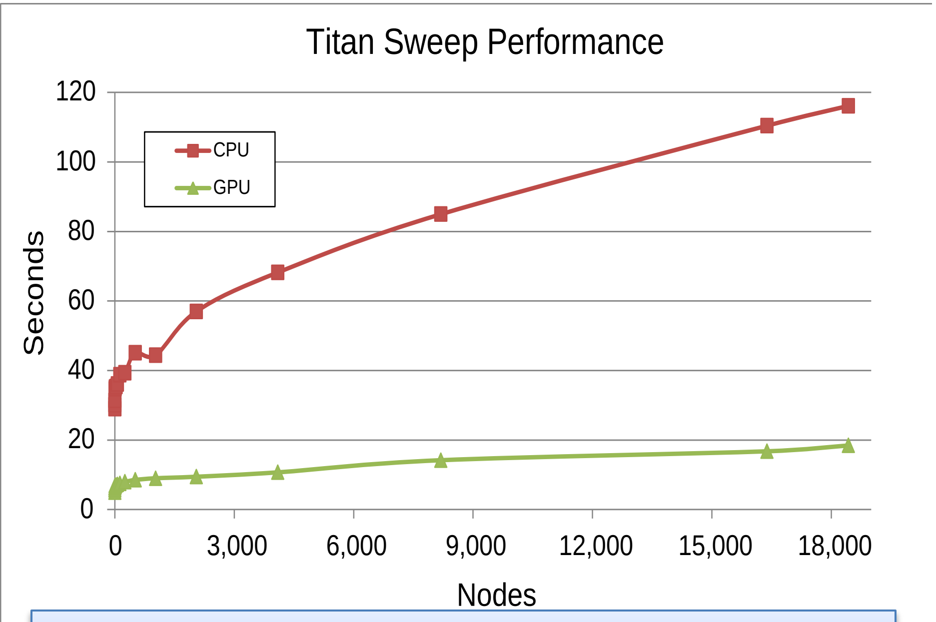
Å Runtime Is **8.6X** faster for the accelerated code

Å Energy consumed Is **7.3X** less

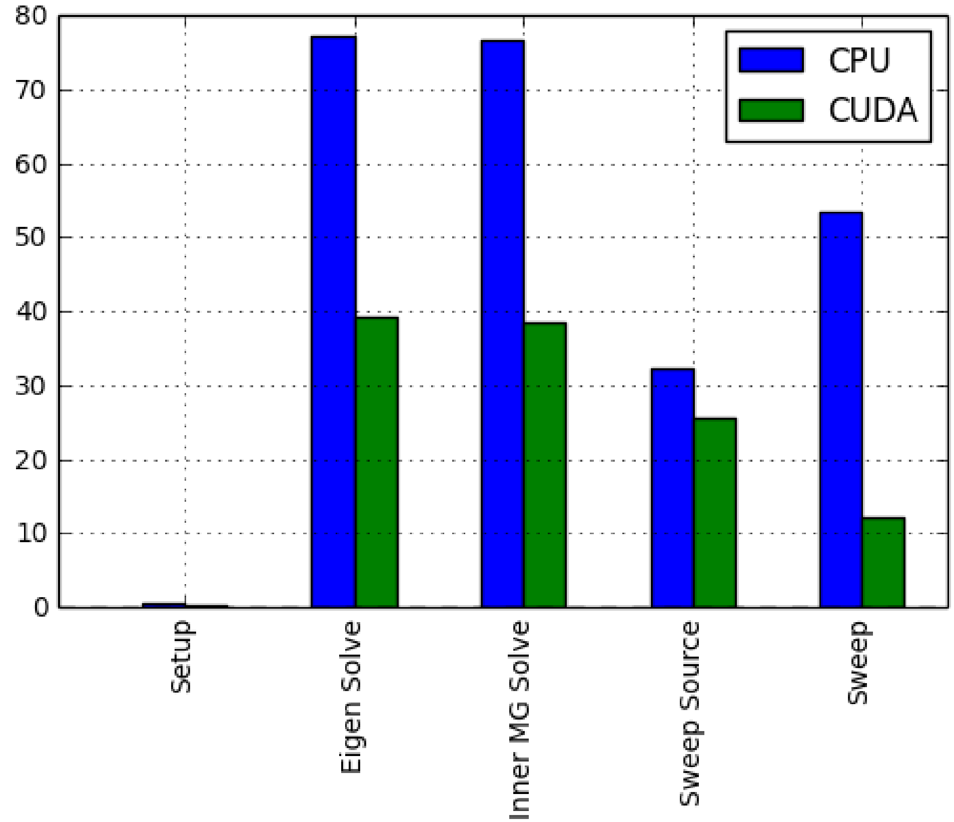
- GPU accelerated code consumed 3,500 kW-hr
- CPU only code consumed 25,700 kW-hr

http://energy.ornl.gov/

Denovo S$_N$ Acceleration



CPU vs. GPU sweeper, Titan, Kepler K20x

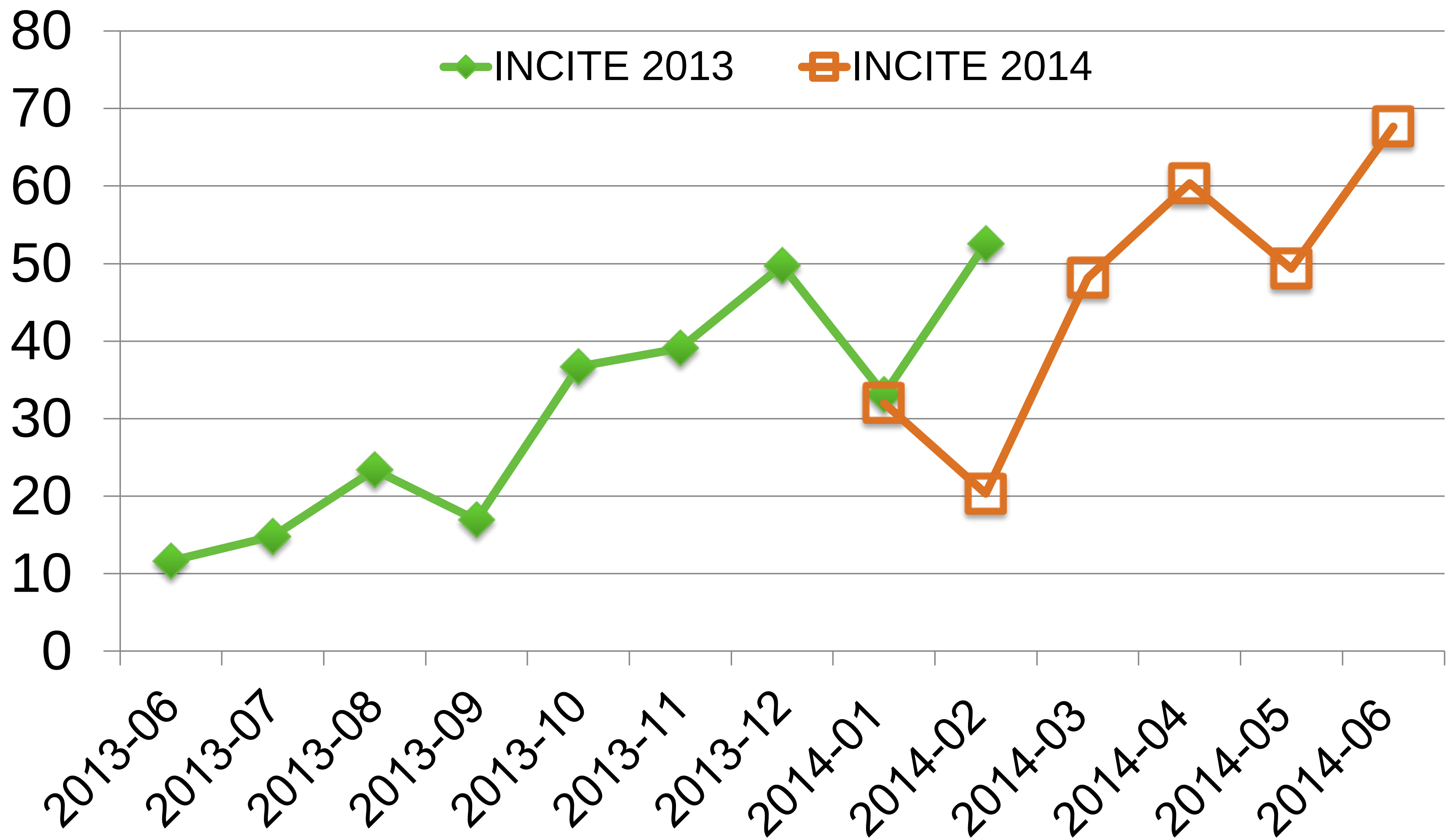


Full Denovo run, CPU vs. GPU sweeper, CPU+GPU vs. CPU only

- SWEEP kernel re-written in C++ & CUDA, runs on CPU or GPU
- Refactored SWEEP is in mainline code
- Titan: SWEEP speedup of 6-7x, Denovo speedup ~3.8x
- Scaling to over 200K cores with opportunities for increased parallelism on G
- Refactored code 2x faster on Cray XT5 (CPU only)

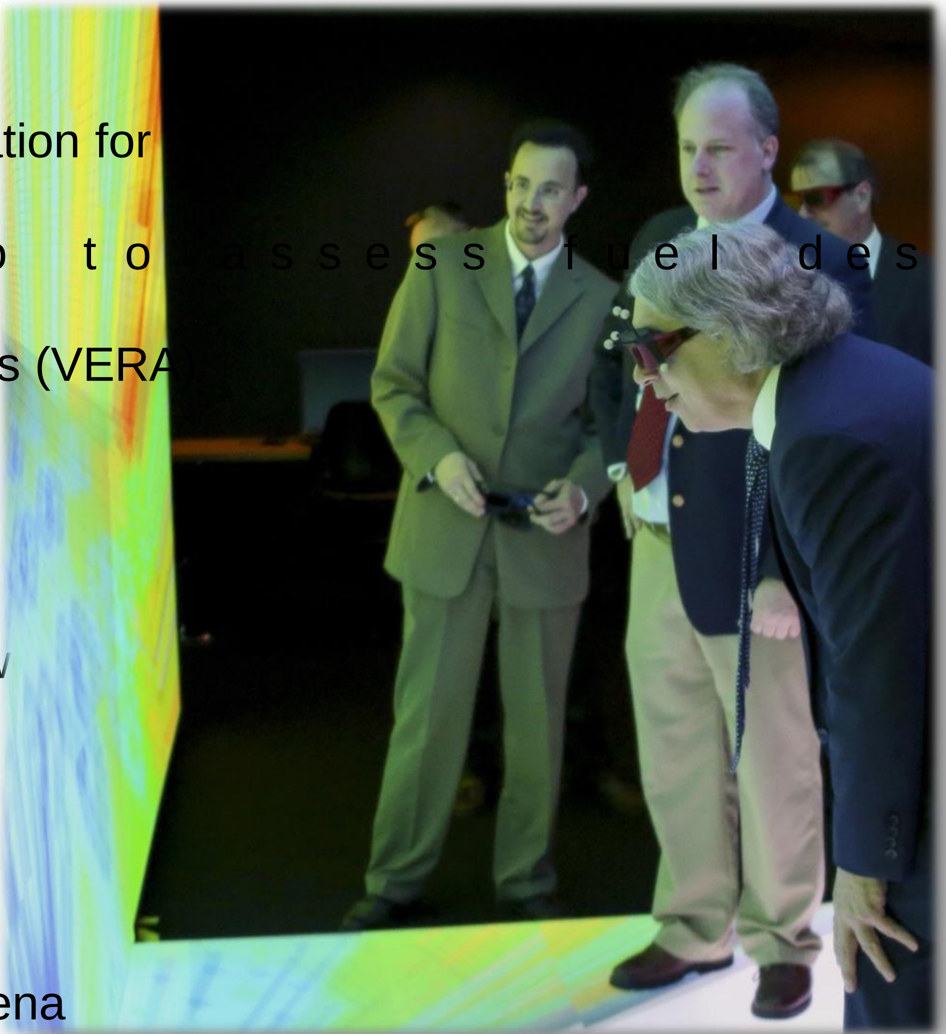
Increasing Requests for GPUs on Titan

Percentage of INCITE Time requesting GPUs



As measured by ALTD against linked libraries

Consortium for Advanced Simulation of Light-Water Reactors (CASL)



Objectives and Strategies

- DOE Innovation Hub on Modeling & Simulation for Nuclear Energy Systems
- Develop "Virtual Reactor" to assess fuel des
  operation, and safety criteria
- Virtual Environment for Reactor Applications (VERA)

Computational Science Areas

- Advanced numerical methods
- Increased coupling of physics
- Increased use of mechanistic models of low length-scale phenomena
- Large-scale software development
  - geographically-dispersed, multi-institutional

Results and Impact

- Advance understanding of key reactor phenomena
- Improve performance of today's commercial power react
- Evaluate new fuel designs to further enhance safety margins

CASL was the first DOE Innovation Hub

A Different Approach

- "Multidisciplinary, highly collaborative te... ideally working under one roof to solve ... technology challenges" – *Steven Chu*
- "Create a research atmosphere with a sense of urgency to deliver solutions" – *Kristina Johnson*
- Characteristics
  - Leadership – Outstanding, independent, scientific leadership
  - Management – "Light" federal touch
  - Focus – Deliver technologies that can change the U.S. "energy game"

Core partners

Oak Ridge National Laboratory

Electric Power Research Institute

Idaho National Laboratory

Los Alamos National Laboratory

Massachusetts Institute of Technol

North Carolina State University

Sandia National Laboratories

Tennessee Valley Authority

University of Michigan

Westinghouse Electric Compa

Contributing Partner

ASCOMP Gmb

CD-adapco

City College of New York

Florida State University

Imperial College London

Rensselaer Polytechnic Institute

Texas A&M University

Pennsylvania State University

University of Florida

University of Wisconsin

University of Notre Da

Anatech Corporation

Core Physics Inc.

G S Nuclear Consulting, I

University of Texas at Au

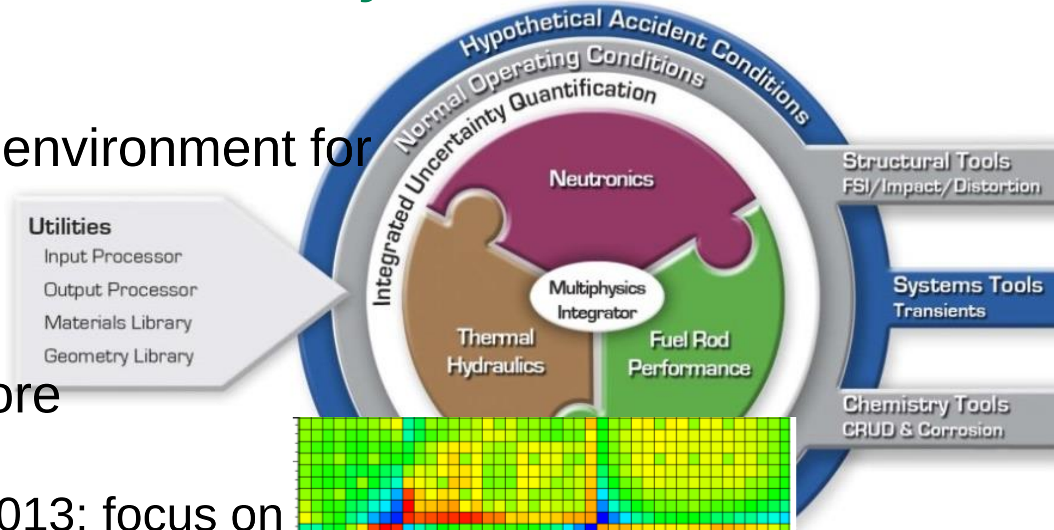
University of Texas at Da

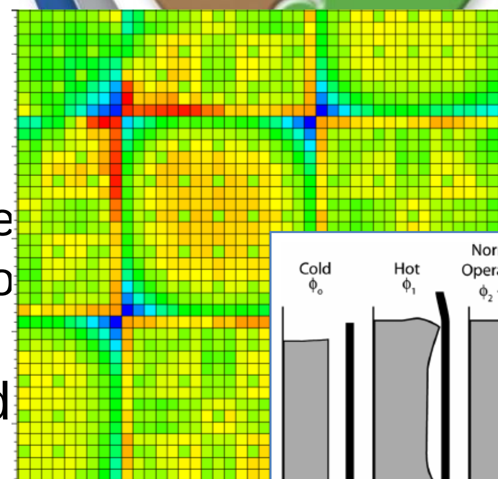
University of Tennessee – Knoxville

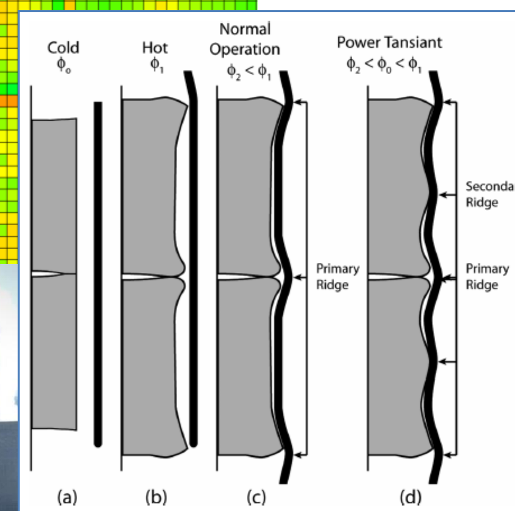
Pacific Northwest National Laboratory

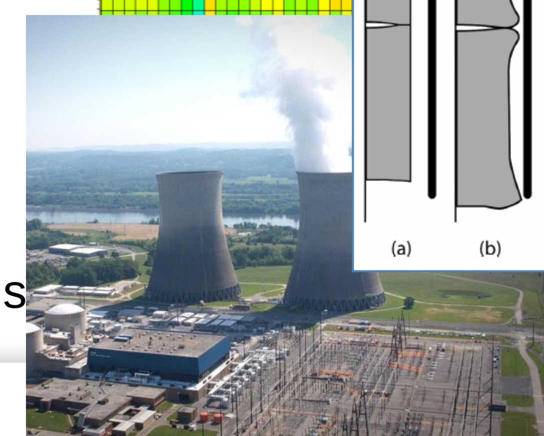
CASL Test Stands provide early deployment to industry.

- Early deployment into industrial environment for testing use, and adoption of VERA to support real-world LWR applications
- Status of initial deployment to core industry partners
  - WEC: Deployment during June 2013; focus on simulation of AP1000 first core startup
  - EPRI: Deployment Dec 2013; fuel performance
  - TVA: Deployment in progress; lower plenum flo anomaly
- Early Test Stand deployment is already prod dividends for CASL and users
  - Better code installation processes
  - Input processing for heterogeneous cores
  - Reductions in user problem setup times
  - Core tilt analysis
  - Analysis of new design features (e.g., tungs







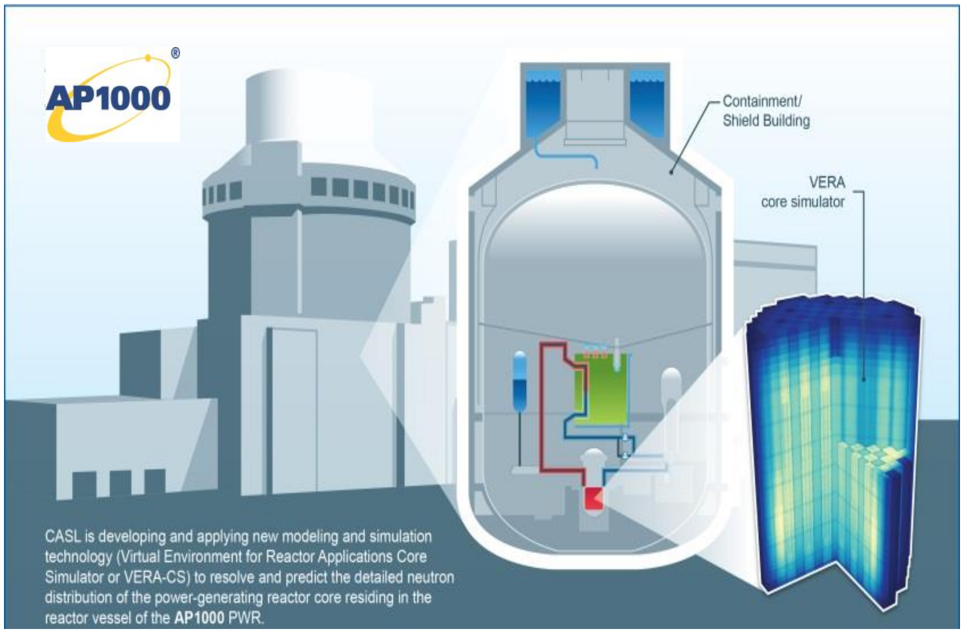


Westinghouse VERA Test Stand

CASL
Consortium for Advanced Simulation of LWRs

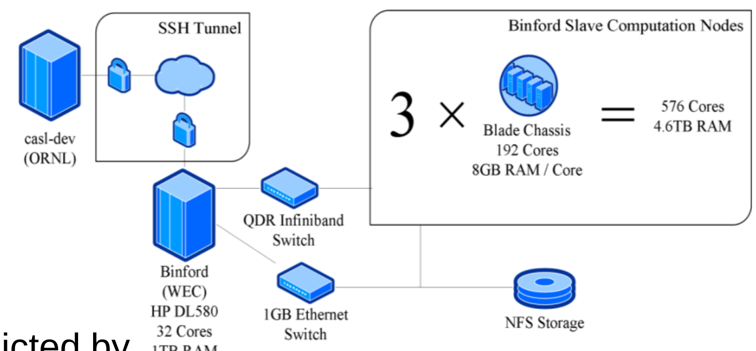
VERA deployed on Westinghouse computer cluster and employed for a high-impact industrial application:

- AP1000 PWR start-up core physics tests simulation



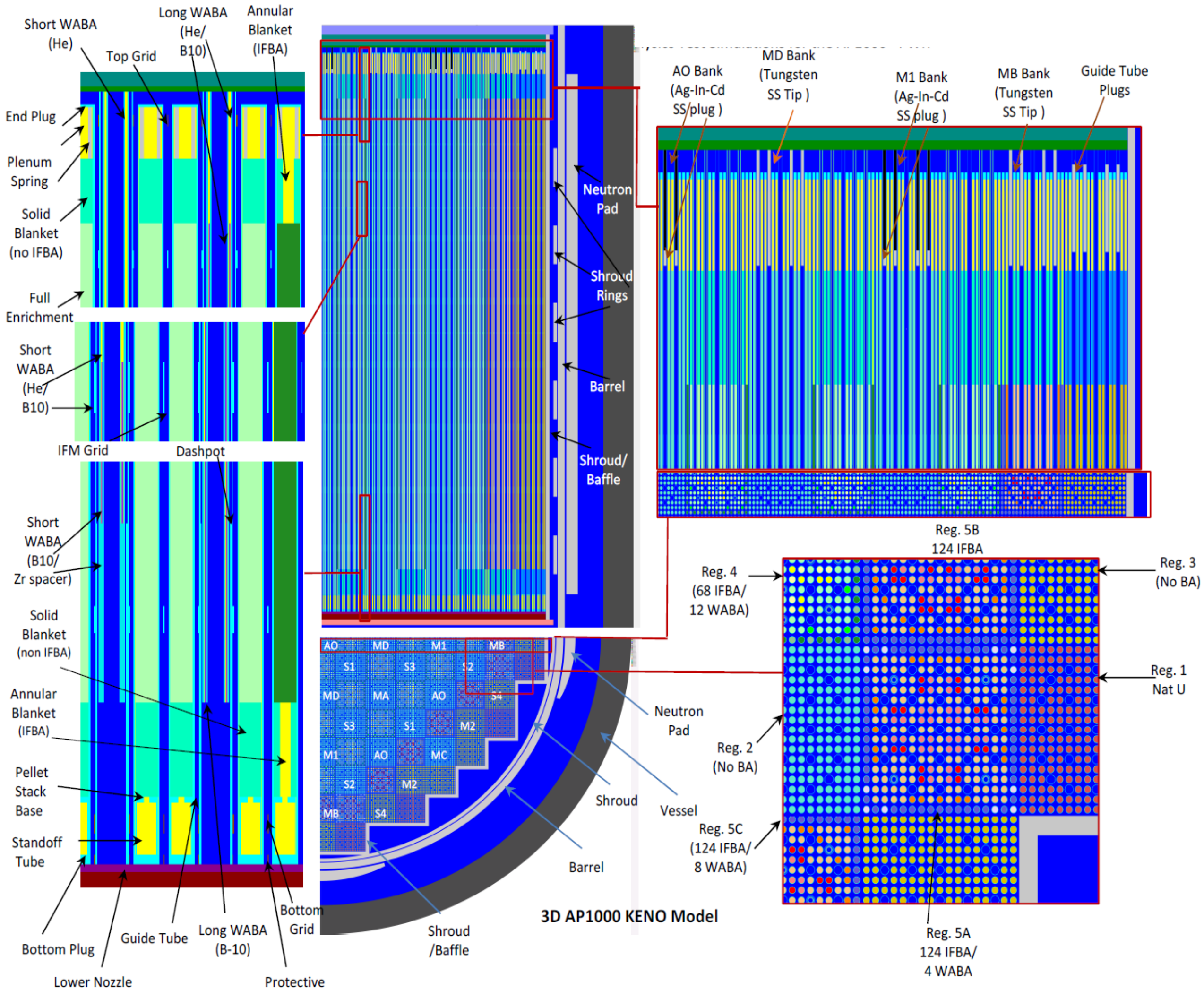
Pictorial of the AP1000 plant with the fission rate distribution predicted by VERA during one of the startup physics tests measurements. It will be possible to compare the VERA predictions against measured data when the first AP1000 unit will come on line in 2015 (Sanmen, China). All the AP1000 units under construction will feature the same startup core modeled by VERA.

VERA-CS simulations performed on a dedicated Westinghouse computer cluster where VERA ha deployed by a CASL-Westinghouse team. The graphic below shows the computer arrangem and communication to automate VERA updates f ORNL central repository as new capabilities are bei added.



The Westinghouse VERA build is operational and exercised by Westinghouse personnel.

AP1000 Advanced First Core Model



The AP 1000 features an advanced first core with sev enrichments and fuel heterogeneities which allow quickly achieve equilibrium cycle after fuel shuffle and reload.

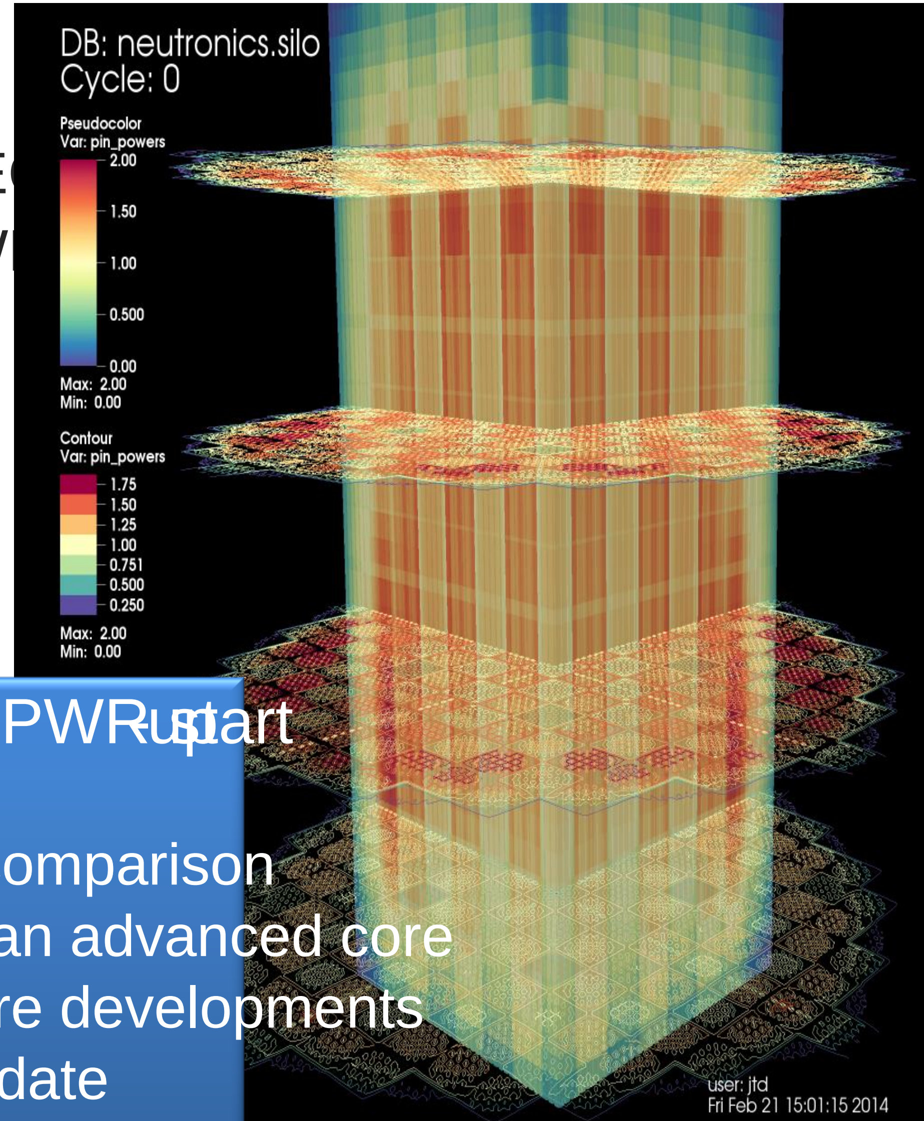
The advanced first core is a major economic advantage since it reduces the number transition cycles before equilibrium but it also poses challenges to the simulation

VERA high-fidelity physics provided an ideal match for simulating this challenging core and gain confidence in the start-up predictions

AP1000 Monte Carlo model sho*Thanks to VERA common input, it has been possible to generate a complete AP1000 core model within a compact and intuitive input*.

Timeline for CASL Westinghouse (WEC) Test Stand

- early 2013: Test Stand discussion
- April 2013: Scope proposed by WE
- June 2013: VERA deployment at W
- July-Nov 2013: Technical analysis
- Jan 2014: Analysis completed and documented (Mar 2014)



Enhanced confidence in AP1000 PWR start-up predictions

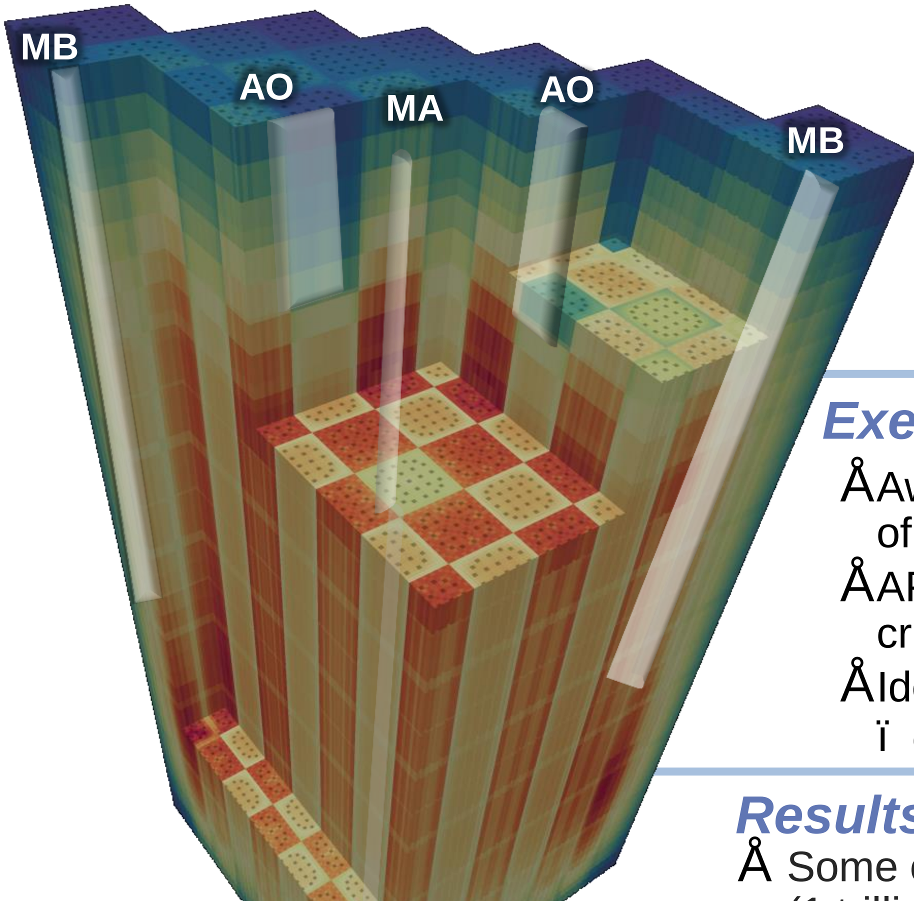
- High-quality benchmarks for code comparison
- Expanded application of VERA to an advanced core
- Feedback from WEC to guide future developments
- Framework for VERA build and update

Approaches to neutronics balance accuracy and computational requirements. As part of Titan Early Science, we compared 3 methods (blue rows below).

Method	Attributes	Code	Cross Sections	Energy	Scattering	Lang., cost	Scalability
Simplified P_N (SP_N)	Cartesian mesh, Lin Syst.	Insilico	pin-homogenized	multigroup	P_N	C++, low	linear solver dependent
Discrete Ordinates (S_N)	Cartesian mesh, Wavefront	Denovo	self-shielded by region	multigroup	P_N	C++, high	>200,000
Method of Characteristics (MOC)	unstruct mesh, Ray tracing	MPACT	subgroup	multigroup	P_N	Fortran, medium	in testing
Monte Carlo	CAD, particle tracking	Keno-V	evaluated	continuous	continuous	Fortran, very high	few hundred
Monte Carlo	CAD, particle tracking	Shift	evaluated	continuous	continuous	C++, very high	>200,000

As part of the OLCF CAAR effort, Denovo was GPU-accelerated via CUDA.

Evaluation of Shift: VERA Continuous-Energy Monte Carlo Quarter-Core Zero Power Physics Test



Monte Carlo prediction of power distribution for an AP1000 with multiple control rod banks inserted (AO, MA, and MB).

Goals

- Compare fidelity and performance of Shift against Keno, SP_N, and S_N (Denovo)
- Generate high-fidelity neutronics solution for code comparison of solutions for predicting reactor sta and physics testing

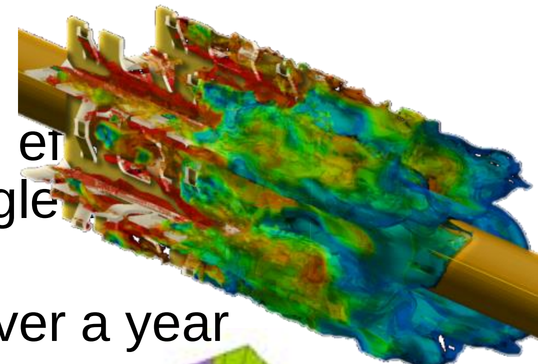
Execution

- Awarded 60 million core-hours on Titan (worth $2M) as part of Titan Early Science program
- AP1000 model created and results generated for reacto criticality, rod worth, and reactivity coefficients
- Identical VERA Input models used for Shift, SP_N, and S_N
  - dramatically simpler than KENO input model

Results

- Some of the largest Monte Carlo calculations ever perform (1 trillion particles) have been completed
  - runs used 230,000 cores of Titan or more
- Excellent agreement with KENO
- Extremely fine-mesh S_N calculations, whi accelerators, are under way

Acceleration efforts are underway or in initial phases for other VERA components



- Hydra-TH (CFD, unstructured finite volume)
  - Performance analysis, code optimization and scaling ef have improved performance and scaling for both single and multiphase (MPI only)
  - Collaboration with NVIDIA has been under way for over a year to exploit hybrid parallelism (MPI + threads)
    - Incorporating NVIDIA's AMG GPU lib, AmgX, into Hydra (https://developer.nvidia.com/amgx)
  - Thread other parts of the code (OpenMP, OpenACC, CUDA)

- MPACT (neutronics, Method of Characteristics)
  - Fortran, so OpenACC is most appropriate path forward
  - NVIDIA staff stationed full-time at ORNL identified to assist team
- Shift (neutronics, Monte Carlo)
  - Core Exnihilo kernels, including Shift, have been extracted and are being released as an open-source mini-app (Profugus)
  - NVIDIA staff have been identified to assist in Shift acceleration
  - Currently awaiting export control ruling for release

DENOVO Shift EXNIHILO INSILICO

Questions?

e-mail: turnerja@ornl.gov

The research and activities described in this presentation were performed using the resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.