# Learning from Each Other: Cooperative R&D with Total at ORNL

**David E. Bernholdt and Oscar R. Hernandez**

Computer Science and Mathematics Division and National Center for Computational Sciences

Oak Ridge National Laboratory

**Kshitij Mehta, Maxime Hugues, Jing Wen, and Henri Calandra**

Total E&P Research & Technology USA

*This work was performed in part at the Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.* This research used resources of the Oak Ridge Leadership Computing Facility

OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY

ORNL is managed by UT-Battelle for the US Department of Energy

w/ contributions from Jack Wells and Suzy Tichenor, ORNL

# What is the Leadership Computing Facility (LCF)?

- Collaborative DOE Office of Science user-facility 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 programs (INCITE, ALCC).
- Projects receive 10x to 100x more resource 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



OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY

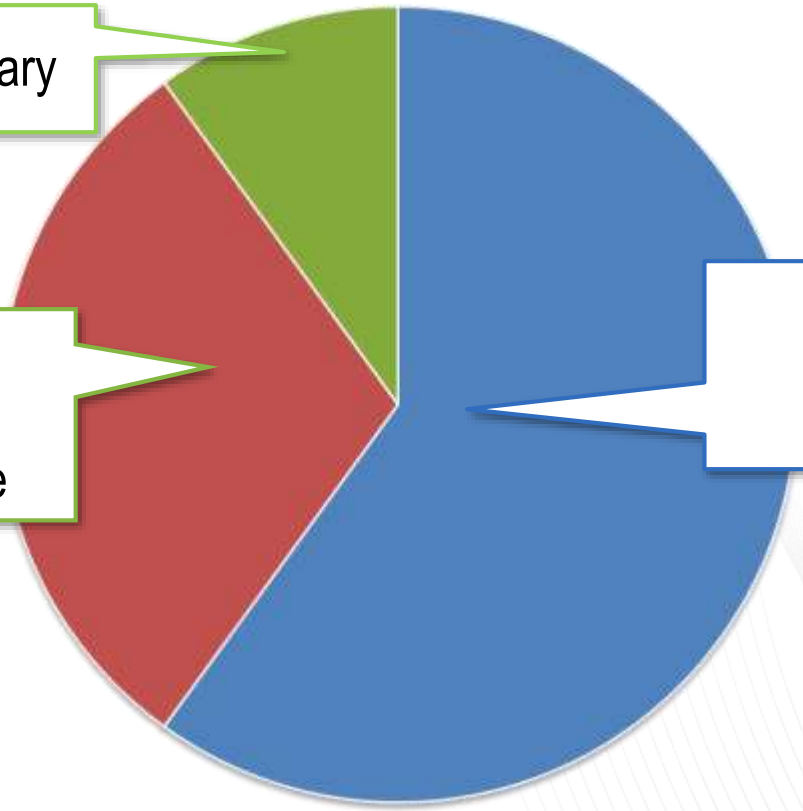
*Courtesy of Jack Wells, ORNL*

# Three Primary User Programs to Access LCF

*Distribution of allocable hours*

User agreements generally require open publication of results. Proprietary use is cost reimbursable

10% Director's Discretionary



**30% ALCC**

ASCR Leadership Computing Challenge

**60% INCITE**



*Courtesy of Jack Wells, ORNL*

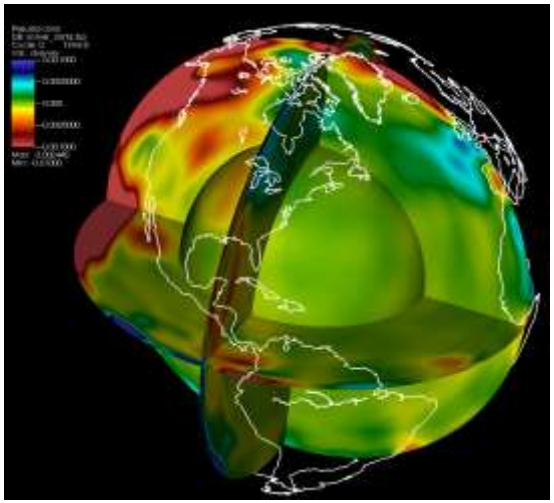
# Global Seismic Tomography

## *Imaging the Earth's Mantle*

INCITE PI: J. Tromp, Princeton
OLCF Liaisons: J. Hill, N Podhorszki, D Pugmire
Hours Used: 52 M

## *Science Objectives and Impact*

- **Overarching Goal:** Map Earth's interior based on seismic imaging.
  - Harness data from thousands of earthquakes
  - Use iterative data assimilation techniques to accurately image Earth's mantle
- **Impact:** Much better resolution of tomographic images at all scales.
  - Essential for better understanding of mantle dynamics and related surface processes
  - Assess seismic hazards in earthquake prone regions and the development of volcanism
- Eventual goal is global depth of 2,900 kilometers.



*Researchers led by Jeroen Tromp are using Titan to map the speeds of waves generated after earthquakes. This process, known as seismic tomography, provides the team with unprecedented temperature and composition (type of rock) estimates for the Earth's interior.*

## *Science Results*

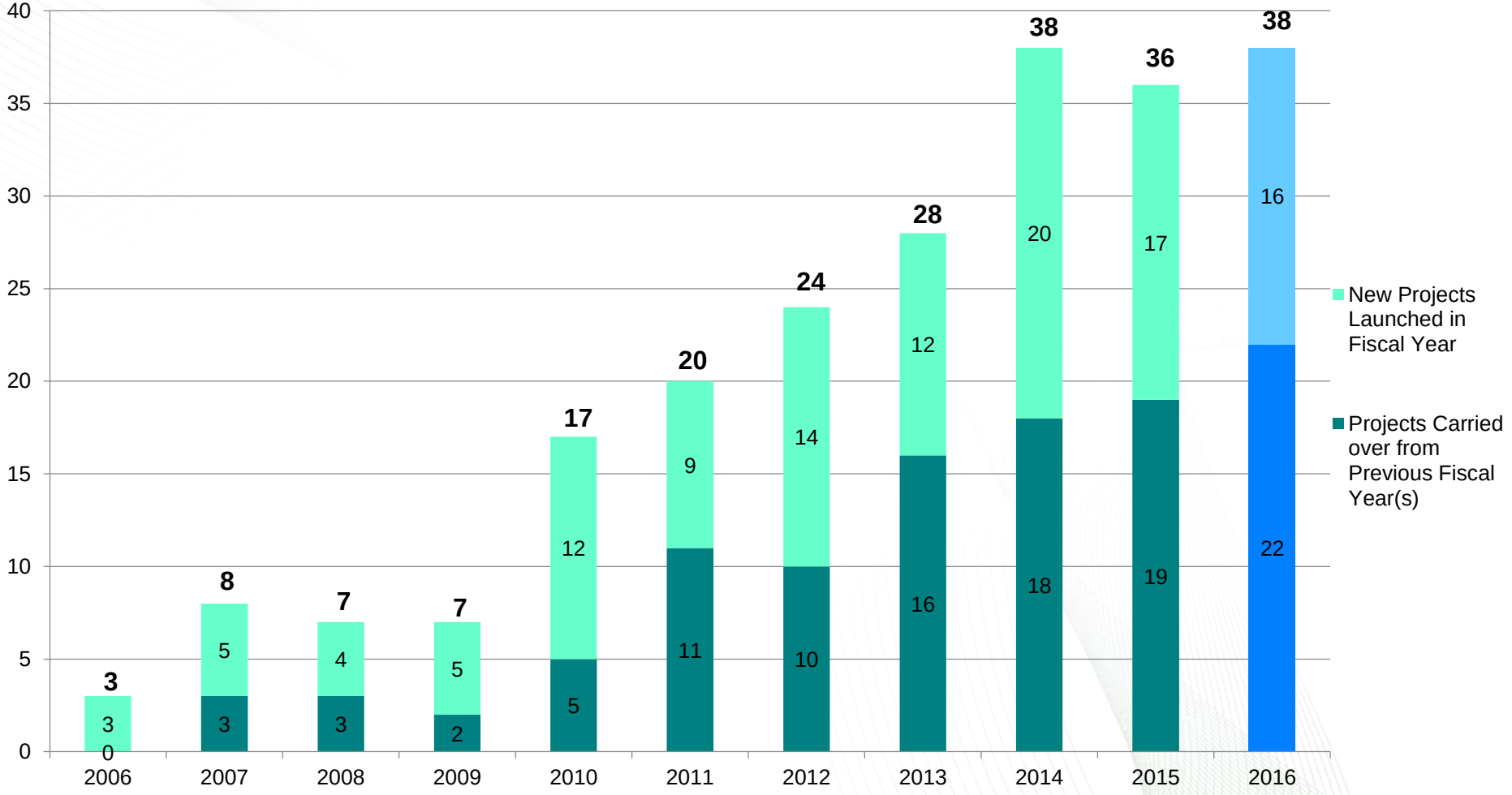
- First global adjoint inversion with a time resolution of 9 seconds resolving smaller scale features in the lower mantle and near the core-mantle boundary.
- Adaptable Seismic Data Format (ASDF): New data format more amenable to large-scale workflows.
- Doubled the speed of SPECFEM3D_GLOBE application.

## *OLCF Contribution*

- Simulations now use the OLCF's ADIOS middleware to write output for data workflow on Titan.
- Developed a parallel data reader for SPECFEM3D_GLOBE simulation output used inside VisIt.
- New visualizations via the new ADIOS data reader.

*Courtesy of Jack Wells, ORNL*

# Number of OLCF Industry Projects by Fiscal Year




*Current through 2016-08-23*



*Courtesy of Suzy Tichenor, ORNL*

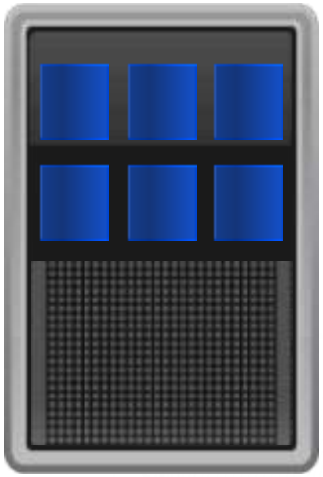
# Who's Been Working with Us?

# Why GPUs? Hierarchical Parallelism

## *High performance and power efficiency on path to exascale*

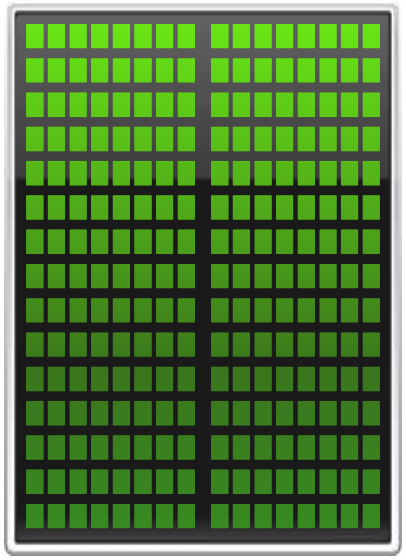
- Expose more parallelism through code refactoring and source code directives
  - Doubles CPU performance of many codes
- Use right type of processor for each task
- Data locality: Keep data near processing
  - GPU has high bandwidth to local memory for rapid access
  - GPU has large internal cache
- Explicit data management: Explicitly manage data movement between CPU and GPU memories

**CPU**



- Optimized for latency and sequential multitasking

**GPU Accelerator**

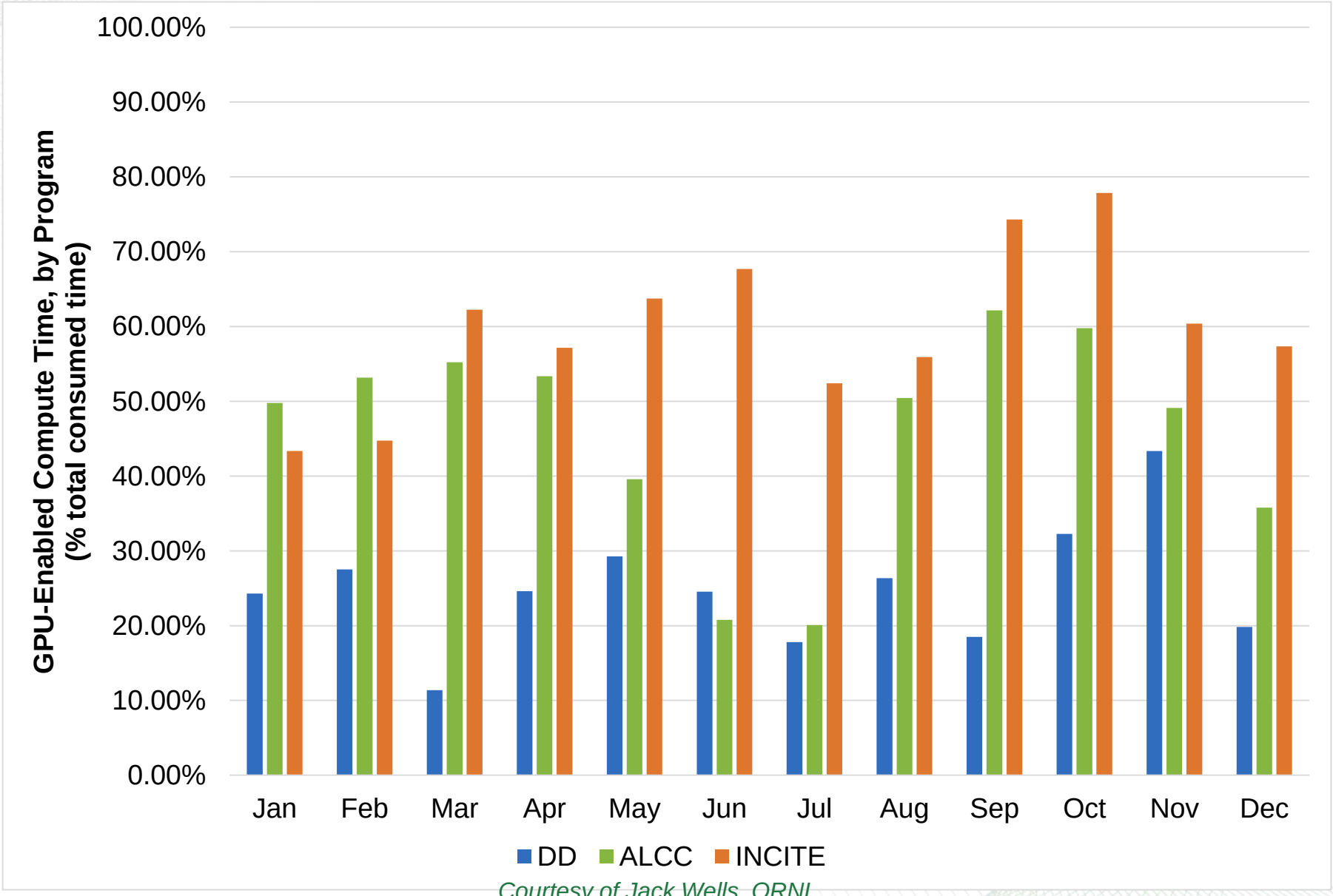


- Optimized for throughput and many simultaneous tasks
- 10× performance per socket
- 5× more energy-efficient systems



*Courtesy of Jack Wells, ORNL*

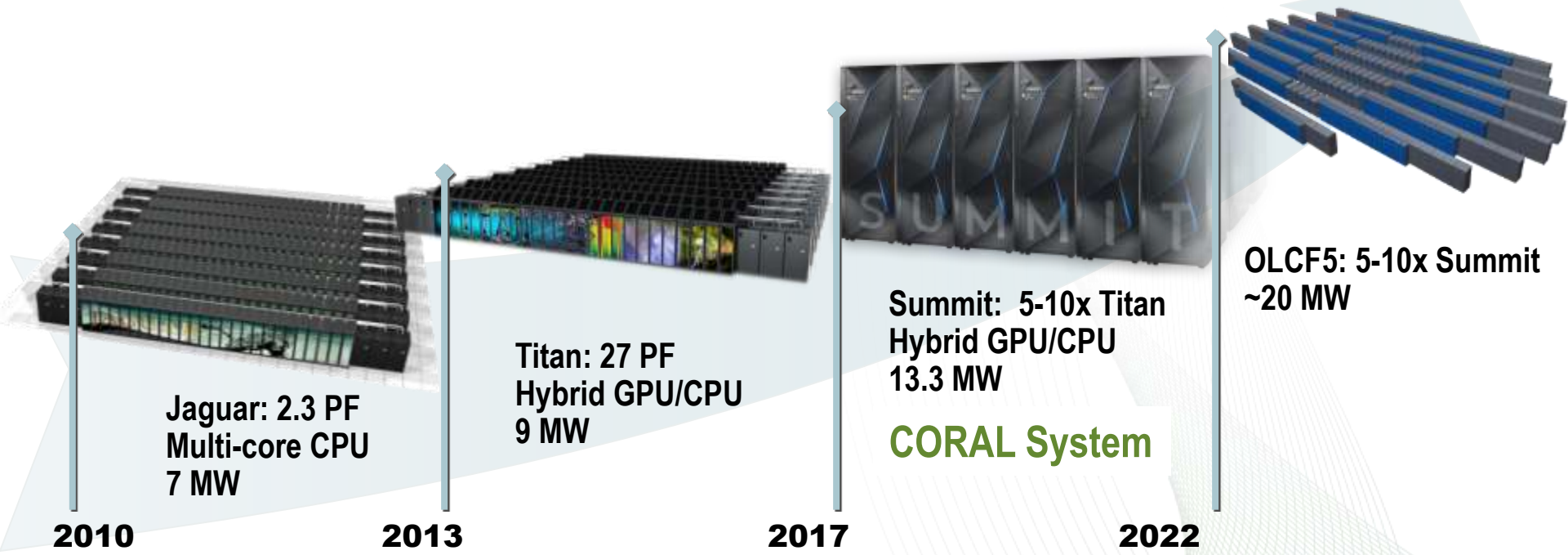
# GPU-enabled percentage of compute time for the DD, ALCC, and INCITE user programs




*Courtesy of Jack Wells, ORNL*

# Our Science requires that we continue to advance our computational capability over the next decade on the roadmap to Exascale

| Since clock-rate scaling ended in 2003, HPC performance has been achieved through increased parallelism. Jaguar scaled to 300,000 cores. | Titan and beyond deliver hierarchical parallelism with very powerful nodes. MPI plus thread level parallelism through OpenACC or OpenMP plus vectors |
|---|---|



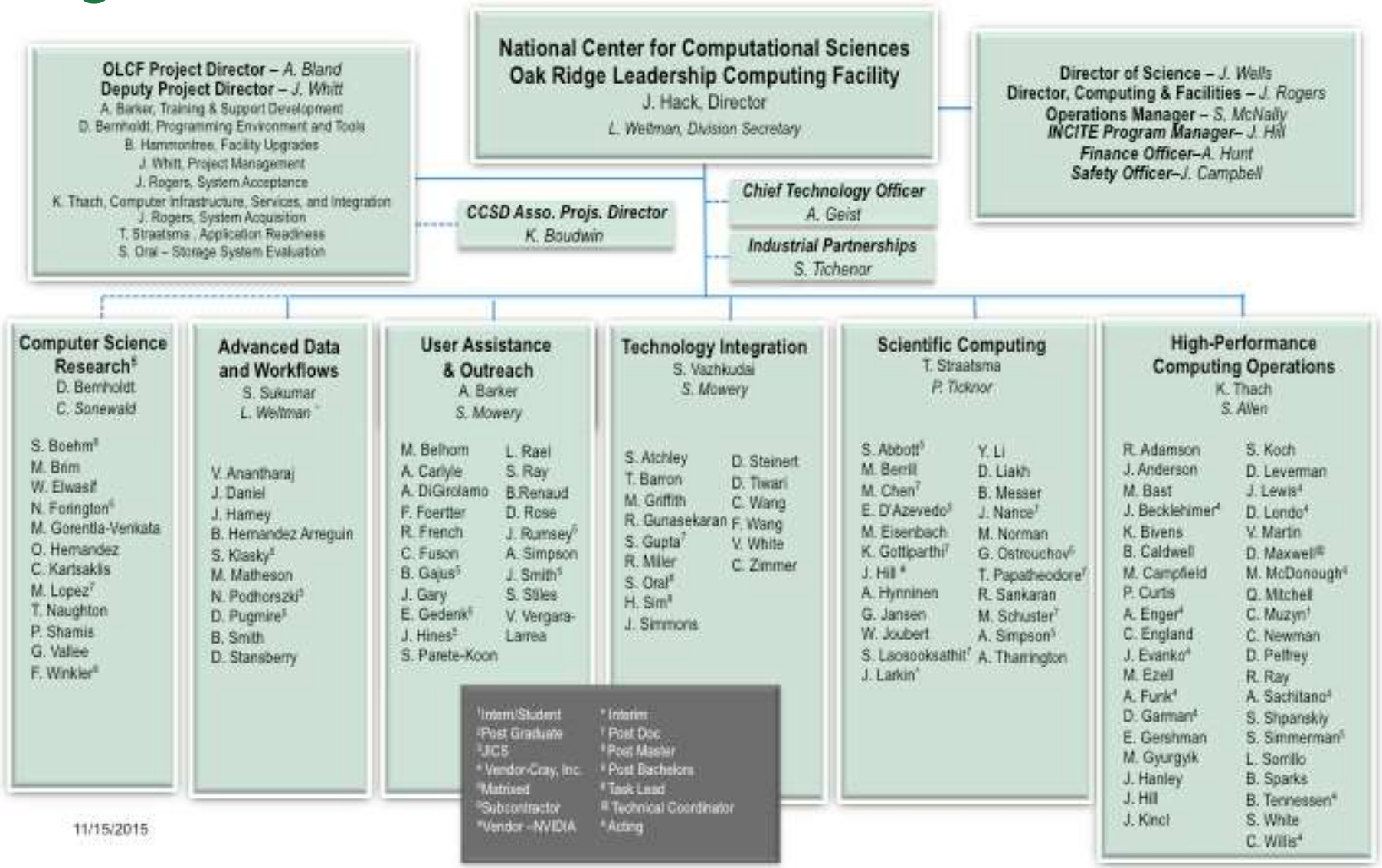



*Courtesy of Jack Wells, ORNL*

# ASCR Computing Upgrades at a Glance

| System attributes | NERSC Now | OLCF Now | ALCF Now | NERSC Upgrade | OLCF Upgrade | ALCF Upgrades | |
|---|---|---|---|---|---|---|---|
| Name<br>Planned Installation | Edison | TITAN | MIRA | Cori<br>2016 | Summit<br>2017-2018 | Theta<br>2016 | Aurora<br>2018-2019 |
| System peak (PF) | 2.6 | 27 | 10 | > 30 | 200 | >8.5 | 180 |
| Peak Power (MW) | 2 | 9 | 4.8 | < 3.7 | 13.3 | 1.7 | 13 |
| Total system memory | 357 TB | 710TB | 768TB | ~1 PB DDR4 + High Bandwidth Memory (HBM)+1.5PB persistent memory | > 2.4 PB DDR4 + HBM + 3.7 PB persistent memory | >480 TB DDR4 + High Bandwidth Memory (HBM) | > 7 PB High Bandwidth On-Package Memory Local Memory and Persistent Memory |
| Node performance (TF) | 0.460 | 1.452 | 0.204 | > 3 | > 40 | > 3 | > 17 times Mira |
| Node processors | Intel Ivy Bridge | **AMD Opteron & NVIDIA Kepler** | 64-bit PowerPC A2 | Intel Knights Landing many core CPUs<br>Intel Haswell CPU in data partition | **Multiple IBM Power9s & multiple NVIDIA Voltas** | Intel Knights Landing Xeon Phi many core CPUs | Knights Hill Xeon Phi many core CPUs |
| System size (nodes) | 5,600 nodes | 18,688 nodes | 49,152 | 9,300 nodes<br>1,900 nodes in data partition | ~4,600 nodes | >2,500 nodes | >50,000 nodes |
| System Interconnect | Aries | Gemini | 5D Torus | Aries | Dual Rail EDR-IB | Aries | 2nd Generation Intel Omni-Path Architecture |
| File System | 7.6 PB<br>168 GB/s, Lustre® | 32 PB<br>1 TB/s, Lustre® | 26 PB<br>300 GB/s GPFS™ | 28 PB<br>744 GB/s Lustre® | 120 PB<br>1 TB/s GPFS™ | 10PB, 210 GB/s Lustre initial | 150 PB<br>1 TB/s Lustre® |

# Organization of the OLCF



# OLCF Programming Environment and Tools

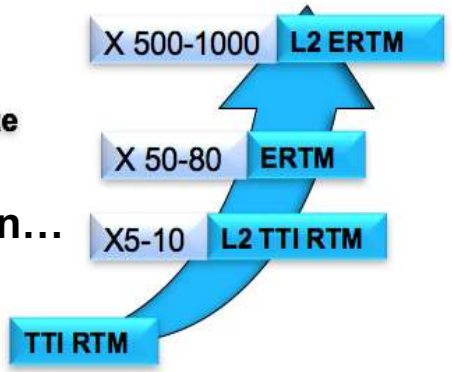
- Programming environment & runtimes
  - Core programming languages: C, C++, Fortran
  - Directive-based languages: OpenACC, OpenMP
  - CUDA
  - MPI
- Tools
  - Performance tools
  - Debugging and correctness tools
- Supported by the Computer Science Research (CSR) group
  - Connect to R&D work in these areas

# CSR Approach to Programming Models

- Providing an evolutionary path
  - Capable of incorporating revolutionary ideas
  - While minimizing need to rewrite apps “from scratch”
- Focus on annotations (directives), libraries, etc.
- Influence standards
  - OpenACC, OpenMP, MPI, OpenSHMEM, OFA, UCX, …
- Use benchmarks
  - Add kernels relevant to OLCF users
  - Understand compiler capabilities, motivate improvement
  - SPEC High-Performance Group, etc.
- Motivated in part by Oak Ridge Leadership Computing Facility (OLCF) requirements
  - Users expect stable, robust, portable, programming env.
  - Research tools rarely meet these expectations

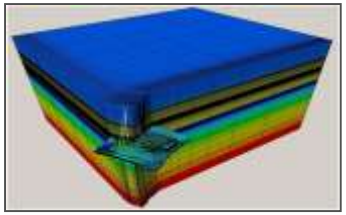
# Moving forward to the post-Petascale ERA

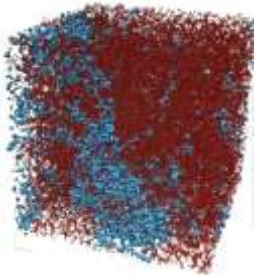
# HPC NEEDS FOR O&G UPSTREAM




- **Seismic acquisition: order of magnitude**
  - more data: multi-component data, dense data, 4D acquisition…
- **Seismic depth imaging methods : 1-3 orders of magnitude**
  - more iterations, more physics, more complex approximations
  - More interactivity, more workflow control, more data integration
- **Reservoir simulation: 1-2 orders of magnitude**
  - bigger model, multi resolution, uncertainties…, more interactions with data, 4D data…


- **New gamers:**
  - Digital Rock Physics,
  - Geo-mechanic
  - Molecular simulation
  - Data Analytics and machine learning

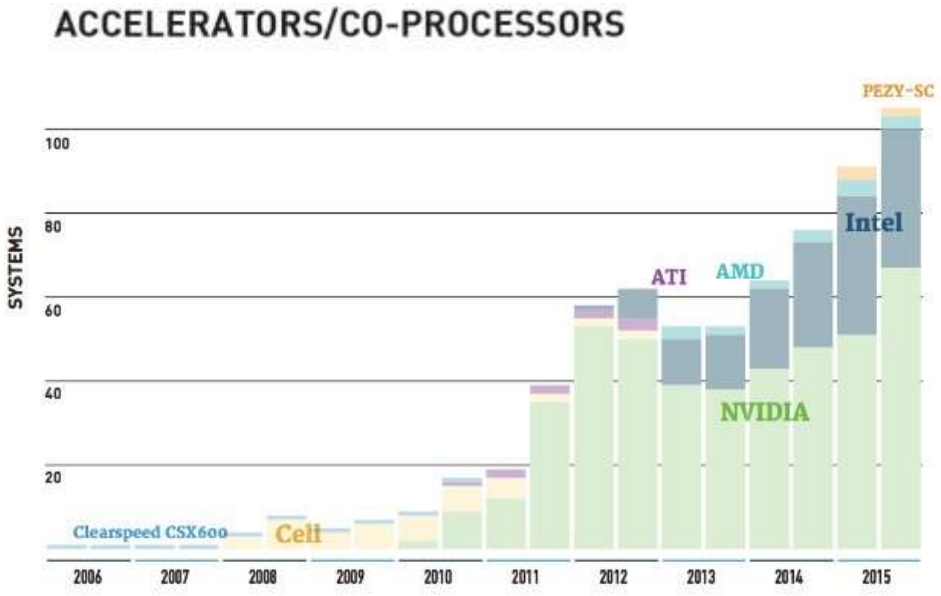


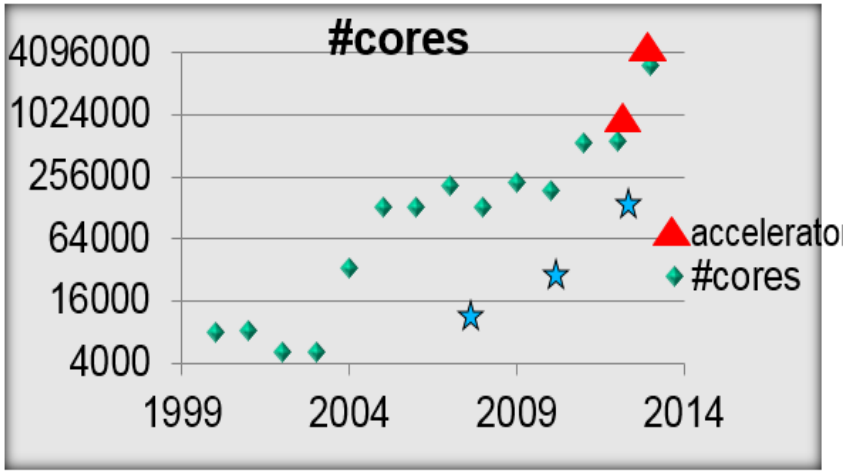




TOTAL

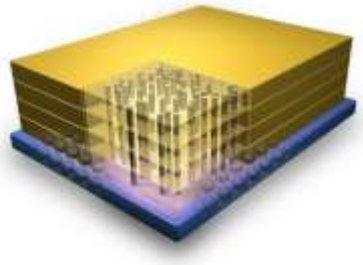
# HPC TECHNOLOGY TRENDS




## ❑Trends

- ✓ Systems with accelerators have grown from few to more than 100
- ✓ First of the 40 greenest computers are using accelerators
- ✓ Performance mainly coming from accelerators
- ✓ New memory type: 3D Stacked Memory / HBM. 20x more bandwidth



## ❑Extreme computer beyond 100 Pflops

- ✓ Require acceptable footprint and energy consumption (< 20MW)
- ✓ Accelerator or Many-Core technology is one potential path to extreme computing

# Seismic Depth Imaging at scale

**Objectives:**

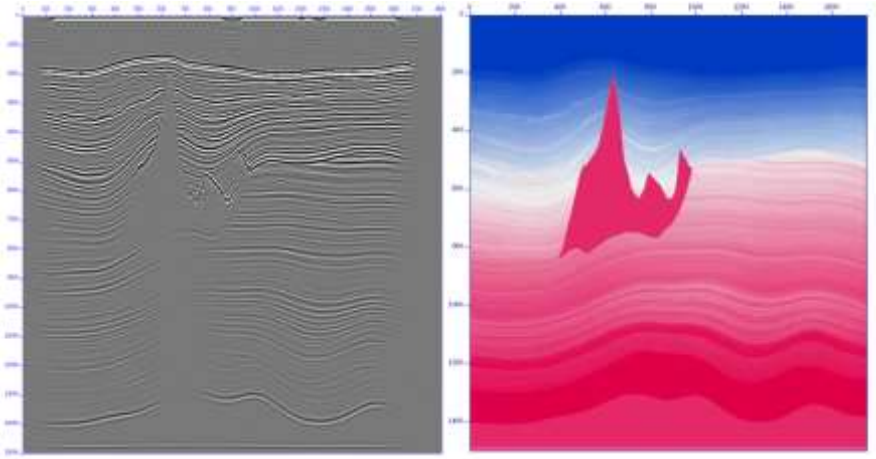
- **Use OpenACC to port Shot Profile migration code on GPU**
- **Run at scale on large Systems: collaboration with Oak Ridge National Lab**



**TITAN:**

- Cray XK7, 2nd most powerful supercomputer in the world
- 18,688 nodes (99% of TITAN resources), each with one K20X NVIDIA GPU

**Processed the 2793 Shot profiles of SEAM Model over 18508 GPUS in 54 minutes**

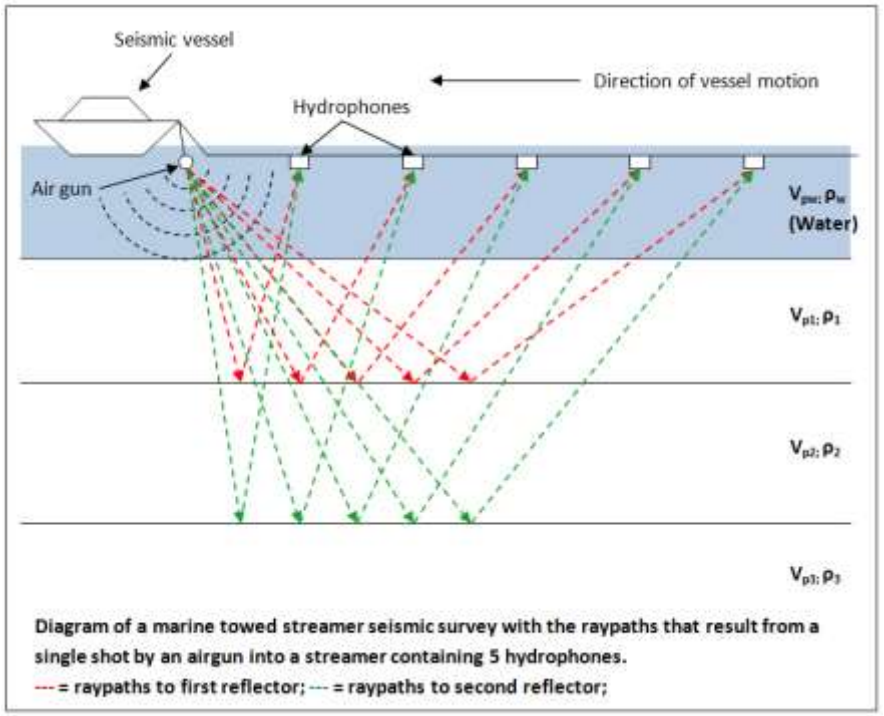


# Cooperative Research & Development Agreements (CRADAs)

- A DOE vehicle for collaborative R&D with the private sector
- Defines access to intellectual property
  - Each party owns what they bring to the table
  - New IP is jointly owned
  - Limited term protection for new IP (eventual public availability)
- Funding is flexible
  - May involve industry funding the DOE lab
  - May be in-kind contributions on either or both sides
    - On lab side, requires other funding aligned with interests of CRADA
- Private sector organizations can also collaborate with the labs via the Strategic Partnerships Project (SPP) Agreement (formerly known as "Work for Others"). Under this Agreement,
  - Private sector funds the lab
  - Work must pertain to the mission of the lab, cannot directly compete with private sector
  - Intellectual property rights generally belong to the private sector funder

# Seismic Data Acquisition

- Ship fires ‘shots’ towards ocean bottom
- Hydrophones record reflections from various layers beneath earth’s subsurface
- Recorded data typically in the order of few hundreds of terabytes

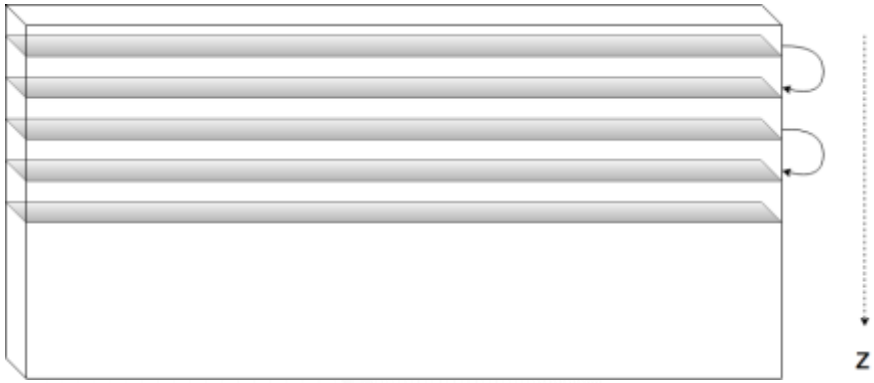


# One-Way Wave Equation Migration (OWEM)

$$k_z \approx \sqrt{\frac{\omega^2}{c^2} - k_x^2 - k_y^2} + \omega\left(\frac{1}{v(x,y,z)} - \frac{1}{c}\right) + \frac{\frac{1}{2}\frac{1}{\omega}[v(x,y,z) - c]\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)}{1 + \frac{1}{4}\frac{v(x,y,z)^2 + v(x,y,z)c + c^2}{\omega^2}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)}$$

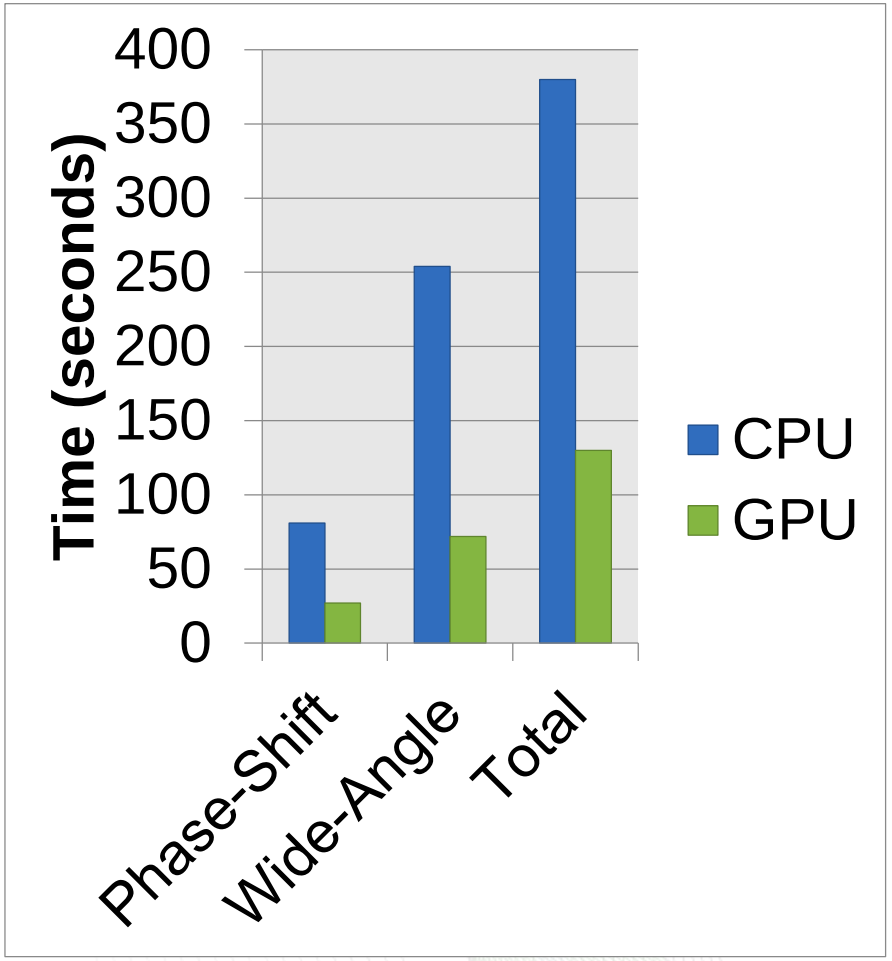
Phase shift | Lens correction | Wide angle correction

- Solve the One-Way Wave Equation in 2D/3D
- Solve in frequency domain
- Use Fourier finite differences approximation
- Approximate wavefield at every depth (*z* dimension)
  - ~75% of total application time



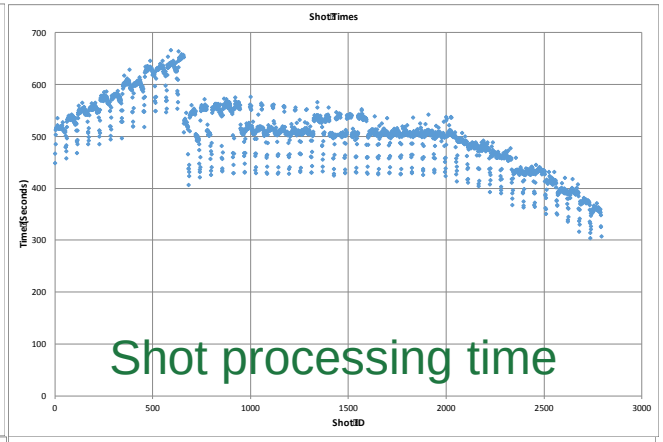
# OWEM Using OpenACC

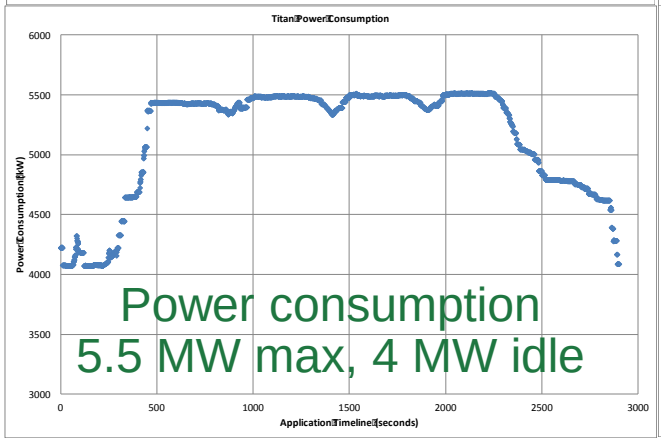
- Comparing
  - 1 NVIDIA K20X GPU on Titan
  - 1x 8-core Intel Sandy Bridge CPU
- ~3X speedup
- Optimization involved…
  - using CuFFT library,
  - optimizing sparse matrix solvers
  - etc.

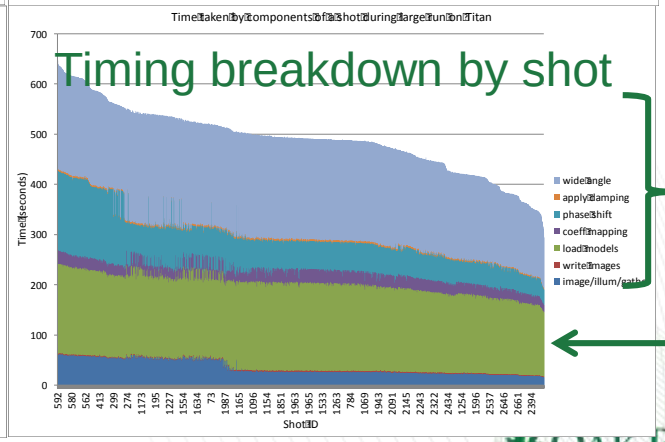


# OWEM at Scale on Titan

- Process a dataset comprising of 2800 shots using ~18,500 GPUs
- First (known) real-world application run at large scale using OpenACC
- 1.2 Petaflops peak using 5.5 MW
  - *Already identified areas for further improvement*

# Currently In Progress: Reverse Time Migration (RTM)

- More sophisticated approach to seismic image processing
  - ~2x more expensive than OWEM
- Accelerated using OpenACC
- At scale on Titan, application writes ~30 PB of data
  - OLCF has two large center-wide Lustre filesystems, with ~14 PB of usable space each
  - Running application in phases to reduce I/O footprint
- Currently working on optimizing I/O using ADIOS

# Lessons Learned About OpenACC at Scale?

- Helped drive OpenACC “deep copy” design
- Helped identify and address implementation quality issues
- Being able to shutdown and re-initialize GPUs is important for large scale runs:
  - Helps save power/energy when the GPUs are not being used (e.g. during I/O)
  - Mitigate silent software errors at scale
    - E.g. OpenACC runtime. Due to buffer memory reuse on the GPUs, memory management, etc.
- Need tools to check for GPU-specific bugs
  - E.g., missing memory synchronizations
- OpenACC interoperability with accelerated libraries
  - Especially aligning data placement and movement between libraries and (OpenACC) program

# Resilience and Correctness Needs

- Need better mechanisms for debugging OpenACC code when memory gets corrupted (software issues)
  - Verify execution and correctness of host ←→ device transfers
  - OpenACC tools API can be used for this
- Can exploit ability of OpenACC to target multiple device types to help detect faults
  - E.g. running a multithreaded version on the host that can verify the results on the GPU
  - CPUs can also be used for managing the uncertainty in the computations

# OpenACC and OpenMP

- Descriptive nature of OpenACC gives compilers more flexibility to map and optimize accelerator code.
  - However this may lead to variation in performance across compiler versions and platforms
  - OpenMP is generally prescriptive
- Translating OpenACC to OpenMP 4.5 is straightforward
  - Finding parallelism and specifying the data scoping is most challenging aspect (same for both OpenACC and OpenMP)
  - When translating to OpenMP4, the programmer explicitly parallelizes loops across teams, threads or SIMD regions
  - Some descriptive constructs such as !$acc loop is still not available in OpenMP 4.5, but will be in OpenMP 5.0

# Summary

- DOE LCF supports computational science and engineering problems that can't be done on other resources
  - Open to national labs, academia, and industry, world-wide
- OLCF provides deep support to users
  - Domain science liaisons, adv. data & workflows, programming environment & tools
- Total was interested tapping OLCF's experience with large-scale GPU-based systems
  - Established CRADA (on-going)
  - Total staff member on long-term assignment at ORNL
  - OLCF Director's Discretion allocation
  - OpenACC implementations of OWEM, RTM
  - Now looking at I/O and ADIOS
- Benefits
  - Largest-scale "real" OpenACC application runs to date
  - Drove implementation quality and standard extensions
  - Identified improvements needed to OpenACC ecosystem