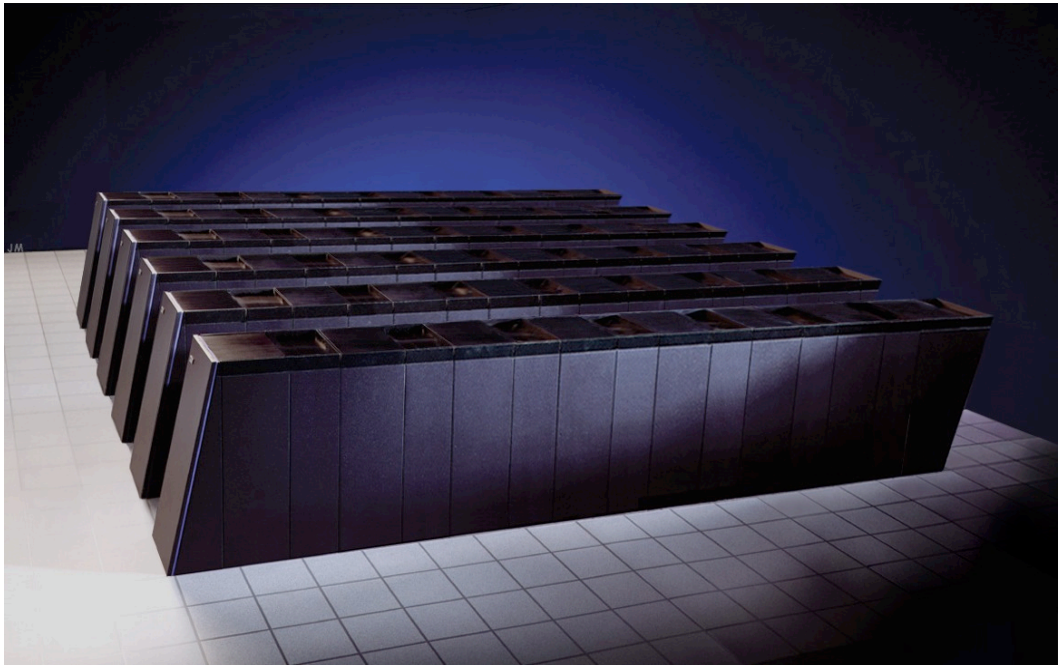
Overview of the Argonne Leadership Computing Facility

Timothy J. Williams

Argonne Leadership Computing Facility

Argonne National Laboratory



CScADS Summer 2011 Workshop 1

- **ALCF Purpose**
- **People & Services**
- **Computers**
  - Hardware
  - Software
  - How to use
- **Future**
- **Science**

Argonne Leadership Computing Facility

- Established 2006 at Argonne National Lab
- One of two DOE national Leadership Computing Facilities (OLCF is other)
- Supports mission of DOE Office of Science Advanced Scientific Computing Research (ASCR)

DOE INCITE Program

Innovative and Novel Computational Impact on Theory and Experiment

- **60% of time at Leadership Facilities**
- **Solicits large, computationally intensive research projects**
  - To enable high-impact scientific advances
  - Call for proposals yearly (closed 6/30/2011)
  - INCITE Program web site: doeleadershipcomputing.org
- **Open to all scientific researchers and organizations**
  - Scientific discipline peer review
  - Computational readiness review
- **Awards large computer time & data storage allocations**
  - Small number of projects for 1-3 years
  - Academic, national lab and industry, with DOE or other support
- **2011 INCITE at ALCF**
  - **30 projects**
  - 732**M core hours**

DOE ALCC Program

ASCR Leadership Computing Challenge

- **30% of time at LCFs**
- **Projects of special interest to DOE**
  - emphasis on high-risk, high-payoff simulations
- **Awards granted in June (review started 2/18/2011)**
  - science.energy.gov/ascr/facilities/alcc
- **2010 ALCC at ALCF**
  - 9 awards
  - 300+ million core hours

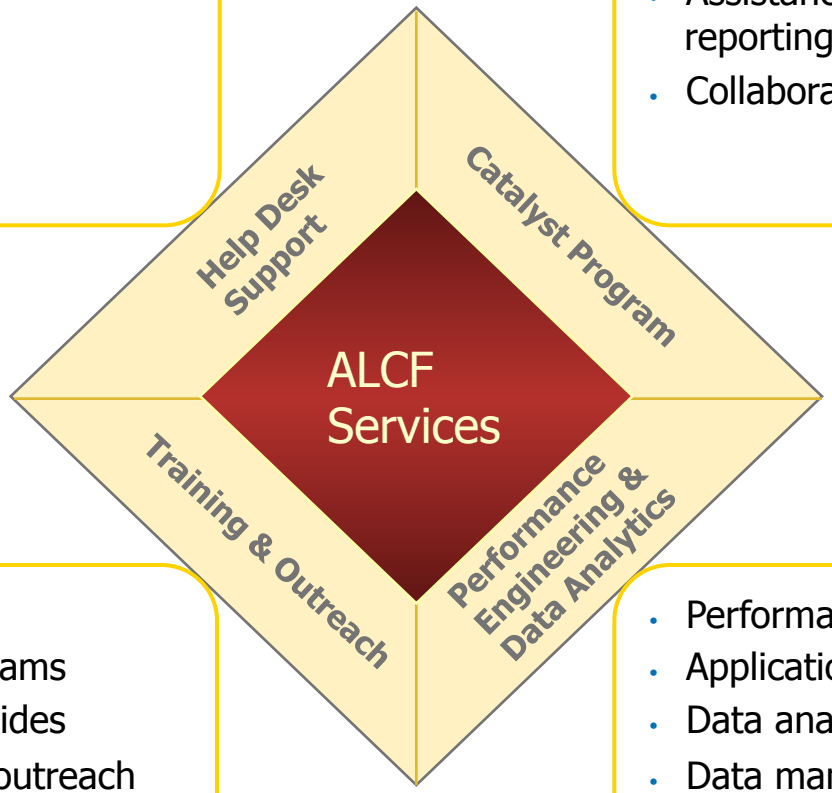
Director's Discretionary Program

- **10% of time at LCFs**
- **ALCF Discretionary allocations provide time for:**
  - Porting, scaling, and tuning applications
  - Benchmarking codes and preparing INCITE proposals
  - Preliminary science runs prior to an INCITE award
- **Apply for time at ALCF allocations page**
  - www.alcf.anl.gov/support/gettingstarted

ALCF Service Offerings

- Startup assistance
- User administration
- Job management
- Technical support

- ALCF science liaison
- Assistance with proposals, planning, reporting
- Collaboration within science domains



- Workshops & seminars
- Customized training programs
- On-line content & user guides
- Educational and industry outreach programs

- Performance engineering
- Application tuning
- Data analytics
- Data management services

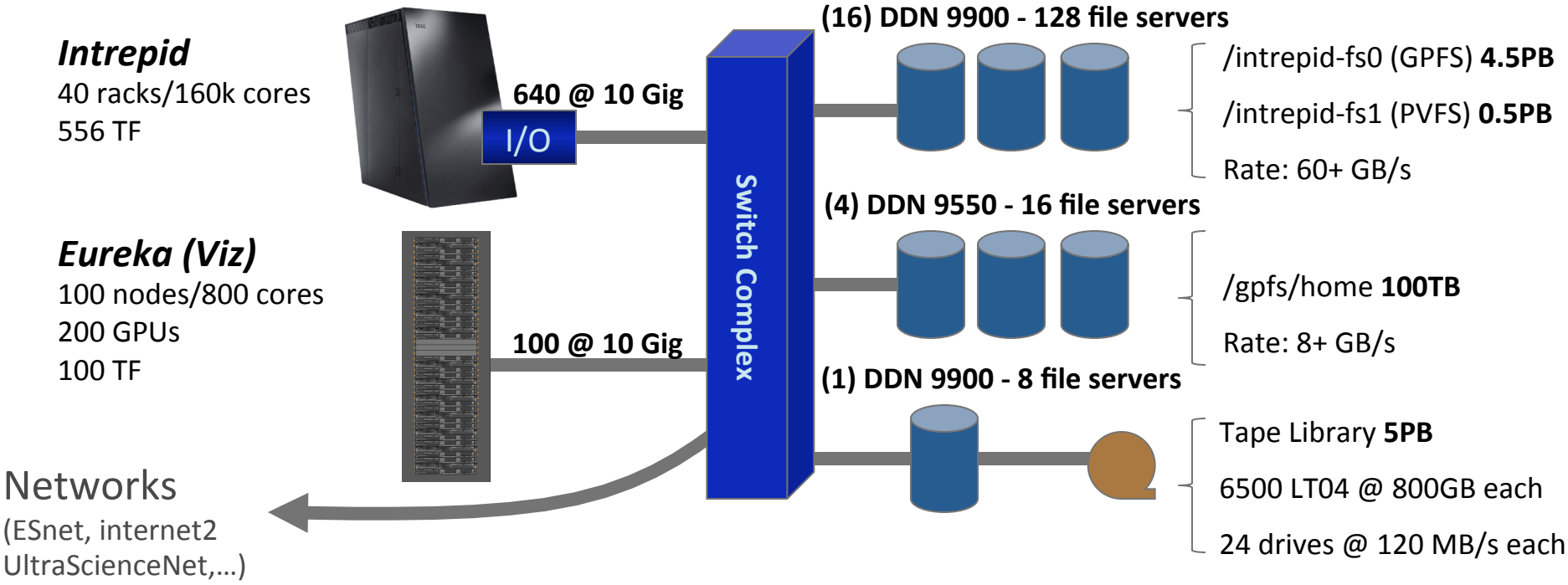
ALCF Hardware

- ***Intrepid* - ALCF Blue Gene/P System:**
  - 40,960 nodes / 163,840 PPC cores
  - 80 Terabytes of memory
  - Peak flop rate: 557 Teraflops
  - Linpack flop rate: 450.3
  - #15 on the Top500 list
  - #1 on Graph500 list
  - #41 on Green500 list

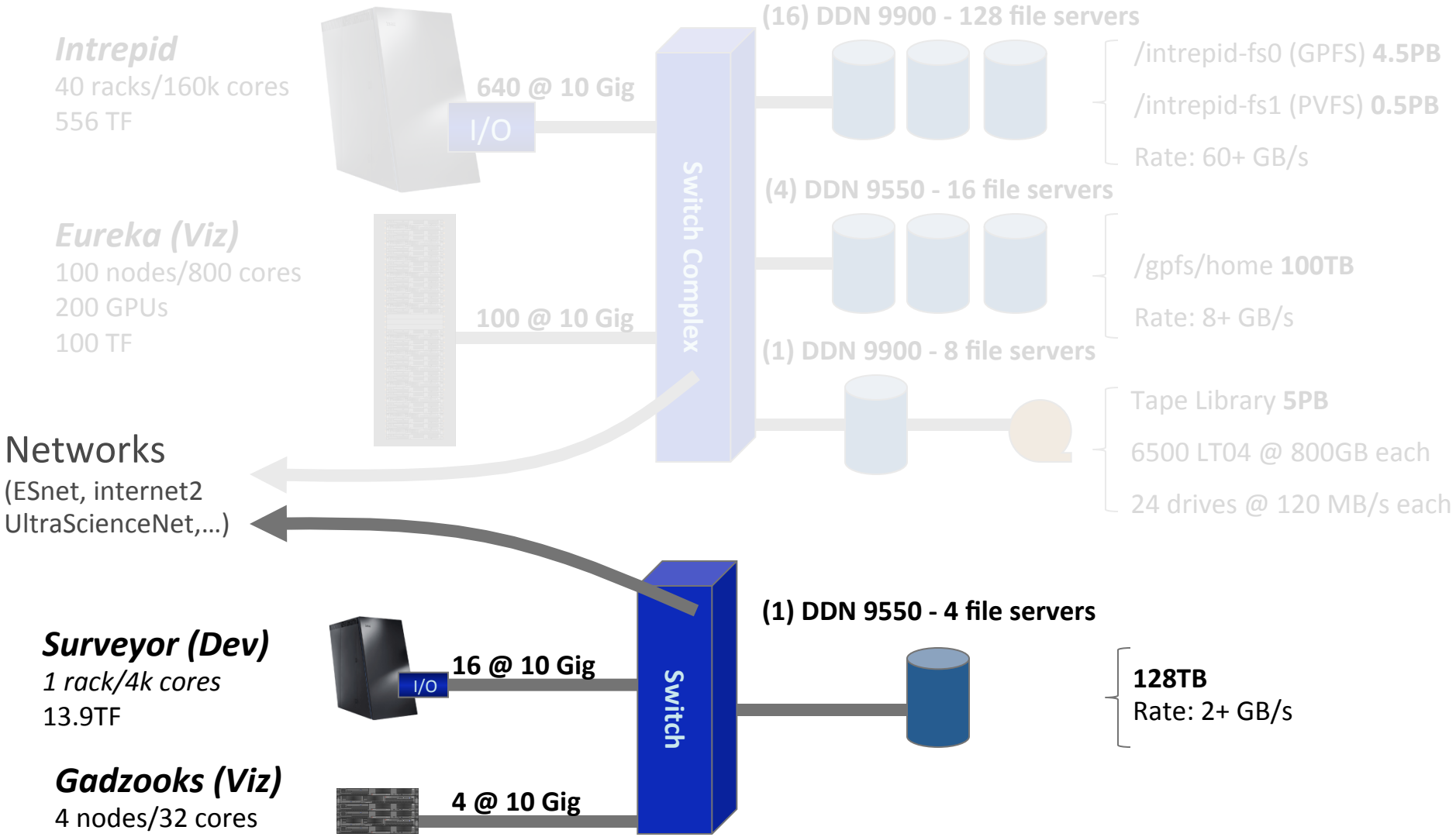
- ***Eureka* - ALCF Visualization System:**
  - 100 nodes / 800 2.0 GHz Xeon cores
  - 3.2 Terabytes of memory
  - 200 NVIDIA FX5600 GPUs
  - Peak flop rate: 100 Teraflops

- **Storage:**
  - 6+ Petabytes of disk storage with an I/O rate of 80 GB/s
  - 5+ Petabytes of archival storage (10,000 volume tape archive)

ALCF Hardware Layout



ALCF Hardware Layout



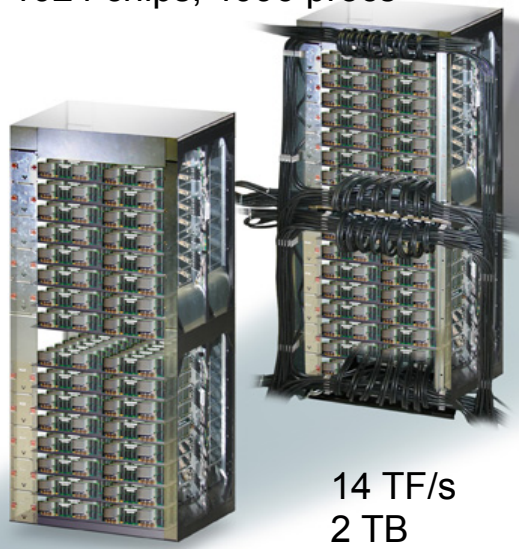
Blue Gene/P Packaging

- 4 850Mhz PowerPC cores per chip
- 1 chip, 2 GB of DDR SDRAM, 5 network interfaces per compute node
- 32 compute nodes per node card
- 32 node cards per rack
- 1,024 nodes total per rack
- 40 rack on Intrepid

Intrepid System
40 Racks

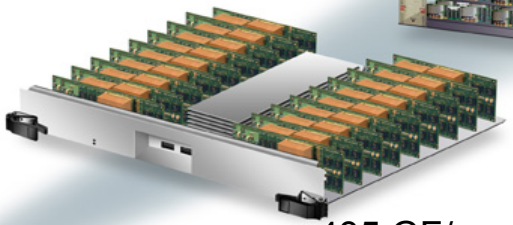
556 TF/s
82TB

Rack
32 Node Cards
1024 chips, 4096 procs



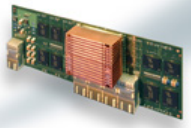
14 TF/s
2 TB

Node Card
(32 chips 4x4x2)
32 compute, 0-2 IO cards



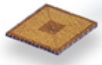
435 GF/s
64 GB

"Node"
(Compute Card)
1 chip, 20 DRAMs



13.6 GF/s
2.0 GB DDR
Supports 4-way SMP

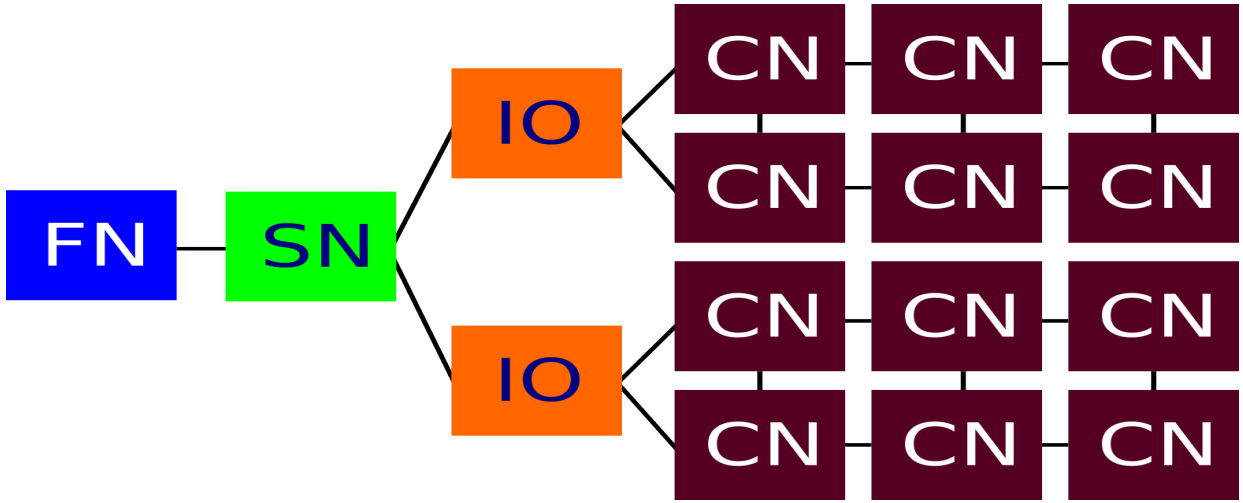
Chip
4 processors



850 MHz
8 MB EDRAM

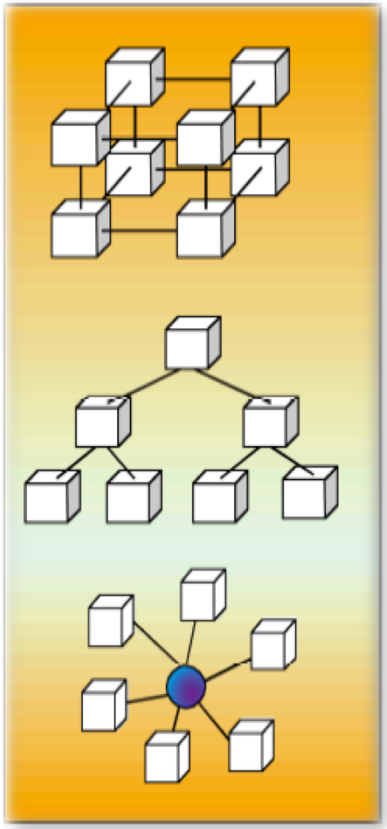
Front End Node / Service Node
System p Servers
Linux SLES10

Blue Gene/P Heterogeneity



- **Front-end nodes (FN): users login, compile, submit jobs**
  - 2.5 GHz PowerPC 970, Linux OS
- **Service nodes (SN): system management services: create and monitor processes, configure partitions, control jobs, store statistics**
- **I/O nodes (IO): OS services: files, sockets, process management, debugging**
- **Compute nodes (CN): run user applications as batch jobs**
  - CNK OS (no shell)

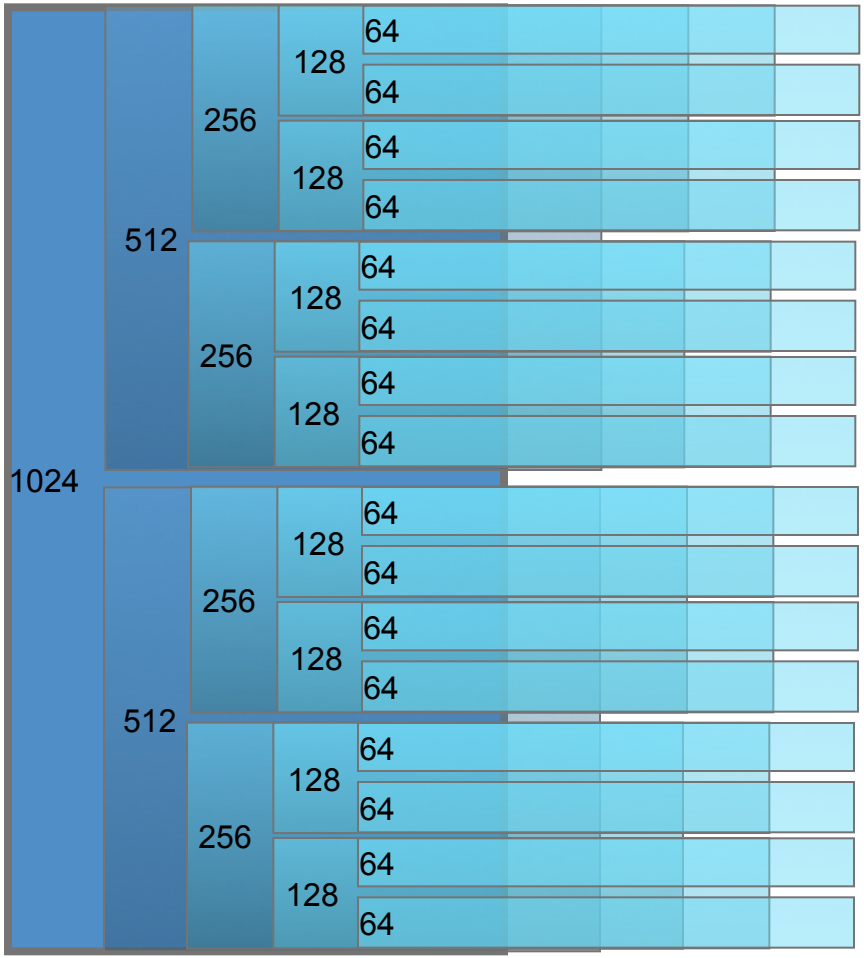
Blue Gene/P Interconnection Networks



- **3 Dimensional Torus**
  - Interconnects all compute nodes
  - Communications backbone for point-to-point
  - 3.4 Gb/s on all 12 node links (5.1 GB/s per node)
  - 0.5 µs latency between nearest neighbors, 5 µs to the farthest
  - MPI: 3 µs latency for one hop, 10 µs to the farthest
- **Collective Network**
  - Interconnects all compute nodes and I/O nodes
  - One-to-all broadcast functionality
  - Reduction operations for integers and doubles
  - 6.8 Gb/s of bandwidth per link per direction
  - Latency of one way tree traversal 1.3 µs, MPI 5 µs
- **Low Latency Global Barrier and Interrupt**
  - Latency of one way to reach 72K nodes 0.65 µs, MPI 1.6 µs

- **10 Gb/s functional Ethernet**
  - Disk I/O
- **1Gb private control (JTAG)**
  - Service node/system management

BG/P Partitions



Partitions on 1 rack of *Intrepid*

- ***Intrepid* compute nodes grouped into partitions ranging from 64 to 40,960**
  - One I/O node for each 64 compute nodes
- **Job gets entire partition to itself**
- **Minimum partition size is 64 nodes**
- **Each partition is its own torus/mesh**
  - Electrically isolated
- **Partitions <512 nodes form mesh network**
- **Partitions >=512 nodes form torus network**

Blue Gene/P at ALCF

Blue Gene/P: Covers Removed

Programming Environment

- **Languages:**
  - Fortran, C, C++, Python
  - IBM XL and GNU compilers
- **MPI:**
  - Based on MPICH2 1.0.x base code:
    - MPI-IO supported
    - One-sided communication supported
    - No process management (MPI_Spawn(), MPI_Connect(), ….)
  - Uses the 3 different BG/P networks for different MPI functions
- **Threads:**
  - OpenMP 2.5
  - NPTL Pthreads
- **Linux development environment:**
  - Compute Node Kernel provides look and feel of a Linux environment
    - POSIX routines (with some restrictions: no fork() or system())
    - BG/P adds pthread support, additional socket support
  - Statically and dynamically linked libraries

Programming Environment

- **Languages:**
  - Fortran, C, C++, Python
  - IBM XL and GNU compilers
- **MPI:**
  - Based on MPICH2 1.0.x base code:
    - MPI-IO supported
    - One-sided communication supported
    - No process management (MPI_Spawn(), MPI_Connect(), ….)
  - Uses the 3 different BG/P networks for different MPI functions
- **Threads:**
  - OpenMP 2.5
  - NPTL Pthreads
- **Linux development environment:**
  - Compute Node Kernel provides look and feel of a Linux environment
    - POSIX routines (with some restrictions: no fork() or system())
    - BG/P adds pthread support, additional socket support
  - Statically and dynamically linked libraries

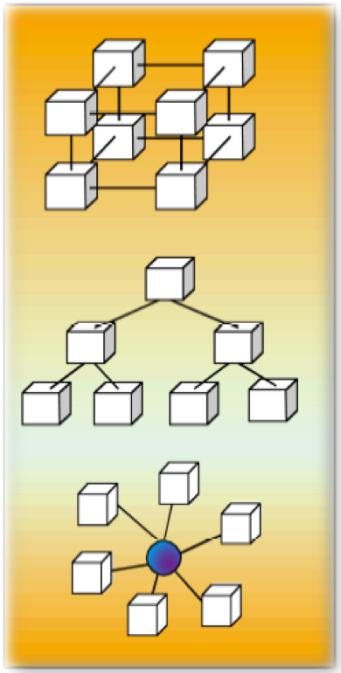
CROSS-COMPILING:
Login and compute nodes have different processor and OS

Programming Environment

- **Languages:**
  - Fortran, C, C++, Python
  - IBM XL and GNU compilers
- **MPI:**
  - Based on MPICH2 1.0.x base code:
    - MPI-IO supported
    - One-sided communication supported
    - No process management (MPI_Spawn(), MPI_Connect(), ….)
  - Uses the 3 different BG/P networks for different MPI functions
- **Threads:**
  - OpenMP 2.5
  - NPTL Pthreads
- **Linux development environment:**
  - Compute Node Kernel provides look and feel of a Linux environment
    - POSIX routines (with some restrictions: no fork() or system())
    - BG/P adds pthread support, additional socket support
  - Statically and dynamically linked libraries

Programming Environment

- **Languages:**
  - Fortran, C, C++, Python
  - IBM XL and GNU compilers
- **MPI:**
  - Based on MPICH2 1.0.x base code:
    - MPI-IO supported
    - One-sided communication supported
    - No process management (MPI_Spawn(), MPI_Connect(), ….)
  - Uses the 3 different BG/P networks for different MPI functions
- **Threads:**
  - OpenMP 2.5
  - NPTL Pthreads
- **Linux development environment:**
  - Compute Node Kernel provides look and feel of a Linux environment
    - POSIX routines (with some restrictions: no fork() or system())
    - BG/P adds pthread support, additional socket support
  - Statically and dynamically linked libraries



Programming Environment

- **Languages:**
  - Fortran, C, C++, Python
  - IBM XL and GNU compilers
- **MPI:**
  - Based on MPICH2 1.0.x base code:
    - MPI-IO supported
    - One-sided communication supported
    - No process management (MPI_Spawn(), MPI_Connect(), ….)
  - Uses the 3 different BG/P networks for different MPI functions
- **Threads:**
  - OpenMP 2.5
  - NPTL Pthreads
- **Linux development environment:**
  - Compute Node Kernel provides look and feel of a Linux environment
    - POSIX routines (with some restrictions: no fork() or system())
    - BG/P adds pthread support, additional socket support
  - Statically and dynamically linked libraries

Runtime Environment

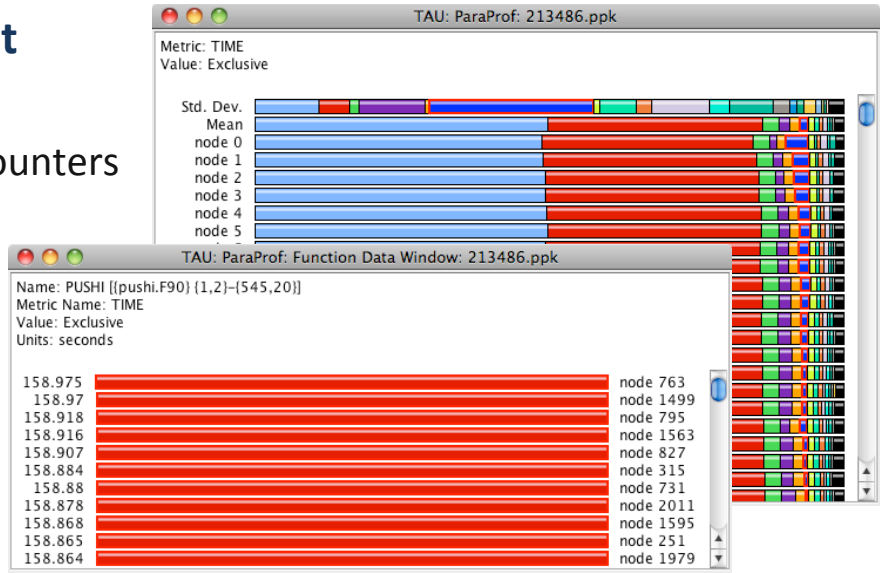
- **Logical partitions, with complete electrical isolation**
  - Partition rebooted between jobs
- **Three modes for processes per node (one MPI rank per process)**
  - SMP
    - 1 processes accessing all node memory (2 GB)
    - up to 12 threads
  - Dual
    - 2 processes accessing half node memory each
    - up to 6 threads each
  - VN
    - 4 processes accessing one quarter of node memory each
    - up to 3 threads each
- **SPMD model:**
  - Normally, compute nodes all run same executable
  - Alternatives: HTC mode, cobalt-subrun
- **No virtual memory**

Performance and Debugging Tools

- **IBM High Performance Computing Toolkit**
  - MPI Profile and Tracing Library
  - HPM Library for hardware performance counters
  - Xprofiler visualization of gprof profiles
- **Rice HPCToolkit:**
  - Sample based profiling of applications
- **TAU – Tuning and Analysis Utilities**
  - MPI Profiling
  - Performance counter data
- **Hardware Counters (Universal Performance Counters – UPC)**
  - Blue Gene/P provides 256 on chip counters for hardware events
  - UPC and HPM libraries provide access to counters
- **Core files – lightweight core files, text format, no full memory dump**
  - Coreprocessor - Generates stack trace from core files
- **Parallel debuggers scaling to thousands of cores**
  - TotalView
  - DDT

Performance and Debugging Tools

- **IBM High Performance Computing Toolkit**
  - MPI Profile and Tracing Library
  - HPM Library for hardware performance counters
  - Xprofiler visualization of gprof profiles
- **Rice HPCToolkit:**
  - Sample based profiling of applications
- **TAU – Tuning and Analysis Utilities**
  - MPI Profiling
  - Performance counter data



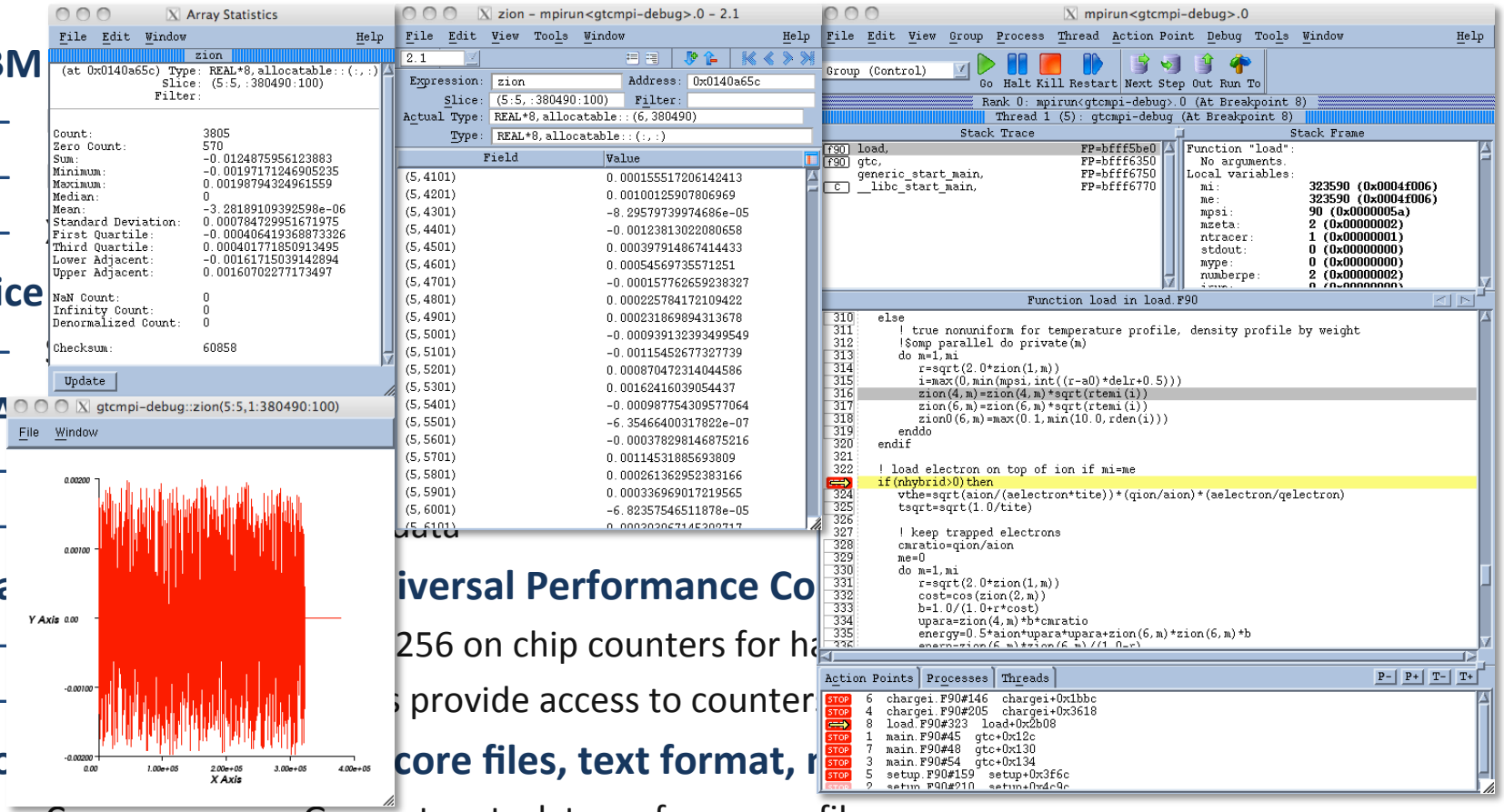
- **Hardware Counters (Universal Performance Counters – UPC)**
  - Blue Gene/P provides 256 on chip counters for hardware events
  - UPC and HPM libraries provide access to counters
- **Core files – lightweight core files, text format, no full memory dump**
  - Coreprocessor - Generates stack trace from core files
- **Parallel debuggers scaling to thousands of cores**
  - TotalView
  - DDT

Performance and Debugging Tools

- **IBM High Performance Computing Toolkit**
  - MPI Profile and Tracing Library
  - HPM Library for hardware performance counters
  - Xprofiler visualization of gprof profiles
- **Rice HPCToolkit:**
  - Sample based profiling of applications
- **TAU – Tuning and Analysis Utilities**
  - MPI Profiling
  - Performance counter data
- **Hardware Counters (Universal Performance Counters – UPC)**
  - Blue Gene/P provides 256 on chip counters for hardware events
  - UPC and HPM libraries provide access to counters
- **Core files – lightweight core files, text format, no full memory dump**
  - Coreprocessor - Generates stack trace from core files
- **Parallel debuggers scaling to thousands of cores**
  - TotalView
  - DDT

Performance and Debugging Tools

- **IBM**
  - –
  - –
  - –
- **Rice**
  - –
- **TA**
  - –
  - –
- **Ha** **iversal Performance Co**
  - – 256 on chip counters for ha
  - – s provide access to counter
- **Co** **core files, text format, r**
  - – Coreprocessor - Generates stack trace from core files
- **Parallel debuggers scaling to thousands of cores**
  - – TotalView
  - – DDT

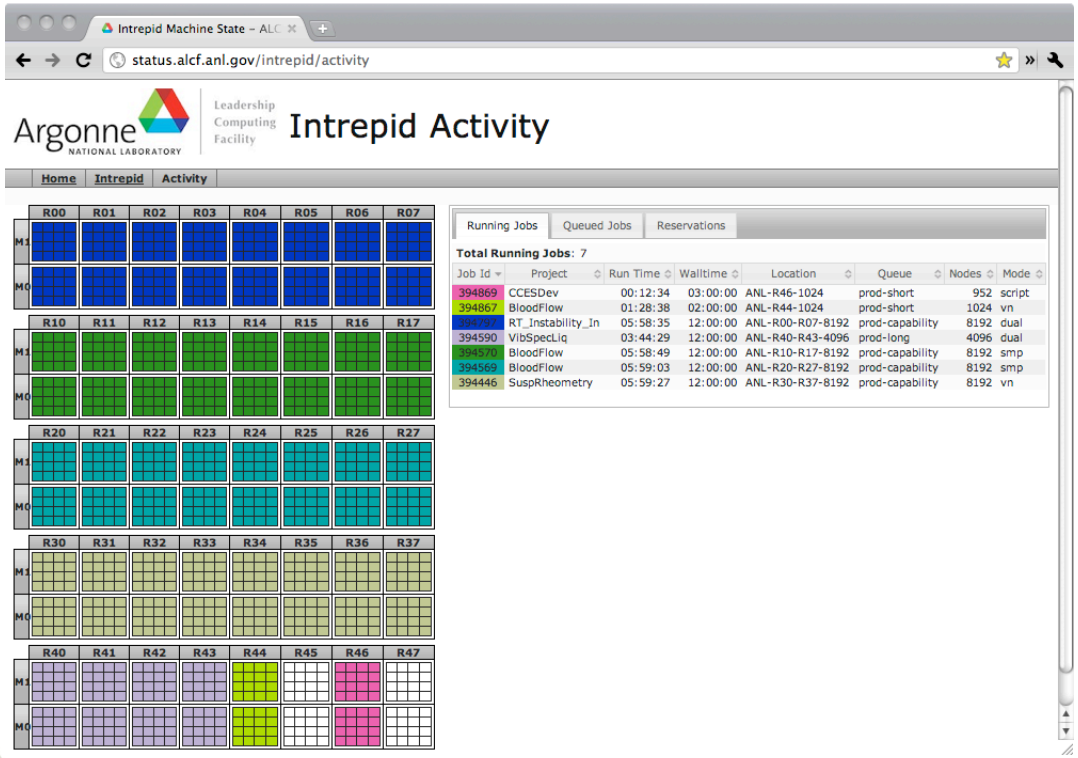


How To Use Intrepid

- **Log into login node (linux)**
- **(Cross) compile your program**
- **Submit a batch job using Cobalt**
  - Much like PBS
  - `qsub -n 1024 -mode vn -t 240 -A MyAllocation ./myexecutable`
  - Monitor with `qstat`

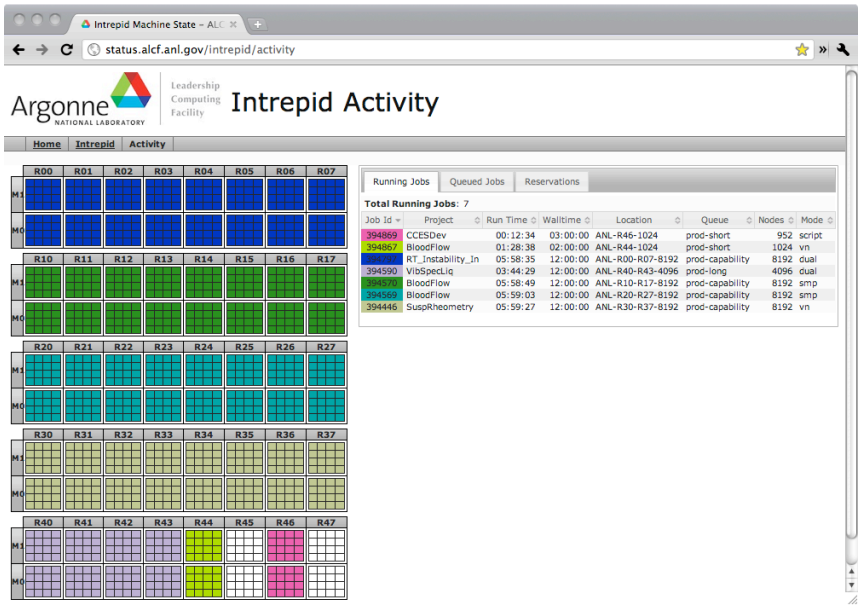
How To Use Intrepid

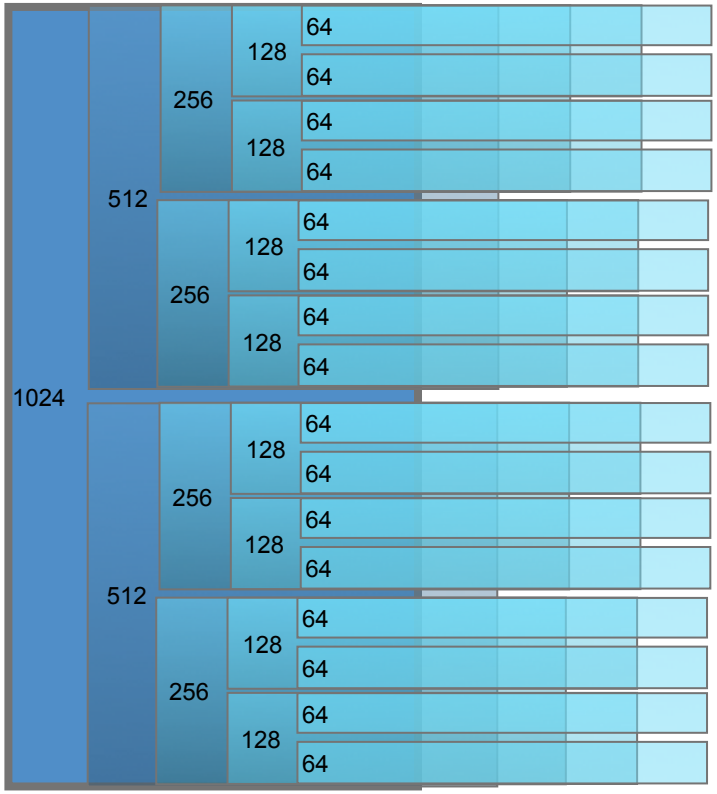
- **Log into login node (linux)**
- **(Cross) compile your program**
- **Submit a batch job using Cobalt**
  - Much like PBS
  - `qsub -n 1024 -mode vn -t 240 -A MyAllocation ./myexecutable`
  - Monitor with `qstat`



How To Use Intrepid

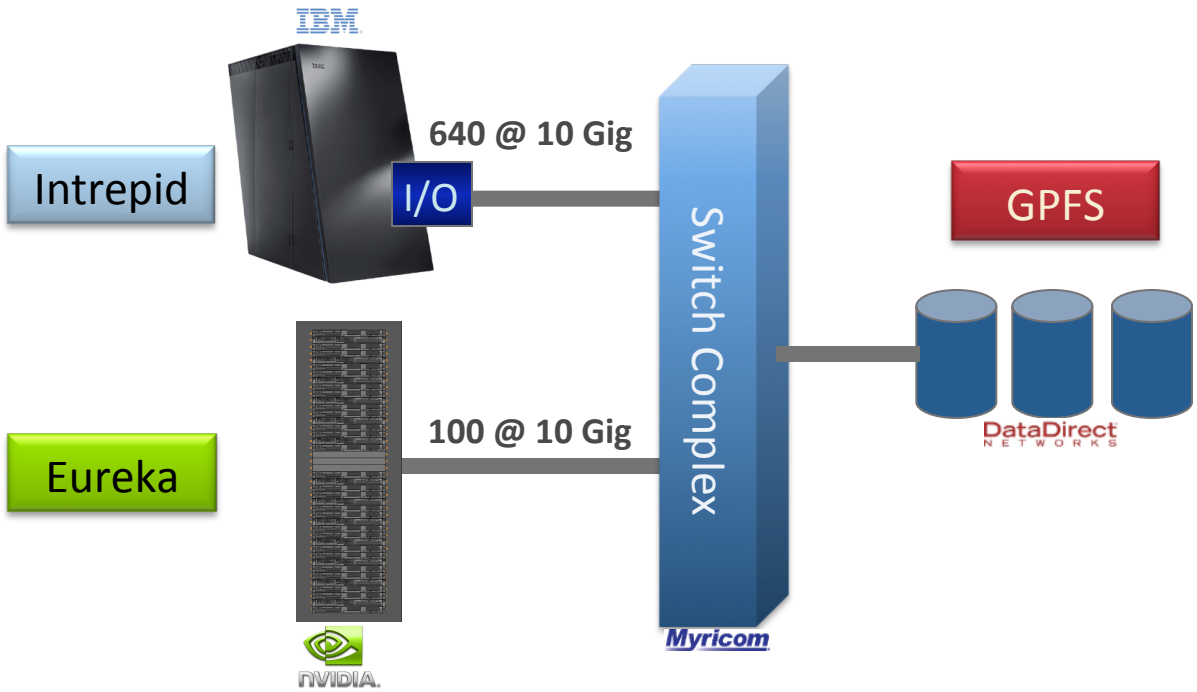
- **Log into login node (linux)**
- **(Cross) compile your program**
- **Submit a batch job using Cobalt**
  - Much like PBS
  - `qsub -n 1024 -mode vn -t 240 -A MyAllocation ./myexecutable`
  - Monitor with `qstat`





Visualization and Data Analytics

- ***Eureka* enables data analytics and visualization at Intrepid's scale**
- **Attached directly to core switch complex with Intrepid**
  - High-bandwidth socket connections between BG/P and analysis nodes
  - Direct access to same parallel file system



Visualization Software

- **Supported software includes:**
  - VisIT
  - ParaView
  - VTK
  - VMD
- **Software stack driven by our users needs**

Future ALCF System: Blue Gene/Q

- **Evolution of the Blue Gene architecture**
  - 16 cores/node
  - 16 GB of memory/node
  - water cooled
- **Coming in 2012: *Mira***
  - 10 petaFLOPS
  - Over 750K cores
  - 800 TB of memory
  - 70 PB of disk
- **BG/P applications should run immediately on the BG/Q**
  - Better performance expected with higher levels of on-node parallelism

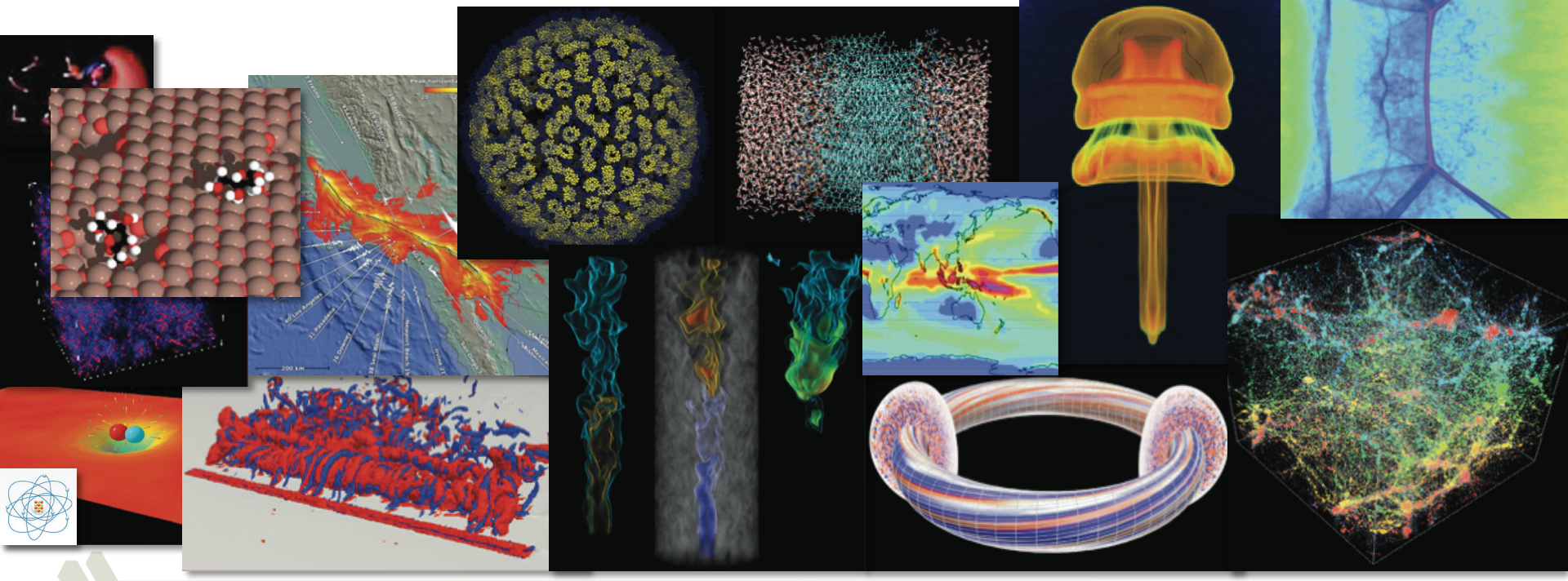
Future ALCF System: Blue Gene/Q

- **Evolution of the Blue Gene architecture**
  - 16 cores/node
  - 16 GB of memory/node
  - water cooled
- **Coming in 2012: *Mira***
  - 10 petaFLOPS
  - Over 750K cores
  - 800 TB of memory
  - 70 PB of disk
- **BG/P applications should run immediately on the BG/Q**
  - Better performance expected with higher levels of on-node parallelism

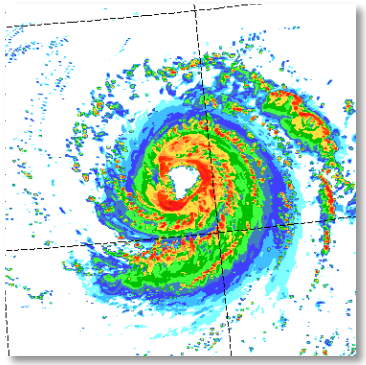
Threads!

First in *Mira* Queue: Early Science Program

- **16 projects**
  - Large target allocation
  - Postdoc
- **Proposed runs between *Mira* acceptance and start of production**
- **2 billion core-hours to burn in a few months**

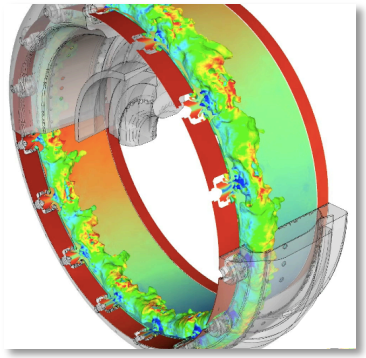


ALCF Projects Span Many Domains



Climate

Predicting hurricane tracks to mitigate risks, hindcasting with climate model data to gauge impact of global change.

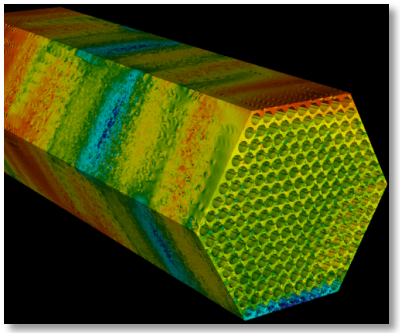


Gas Turbines

Modeling two-phase flow and combustion for the design of more efficient aircraft engines.

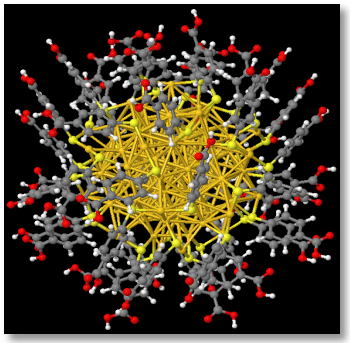
Nuclear Energy

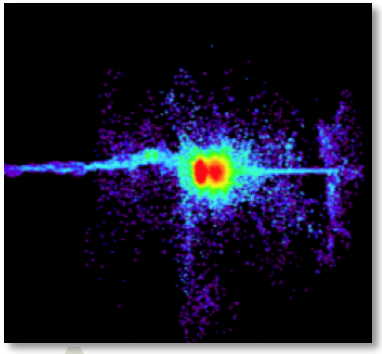
High-fidelity fluid flow and heat transfer simulation of next-generation reactor designs, aiming to reduce the need for costly experimental facilities.



Nano Catalysts

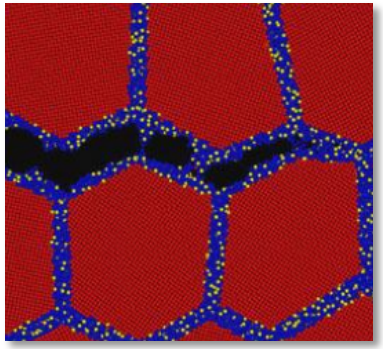
Mapping out properties of gold nanoparticles to design catalysts for fuel cells and methane conversion.





Fusion Energy

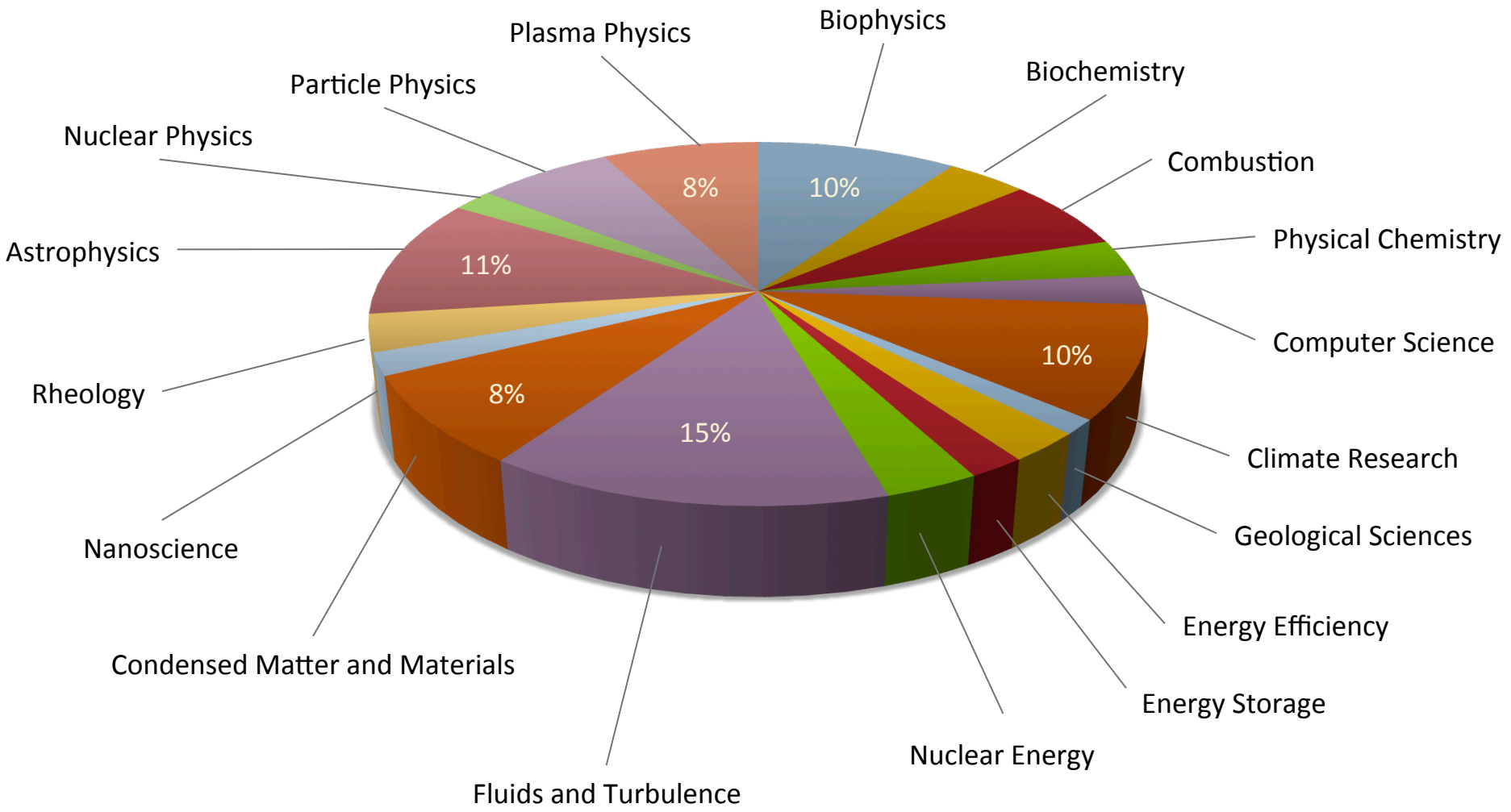
Understanding the detailed physics of Fast Ignition inertial confinement fusion.



Materials Science

Molecular simulation of fracture dynamics in structural materials in next-generation nuclear reactors.

2011 INCITE Allocations at ALCF



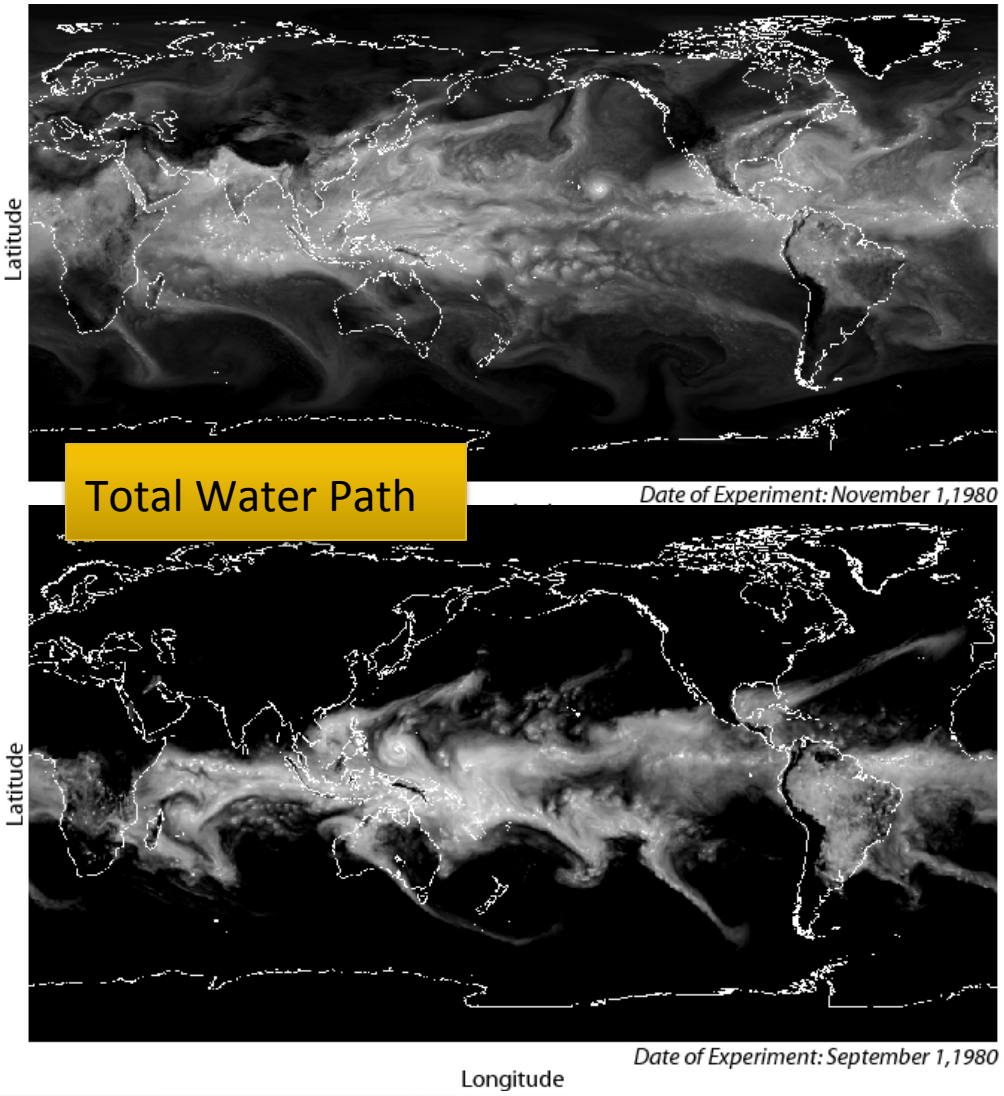
Toward Ultra-High Resolution Climate-Weather Modeling

S-J Lin, Geophysical Fluid Dynamics Laboratory

- **HIRAM global atmospheric code**
  - Cubed sphere grid

Hydrostatic:
12 km resolution

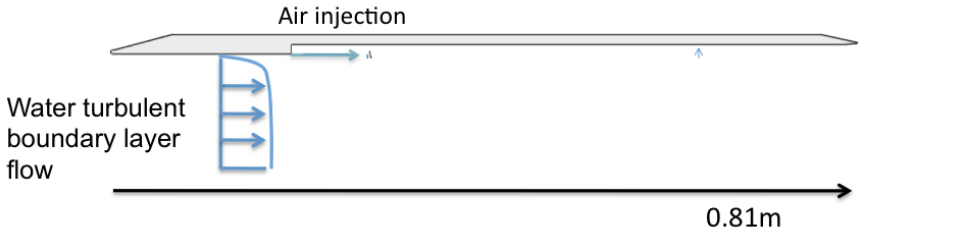
Non-Hydrostatic:
3.5 km resolution
Global cloud resolving

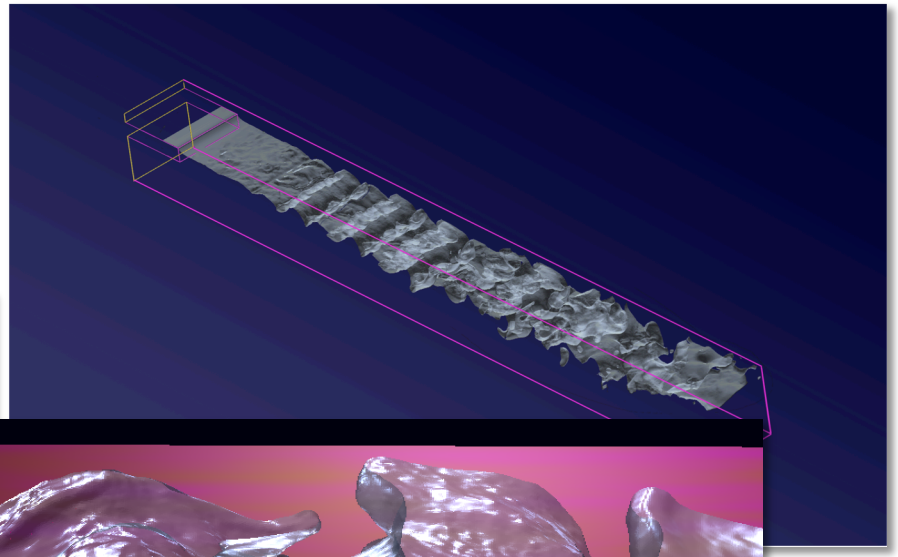


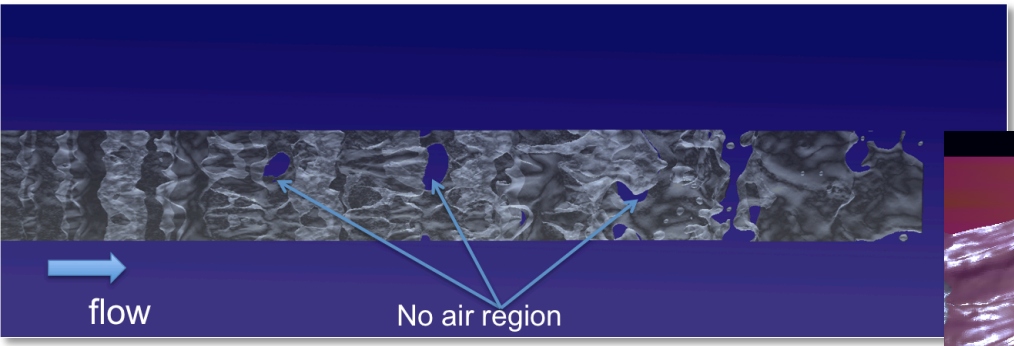
Air-Layer Drag Reduction

Parviz Moin, Stanford University

- **3D LES simulation of turbulent multiphase flows on a flat plate**







Breakup of air layer →
Water spots→
Much larger skin friction

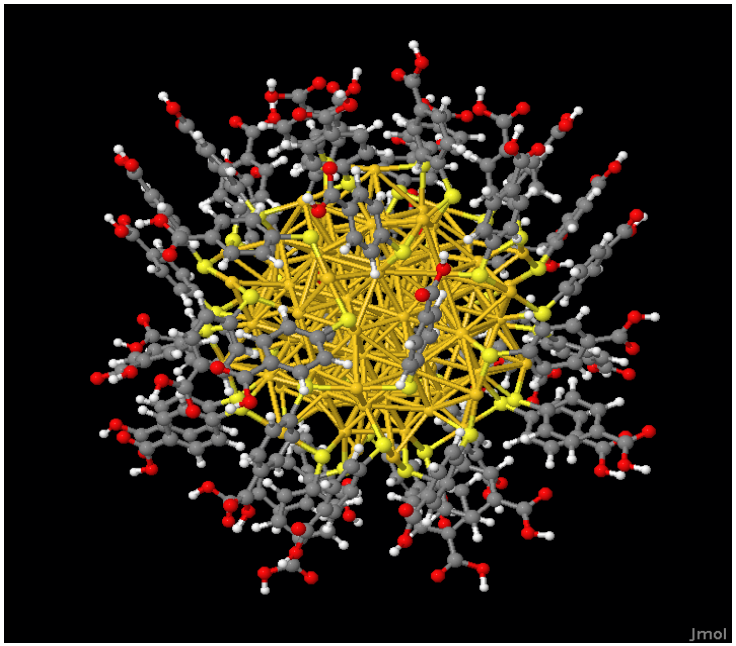
Kelvin-Helmholtz instability

Catalytic Nanoparticles

Jeff Greeley, Argonne National Laboratory

- **Code: GPAW (DFT)**
- **Gold has exciting catalytic abilities, still poorly understood**
  - e.g. CO into CO_2
- **Simulated nanoclusters of 13-923 atoms**
  - 32 racks on *Intrepid* (80% of system)

ALCF Catalyst:
- Reduced memory footprint
- New layer of parallelization
- Expanded production from 128 cores to 16,384

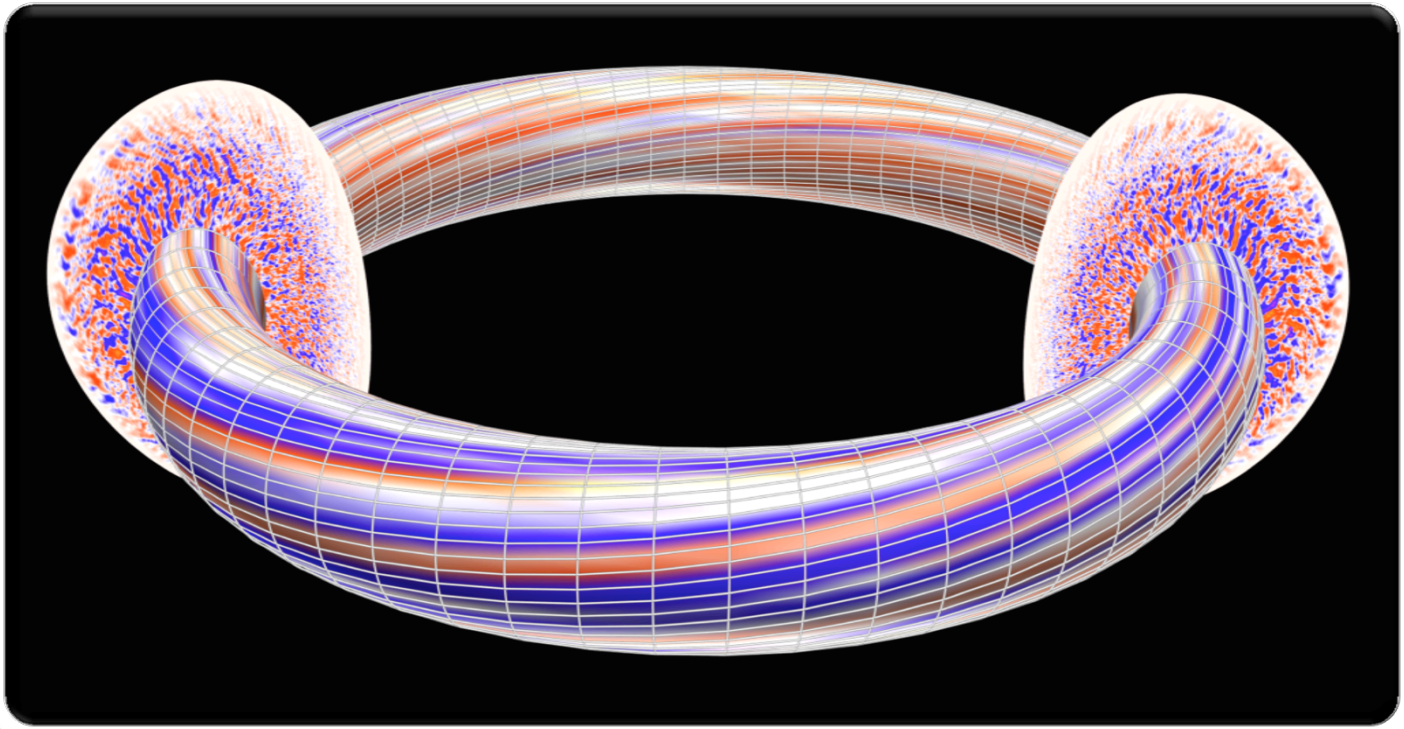


Microturbulent Transport in Tokamaks

William Tang, Princeton Plasma Physics Laboratory

- **Code: GTC**
- **Particle-in-cell simulation of plasma**
  - Study energy loss through turbulence
  - Trying to validate key assumption about scaling in ITER

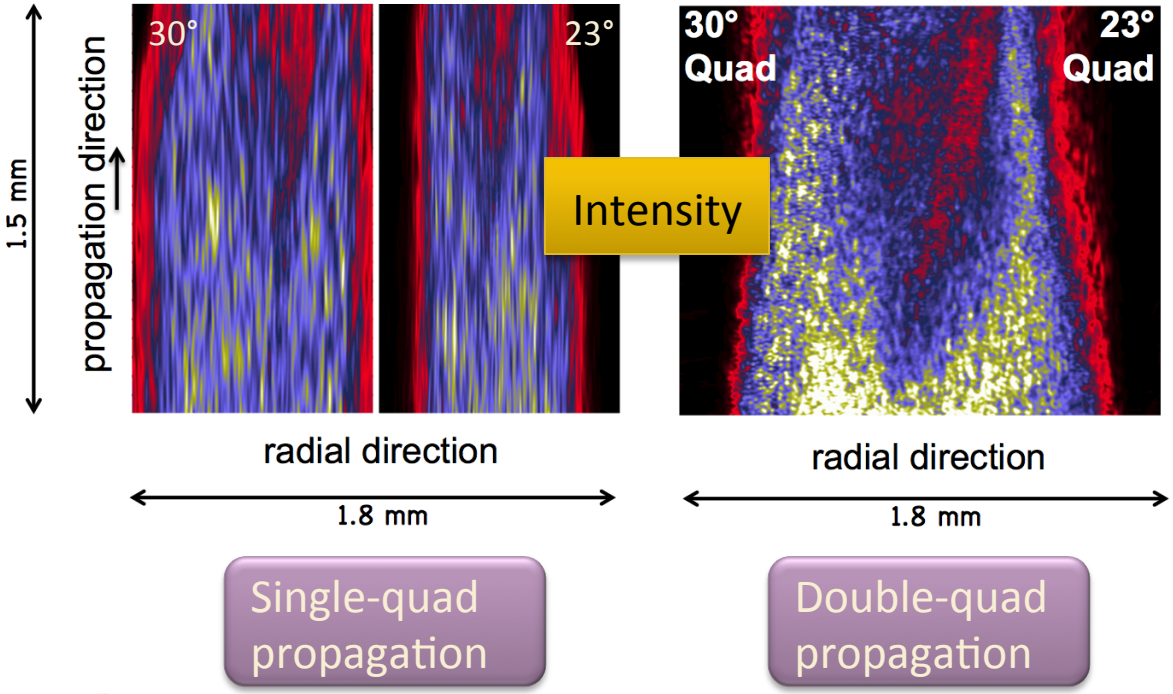
$O(10^9)$ particles
$O(10^8)$ grid cells

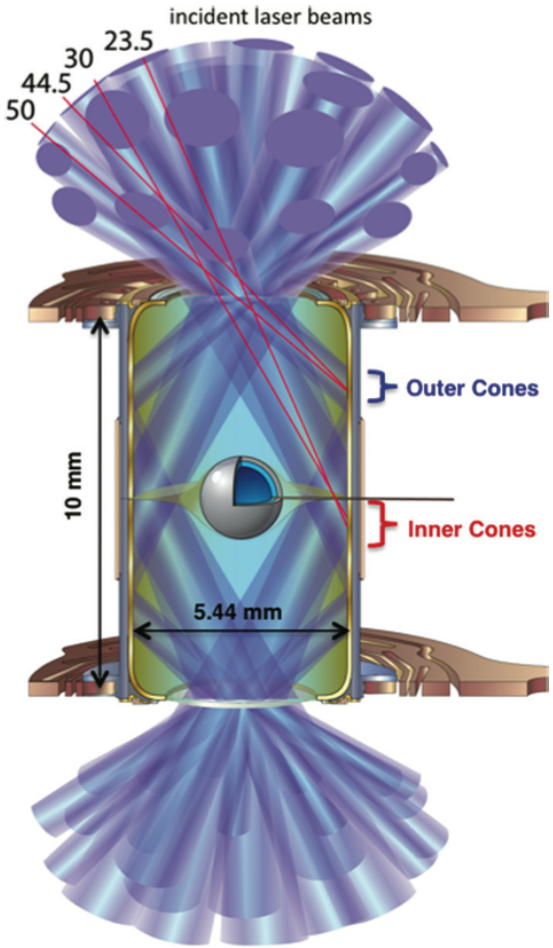


Laser-Plasma Interaction in ICF

Denise Hinkel, Lawrence Livermore National Laboratory

- **Code: pF3D (radiation & hydrodynamics)**
- **Laser backscatter**
  - Problem for indirect-drive inertial confinement fusion
  - Overlap of beams important!



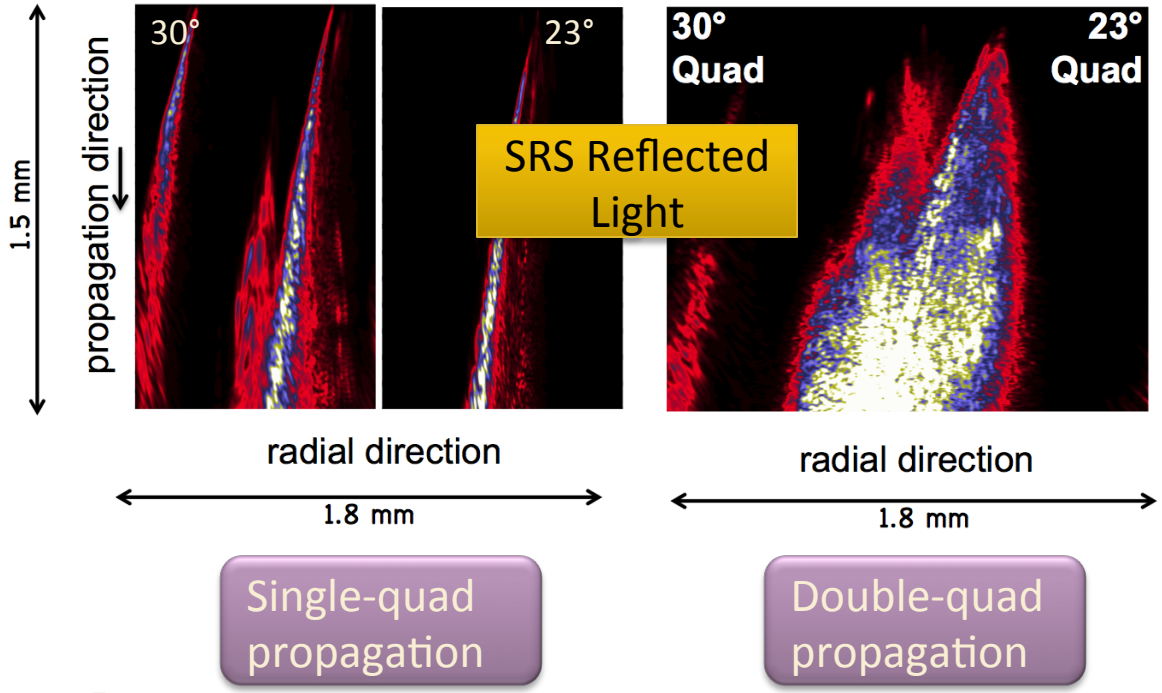


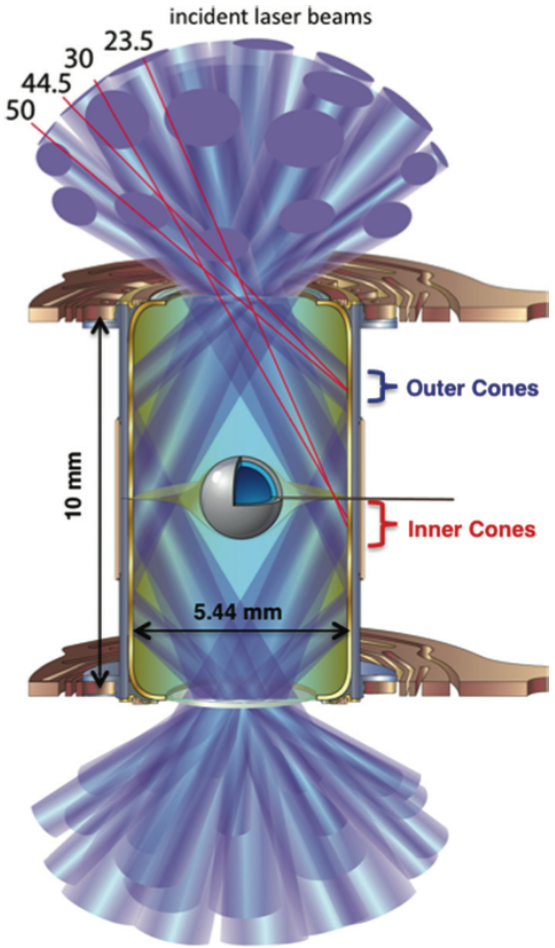
Physics of Plasmas 18, 056312 (2011)

Laser-Plasma Interaction in ICF

Denise Hinkel, Lawrence Livermore National Laboratory

- **Code: pF3D (radiation & hydrodynamics)**
- **Laser backscatter**
  - Problem for indirect-drive inertial confinement fusion
  - Overlap of beams important!

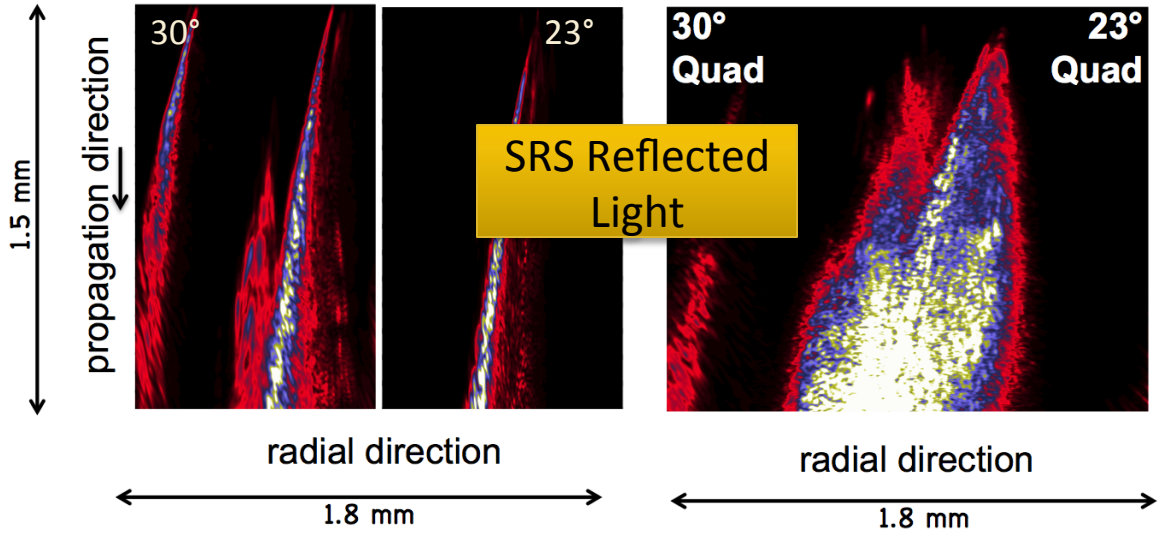




Laser-Plasma Interaction in ICF

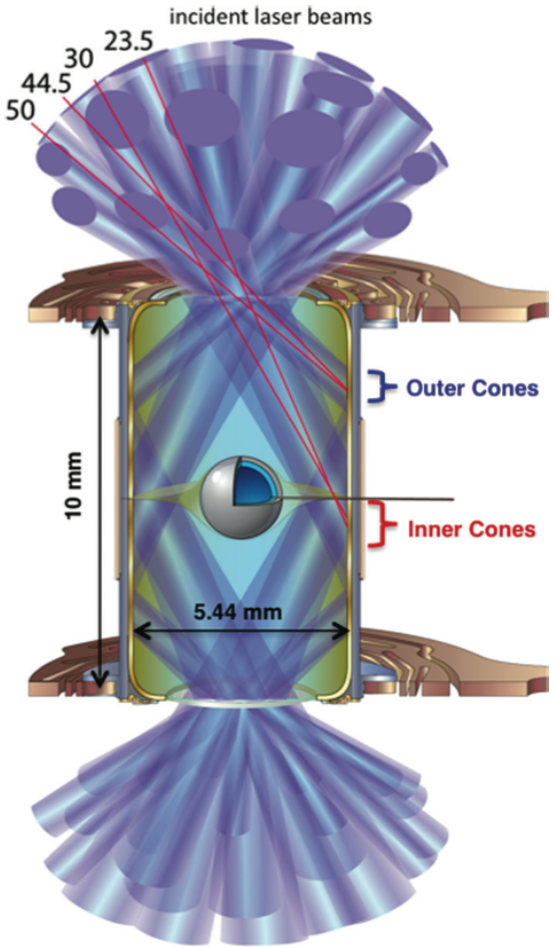
Denise Hinkel, Lawrence Livermore National Laboratory

- **Code: pF3D (radiation & hydrodynamics)**
- **Laser backscatter**
  - Problem for indirect-drive inertial confinement fusion
  - Overlap of beams important!



Single-quad propagation

Double-quad propagation



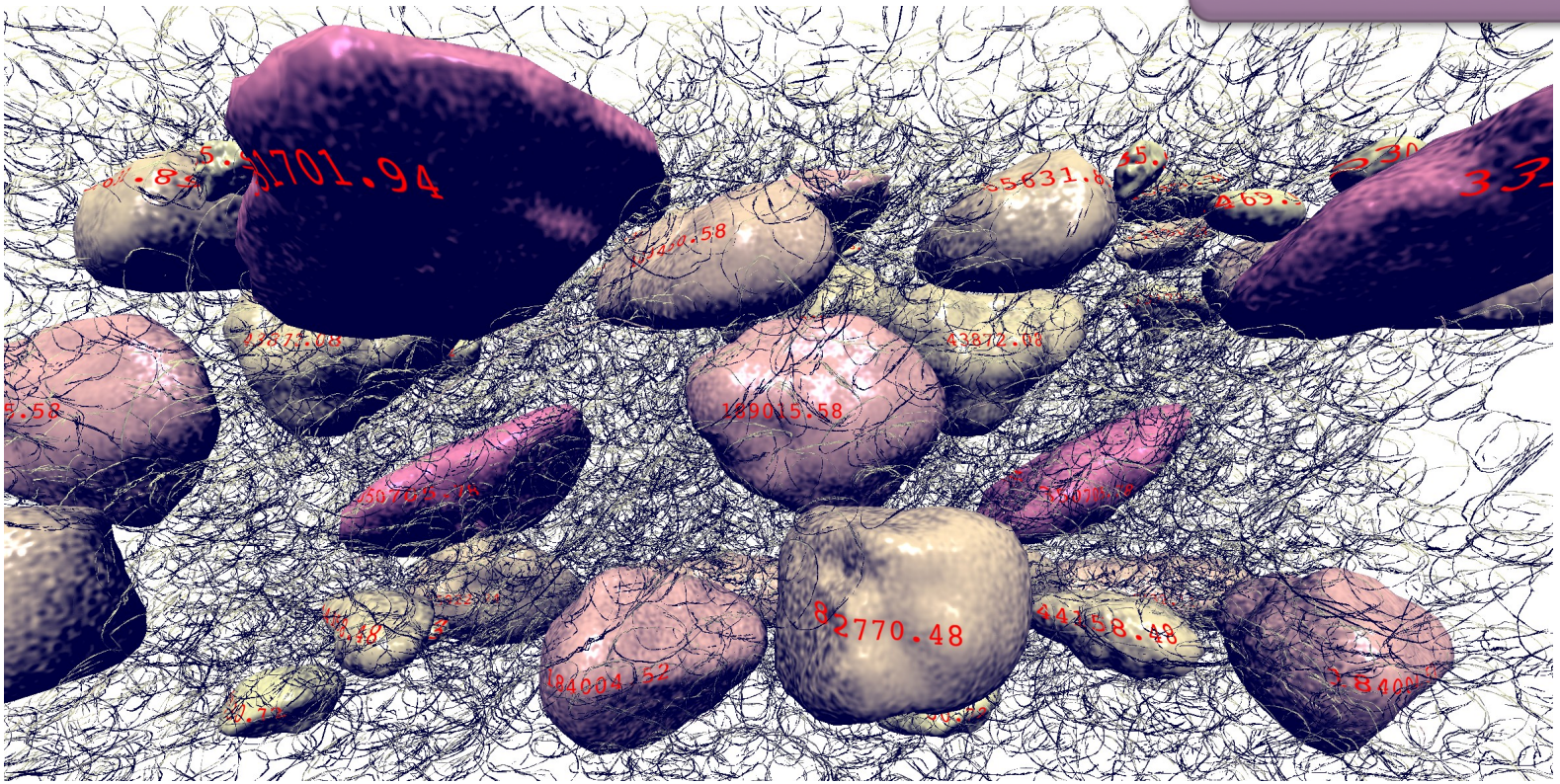
32K *Intrepid* nodes (132k cores)

Rheology of Concrete

William George, National Institute of Standards and Technology

- **How do shape and size of aggregate impact behavior?**
- **Simulate how stress propagates through the fluid.**
  - Non-Newtonian matrix fluid
  - Code: smoothed particle hydrodynamics (SPH)

Stress level "painted" on surfaces

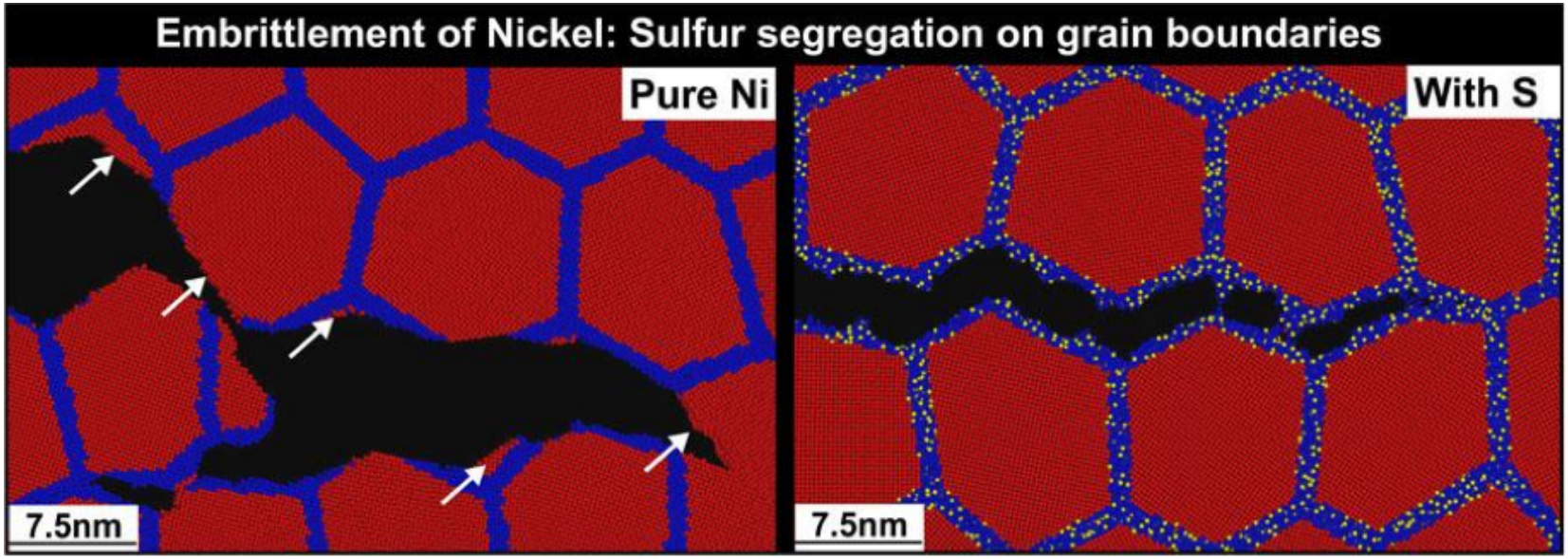


Reactive MD Simulation of Nickel Fracture

Priya Vashista, University of Southern California

- **Sulfur atoms in nickel grain boundaries embrittle the nickel**
  - Phys. Rev. Letters 16 Apr 2010
- **65K cores on Intrepid BlueGene/P**
  - 48 million atoms
  - Chemically reactive molecular dynamics
  - Fracture mechanics modeled

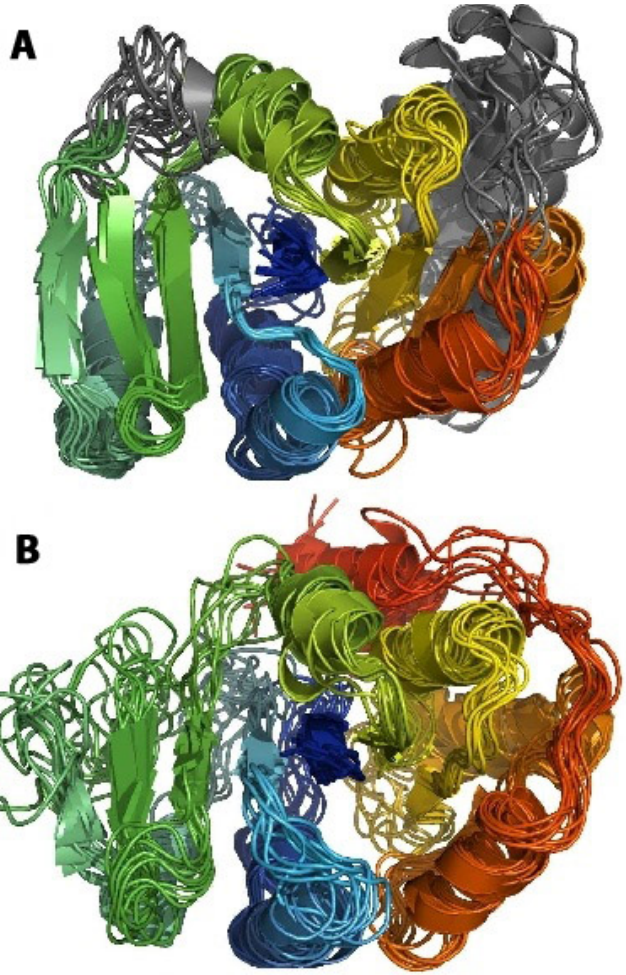
Relevant to next-gen nuclear reactors.



Computational Protein Structure Prediction and Protein Design

David Baker, University of Washington

- **INCITE, PI David Baker (U. Washington)**
  - Rosetta protein folding application
    - *Science* **327**, 1014 (2010)
      NMR Structure Determination for Larger Proteins Using Backbone-Only Data
  - Structure of ALG13
    (A) computationally
    (B) experimentally by NMR



Systems with > 150 amino acids
- Full experimental NMR difficult
- Backbone data easier:
  Input to Rosetta

CScADS Workshop Accounts & Allocation

- **If you don't have an account on an ALCF BG/P system (*Intrepid or Surveyor)* you can apply for a workshop account**
- **To apply:**
  - Go to the URL: https://accounts.alcf.anl.gov/accounts/request.php
  - Select "Proceed with Account Request" at the bottom of the page
  - Select the project 'CScADS'
  - Foreign nationals require 593 forms which can take a while
- **Running under the 'CScADS' project**
  - User with account can run using the 'CScADS' project
    - qsub –A CScADS …

If you want to know more…

- **ALCF web site: www.alcf.anl.gov**
  - Information on ALCF system and activities
  - Information on applying for accounts
- **Getting Started Guide:**
  - https://wiki.alcf.anl.gov/index.php/Quick_Reference_Guide
- **ALCF Support Wiki: wiki.alcf.anl.gov**
  - Documentation
  - FAQ
- **Support email address: support@alcf.anl.gov**
  - Any question: account, technical, etc.


IBM