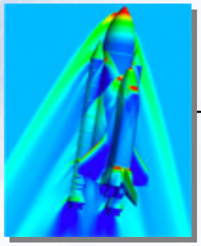
Petascale Computing: Impact on Future NASA Missions

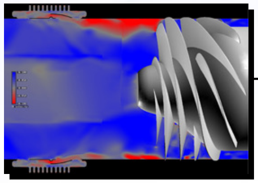
Architectures and Algorithms for Petascale Computing
Schloss Dagstuhl, Germany
February 12-17, 2006

Rupak Biswas
Chief (Acting)
NASA Advanced Supercomputing Division
NASA Ames Research Center, Moffett Field, California
http://www.nas.nasa.gov

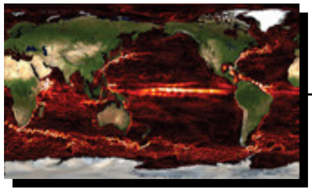
NASA’s Engineering and Science Applications



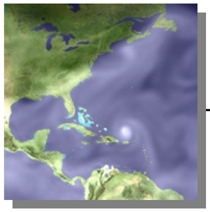
- Aerospace Analysis and Design



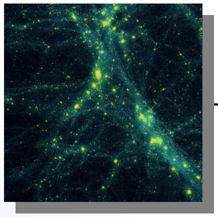
- Propulsion Subsystem Analysis



- Climate Modeling



- Hurricane Prediction



- Astrophysics and Cosmology

Columbia – World Class Supercomputing

- **Provide supercomputing & storage environment to maximize NASA user productivity and achievement**
- **Fast build: Order to full ops in 120 days; ~10X faster than similar systems; dedicated Oct. 2004**
  - Unique partnership w/ industry (SGI, Intel, Voltaire)
  - Built from trusted components (Altix)
- **Immediate impact: Full production and increased capacity during build**
- **Leadership capability: 4th fastest supercomputer in world: 62 Tflops peak**
  - Linpack runs at 51.9 Tflops
- **Effective architecture: Easier application scaling for high-fidelity, shorter time-to-solution, higher throughput**
  - 20 x 512p/1TB shared memory nodes
  - Some apps scaling to 2048p and above
- **Supporting all Mission Directorates**
  - >160 projects; >900 accounts; ~150 simultaneous logins
  - Users from across and outside NASA
  - 24x7 support; high Quality of Service
- **Deep and Broad Mission Impact!**

Systems: SGI Altix 3700 and 3700-BX2
Processors: 10,240 Intel Itanium 2
Global Shared Memory: 20 Terabytes

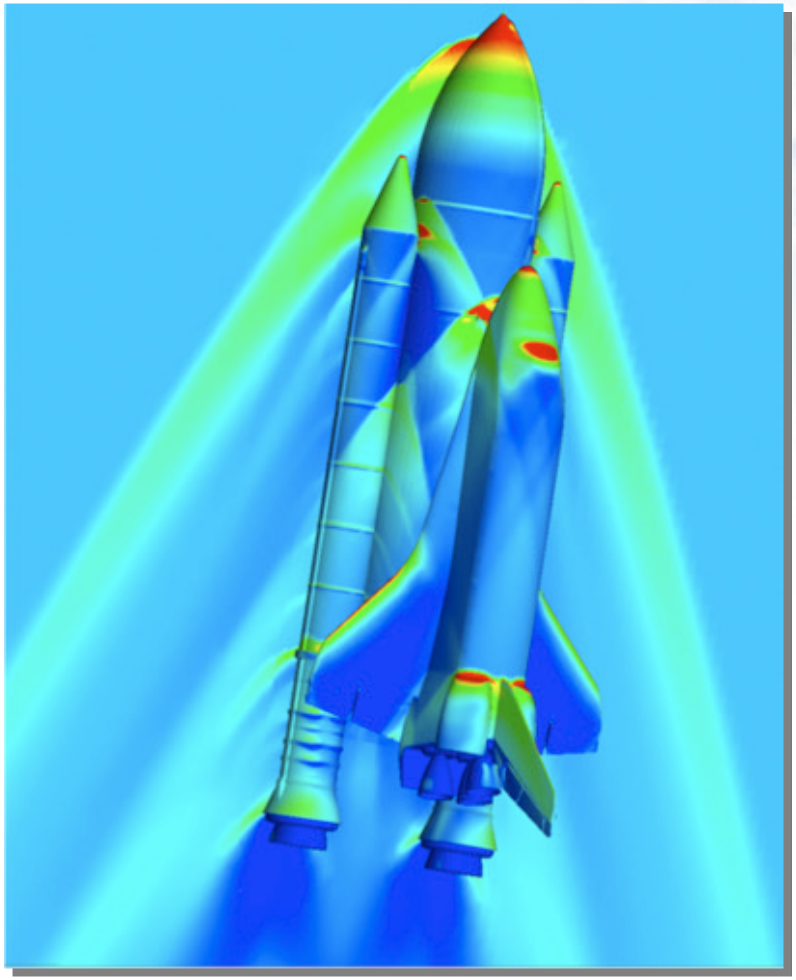
Front-End: SGI Altix 3700 (64 proc.)
Online Storage: 440 Terabytes RAID
Offline Storage: 6 Petabytes STK Silo

Internode Comm: Infiniband
Hi-Speed Data Transfer: 10 Gigabit Ethernet
2048p subcluster: NUMAlink4 interconnect

Aerospace Analysis and Design

Relevance to NASA missions:

- High-fidelity CFD techniques are developed and applied to many NASA aerospace analysis & design problems:
  - Full Space Shuttle Launch Vehicle (SSLV) configuration including orbiter, external tank, solid rocket boosters, and fore / aft attached hardware
  - Exploration vehicles (e.g. CEV, CRV, HLV)
  - Solid Rocket Booster (SRB) blast within Vehicle Assembly Building (VAB)
  - Flame trench analysis for Launch Pad



Pressure contours around full SSLV configuration

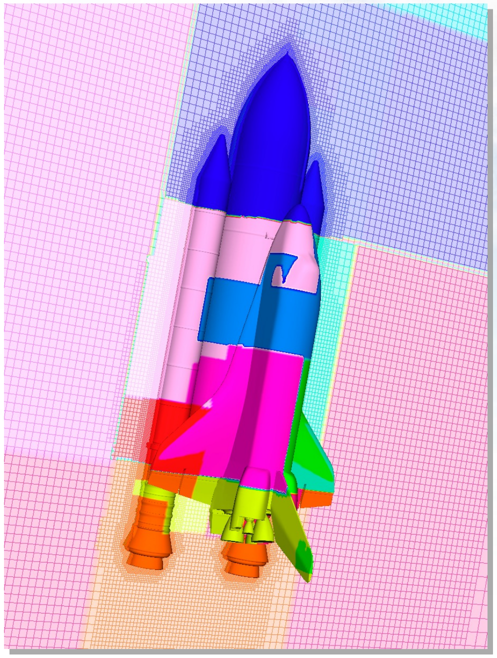
Numerical Methods of Research:

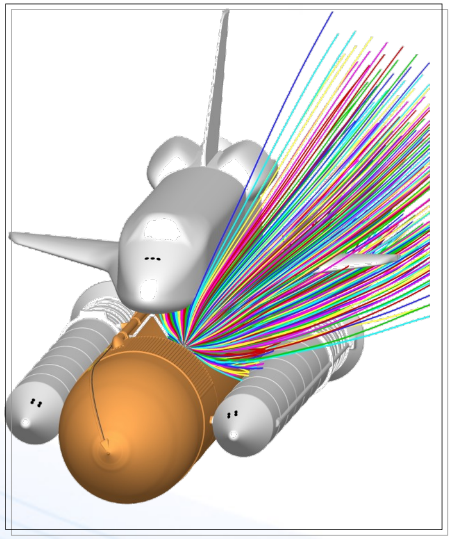
- Two high-performance aerodynamic simulation packages were used on Columbia:
  - *Overflow:* To perform analysis at the most important flight conditions and drive the high-fidelity design optimization procedure
  - *Cart3D:* To validate the new design over a broad range of flight conditions
- Combination of *Overflow* & *Cart3D* enables high-fidelity characterization of aerospace vehicle design performance over entire flight envelope

POC: *Michael Aftosmis, NASA Ames Research Center*
(650) 604-4499, Michael.J.Aftosmis@nasa.gov

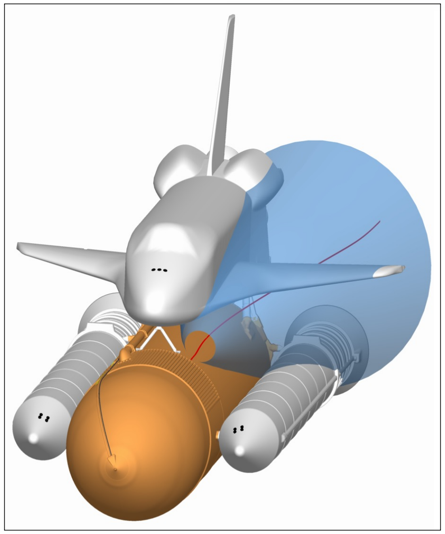
Cart3D

- Inviscid analysis package insensitive to geometric complexity
- Aimed at aerodynamic database generation via parametric studies
- Includes modules for surface modeling, mesh generation, data extraction; highly automated
- Unstructured (cut-cell) Cartesian, finite-volume upwind, multigrid
- Code widely disseminated: NASA, DoD, DOE, intel Agencies, US aerospace industry
- For Shuttle RTF, computed unsteady moving-body 6-DOF simulations of isolated pieces of debris to develop drag models and cross-range in supersonic flow

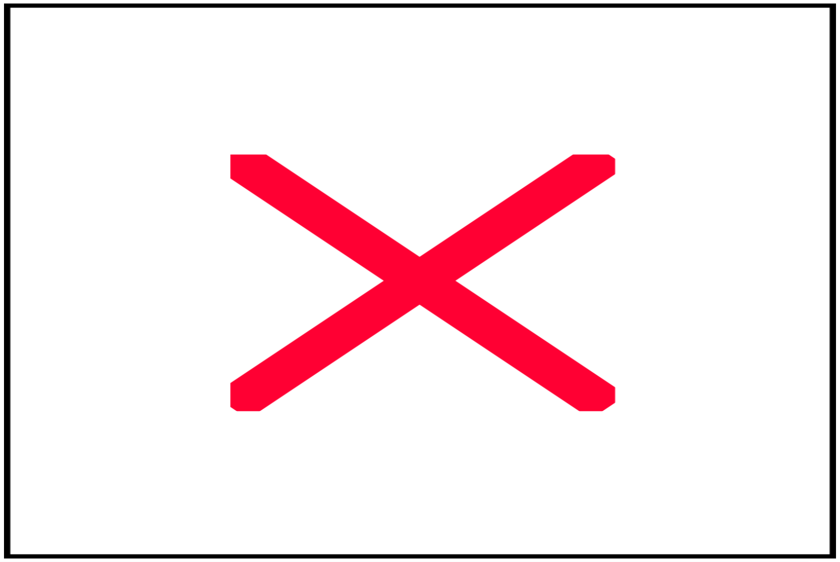




Deterministic
Zero Lift Trajectory +
Range of Initial Velocities



Probabilistic
Zero Lift Trajectory +
Crossrange Cone



Results were critical in returning Shuttle to flight

Parallelization Strategy

– Space-Filling-Curve based partitioner and mesh coarsener

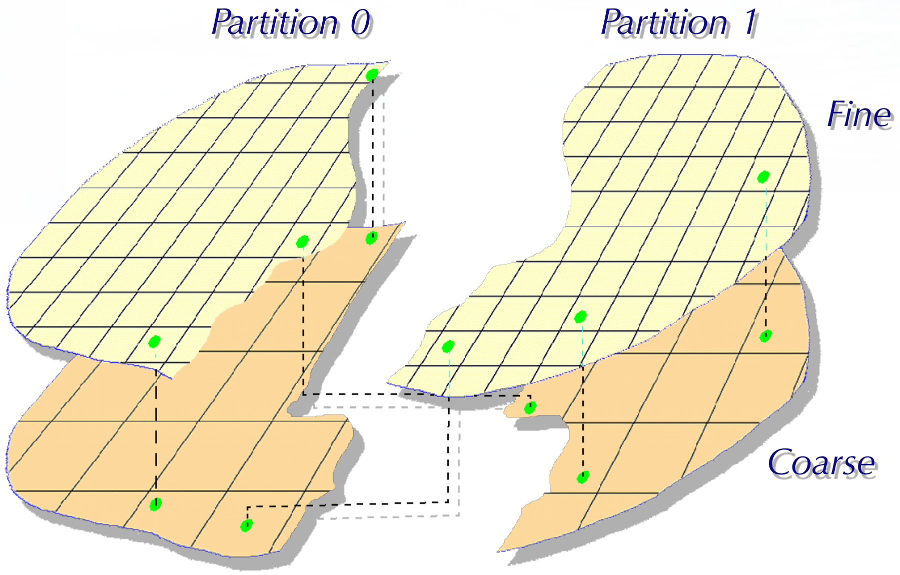
– Each subdomain resides in processor local memory

– Each subdomain has own local grid hierarchy

– Good (not perfect) nesting;
  favor load balance at each level

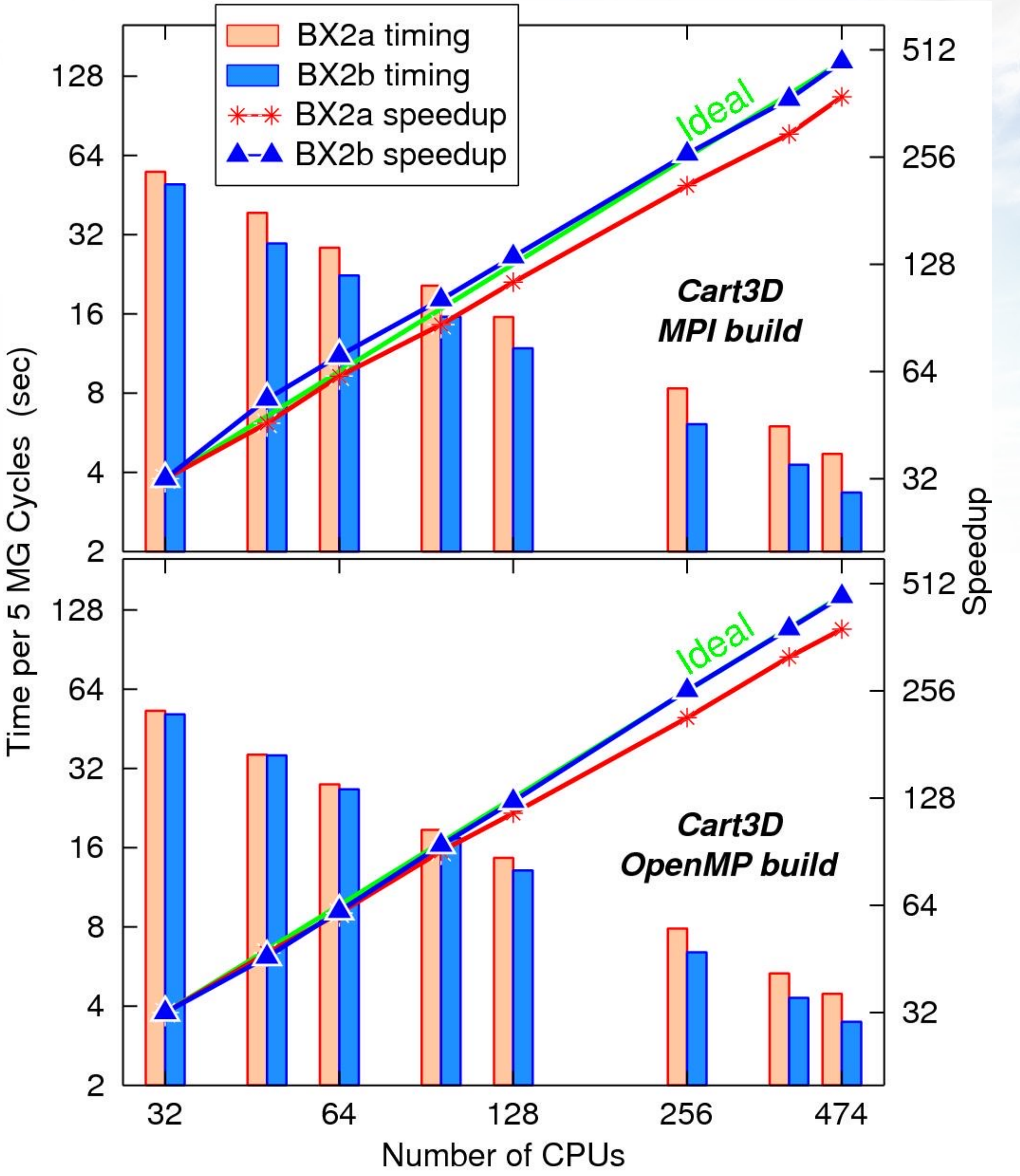
– Restrict use of naïve OpenMP constructs:
  use MPI-like strategy

– Exchange via structure copy (OpenMP),
  send/receive (MPI)

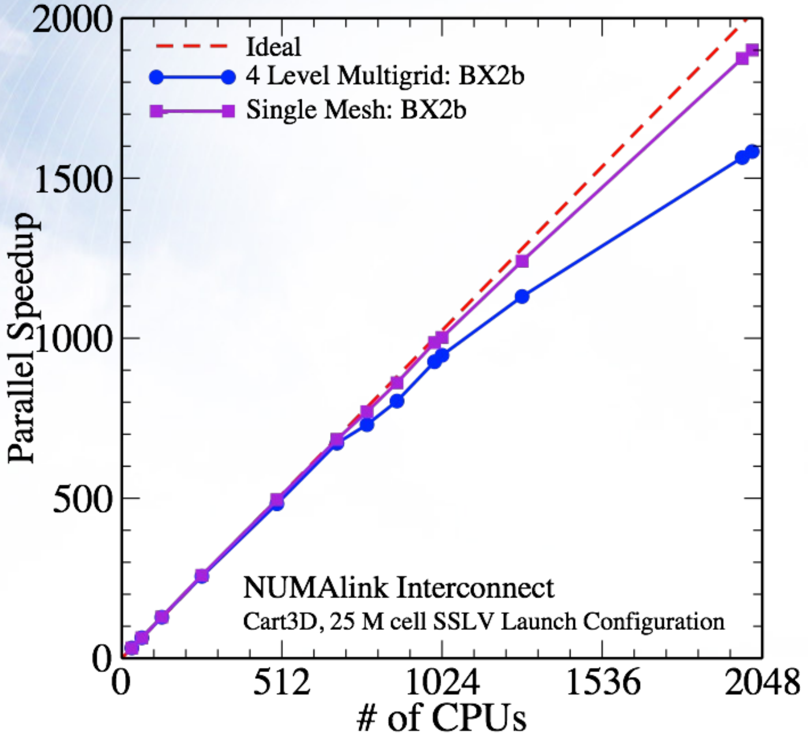


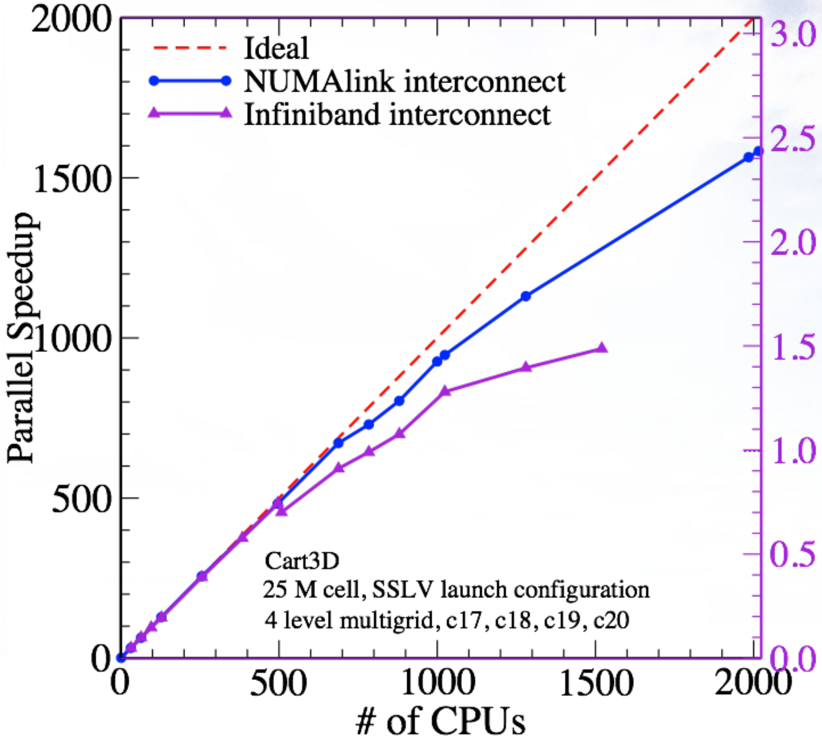
Cart3D Columbia Results: Single Node

- Columbia has 3 types of nodes
  - 3700 (1.5 GHz, 6 MB, 3.2 GB/s)
  - BX2a (1.5 GHz, 6 MB, 6.4 GB/s)
  - BX3b (1.6 GHz, 9 MB, 6.4 GB/s)
- Parallel speedup on 474p
  - BX2b: 473.8 (MPI), 472 (OpenMP)
  - BX2a: 380 (both MPI & OpenMP)
- Same interconnect, but BX2b has faster CPUs
  - BX2 routers can provide sufficient bandwidth to handle increased data consumption rates of faster CPUs
  - BX2a should have better scalability and BX2b better raw timings
  - BX2a shows poorer scalability due to slight load imbalance (weights set experimentally for given platform)
- Scalability and runtimes on BX2 consistently better than that on 3700



Cart3D Columbia Results: Multiple Nodes





- NL on BX2b using MPI
- 32-496p on one node; 508-100p on two nodes; 1000-2016p on 4 nodes
- Reducing number of multigrid levels de-emphasizes communication
- Parallel speedup on 2016p:
  1900 (single grid), 1585 (multigrid)
- Coarsest mesh in 4-level multigrid has only ~16 cells per partition

- 4-level multigrid using MPI
- No inter-node comm up to 496p
- IB runs end at 1524p due to limited number of connections
- IB performance lags due to reduction in delivered bandwidth
- Bandwidth drops again when going from 2 to 4 nodes
- NL on 2016p achieves 19% of peak

Aerospace Analysis and Design with Petascale Computing

What can be accomplished with petascale computing?

- Get closer to reaching the full potential of what high-fidelity numerical dynamic simulation techniques have to offer: deliver more optimal designs (larger parametric analysis) and accelerate design cycle (reduced time-to-solution)
  - Aerospace vehicles could potentially be "flown" through the database by guidance and control system designers to:
    Explore issues with stability and control
    Determine vehicle's suitability for various NASA mission profiles

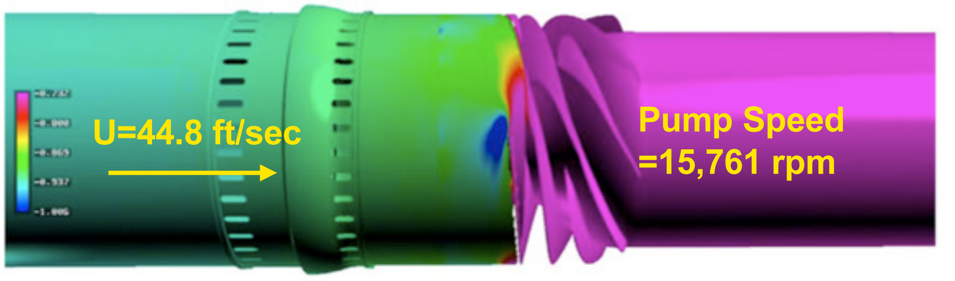
What are the architecture and algorithm bottlenecks?

- Bandwidth is the biggest problem facing CFD solvers today
  - Important also when synchronizing calculation over large number of processors
- Current high-fidelity NASA CFD codes are processor-speed-bound on Columbia
  - Runs utilizing many hundreds of processors typically use small fraction of available memory
- Significantly higher grid resolution required to improve accuracy of CFD methods
- Investment in scalable solution techniques must be made to replace current block tri- and penta-diagonal solvers used in NASA's production codes
  - Program models which emphasize good scalability for message passing architectures need to be developed
  - Fundamental change in solution strategy and algorithm are required to make use of petascale systems (multiple cores, high clock rate, huge concurrency)

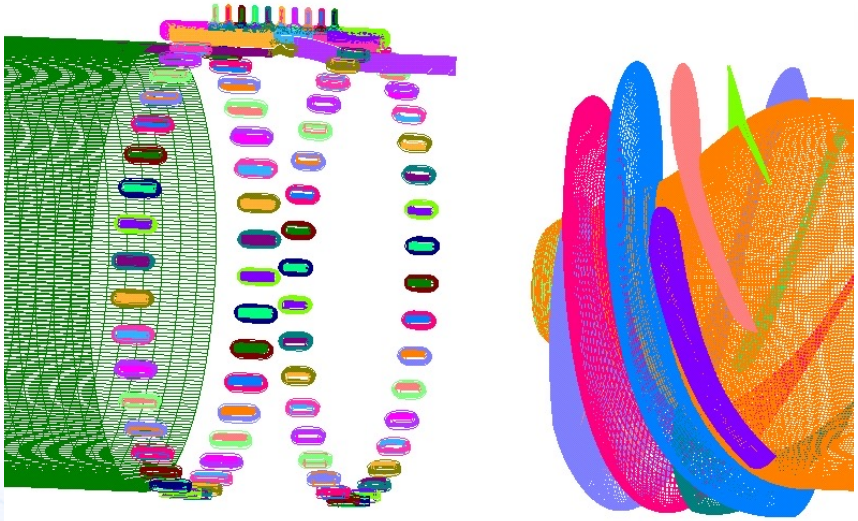
Propulsion Subsystem Analysis

Relevance to NASA missions:

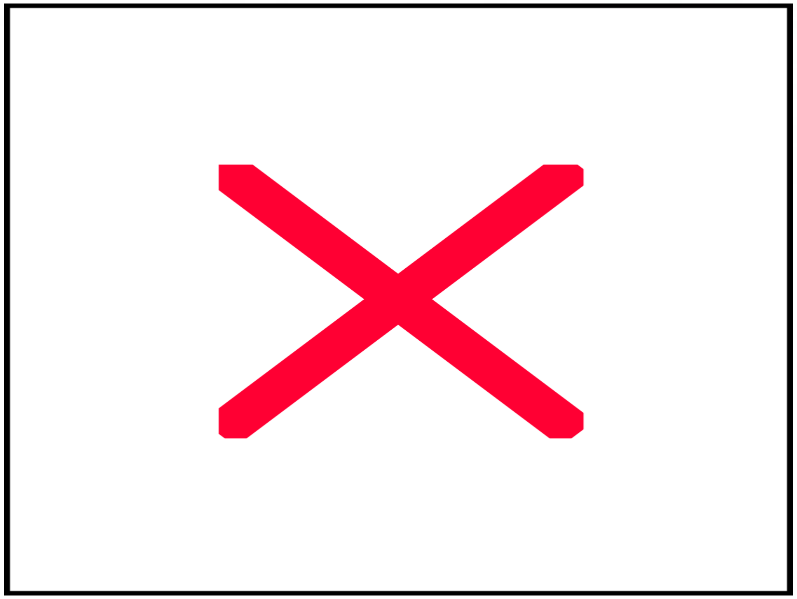
- High-fidelity unsteady flow simulation techniques for design and analysis of propulsion systems for NASA missions (Shuttle - retrofit, CEV - new design)
  - Liquid rocket engine flowliner analysis for Space Shuttle Main Engine (SSME)
  - Reduces cost of space flight—make design decisions early to improve efficiency, performance, and reliability based on computational models of propulsion systems



Instantaneous surface pressure contours on inducer and flowliner (flow unsteadiness cause flowliner cracks)



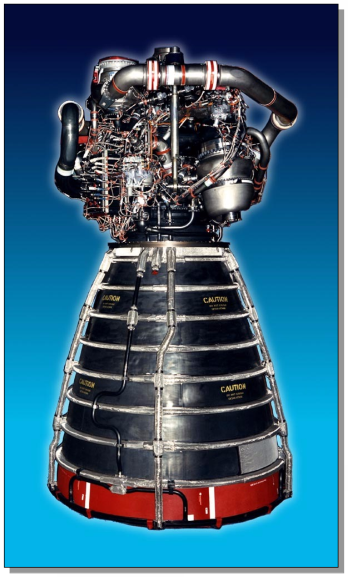
Flowliner overset grid system: 264 grid blocks of various sizes, total 66 M grid points



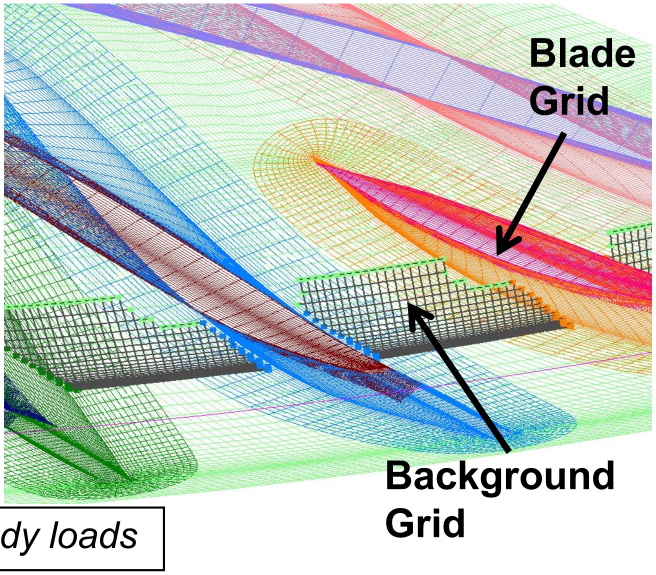
Computed particle traces colored by axial velocity direction (Blue: Forward; Red: Backward)

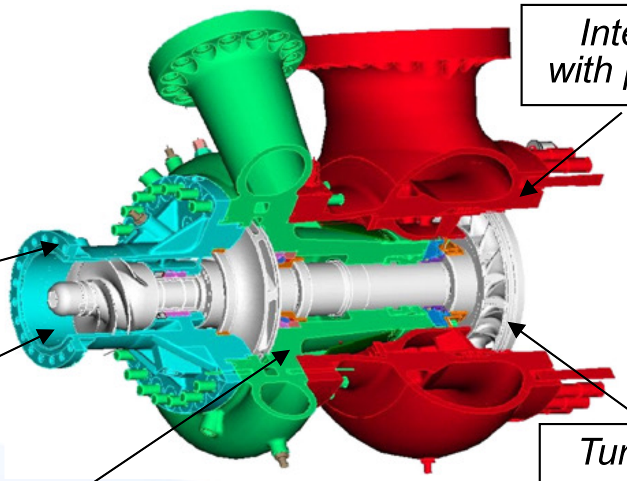
POC: *Cetin Kiris, NASA Ames Research Center, (650) 604-4485, Cetin.Kiris@nasa.gov*

INS3D



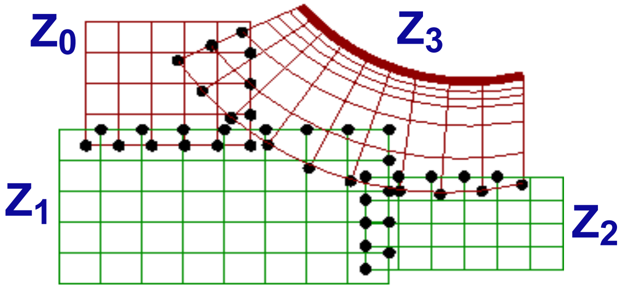
- Incompressible Navier-Stokes solver for complex configurations
- Based on method of artificial compressibility
- Multi-block structured overset grids w/ Chimera-style domain decomposition
- Spatial / temporal finite-differencing
- Moving grid capability
- Curvilinear body-fitted near-field grids embedded in Cartesian off-body grids
- Steady-state and time-accurate formulations
- 3rd and 5th-order flux difference splitting for convective terms
- Central differencing for viscous terms
- One- / two-equations turbulence models

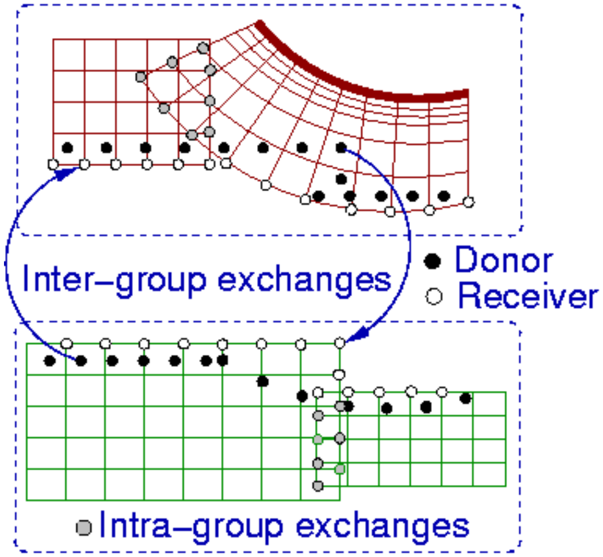



Interaction with preburner
Cavitation induced environments
Interaction between pump and feedline
Impeller unsteady loads
Turbine unsteady loads

Parallelization Strategy

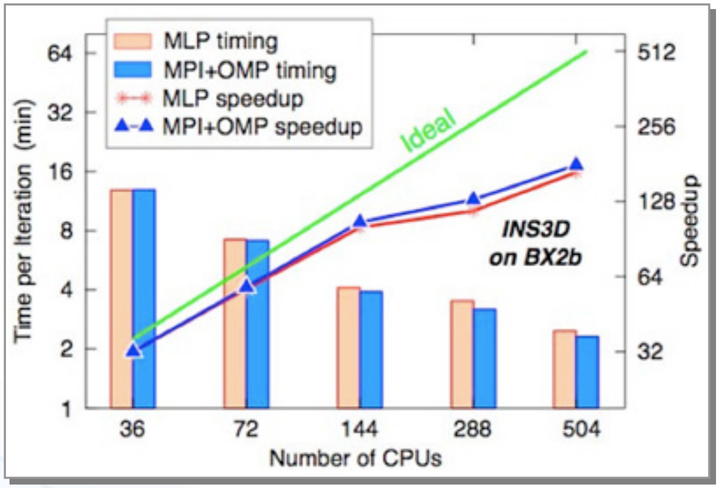
- Cluster grids into groups; assign each group to a processor
- Split large grids into subgrids; may lose implicitness
- Interpolation and exchange of boundary data at every time step
- Intra-group: local update / exchange within processor
- Inter-group: OpenMP, MLP shared-memory copy, or asynchronous MPI
- Two-level hybrid programming paradigm extends solutions for a given overset grid system to larger processor counts
  - **Level 1**: coarse-grained parallelism based on MPI or MLP
  - **Level 2**: fine-grained parallelism based on OpenMP nested within level 1
- Inter-process communication
  - **MPI:** non-blocking send/receive relaxes communication schedule to hide latency
  - **MLP:** shared memory copy reduces latency and buffering



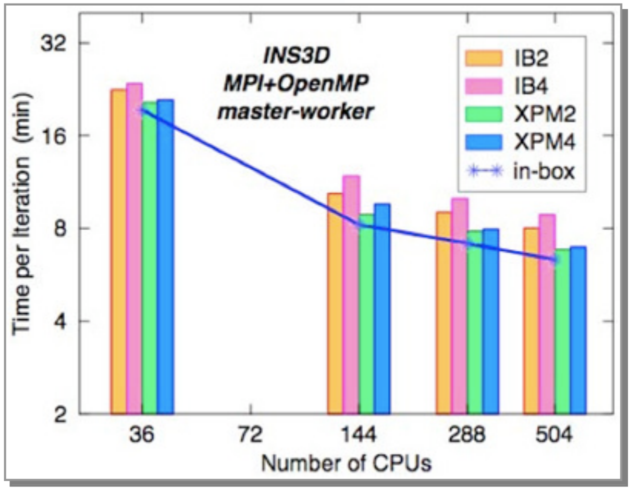


INS3D Columbia Results

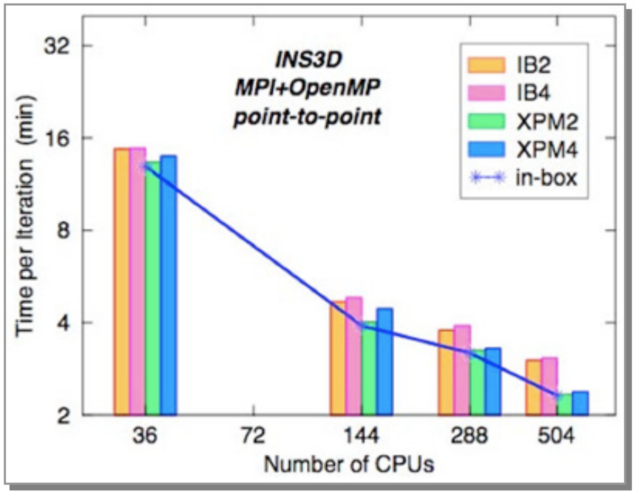
- Single-node computations indicated that MPI+OpenMP version performs slightly better than MLP version, due to local copies of connectivity arrays
  - Scalability decays beyond 8 threads due to remote memory accesses
  - Scalability can be improved by increasing MLP or MPI processes, but convergence (and eventually workload balance) deteriorates
- Multi-node results showed point-to-point implementation of MPI+OpenMP code performs more efficiently than "master-worker" version due to better utilization of network bandwidth
  - NUMAlink ~20% better than InfiniBand; ~10% performance hit when using multiple nodes



MLP and MPI+OpenMP performance within a single 512p node (# MLP or MPI processes fixed at 36)



Performance across multiple BX2b nodes using NUMAlink and InfiniBand (master-worker paradigm)

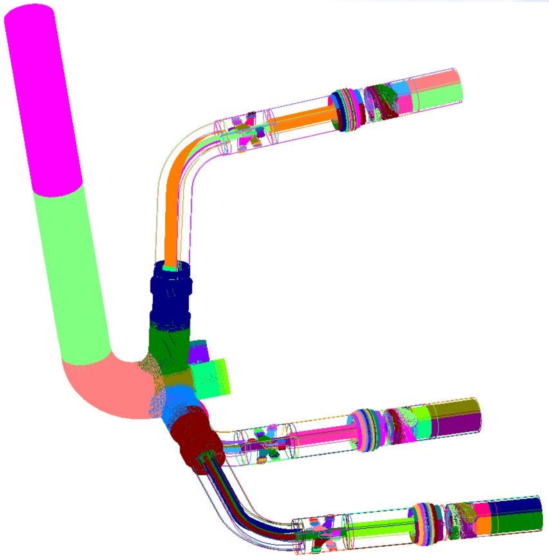


Performance across multiple BX2b nodes using NUMAlink and InfiniBand (point-to-point communication)

Propulsion Subsystem Analysis with Petascale Computing

What can be accomplished with petascale computing?

- Increase the fidelity of the current propulsion subsystem analysis to full-scale, multi-component, multidisciplinary propulsion applications
  - Extend single feedline and 1-SSME capability to 3-SSME and simulate Shuttle flight condition
  - Analyze engine integration for CEV to predict flow-induced vibration
- Model propulsion systems of new / existing launch vehicles to reach full flight rationale:
  - Modeling turbulent combustion in solid rocket boosters
  - Cavitating hydrodynamic pumps in the space shuttle main engine



What are the architecture and algorithm bottlenecks?

- Improvements to numerical and domain decomposition algorithms for overset grid systems so as to enhance scalability while retaining robustness and convergence
- Scalability bottlenecks inherent in current H/W archs, particularly if built out to 100K procs
  - Require faster interconnection technology
- Development of multi-phase and multi-fluid flow models
  - Very small time scales, highly nonlinear and unsteady, require more physical time steps: implies longer runtimes and lead to intractable problems
- Grid generation (moving & stationary), dynamic adaptation (multiple length and time scales), numerical stability

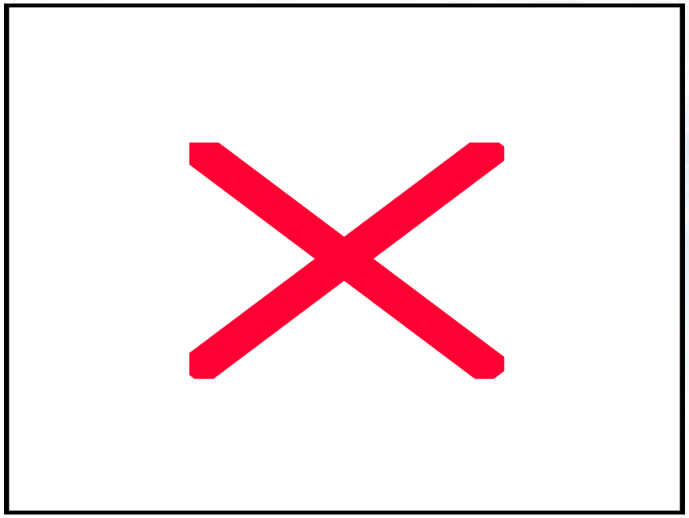
Climate Modeling

Relevance to NASA mission:

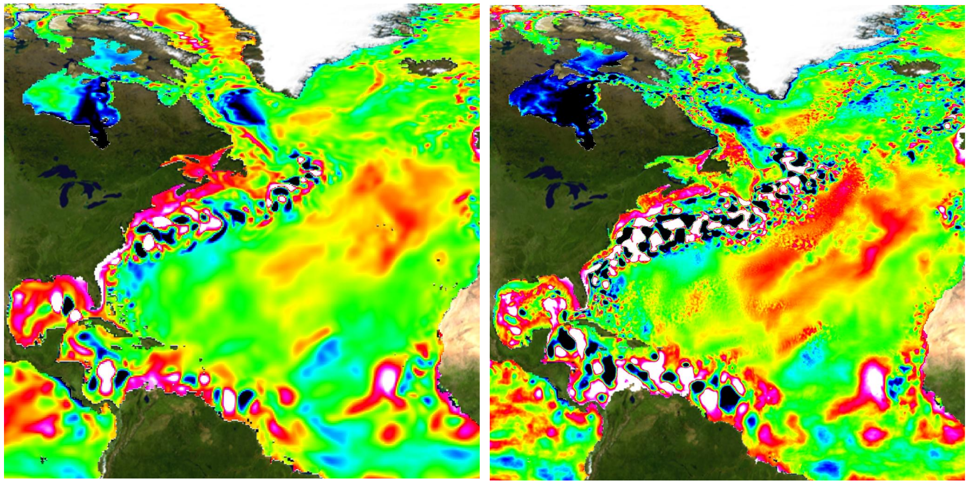
- Accurately monitoring the evolving state of ocean+sea-ice system helps NASA achieve its mission to establish a better understanding of Earth
  - More energy in top few meters of ocean than entire atmosphere
  - Estimating the Climate and Circulation of the Ocean (ECCO)

Method of Research:

- Ensure integrity of computation (using high-resolution models) and consistency with observed data
  - Data assimilation with observations: remote (host of NASA satellites) and in-situ
  - Numerical simulation using *MITgcm* code
- Designed to study atmosphere, ocean, climate
- Non-hydrostatic formulation enables code to simulate fluid phenomena over range of scales
- Fluid isomorphism allows one hydrodynamical kernel to simulate flow in both atmosphere and ocean



Decadal run on cube sphere grid

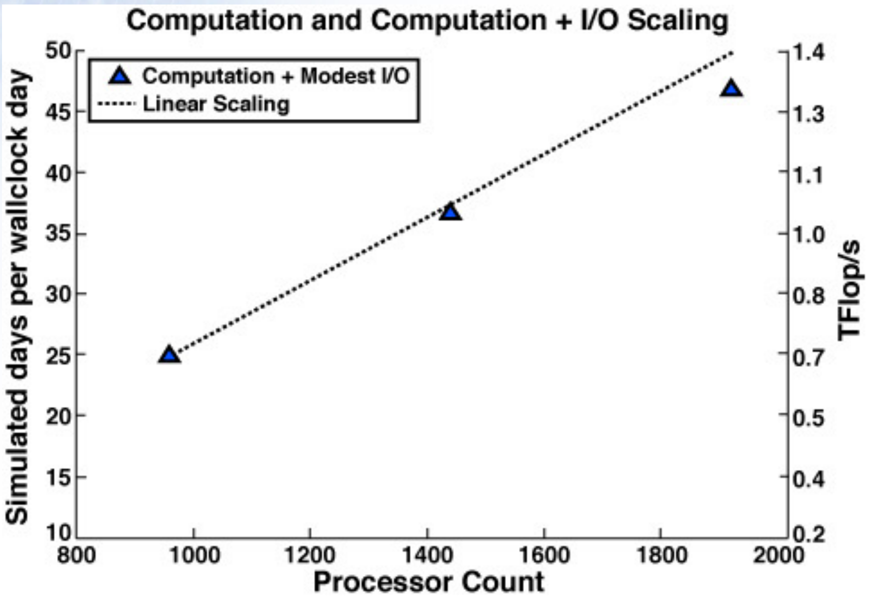


1-month sea surface height difference in Gulf Stream region w/ full-depth, global ocean, sea-ice simulations (Left: 1/4 deg; Right: 1/16 deg)

POC: *Chris Hill, Massachusetts Institute of Technology*
(617) 253-6430, cnh@mit.edu

ECCO Columbia Results

- Good scalability out to 1920 CPUs



Overall scaling and performance of 1/16 deg resolution simulation on 960, 1440, 1920p

Shown here is Ocean Planetary Boundary Layer (PBL) in meters which is the mixing layer depth.

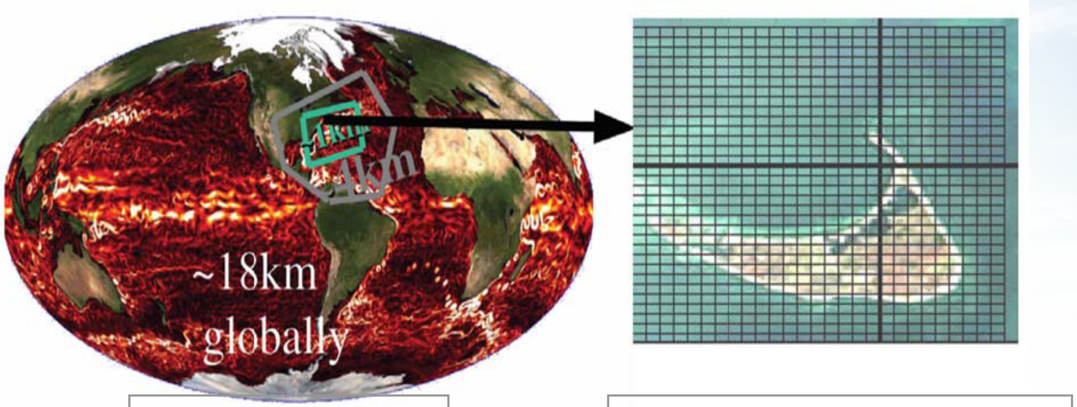
White presents shallow mixing layer due to warm, buoyant water from solar heating. The shallow layer is a stratification that keeps the wind from affecting the water below. Note that it follows the diurnal cycle. Darker blue indicates deeper, colder mixing layers that change on a slower timescale.

- Concurrent visualization made every integration time step available without overwhelming I/O time or space requirements
  - Traditional batch compute and post-process visualization strategy will not scale to ultra-high resolutions and long-duration physics simulations

Climate Modeling with Petascale Computing

What can be accomplished with petascale computing?

- More accurate monitoring of sea-ice extent, thickness will lead to better volume estimates
- Increased accuracy in monitoring air-sea fluxes of heat, freshwater, CO_2, etc.
- Would allow global scale monitoring with sub-regions at non-hydrostatic resolutions
- Would have the ability to include different physics/chemistry in embedded models, as appropriate



global simulation

Nantucket sound at ~1km

What are the architecture and algorithm bottlenecks?

- Dynamic, multi-scale embedded models will be required
- More sophisticated programming paradigm (e.g. to enable dynamic load distribution)
- Hierarchical storage and analysis needed to handle potential petabyte per day outputs
- Boundary layers are not well resolved
- Hardware and software in current 8,000+ CPU runs are prone to faults (longer MTBF)

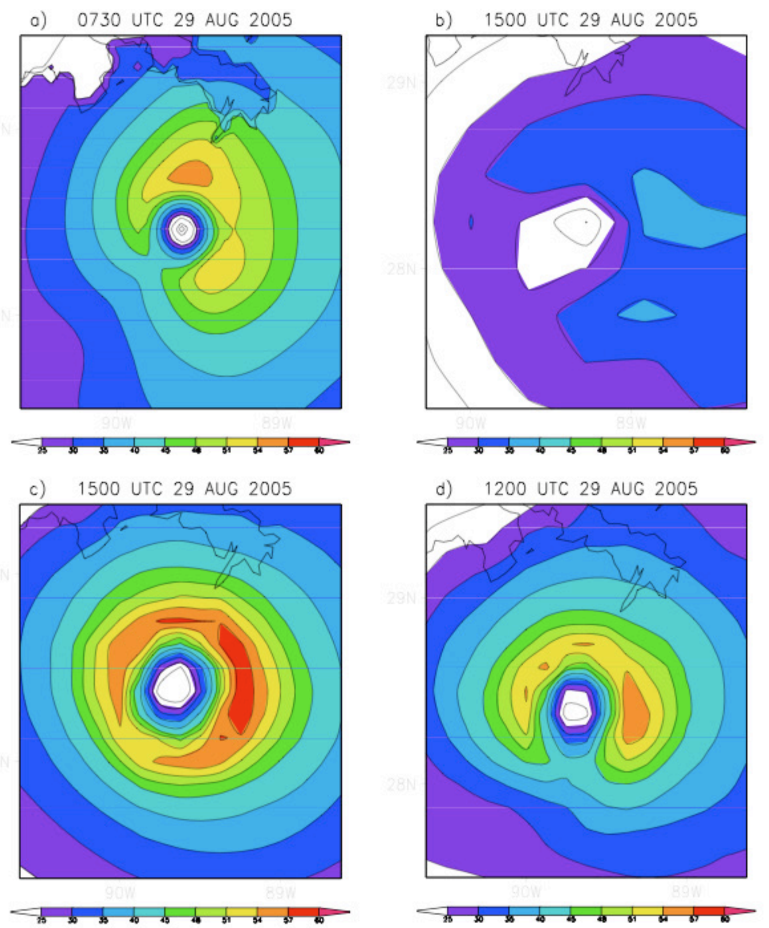
Hurricane Prediction

Relevance to NASA mission:

- Hurricane track and intensity predictions help provide early warning to people in the storm's path, thereby saving life and property
- NASA launches many high-resolution satellites, but 1/8 deg *fvGCM* is one of few global circulation models with comparable resolution to satellite data (e.g. QuikSCAT)
- This model and its cloud-resolving version could provide unprecedented opportunities to compare satellite data and model outputs

Method of Research:

- Ultra-high resolution *fvGCM* is based on a finite volume dynamical core and the community built physical parameterizations and land surface models
  - Lagrangian control volume, vertical discretization of basic conservation laws
  - 2D horizontal flux-form semi-Lagrangian, genuinely conservative, Gibbs oscillation free

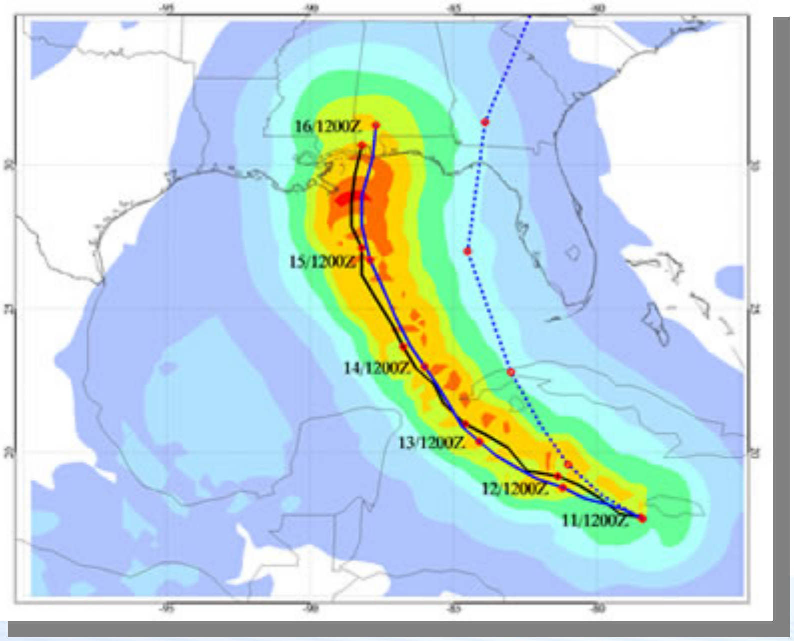


Comparison of wind distribution near the eye of Hurricane Katrina in a 2x2 degree box between a) high-resolution (0.0542^0) surface wind analysis data, and fvGCM simulations for b) 0.25^0, c) 0.125^0, and d) 0.125^0 without convection parameterization resolutions

POC: *Robert Atlas, National Oceanic and Atmospheric Administration (NOAA), (305) 361-4300, robert.atlas@noaa.gov*

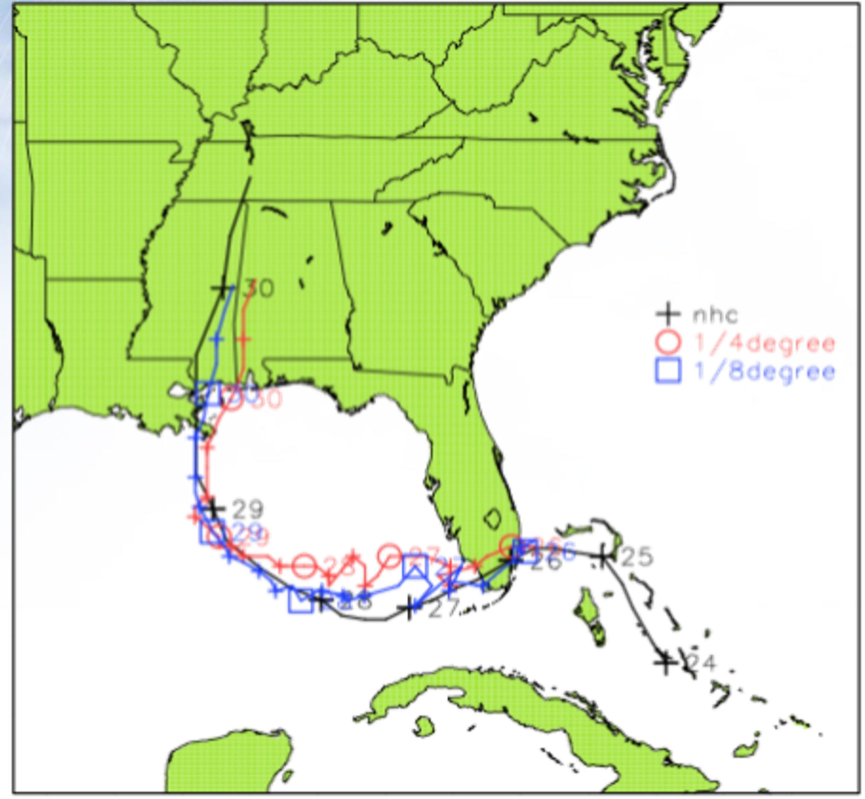
fvGCM Columbia Results

Simulations of Hurricane Frances and Typhoon Songda simultaneously in the Atlantic and Pacific Oceans (1/8 deg resolution)



Hurricane Ivan's track and intensity as forecasted by fvGCM 5 days before landfall (solid black line), the official National Hurricane Center (NHC) forecast (dashed blue line), and the NHC observed positions (solid blue line)

fvGCM Columbia Results



Comparable track predictions of Hurricane Katrina at different resolutions from 5-day forecast

Compute time approx 45 mins on 480 p (1/8 deg)

Simulations showing excellent prediction of landfall time and location (1/12 deg resolution)

Compute time approximately 3 hours on 480p

Weather Modeling with Petascale Computing

What can be accomplished with petascale computing?

- Reliable longer-duration weather forecasts
- Global non-hydrostatic Earth modeling system, including eddy-resolving oceans, cloud-resolving atmosphere, and land models, coupled with chemical and biological components

What are the architecture and algorithm bottlenecks?

- As resolution increases significantly, horizontal scales become smaller and hydrostatic assumption therefore no longer valid
  - Non-hydrostatic dynamics must be included before trying 4km resolution runs
  - New schemes to better represent physical processes at 1-10km resolutions may be needed
  - Assumptions of physical parameterizations needed at coarse resolutions but may be invalid at resolutions of 10 km and less (e.g. clouds)
- New grid systems (e.g., geodesic grid) are needed due to inefficiencies of (non-uniform) latitude-longitude grid at ultra-high resolutions
- Earth modeling systems typically both computation- and memory-intensive; thus, faster processor (multi-core) and larger cache / local memory would be beneficial

Computational Astrophysics and Cosmology

Relevance to NASA mission:

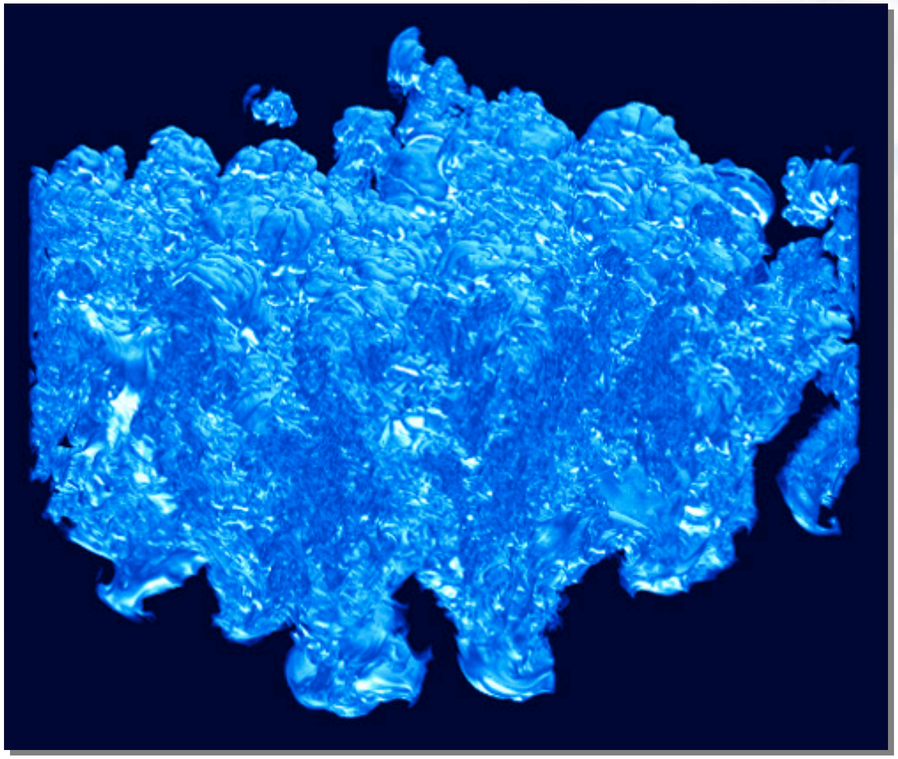
- The study of computational astrophysics accelerates NASA's understanding of the evolution of the universe—one of the agency's primary missions

Method of Research:

- Extended previous 2-D reactive Rayleigh-Taylor study to 3-D, focusing on the flamelet regime to understand the nature of flame-generated turbulence in 3D
- All of the calculations are fully resolved - no flame model is used
- These large calculations encompass over 170 million zones at the end

Current Results:

- Successfully modelled a 3D Rayleigh-Taylor unstable flame and demonstrated that the turbulence is Kolmogorov
- Ran first set of turbulent flame interaction calculations
- The application code runs ~5x faster per CPU on Columbia than on the Seaborg machine at the National Energy Research Scientific Computing Center (NERSC)



Animation of the carbon mass fraction.

POC: *Stan Woosley, University of CA, Santa Cruz*
(831) 459-2976, woosley@ucolick.org

Low Mach Number Hydrodynamics

(Bell et al. 2004 JCP 195, 677)

- Low Mach number formulation projects out the compressible components
  - Pressure decomposed into thermodynamic and dynamic components

$$p(x,t) = p_0(t) + Mp_1(t) + M^2\pi(x,t)$$

  - Elliptic constraint provided by thermodynamics.

$$0 \equiv \frac{Dp}{Dt} = \frac{\partial p}{\partial \rho}\frac{D\rho}{Dt} + \frac{\partial p}{\partial T}\frac{DT}{Dt} + \sum_k \frac{\partial p}{\partial X_k}\frac{DX_k}{Dt}$$

$$\nabla \cdot U = \frac{1}{\rho\frac{\partial p}{\partial \rho}}\left(\frac{\partial p}{\partial T}\frac{DT}{Dt} + \sum_k \frac{\partial p}{\partial X_k}\frac{DX_k}{Dt}\right)$$

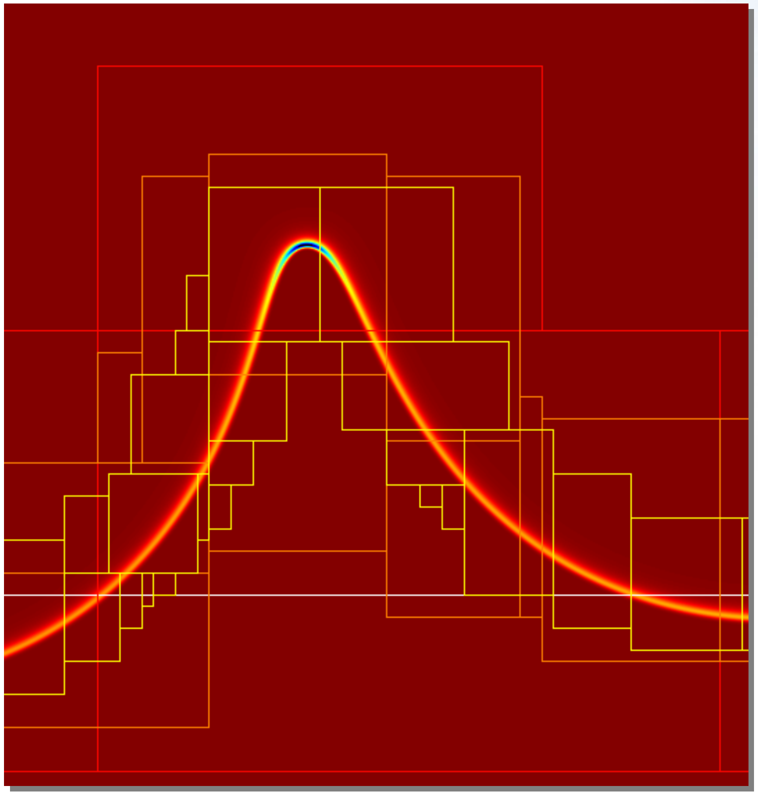
  - Advection/Projection/Reaction formulation solves system.
  - Timestep limited by lvl and not lvl + c.

Simulation Method

(Bell et al. 2004 JCP 195, 677)

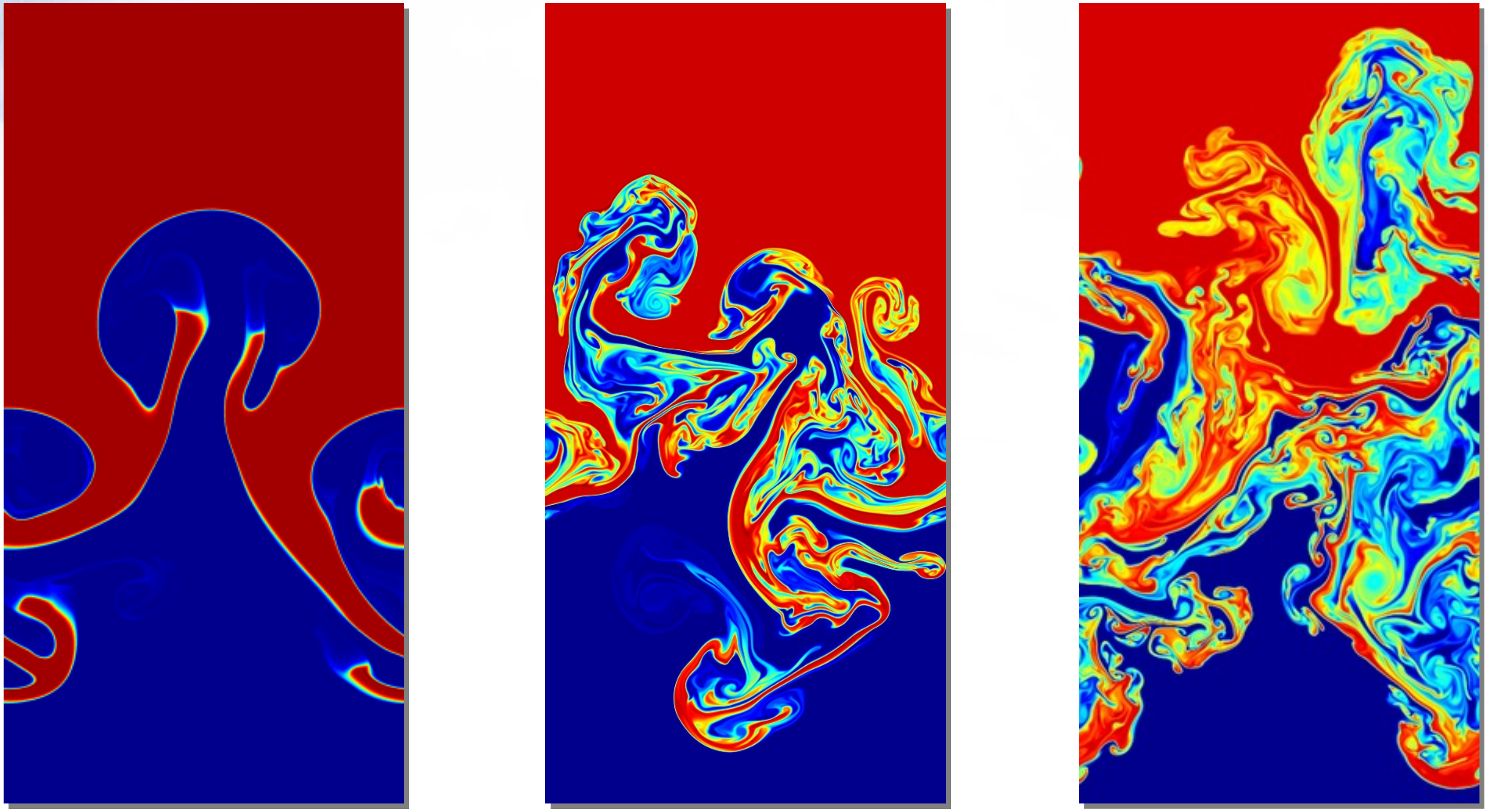
- Low Mach number hydrodynamics
  - Advection/projection/reaction
  - Block structured adaptive mesh
  - Timestep restricted by lvl not lvl + c
  - Degenerate/Relativistic EOS used
  - Single step $^{12}C+^{12}C$ rate

- Initialized by mapping 1-D steady-state laminar flame onto grid
  - 5-10 zones inside thermal width



Transition to Distributed Burning

(Bell et al. 2004, ApJ, 608, 883)



← □

- As □ decreases, RT dominates over burning
- At low □, flame width is set by mixing scale

Astrophysics and Cosmology with Petascale Computing

What can be accomplished with petascale computing?

- Model the evolution of the physical universe—from first few minutes after inflation and Big Bang, until now—and into the next 10 billion years, including studies of:
  - Cosmic microwave background radiation
  - Structure formation (first stars, galaxies, clusters)
  - Explosion of supernovae, all types

What modifications (HW, SW, architecture) are necessary to accomplish this?

- Algorithm components such as the following would have to be developed:
  - Low Mach Number/All Mach Number (to Ma = 1) hydrodynamics
  - Radiation transport (multi-D and non-LTE for SN spectra)
  - Radiation transport (multi-D, multi-group)