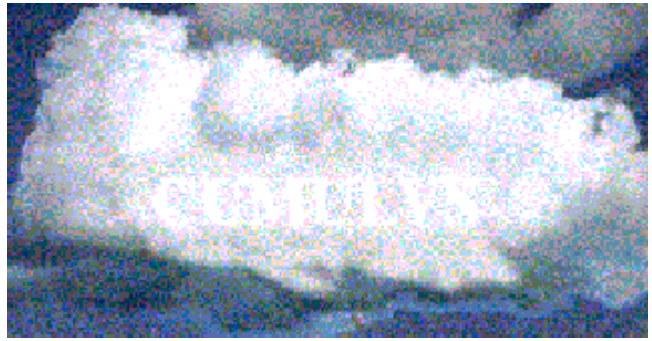
DOE 2000

CUMULVS

James Arthur Kohl
Oak Ridge National Laboratory

ACTS Toolkit Review
January 5, 1999

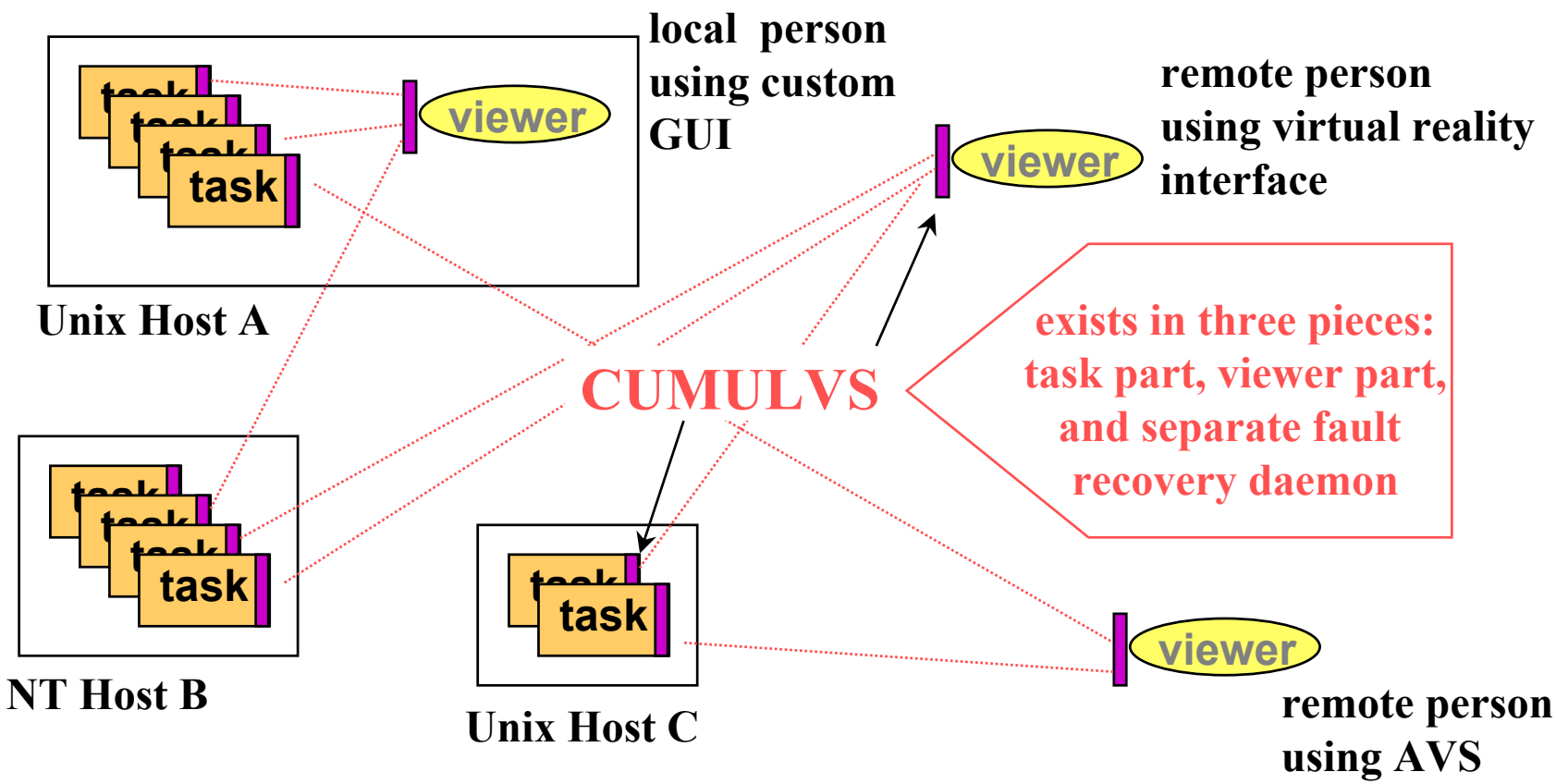
Research sponsored by the Mathematics, Information and Computational Sciences Office
U.S. Department of Energy



- Collaborative Infrastructure for Interacting with Scientific Simulations
  - ⇒ Run-Time Visualization by Multiple Viewers
  - ⇒ Coordinated Computational Steering
  - ⇒ Application-Directed Fault Tolerance
    - Heterogeneous Checkpointing for Automatic Fault Recovery and Task Migration
  - ⇒ Coupled Models…

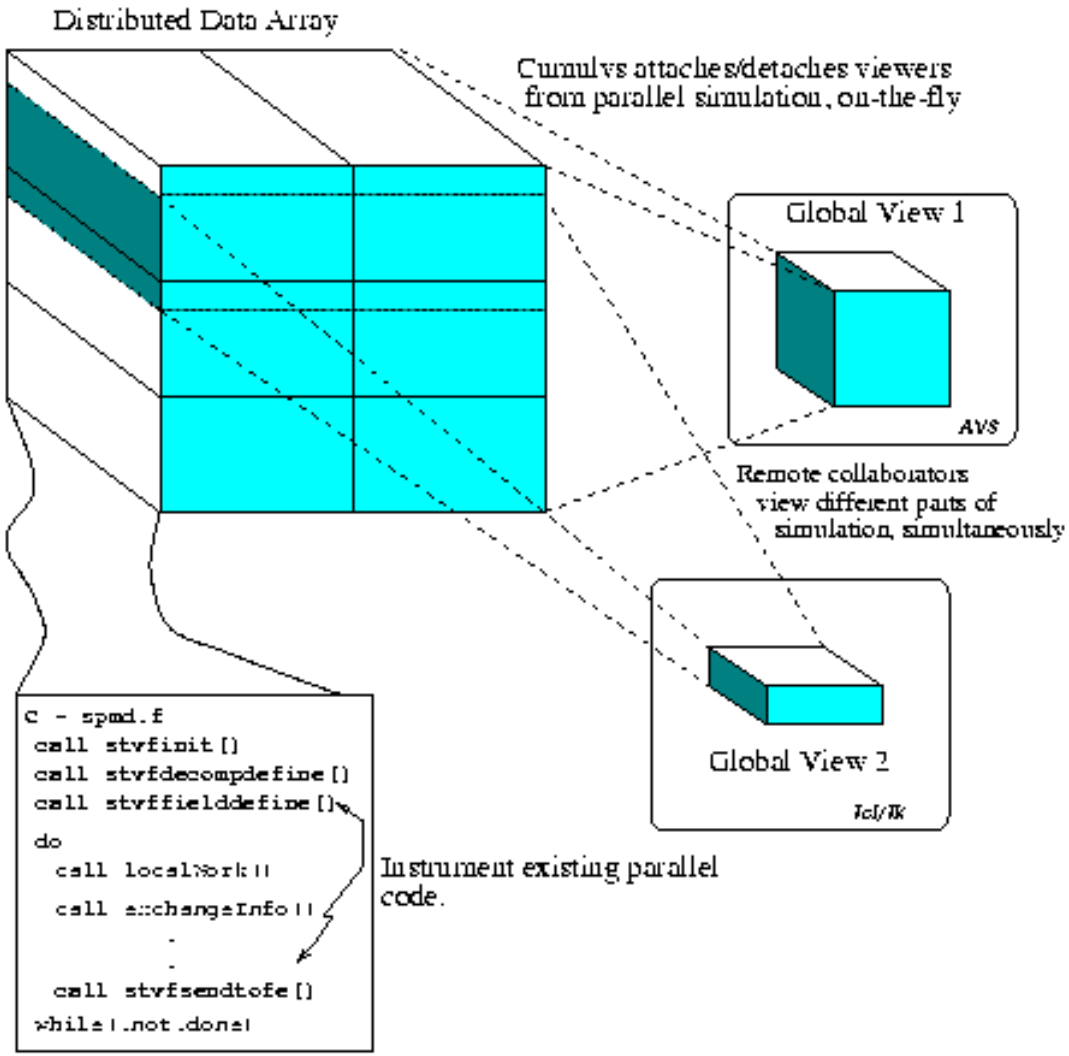
CUMULVS

coordinates the consistent collection and dissemination of information to/from parallel tasks to multiple viewers



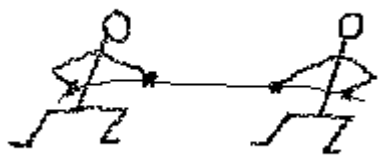
distributed parallel application or simulation

Multiple Distinct Views



Coordinated Steering

- Multiple, Remote Collaborators
- Simultaneously Steer Different Parameters
  - ⇒ Physical Parameters of Simulation
  - ⇒ Algorithmic Parameters ~ e.g. Convergence Rate
- "What If?" Analyses
  - ⇒ Explore Non-Physical Effects
- Efficient Experimentation Cycle
  - ⇒ Keep Simulation On Track
  - ⇒ Crop Off Experiments Gone Awry…

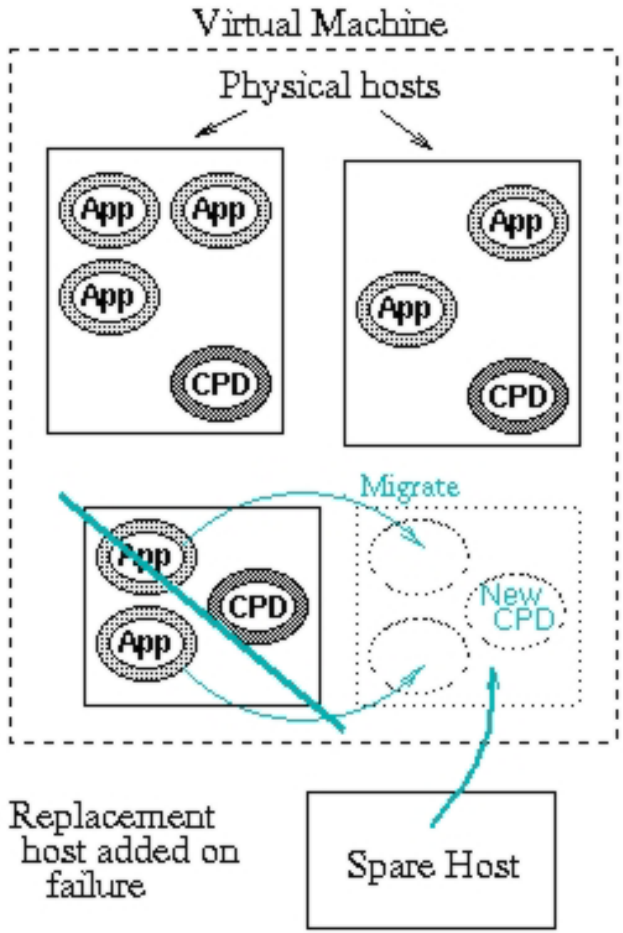


Heterogeneous Checkpointing

- Application Defines Its Own State
  - ⇒ Provide CUMULVS with Semantic Information
  - ⇒ CUMULVS Handles Checkpoint Collection
  - ⇒ Automatic Fault Recovery System
- Efficient & Flexible Checkpoints
  - ⇒ Not Full Core Image
  - ⇒ Semantics → Heterogeneous Restart & Migration
  - ⇒ On-The-Fly Reconfiguration Also Possible…

Run-Time System Architecture

- One Checkpointing Daemon (CPD) Per Host.
  - ⇒ Ckpt Collector / Provider.
  - ⇒ Run-Time Monitor.
  - ⇒ Console for Restart / Migrate.
- CPDs Comprise Fault-Tolerant Application…
  - ⇒ Handle Failure of Host / CPD.
  - ⇒ Coordinate Redundancy.
  - ⇒ Ring Topology.



Manual Software Instrumentation

- SPDT 98 Case Study ~ SW Instrumentation Cost

Instrumentation:	Seismic:	Wing Flow:
Original Lines of Code	20,632	2,250
Vis / Steer System Init	3	3
Vis / Steer Variable Decls	48	73
CP Restart Initialization	21	12
CP Rollback Handling	41	34
Total Instrumentation	204 ~ 1.0 %	188 ~ 7.7 %

Checkpointing Efficiency

- SPDT 98 Case Study ~ Execution Overhead

Seconds per Iteration

Experiment:	SGI:	Cluster:	Hetero:
Seismic - No Checkpointing	2.83	6.23	9.46
Seismic - Checkpoint for Restart	2.99	6.50	10.76
Seismic - Checkpoint for Rollback	3.03	6.66	10.90
Wing - No Checkpointing	0.69	1.58	6.14
Wing - Checkpoint for Restart	0.77	1.71	7.10
Wing - Checkpoint for Rollback	0.79	1.71	7.30

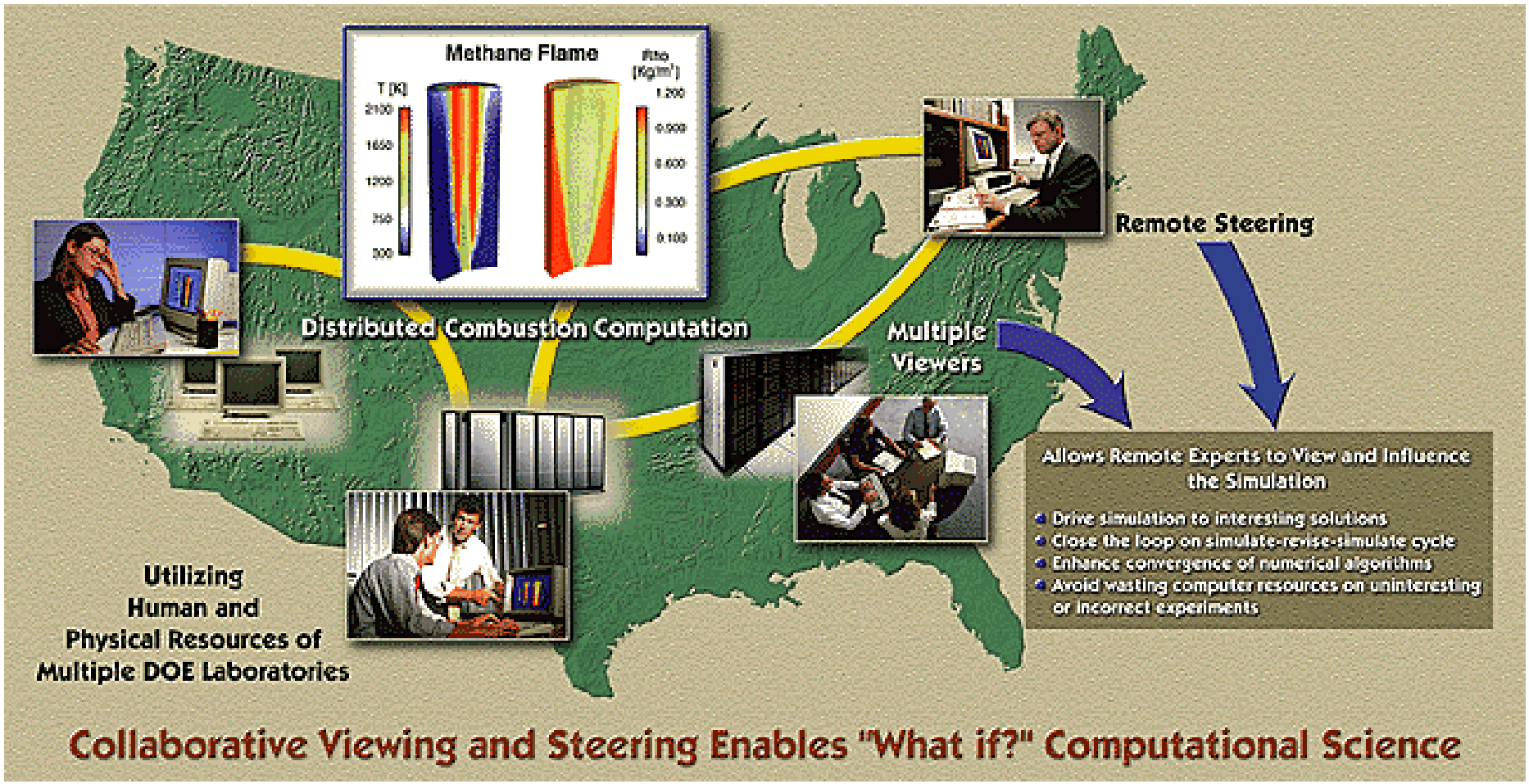
(Checkpointing Every 20 Iters.)

Seismic Overhead: 4-14% Restart, +1-3% Rollback.
Wing Overhead: 8-15% Restart, +0-2.5% Rollback.

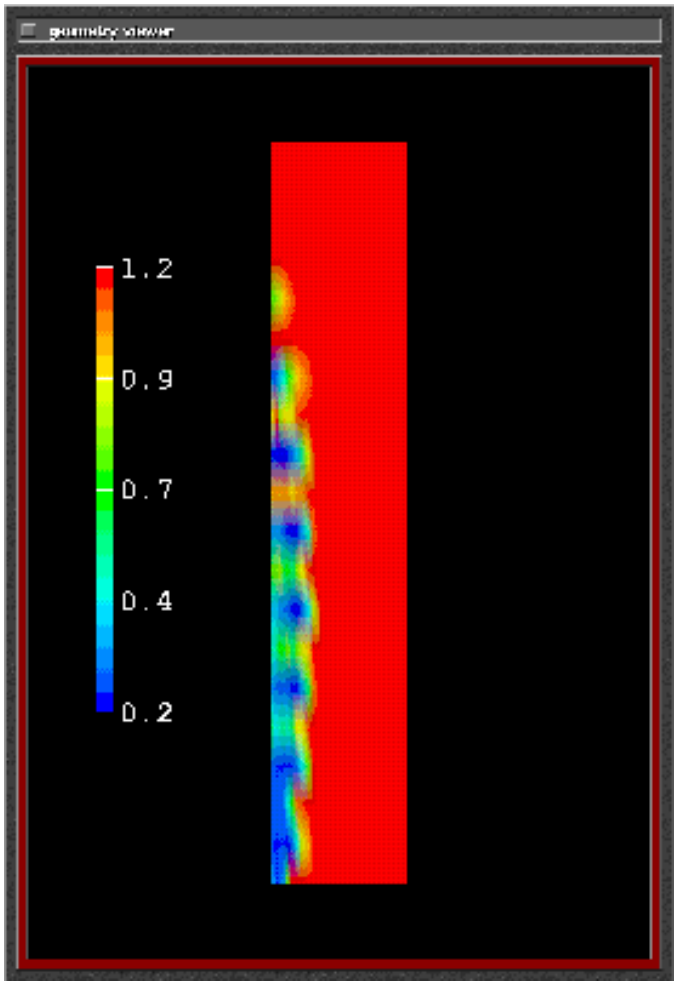
Integration With InDEPS (POET/PD++)

- Merge With Development Environment
  - ⇒ Automates Instrumentation, Built-In Support
- Collaboration with ORNL & Sandia
  - ⇒ Kohl, Semeraro, Armstrong, Wyckoff ~ SC 97
- Instrumented Applications:
  - ⇒ Combustion Simulations
  - ⇒ ASCI Applications:
    - Material Impact and Deformation (CTH)
    - Smooth Particle Hydrodynamics (SPH)

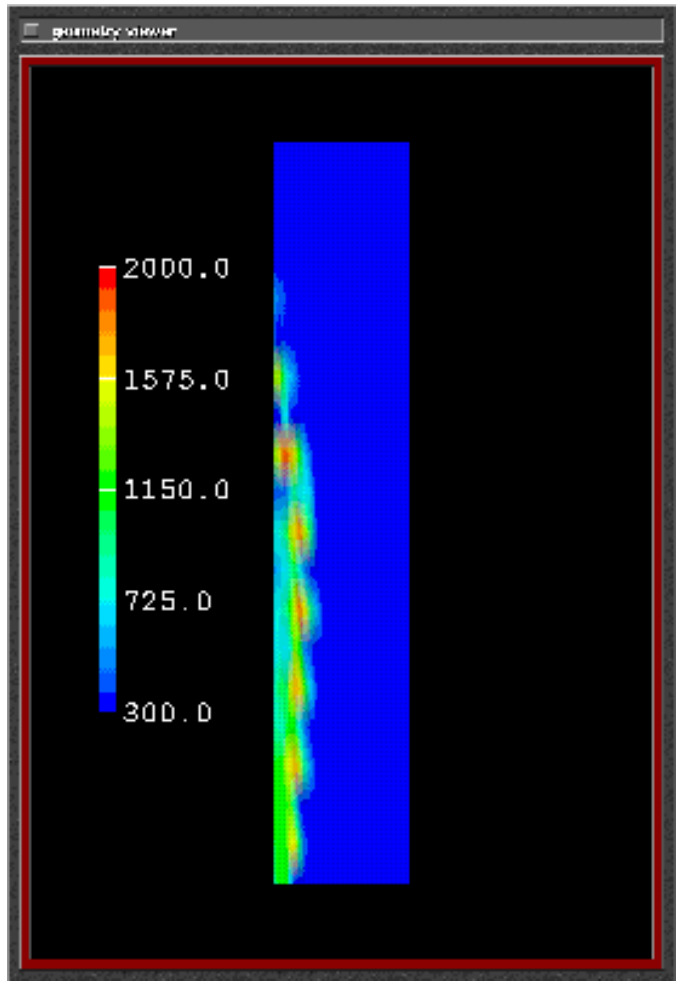
Collaborative Combustion Simulation



Multiple Simultaneous Views

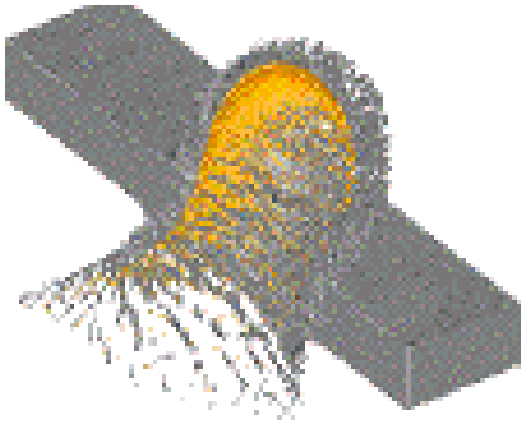


Density



Temperature

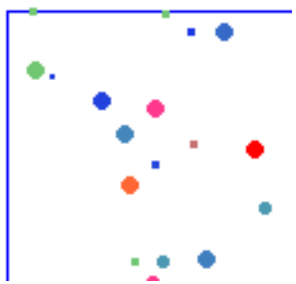
Material Impact and Deformation (CTH)



- CUMULVS Integrated Into Shock Physics Application
- Cross-Country Production Runs Between ER & DP Labs
- Utilized Low-Level ATM Network (AAL-5 Data Transfer)

CUMULVS Particle Handling

- Particle Data Fundamentally Different
  - ⇒ Nested Data Fields, Explicit Coordinates
  - ⇒ Required for SPH Viewing
- ORNL & Sandia ~ Kohl & Armstrong
  - ⇒ Particle-Based Decomposition API
    - Particle Field Array Handling
    - User-Defined, Vectored Accessor Routines
  - ⇒ Viewing Particle Data
    - AVS Module Extensions
    - Tcl/Tk Slicer Particle Mode



Remotely Monitor T3E Applications

- Collaboration with LBL & LLNL
  - ⇒ Users Request Real-Time Visualization of T3E Computation from Remote Desktop
  - ⇒ Efficiently Utilize Standard Networks
  - ⇒ CUMULVS Coordinates & Collects Data
- Collaborators:
  - ⇒ Youngbae Kim @ NERSC / LBL
  - ⇒ Alice Koniges @ LLNL
  - ⇒ Shrihari Gopalakrishna @ UC Davis

Utilizing Optimized Cray PVM on T3E

- Not Full Implementation of PVM API
  - ⇒ CUMULVS Library Expanded to Cover Missing Functionality
    - Name Service for Application Discovery
- Requires use of "Distributed Mode"
  - ⇒ Allow Off-Host Communication from T3E and Dynamic Groups

Collaboration with NCSA

- Tcl/Tk Language Bindings for CUMULVS
  - ⇒ Library Calls & Objectized Data Pointers
- CUMULVS Viewers
  - ⇒ Coordinated Viewers via VTK & Tango (AFRL)
  - ⇒ Virtual Reality / OpenGL Viewer ~ ImmersaDesk & Cave
  - ⇒ Windows-Based Viewers / MFC
- Applications Development
  - ⇒ Chesapeake Bay Simulation
  - ⇒ Neutron Star Collision (SUNY)
  - ⇒ DOD Codes (ASC)

Supporting MPI Applications

- Vendor-Supported Standard
- Growing Application Base
- Selected for ASCI Applications
- MPI Built for High Performance
  - ⇒ Static Model, Minimal Operating Environment
  - ⇒ No Task Spawn (yet), Name Service / Database, Fault Tolerance / Notification?
- CUMULVS Solution:
  - ⇒ Applications Communicate Using MPI
  - ⇒ CUMULVS Attaches Using PVM
- Possible “Reduced-Functionality” MPI Version...

MPI_Connect / PVMPI

Goal: Provide tool that allows different vendor's MPI implementations to inter-operate.

Currently supports interoperation between:

MPICH, LAM 6, IBM MPIF, and SGI MPI

MPI_Connect provides two programming models:

→ Completely transparent single MPI_COMM_WORLD model

→ MPI pt-to-pt communication between uniquely identified groups

Presently in use and supported at:

DOD MSRC Sites (CEWES, ASC, ARL)

Jeff Brown @ LANL (ASCI Applications)

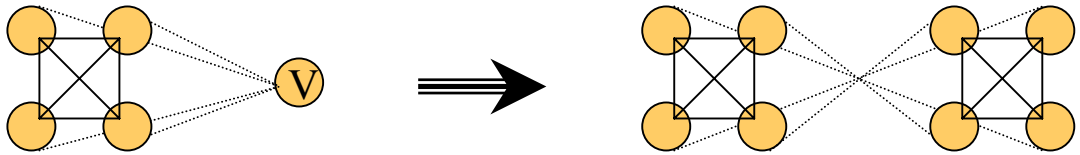
(Software Release ~ January 1999)

Future CUMULVS Plans (1 of 3)

Coupling Data Fields in Simulation Models

- Natural Extension to Viewer Scenario

  ⇒ Promote "Many-to-1" → "Many-to-Many"

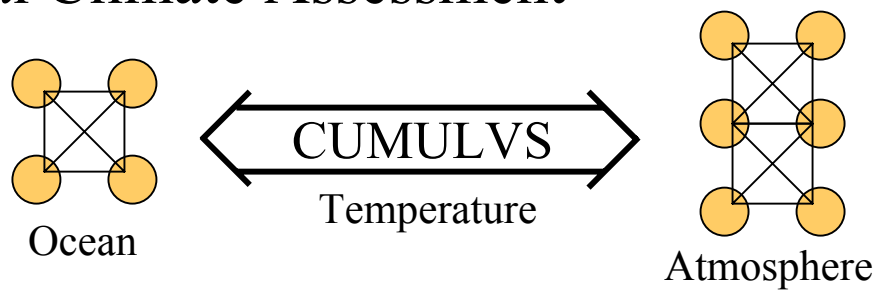


- Translate Disparate Data Decompositions

  ⇒ Complements PAWS Coupling Work

  ⇒ Builds on CCA Foundation Technology

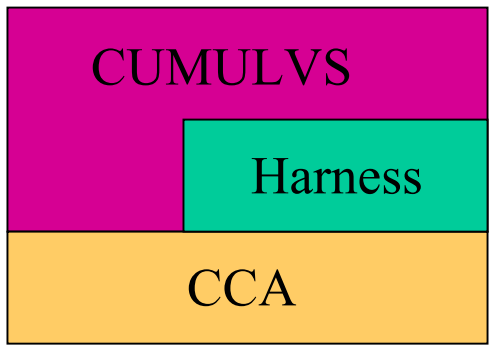
E.g. Regional Climate Assessment



Future CUMULVS Plans (2 of 3)

Building on Harness

- Harness ~ Next Generation HDC Environment
  - ⇒ Pluggable Virtual Machine, Distributed Control
    - Follow-on to PVM...
  - ⇒ Inspired by CUMULVS Coupling Needs
  - ⇒ ORNL, UT and Emory (Basic Research)
- Reinforces Need for CCA
  - ⇒ Harness Pluggability Builds on CCA Effort



Future CUMULVS Plans (3 of 3)

- Application Interface:
  - ⇒ Assist Manual Instrumentation of Applications
    - GUI, Pre-Compiler...
- Checkpointing:
  - ⇒ Checkpoint Efficiency
    - Tasks Write Data in Parallel / Parallel File System?
    - Redundancy Levels, Improve Scalability
- Portability:
  - ⇒ Other Messaging Substrates
    - Nexus, Reduced Functionality for MPI

CUMULVS Summary

- Toolkit Integration Efforts:
  - ⇒ InDEPS / Particle Fields ~ Sandia, ORNL
  - ⇒ Cray T3E ~ NERSC/LBL, LLNL, UC Davis, ORNL
  - ⇒ MPI / MPI_Connect ~ UT, ORNL, ANL, OSC, IBM, SGI
  - ⇒ Data Field Coupling / Harness / CCA ~ ORNL, UT, Emory
  - ⇒ Tcl/Tk, VTK, OpenGL, Tango ~ NCSA, AFRL, ASC, SUNY
- Future Opportunities
  - ⇒ Integration with Other Frameworks via CCA
  - ⇒ Application Instrumentation GUI / Pre-Compiler

http://www.csm.ornl.gov/cs/cumulvs.html