# *Computational Science in the Office of Biological and Environmental Research*

Dave Bader

# *BER Needs for Computation and Simulation*

- Complex Environmental Systems that Are Impossible to Study in a Laboratory
- Massive Volumes of Biological Data that Need to Be Searched and Analyzed (Genome, Protein, Molecular, etc)
- Molecular and Biomolecular Simulation for Environmental and Biological Applications

# *BER Approach: Interdisciplinary and Multi-disciplinary Partnerships*

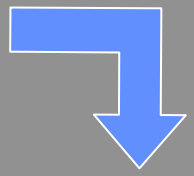
- State-of-the-Art Computational Capabilities Required for BER to Meet Its Mission
- Historically Good Relationships with ASCR and Other Agencies (Molecular Science, Climate Simulation, Genome, Structural Biology)

# *BER Partnerships with ASCR Computational Facilities*

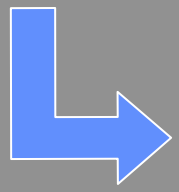
- National Energy Research Supercomputer Center - Shared by all Office of Science Programs - 512 Proc IBM SP (.41 TFLOP/s peak); 644 Proc Cray T3E-900 (.62 TFLOP/s peak)
- Oak Ridge National Laboratory - 20 % BER Investment in 736 Proc IBM SP (1.0 TFLOP/s peak)
- Los Alamos National Laboratory - 18 % BER Investment in 2048 Proc SGI Origin 2000 (1.0 TFLOP/s peak)
- EMSL Computing Facility ****

# How does biology "happen"?

–(Picture removed)

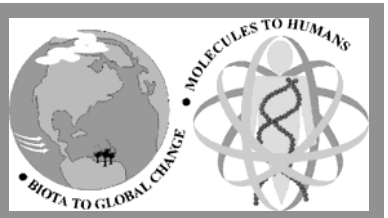


- **Instructions for making proteins**
- **On/off switches for genes**
- **Coordination of gene expression**
- **Protein function/behavior/interaction**



**Biology**

# Bringing the genome to life

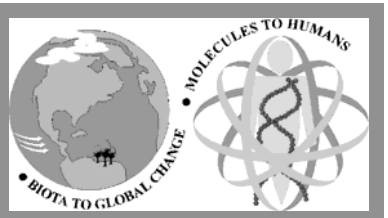



–Pictures removed

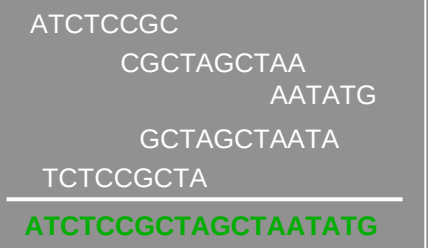
- Predicting health risk
- Cleaning up the environment
- Removing excess $CO_2$
- Clean energy

$e = mc^2$

# *Bringing the Genome to Life Requires Many Areas of Computational Biology*




**Assembly and analysis of genome sequences**




**Consensus DNA sequence from shotgun sequencing**

**Prediction of protein structure**

–Picture removed

**Predicted structure for DNA clamping protein**

**Molecular modeling of biochemical reactions**

–Picture removed

**Enzymatic activation of environmental mutagen**

**Process simulation of biological systems**

–Picture removed

**Portion of amino acid synthetic pathways**

**Each of these types of modeling requires advances in simulation techniques and computer hardware**

# *High Performance Computers Will be Needed for Each Component of Cell-level Simulations*

**Genome assembly:**

>10 TeraFlops sustained speed required to keep up with expected sequencing rates

**Protein structure prediction (using threading):**

>100 TeraFlops speeds required to model all proteins in a microbial genome in a day

**Classical Molecular Dynamics (molecular mechanics force field):**

100 TeraFlops for 1 day to simulate DNA-protein interaction (20000 atoms for 2 ms)

**First Principles Molecular Dynamics (Quantum mechanical force field):**

1 PetaFlops for 1 day to simulate reaction in enzyme active site (200 atoms for 1 ns)

**Simulations of biological networks:**

>10 TeraFlops speeds required to perform simple correlation analysis for small biological network (100 genes, 500 molecular components,10 compartments)

**Computational biology is characterized by its need for continuous high performance computing, rather than periodic large scale simulations**

*William R. Wiley*
## *Environmental Molecular Sciences Laboratory*

*Richland, Washington*

–Picture removed

### Research Mission

**. . . conduct fundamental research on the physical, chemical and biological processes that underpin environmental issues.**

**. . . advance molecular science in support of the long-term missions of the U.S. Department of Energy.**

### National Scientific User Facility Mission

**. . . provide unique research and computational resources to scientists engaged in research on environmental problems.**

**. . . educate young scientists in the molecular and computational sciences to meet current and future environmental challenges facing DOE and the Nation.**

*William R. Wiley*
*Environmental Molecular Sciences Laboratory*

*Richland, Washington*

# Research Capabilities in the EMSL

- **Focus on developing a molecular-level understanding of the physical, chemical, and biological processes that underlie:**

  **-- contaminant fate and transport in the subsurface**
  **-- processing and disposal of stored wastes**
  **-- cellular response to environmental contaminants**
  **-- atmospheric chemistry**

- **Capitalize on experimental, theoretical, and computational advances to address the complex molecular-level interactions critical to solving the problems described above.**

- **Develop integrated research programs that take advantage of the full complement of EMSL capabilities to address DOE needs.**

*William R. Wiley*
*Environmental Molecular Sciences Laboratory*

*Richland, Washington*

–Pictures removed

## Molecular Science Computing Facility

- **The Molecular Science Computing Facility (MSCF) provides EMSL users and staff with production computing needs, advanced molecular modeling software, and production facilities for visualization and analyses of complex data sets.**
- **The facility contains a 512-processor IBM RISC System/6000 Scalable POWERparallel computer with 262 gigabytes of memory and 5 terabytes of online disk, a 20-terabyte EMASS data storage system, and a 96-processor IBM SP for software research.**
- **EMSL staff developed new molecular modeling software (the Molecular Sciences Software Suite - MS$^3$) for use on high-performance parallel computing systems. MS$^3$ includes NWChem, the Extensible Computational Chemistry Environment, and ParSoft.**

*William R. Wiley*
*Environmental Molecular Sciences Laboratory*

*Richland, Washington*

# Molecular Science Software Suite ($MS^3$)

- **Northwest Computational Chemistry Software (NWChem)**

  **Molecular modeling software capable of single or combined quantum mechanics and molecular-mechanics simulations. Scaleable to problem size and machine size on a wide range of parallel computing systems.**

- **Extensible Computational Chemistry Environment (Ecce)**

  **A domain-encompassing, problem-solving environment for molecular modeling, analysis and simulation. Functions include integrated use of computational chemistry methods, molecular computation data management, and molecular property data analysis and visualization.**

- **Parallel Software Development Tools (ParSoft)**

  **Provides high-performance parallel computing libraries and tools to enable NWChem to run on a wide variety of parallel computing systems. Includes the Global Array toolkit.**

*William R. Wiley*
*Environmental Molecular Sciences Laboratory*

*Richland, Washington*

# *Computational Research at the EMSL*

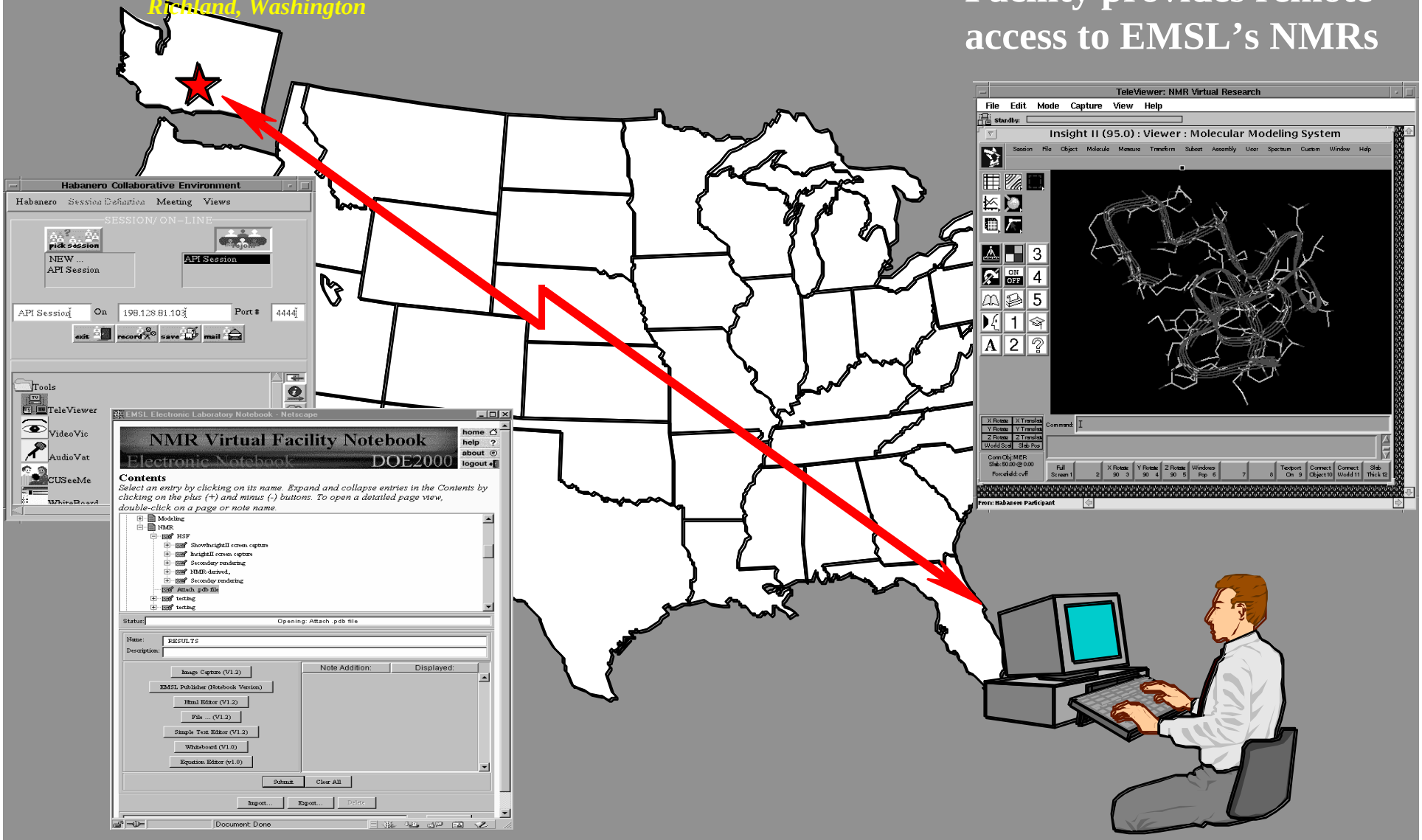
- **Density Functional Theory Used to Show How the Ligand Tetramethoxycalix[4]arene Selectively Binds to Cesium-137 and Provides Criteria for Enhanced Ligand Design for Use in Tank Waste and Groundwater Remediation.**
- **Calculations Establish a Key Mechanism in the Radiolysis of Nitrite Ion in Alkaline Solutions, and Provide Insights to Reactions in High Level Wastes.**
- **High Level *ab initio* Calculations Show that C -H···O=C Hydrogen Bonding Interactions Contribute to Protein Conformation.**
- **Dynamical Nucleation Theory Used to Predict Nucleation Rates of Molecules involved in the Formation of Atmospheric Aerosols.**
- **Computational and Experimental Studies Provide Insights into the Reaction Mechanisms of Important Catalysts for Hydrocarbons (Zeolites).**

*William R. Wiley*
*Environmental Molecular Sciences Laboratory*

*Richland, Washington*

# EMSL Collaboratory

## The Virtual NMR Facility provides remote access to EMSL's NMRs



*William R. Wiley*
*Environmental Molecular Sciences Laboratory*

*Richland, Washington*

## Web Addresses

- **Theory, Modeling and Simulation Directorate**

  **http://www.emsl.pnl.gov:2080/homes/tms/**

- **Molecular Science Computing Facility (High Performance Computing Center, Molecular Sciences Software Suite, Graphics and Visualization Laboratory)**

  **http://www.emsl.pnl.gov:2080/capabs/mscf/index.htm**

## Points of Contact

- **EMSL Director - Jean Futrell**

  **(509) 376-0226 jean.futrell@pnl.gov**

- **Associate Director for Theory, Modeling and Simulation - David Dixon**

  **(509) 372-4999 da_dixon@pnl.gov**

- **Molecular Science Computing Facility (MSCF) Operations - Rob Eades**

  **(509) 375-2279 robert.eades@pnl.gov**

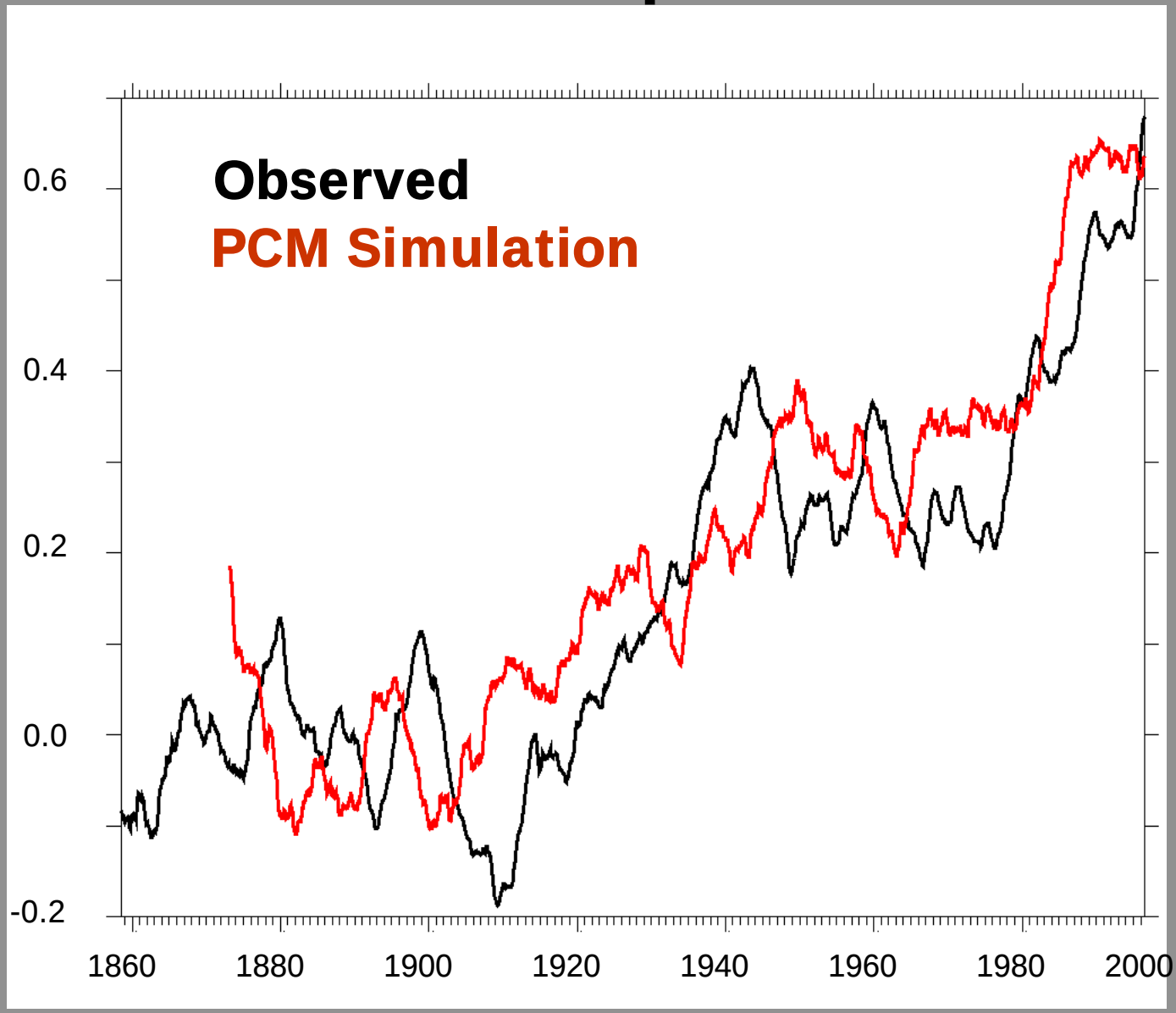
# Demands on Climate Models

- 1990-1992. Bounding global climate change: Is there a problem?
  - A few, coarse resolution equilibrium model runs
- 1993-1998. Estimating future climate change: How big is the problem?
  - ~10 medium resolution time-dependent runs
- 1999-2003. Predicting details of greenhouse gas scenarios: How to manage the problem?
  - Requires many (~1000) high resolution, time-dependent runs

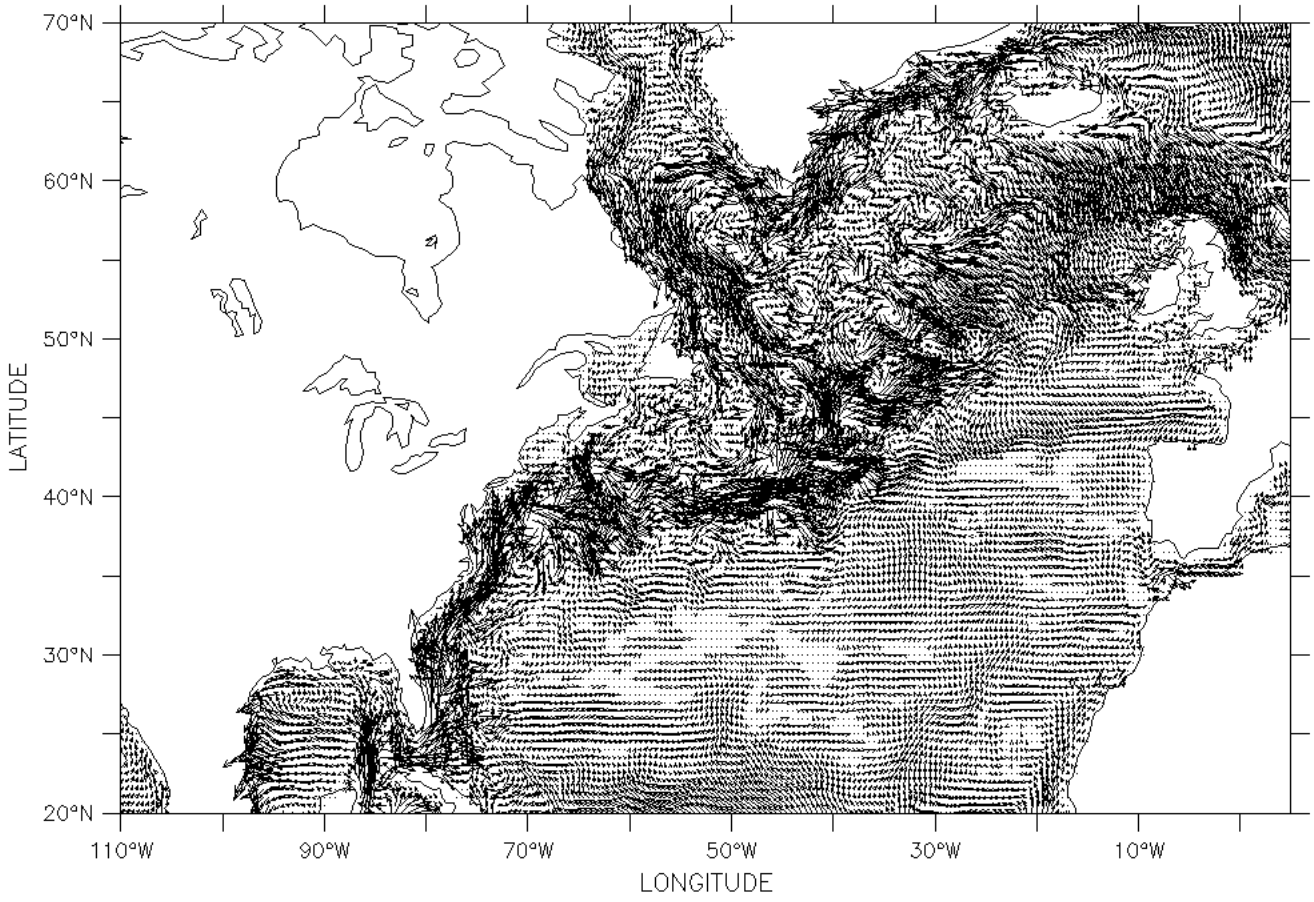
# Climate Prediction Products

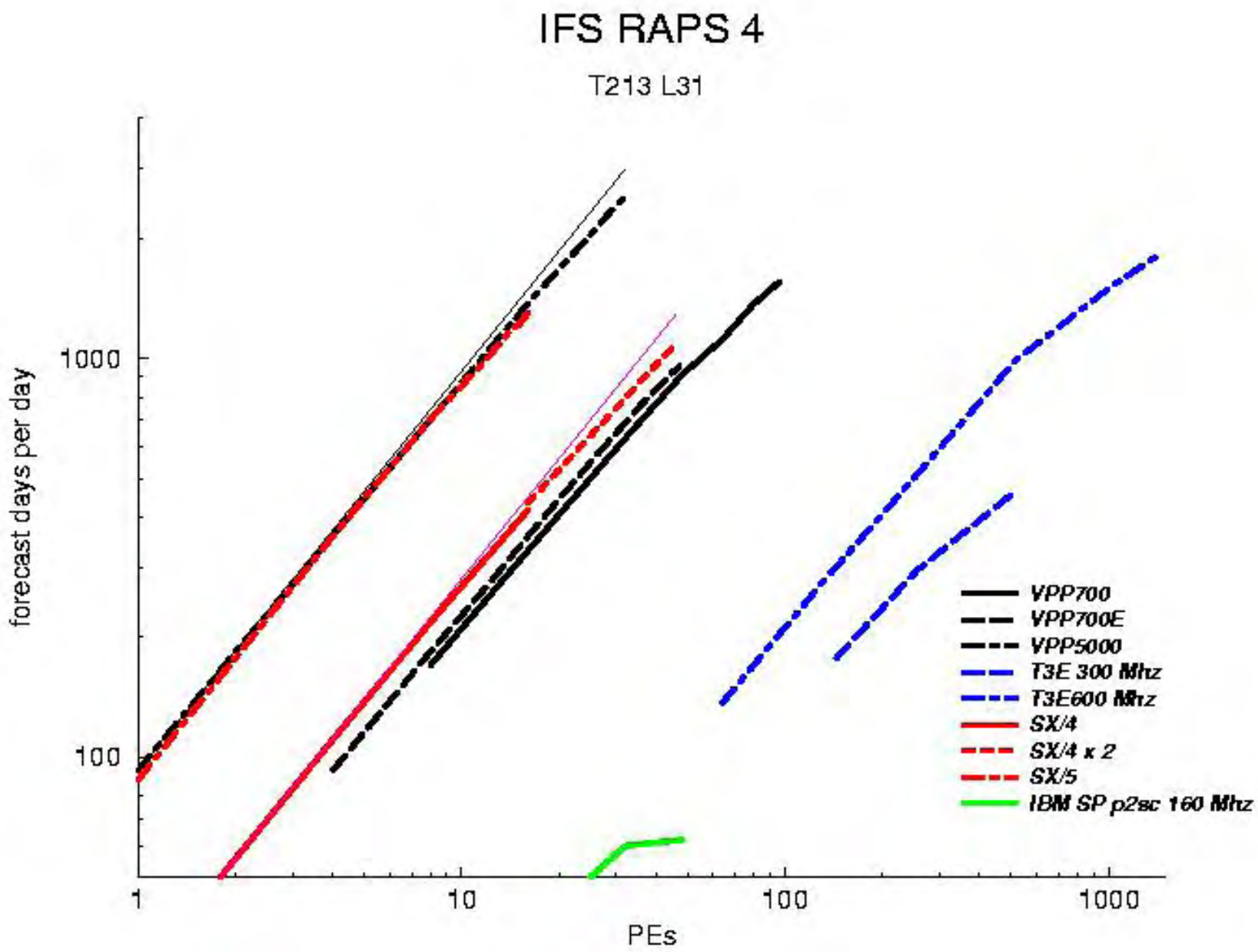
- Accurate and verifiable projections of climate change at regional resolution
- Statistically meaningful measures of natural variability
- Multiple scenarios to evaluate emission reduction strategies
- Completion of IPCC assessment by 2005

# Global Surface Temperature Anomaly



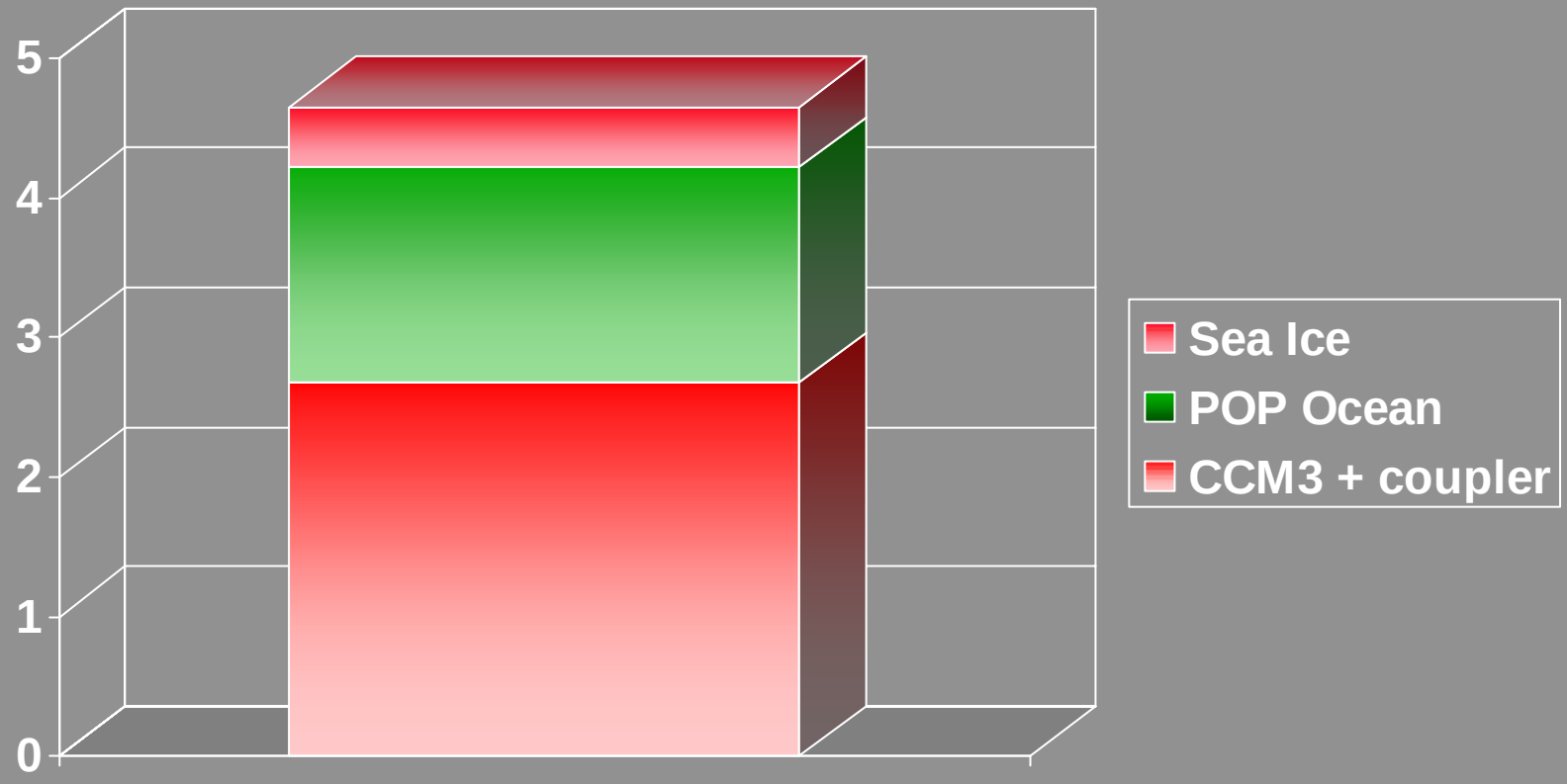
# Ocean Surface Currents




IFS RAPS 4
T213 L31
forecast days per day
1000
100
1
10
100
1000
PEs
VPP700
VPP700E
VPP5000
T3E 300 Mhz
T3E600 Mhz
SX/4
SX/4 x 2
SX/5
IBM SP p2sc 160 Mhz

# PCM Total Wall Time

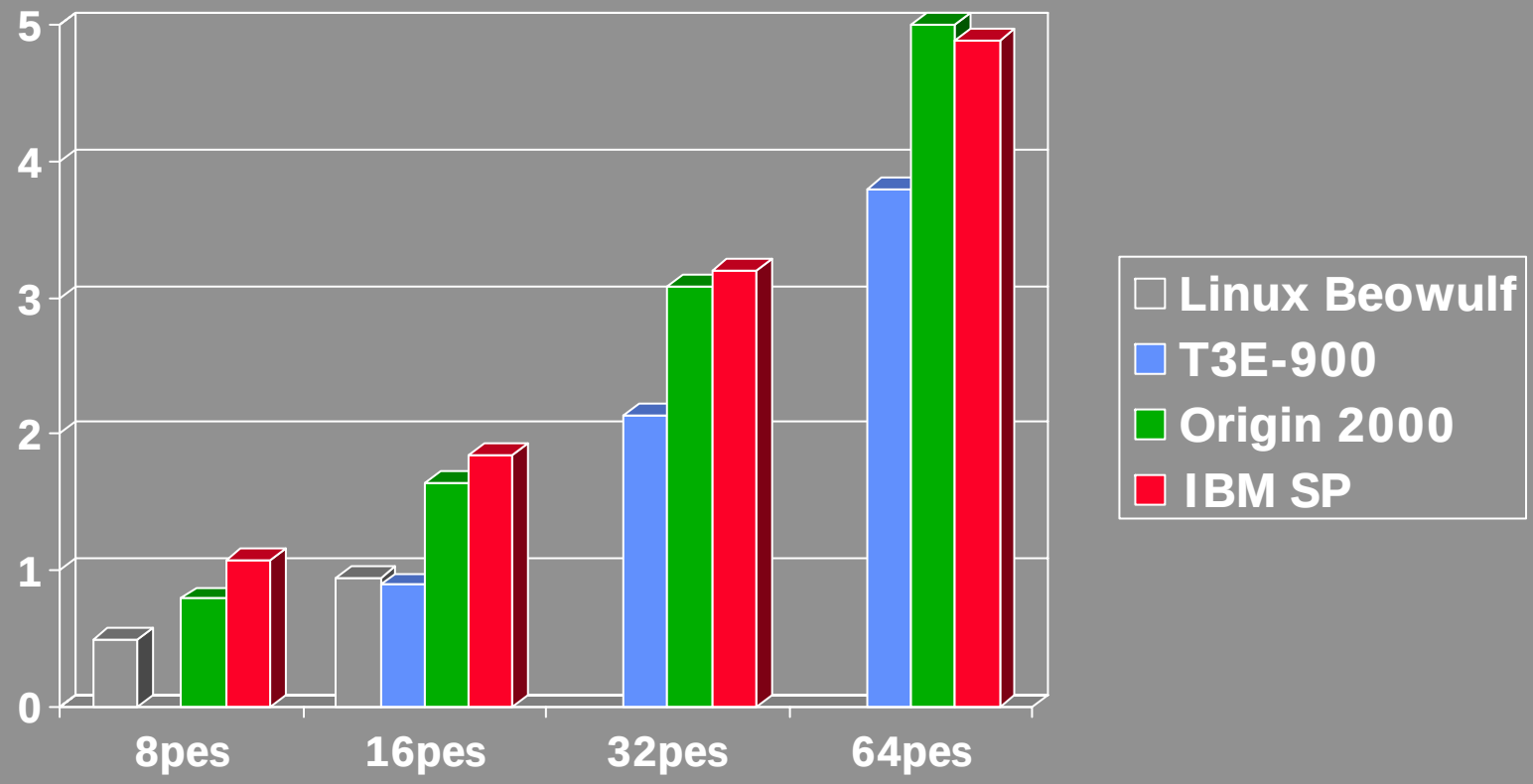
*64pes SGI Origin 2000*




**Wallclock Time (hours) per Simulated Year**

# PCM Performance

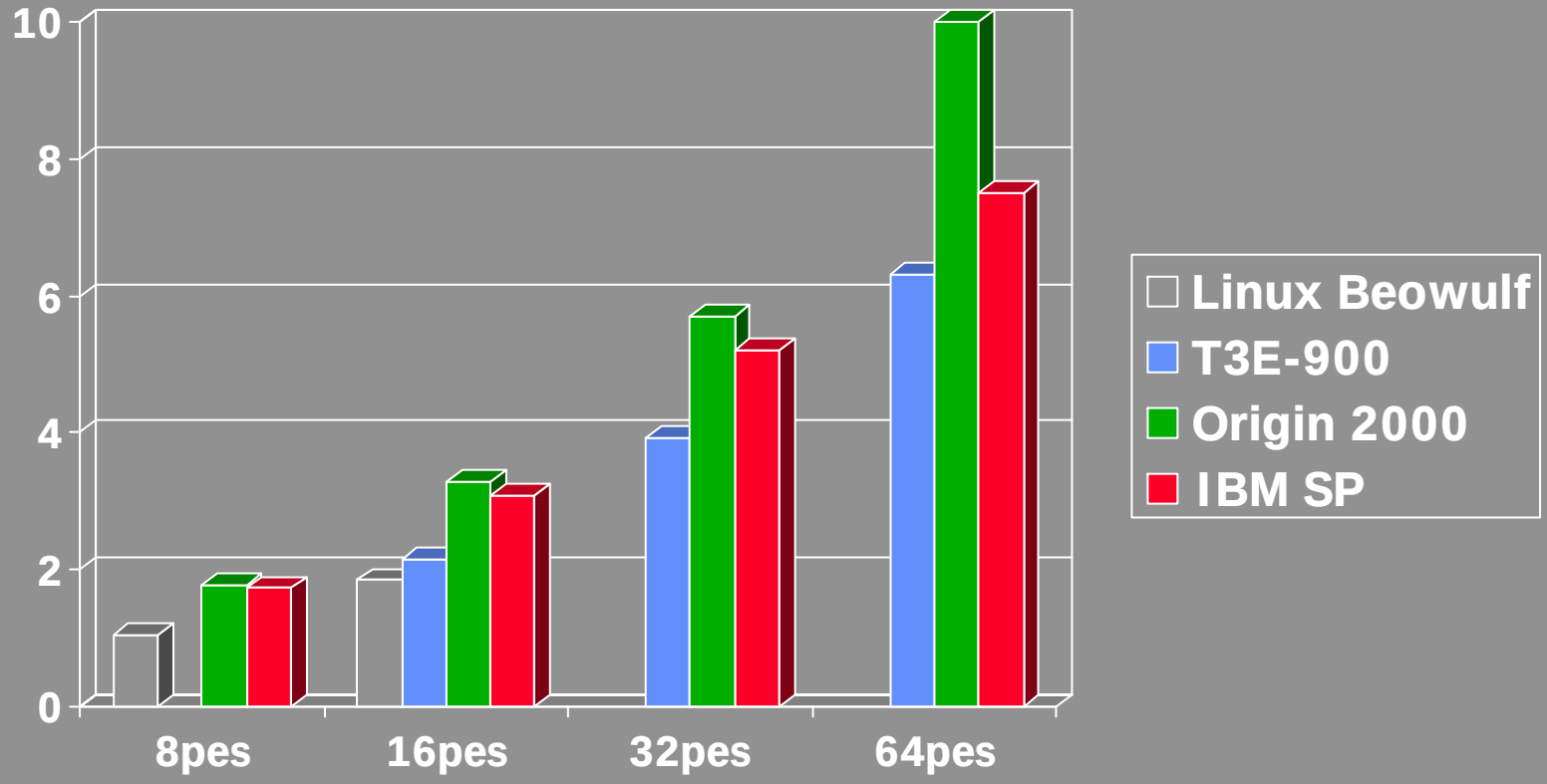
## Full Model




**Simulated Years per Wallclock Day**

# PCM Performance

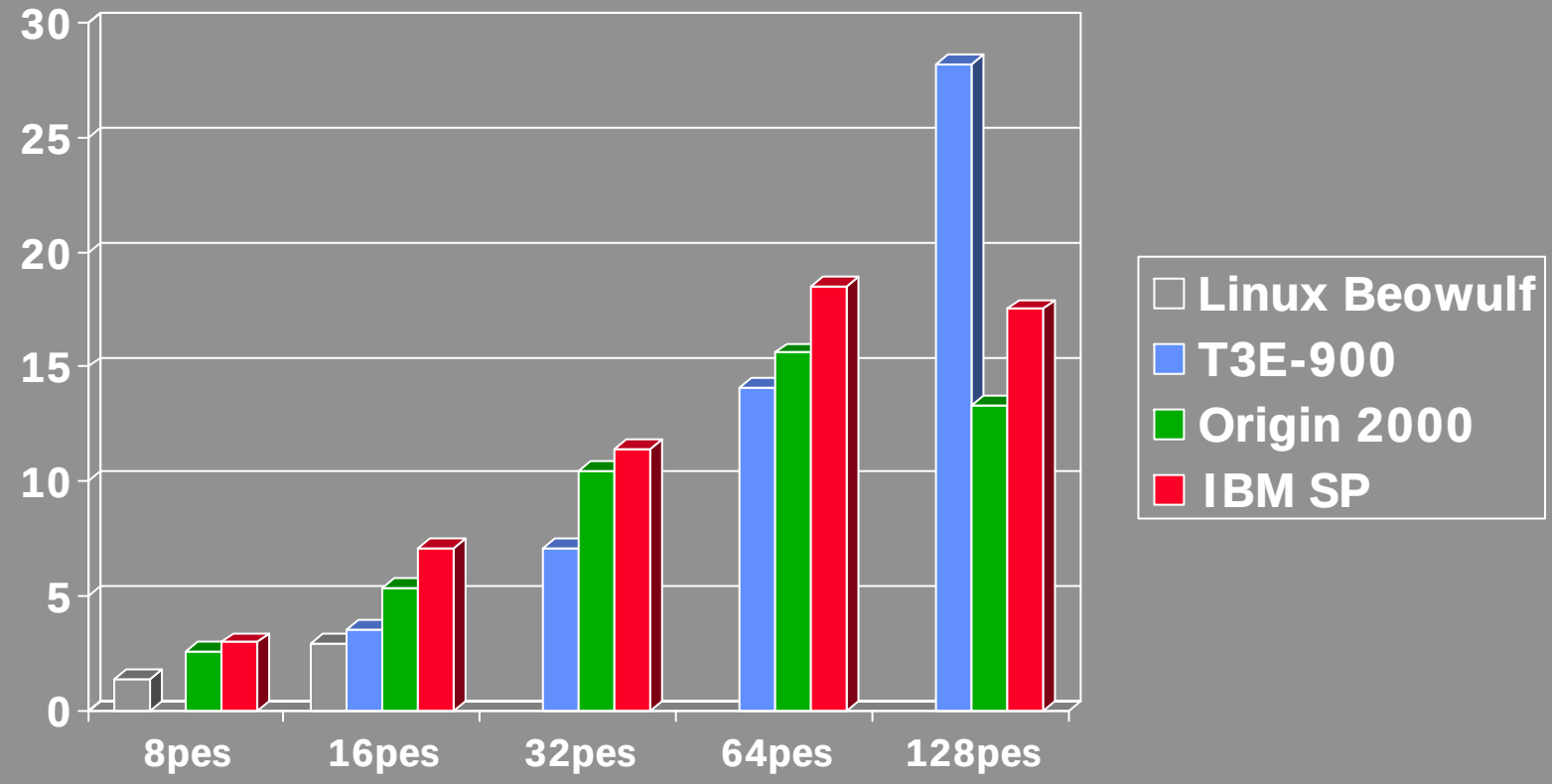
## CCM3 plus coupler




**Simulated Years per Wallclock Day**

# PCM Performance

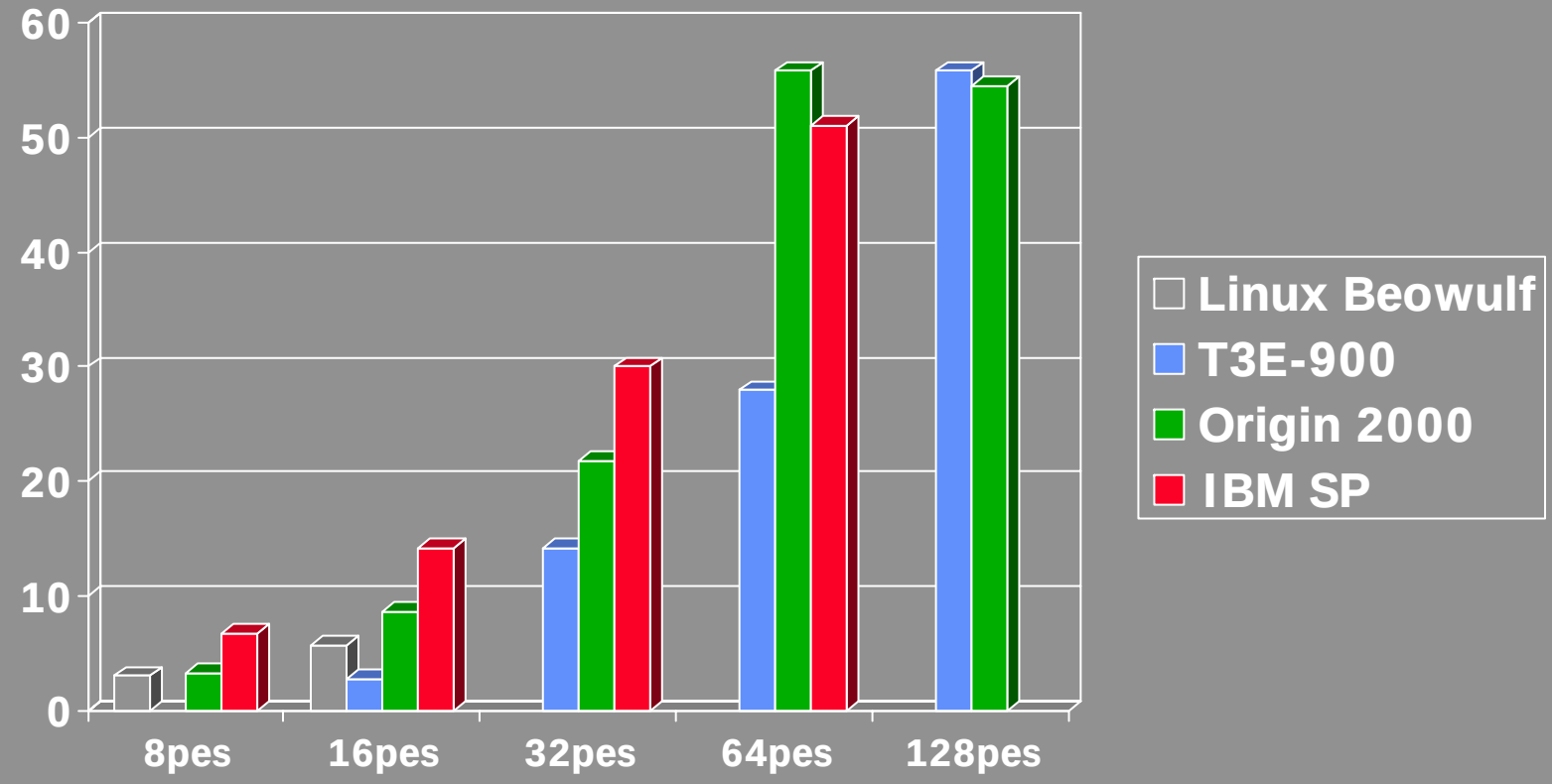
## POP Ocean




**Simulated Years per Wallclock Day**

# PCM Performance

## Sea Ice




**Simulated Years per Wallclock Day**

# *Initial BER SciDAC Application - the Accelerated Climate Prediction Initiative (ACPI)*

## *Goals*

- To accelerate progress in climate model development and application;
- To substantially reduce the uncertainties in decade-to-century model-based projections of climate change; and
- To increase the availability and usability of climate change projections to the broader climate change research and assessment communities.

# *Estimated Computing Needs: 10-40 TFLOPS Sustained*

| ***Weather Scale Resolution*** | | ***Ensemble Runs*** |
|---|---|---|
| Ocean | 10 km | 10 scenarios |
| Atmosphere | 30 km | 10 realizations/ |
| Land surface | 1 km | scenario |

***Processes***

Cloud-radiation interaction

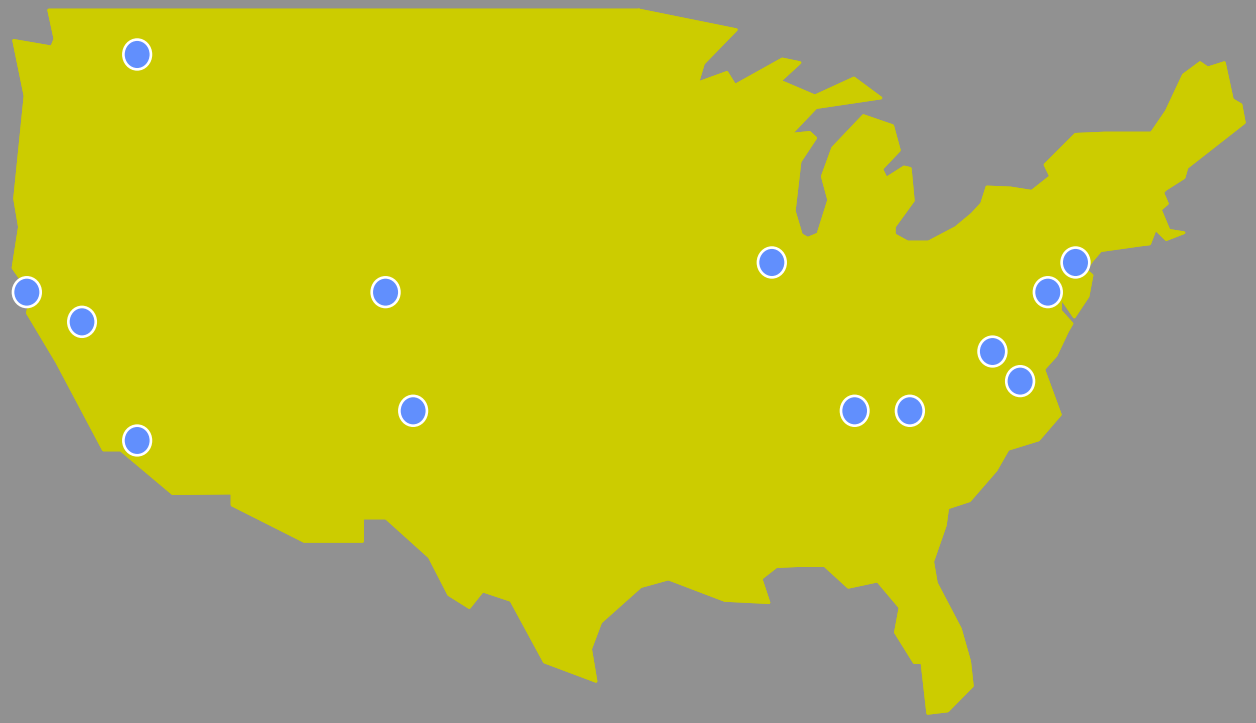
Biogeochemistry and hydrology

Aerosols and non-$CO_2$ greenhouse gases

Surface-atmosphere exchange

# ACPI Technical Approach

Create Multi-Agency, Multi-Institution Model Development Collaboration linked by Network Fabric



# *Partnerships through SciDAC will provide Infrastructure and Environments*

- Virtual Laboratory Collaboration Fabric
  - Advances in storage, analysis and visualization of petabyte datasets
  - Multi-institutional problem solving environment
  - Regional collaboratories for assessment and impacts research
- Terascale Computing Systems
  - Better numerical and computational methods for parallel systems
  - Breakthroughs in scaling for parallel systems

# *ACPI FY 2000 "Quick-Start" Activities*

- Pilot End-to-End Project led by Tim Barnett
- Computational Science/Engineering Collaboration on the NCAR CCSM

# *Relationship Between CSM and CCPP*

- Common Interests
- Cooperation on model development at the PI and Project Leader level
- CCPP owns computer resources
- CCPP focus on Dec-Cen and ACC simulations
- CCPP focus on advanced computing systems and computational science
- Some friction is inevitable and healthy

# *Characteristics of Model Consortium*

- Requires true collaboration among participants - not dominated by any one discipline or organization (science, computational science, computer science)
- Little precedent in climate and weather, has worked in other disciplines, e.g. computational aerodynamics
- Must work toward identified milestones in cooperation with other SciDAC elements