# CyRDAS

## Cyberinfrastructure for Research, Development and Education in the Atmospheric Sciences

### A Community Planning Activity

# Background

Implementing the strategy for accelerating progress in research and education in the geosciences will require "…a commitment to improve and extend facilities to collect and analyze data on local, regional, and global spatial scales and appropriate temporal scales…"

- *NSF Geosciences Beyond 2000*

*Understanding and Predicting Earth's Environment and Habitability*

" … a new age has dawned in scientific and engineering research, pushed by continuing progress in computing, information and communication technology, and pulled by the expanding complexity, scope, and scale of today's challenges."

- Atkins Report, 2003

Blue Ribbon Panel on Cyberinfrastructure

*Revolutionizing Science and Engineering Through Cyberinfrastructure*

- The Atkins report recommended an additional $1B/year support of cyberinfrastructure (CI) to support numerical investigation

# Drivers – Atmospheric Science

- Huge growth in observational and model data volumes and data streams, e.g., constellations of research satellites and use of GPS occultation
- Rapid expansion of application of data assimilation as a tool for providing homogeneous data sets and extracting maximum information content from observations
- Worldwide collaboration – the atmosphere is global – necessitates better tools for communication and cooperation

# Drivers – Atmos Sci Education

- Integration of collaboration tools and distributed data capabilities into the atmospheric sciences curriculum can fundamentally change the way meteorologists are taught and trained
- Real-time weather data and archives of climate data now commonly available in the classroom
- Digital libraries – assemblies of high-quality teaching and learning resources with tools to enable exploration of large atmospheric science data sets
- Services available to support distance learning, publishing on the Web and multi-disciplinary educational activities

# Drivers – Computer Hardware

- Rapid advances in processor performance (Moore’s Law) enable very large, complex computations
- Rapid advances in storage capacity enable online storage and interactive processing of enormous volumes of data
- Growth of broadband communications  enables the transmission and distribution of very large data sets over wide area networks → remote computing, possibly with loosely connected heterogeneous assemblies of processors (GRID computing), and remote data access, analysis, and visualization
- Wireless communications offers the prospect of “ubiquitous computing”

# Drivers – Computer Software

- Application of software engineering principles, e.g., modular design, software reuse, and platform-independent software
- Development of multi-thread codes to take advantage of symmetric multi-processor computing platforms
- Developed of distributed memory, message-passing codes to take advantage of tightly or loosely coupled networks of commodity processors
- Data distribution software to take advantage of higher performance broadband communications
- Visualization software that can graphically represent the information content of large data volumes

# CyRDAS – CI Planning for the Atmos Sci Community

- Seek broadest possible representation of the atmospheric sciences community
- Begin process for integrating CI planning for the atmospheric sciences into the larger planning processes for geosciences and environmental sciences
- Seek broadest possible dissemination of findings and recommendations

- NSF grant provided for one-year fact-finding and report

# CyRDAS – Activities

- Assess:
  - Opportunities for advancement in atmospheric sciences research and education made possible by current or anticipated advances in IT and CS
  - Opportunities for advances in atmospheric sciences research and education  that might result from collaborative research between atmospheric scientists and computer scientists
  - Ways in which CI can contribute to formal and informal education in atmospheric science
- Recommend:
  - Strategies that will help the academic research community to exploit the opportunities identified
- Develop:
  - An implementation plan for a distributed CI that will meet the needs of the academic atmospheric science research community and which includes the flexibility to grow smoothly as that research advances and CI needs grow

# CyRDAS – Committee

Jim Kinter (COLA) – Chair

D. Bader (Dept. Energy)
E. Barron (Penn State)
J. Bredekamp (NASA HQ)
G. Carmichael (Univ. Iowa)
C. DeLuca (NCAR - ESMF)
K. Droegemeier (Univ. Oklahoma)
T. Gombosi (Univ. Michigan)
J. Hansen (MIT)
J. Holt (MIT)
W. Matthaeus (Univ. Delaware)
M. Marlino (UCAR - DLESE)
M. Meehl (NCAR - SCD)
M. Ramamurthy (UCAR - Unidata)
R. Wilhelmson (Univ. Illinois)

# CyRDAS – 2003

- Engage the atmospheric sciences community
- Gather information from atmospheric sciences and computer sciences communities
- Develop a report that defines CI, identifies CI opportunities for the atmospheric sciences research and development and education community and makes recommendations for strategies to best to take advantage of developments in CI for more rapid advances in the atmospheric sciences

# CyRDAS Focus Groups (Conducted in Fall ’03 and Spring 04)

- Mountain region, NCAR
- Midwest region, NCSA
- Northeast region, ACCESS (NSF)
- Southwest region, SDSC
- Southeast region, Georgia Tech
- Northwest region, Univ. Washington
- Educators focus group, UCAR/DLESE

# CI Goals for the Atmospheric Sciences

- How can cyberinfrastructure lead to more rapid and more substantial progress in research and more effective education?
- What cyberinfrastructure barriers are impeding progress?
- What are the central issues that atmospheric scientists, educators and technologists consider most important, from their individual perspectives, to help them achieve what they hope to accomplish.

# CI areas covered during focus group meetings

- Social and cultural (workforce) issues
- High-end computing issues
- Data issues
- Software issues

# CyRDAS – Questions

## Social and cultural (workforce) issues

- What are the CI-related atmospheric science work force needs of the Nation?
- How can we ensure that students who intend to pursue careers in atmospheric sciences are aware of and know how to use the appropriate hardware and software tools?
- How best can computer scientists and atmospheric scientists be encouraged to work together?
- Is it reasonable to assume that the CI can stimulate closer relationships between the research and education communities?
- How should the intellectual property rights culture in the atmospheric research and education community be structured?
- What is a reasonable level of cost to ensure universal access to CI?

# Findings

## Social and cultural (workforce) issues

- Need ways to change academic culture to recognize achievement in CI
  - Count code like papers and code usage like citations for reward metrics
  - Peer-reviewed journal for CI in AS (GEO?)
  - Sabbaticals with major CI component
- Need basic programming support, not just collaboration with CS
- Need ways to increase trust between AS and CS
  - Antagonism between AS and CS, due to perceived competition for funding
  - Small teams of software "craftspersons"
- Programming languages and networking technology change faster than domain sciences → universities can't stay current
  - Need "workplace of the future" workshops
- Students are unprepared to understand computational physics, work with complex programming environments and interpret large volumes of data → Should have students demonstrate competency in computational physics, programming and data analysis, analogous to math proficiency
  - Computational physics courses, including interpreting incomplete, erroneous data
  - UG programming for scientists & engineers (incl. practical tools: debuggers, sccs)
  - CS community likewise needs to be more educated about AS

# CyRDAS – Questions

## High-end computing issues

- Are the high-end computing solutions (architectures, capability, and capacity) that are currently available commercially, and those that may reasonably be anticipated to be available in the near future, meet the needs of atmospheric sciences research?
- What are the requirements in the next decade for implementing more accurate, and better spatially resolved ("larger") problems, and problems with more realistic physical parameters, longer duration and ensembles?
- What demands will the atmospheric sciences community have for different types of computing facilities, including large shared-memory multiprocessor supercomputers, local Beowulf clusters, and distributed grid computing capabilities?
- How should resources be distributed among individual PIs, small centers and large centers?

# Findings

## High-end computing issues

- Opinion divided on best architecture: vector vs. massively parallel
  - Large-model advocates
    - scalability of hydrodynamic codes (highly vectorizable) is limited by Courant condition
    - need high resolution to take advantage of many of processors
    - impractical for long climate integrations
  - Moderate-model advocates
    - economics argue strongly in favor of distributed-memory architectures
    - massive parallelism could serve AS needs, e.g., large ensembles, parameter exploration
    - need codes designed to adapt to constraints of distributed memory
- Scientists want to control computing resources
  - Better turnaround time
  - High-risk calculations
  - Trend is toward more campus-based cluster computing
- Broadly-held skepticism about Grid computing (idea far ahead of reality)
  - People get local resources for control … why will they give up control to Grid?
  - Technical challenges, e.g., heterogeneity and reproducibility, seem daunting
- Setting up automated mass storage is harder, more expensive than FLOPS
- Modest resolution models should get priority in climate modeling for larger ensembles and broader investigation of parameter space

# CyRDAS – Questions

## Data issues

- What are the requirements of the atmospheric sciences research community for data archiving and dataset organization, management and access for the vast datasets that will be accumulated using multi-point measurements and multi-station observational datasets (weather stations, spacecraft, etc)?
- What is the proper balance between archiving data at a central location, such as NCAR, and at distributed sites such as observatories or PI institutions?
- Are the ongoing efforts in metadata generation by digital library communities sufficient to address the needs of the entire atmospheric sciences community?
- Are scientific data mining tools being used in current atmospheric sciences research?

# Findings

## Data issues

- Data release needs to be more timely, useful
  - Data should be distributed ASAP after suitable publication time has elapsed
  - Provide funds for costs of data release (documentation, metadata standardization, distribution etc.)
  - Impose penalties for not releasing in a timely manner
- Data protection is critical
  - Save all unique data that cannot be recreated in multiple locations with plans for ongoing media migration, ongoing testing of back-up and recovery plans, and on-site data stewards
- Individual data sets are not very valuable any more
  - Need many interoperable data sets, including catalog interoperability
  - Metadata standardization is major issue; also need standards for mass storage and data transport among large centers
  - Most important technical barriers to attack are those that inhibit rigorous comparison between models and observations
- Data mining is part infrastructure and part research problem → need way to transfer technology from research to CI

# CyRDAS – Questions

**Software issues**

- What software practices and design methods are most critical for smaller projects and large efforts in the atmospheric sciences, and how can those skills be taught?
- What can we learn from the commercial domain and other fields about the management of software projects?
- What role do frameworks play in enabling new atmospheric science? How can frameworks be designed to accommodate scientific flexibility and technical innovation?
- How important are geographic information systems (GIS) tools and services in current and future atmospheric sciences education and research activities? Would there be a benefit and what specific steps are needed to bridge the gap between GIS tools and scientific analysis and display applications?

# Findings

## Software issues

- Open source should be encouraged/required in software/tool development
- Development environment: US-born students learn in Microsoft environment, but scientific systems are Linux/Unix-based
- Need robust process for making standardized libraries available
- Need to think of CI as base for frameworks activity
  - frameworks are boring CS research once they become useful
- Should not invest solely in community models; need variety of targeted models
- Steep learning curve of many tools makes them impractical in educational setting
- GIS is not a mainstay of AS; however, it has the potential to offer a computational framework for integrating weather, climate, environmental, and socio-economic data

# Findings: General issues

- Many departments are not adequately provided with computing professionals, computing resources or wide area network bandwidth
- Many scientists and educators in the AS are not aware of or engaged in advances in CI
  - Need to disseminate CI information (newsletters etc.)
- CI needs to be as transparent as possible
  - More closely integrated in domain sciences
  - Large centers should be domain-centric with local expertise in that domain
- High-end computing is self-selecting and self-limiting → always looking for next 10X in capability
  - Report should include a "catalog" of AS challenges that motivate/require the next 10X in computing power
- Proportion of resources should shift from centralized to medium-sized creative projects
- Virtual observatories & collaboratories: "build it & they will come" is backwards → collaboratories should be for those who want to collaborate
- Concentrating resources <u>exclusively</u> into centers is antithetical to CI; centers' success metric is utilization, which increases queue time and thrashing

# Recommendations

The committee identified five main areas for its recommendations to NSF, organized into findings and short-term (6 months to 2 years) and long-term (5-10 years) recommendations

- General issues
- Social and Cultural issues
- Data
- Computing infrastructure and capacity building
- Software

# General Issues

- Human resources, an essential element of CI, have been significantly underemphasized
- Two unique aspects of Atm. Sci. are its operational and real-time elements
- The software life cycle, including training, support and maintenance, has been neglected
- The diversity of CI activities threatens to lead to inefficiency, duplication of effort, and retardation of progress

# Social and Cultural Issues

- The academic culture inhibits development/application of CI
- Many departments are inadequately provided with computing professionals
- Inadequate education and training
- Cultural clash between IT research and applications (CS researchers and atmospheric scientists)
- AS research and education is increasingly complex with many diverse data streams and sources
- Collaboratories have not fulfilled their potential

# Data Issues

- Data and metadata from diverse sources must be universally available in a timely manner and seamlessly interoperable
- Utility of data must be improved
- Scientists must be able to publish and distribute their analysis, findings and results, along with the underlying data
- Valuable, unique data must be curated for the long term

# Computing Infrastructure and Capacity Building

- AS will continue to drive the high-end computing requirements of the Nation (but it will not necessarily drive the computing infrastructure market
- Distribution of computational resources is not adequate, balanced, or seamless
- The proliferation of architectures and computing paradigms and the related lack of effective systems-level tools presents difficulties (opinion sharply divided on best architecture for AS)

# Software

- A proliferation of specialized software tools limits productivity with steep learning curves and a fog of ignorance
- Many community codes suffer from poor performance due to slow adoption of open source model for software development
- Software lifecycle: many software frameworks and codes suffer from the absence of sustained funding