Cyberenvironments

James Myers
Associate Director, Cyberenvironments

National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign
Cyberenvironments Directorate

• Partnering with communities
• To create and operate end-to-end environments supporting their research, education, and applied goals, and
• Building such environments as infrastructure
Data + Computation + Community = Knowledge

- Distributed data/metadata with global IDs and provenance
  - Source to Preserved Reference
- Workflow supporting modeling and analytics on streaming and reference data
  - Generates provenance going from raw data to discovery to application
  - Application-style interfaces
- Active Publication of Data and Models
  - Web 2.0 and desktop interfaces to run models and explore data
  - Community contributions of data, models, annotations, citations → expertise
Cyberenvironments

Mosaic and Cyberenvironments

- **Mosaic**
  - By early 1990s, the internet had a wealth of resources, but they were inaccessible to most scientists
  - Individual publishing
  - Browsing versus retrieving
  - See “Web 2.0 ... The Machine is

- **Cyberenvironments**
  - By the early 2000’s, the internet and grid had a wealth of interactive resources, but they were inaccessible to most scientists
  - Individual information models
  - Fusion versus gathering
  - Evolving towards end-to-end support for creating, organizing, and using information resources...

See “The Machine is Us/ing Us”! Michael Wesch
MAEViz: Consequence-Based Risk Management for Seismic Events

- Engineering View of MAE Center Research
- Physical through Socio-economic Analysis
- A “Cyberinfrastructure Aware” Application
Tennessee: Over 250,000 buildings are moderately or more severely damaged, over 260,000 people are displaced and well over 60,000 casualties (injuries and fatalities) are expected. Total direct economic losses surpass $56 billion.

Missouri: Well over 80,000 buildings are damaged leaving more than 120,000 people displaced and causing over 15,000 casualties. Total direct economic losses in Missouri reach nearly $40 billion.

Kentucky, Illinois, Arkansas, Mississippi, Indiana, Alabama: $45B, $35B, $19B, $9.5B, $1.5B, and $1.0B billion, respectively.

…the total economic impact of a series of NMSZ earthquakes is likely to constitute by far the highest economic loss due to a natural disaster in the USA.
“I have sensitive data I won’t distribute”

Network Aware

- Secure Enterprise Data
- Desktop
- Data/Metadata
- Public Reference Data
- Computation

- WebDAV, JCR, RDF, SAM, Tupelo
“Understanding the Scientific Basis of Decisions is Critical”

→ Process Aware

- Process Capture
- Discover
- Execute
- Report

• Workflow, Provenance, OPM/RDF
“Developing a Scenario requires a wide range of expertise”

Group Aware

- Plan, Coordinate, Share, Compare

- Collaboratory, Portal, …

- Wiki
- Task List
- Chat
- Document Repository
- Scenario Repository
- Training Materials

SSO
“My results could impact how we prepare for the next event”

→ Dynamic

New Third-Party Analyses

Compare, Contrast, Validate

Auto-update

MAEviz
GIS
Workflow
Data
Eclipse RCP
Plug-in Framework

• Plug-ins, Provenance, Environment
Digital Observatories

Observe

Model

Explore

Understand

From Basic Research to Societal Impact

Researchers

Policy Makers

Students

Citizens
Addressing System Issues - Scaling

Non-linearity & state shifts

Hot Spots and Hot Moments
National Cyberinfrastructure For Community Data Synthesis, Modeling, and Collaboration

WATERS Site Observations
Nested, high spatial and temporal frequency, broad variable coverage

Coupled Model Development
Enhancement of current models to explain high-frequency observations at sites and broad geographic and temporal correlations

National Survey and Classification
Synthesis of data from multiple networks to provide a one-time/low frequency characterization of large regions

Statistical Model Validation
Broad optimization and sensitivity analyses identifying correlations of predictions with variables to identify missing/incorrect process modeling

Policy Makers
The Public
Education and Training
Closing the Loop: An Advanced Information System for Real-Time Decision Making

Data Fusion

Virtual Sensors

Model & Tool Integration

Automated Provenance

Data Robots

Archives

Visualisation

Executive Dashboard

Background Cloud & Grid Computing

Calibration
Tupelo II: Semantic Content Management

- Web Protocol to
  - Authenticate
  - Get/Set Data
  - Get/Set Metadata
- Flexible Global Identifiers
- Secure Context Management
- Extensions to support specific ontologies (provenance, data streams, GIS, …)
- Extensions to support computational inference (spatial indexing, provenance, extracted metadata…)

Data Streams
Data Files, Documents
GIS Structures, Images, graphs
Metadata, Provenance

Local and Distributed Data Sources

Tupelo Semantic Content Middleware
Time-Annotated RDF – extensions for streaming

- Streaming data → content + metadata
- Time Indexing
- Metadata and query constructs to allow stream-oriented access: the last 24 hours of data from sensor X:

```
SELECT ?temp ?s.t
WHERE {
  <urn:Chicago> <urn:hasSensor> ?s .
  ?s[LAST] <urn:hasValue> ?temp .
}
```
The Open Provenance Model

- A directed graph representation with three node types
  - **Artifact** - Immutable piece of state - a physical or digital object
  - **Process** - Action or series of actions performed on (data flow) or caused by (control flow) artifacts resulting in new artifacts.
  - **Agent** - Contextual entity acting as a catalyst of a process, enabling, facilitating, controlling, affecting its execution.
- **Nodes connected by edges**
  - used(R), wasGeneratedBy(R), wasControlledBy(R), wasTriggeredBy, wasDerivedFrom
- **Edges relating to processes can specify a Role (R)**
- **Nodes can be annotated with properties**
CyberIntegrator Workflow System

- Exploratory workflow (macro-recording)
- Simple integration with Matlab, Excel, Fortran, etc.
- Provenance tracking
- Distributed, shared data access (HIS, WebDAV, …)
- Remote Execution
- Workflow/model publication
- Metadata and Annotation of data, modules, workflows
- All data, metadata, configuration information stored as semantic content
Core DSF Concept

Web Inputs
- Parameters
- Input Streams
- Trigger Conditions
- Visualization
- Provenance and Annotation Options

Workflow Execution Service

Visualized Outputs

Desktop Exploration

Publish

Semantic Content Repository & Provenance Store

VM Farm / Compute Cloud

Annotations
- Spatial Distribution of Snow
- Historical Average Temperature
- I remember the Holiday Blizzard!
Institute for Genomic Biology and 4-H: Plant Growth Model

Making state-of-the-art plant growth models available for 4-H/student use

Integrating sophisticated modeling into “Seeds and Soils” 4-H activities
Virtual Sensor Result 2: Polygon-Based Real-Time Rainfall Data

Applications, Services, Repositories, Portals, Middleware

- WebDAV Drive
- Semantic DocLib
- Workflow (CI)
- Eclipse RCP Data Catalog
- Streams/events
- Java Beans
- Web Page Dev.
- Translation/extract ion plug-ins
NCSA CyberCollaboratory

- Liferay Portal-based
- Group Spaces
  - Document Library
  - Forum discussion
  - wiki, blog
  - Announcements
- Simplified email invitation
  - To register or join a group
- Email integration with Forum
- Provenance and event tracking
  - Social network analysis
- Single-Sign-On (SSO)
- Deployed into multiple VMs for communities
  - NSF HydroSynthesis Project
  - NSF WATERS Planning Office
  - Illinois Math and Science Academy CoolHub portal
  - Office of Naval Research TRECC Ed KC portal
  - ..........
Medici – Multimedia Resource Environment

- Desktop and Web Interfaces to a Lustre-backed Tupelo store...

**Processing**
- Hooks to existing imaging software
- Provenance

**Ingestion**
- People + Lab Instruments
- Single file / batch ingestion
- Flexible / incremental metadata
- Automatic preprocessing
- Click and Drag

**Retrieval**
- Large image preview
- Desktop and web clients
- Portable data and metadata
- Search

**Social Annotation**
- Tags
- Comments
- Ratings

**Sharing**
- Copyright
- Attribution
- Science Commons Data Licenses
- Citable persistent URLs
- Access control

**Scalability**
- Millions of images
- Large images (200MB+)
- Virtual Machine hosted
- Distributed Image Collections

Imaginations unbound
Virtual Machine/Cloud Hosting

- NCSA is developing a VM infrastructure and cyberenvironment / data services appliances

Diagram:
- Desktop
- Cyberenvironment and Data Services
  - Load Balanced
  - Replicated
  - Monitored
  - Managed
- HPC Resources
- Cloud Resources
- Digital Libraries
Cyberenvironment Directions

‘Data-Intensive’ E-Science (CS)
Scaling in terms of computation, data, and size of community, statistical analyses, multiscale representations
Age and curriculum appropriate custom interfaces to research data and modeling capabilities

‘MAEviz-style’ E-Science
Additional science and engineering environments

‘Preservation’/Digital Scholarship (GSLIS/Library)
Providing long-term access to data and modeling capabilities to support ongoing research
Packaged products, Virtual machine farms, standardized processes, customer support

Cyber-Education (Education)

Operations (CITES/CIO)
Conclusion

• There are significant opportunities to support more of the scientific lifecycle and to integrate active research and community / reference services

• In doing so one must recognize that research involves heterogeneous and evolving data and processes… and design accordingly

• Semantic Content Management, Workflow and Provenance, VO Contexts, and Virtual Machines provide a coherent platform for development of such scalable and evolvable data services
Questions?

More info:  
http://cet.ncsa.uiuc.edu/  
jimmyers@ncsa.uiuc.edu

Acknowledgements: NCSA  
Cyberenvironments and Technologies Staff  
Office of Naval Research

National Center for Supercomputing Applications  
University of Illinois at Urbana-Champaign