Carbon Dioxide Flux Data Computing and Data Warehouses Using Grid Techniques

Primary Authors:
Cheng-Hsin HSU ; Wei-Long UENG
Academia Sinica Grid Computing
Taiwan, R.O.C.

Co-Authors:
Hsin-Yen CHEN¹ ; Eric Yen¹ ; Simon LIN ¹ ; Chau-Chin LIN ² ; Yue-Joe HSIA ³
1.Academia Sinica Grid Computing
2.Taiwan Forestry Research Institute
3.Institute of Natural Resources, National Dong Hwa University
Outline

- Background
- Objective
- Carbon Flux Data Warehouse
- Carbon Flux Data Processing
- Conclusion & Future Work
Background: Global Carbon Flux Network
Background: Global Carbon Flux Network
Background: FLUXNET

- a "network of regional networks," coordinates regional and global analysis of observations from micrometeorological tower sites.
Background: FLUXNET

Growth of Fluxnet
543 Towers as of March 30, 2008

- Africa (34)
- Asia (92)
- Australia/Oceania (17)
- Europe (170)
- North America (200)
- South America (30)

Year

Number of Towers

Research challenge

- **Archival storage:**
  - **A huge amount of data:**
    data receiving continuously in long-term time scale
  - **Metadata Management Issue:**
    accompany heterogeneous and complexity ecological data

- **Data Computing:**
  - **Computation Issue:**
    performs long-duration computations
  - **Computing Model**
    tower sites in Asia that locate at heterogeneous and complex land cover may need different analysis model
Objective: Carbon Flux Application Data Warehouse

- Utilizing the Advantage of EGEE infrastructure to:
  - Data Archived
  - Data Management
  - Data analysis
- Minimize the barrier to use the EGEE infrastructure:
  - Grid Application Platform Integration
  - Grid Portal Data Warehouse
  - Metadata Service for Data Discovery
- Offer a flexible Workflow Management Approach:
  - Workflow Composition and Management
  - Workflow Reusability
- Develop sharing mechanism:
  - Data Sharing
  - Data processing and analysis model workflow Sharing
Carbon Flux Data Warehouse Framework

- Integrate and Synthesize
- Share Information
- Information Management
- Data Collection

- Workflow
- Grid Portal
- Metadata
- Data Grid

From Dr. John Porter
Overview of Flux Application Data Warehouse

- **Sensor Network Layer**: measurement tower → field station → lab archive
- **Information Synthesis Layer**: field station data transfer → data warehouse and application → EML metadata
- **Information Management Layer**: metadata server
- **Computing Grid**: grid portal

Diagram elements:
- **Lab Archive**: for data storage
- **Data Grid**: for data management
- **Kepler**: software for data visualization
- **Grid Portal**: interface for data access
- **Metadata Server**: for metadata management
- **Field Station**: for collecting data
- **Wireless**: for data transmission
- **Computing Grid**: for processing data
- **CE**: for grid computing

The diagram illustrates the flow of data from the sensor network layer to the information synthesis layer, through the information management layer, and into the computing grid.
Carbon Flux Data Warehouse

- **Data Archived**
  - Automatic or semi-automatic archived data to Data Grid
  - By Kepler integrated GAP framework
  - By Grid Portal upload dataset

- **Data Management**
  - Used *Ecological metadata language (EML)* as the standard to describe data
  - To discovery and access data, we utilized the Grid Portal integrated with Metadata Server and GAP framework
The Whole Process

Data Archived

Collect data set

Kepler

Morpho (EML editor)

Create Metadata

Metadata System

MetaCat

Data Grid

SE

Data management & discovery

Grid Portal User Interface

Data management & discovery
Carbon Flux Data Archived Flow

- Automatic
  - Field Station
  - Level 0 or Level 1
  - Wireless

- Semi-automatic
  - Grid Portal

- Data Grid

- SE

- Semi-automatic
Carbon Flux Metadata Register Flow

- Generate Dataset Metadata

Morpho

- Combine Dataset and Metadata and Assign Dataset permission
Carbon Flux Data Processing

- **Data Processing**
  - Process raw data to 4 standard products
  - From Level 0 to Level 1, 2, 3, 4

- **Workflow Management**
  - Customized workflow
  - Workflow management and sharing

- **Outcome Display**
  - Produce diagram
Scenario of Carbon Flux Data Computing

- **Pre-requisites**
  - Security issues: certificate & Proxy

- **Discovery dataset and Submit Job**
  - Assign Target Dataset
  - Create computing workflow
  - Job Submit and monitor

- **Production**
  - Diagram
  - Synthesize result
Carbon Flux Data Processing Flow

**Raw data acquisition**
- De-spike
- Absolute limit

**Flux calculation**
- Averaging interval
- Coordinate rotation
- Time-lag correction
- WPL correction
- QA/QC: ITC, stationary test
- Flux calculation

**Quality control**
- 4 classes (0, 1, 2, 3)

**Data analysis**
- Energy balance closure
- Spectrum
- Meteorological parameter
Overview: Integration of GAP and Kepler

User

EGEE infrastructure
Carbon Flux Data Processing Flow

Get EML (Metadata)

generate R-script

Component:
Authentication
Job Submission
Job Monitor
Carbon Flux Data Grid Computing: Grid Portal

Data discovery

Workflow Compose

Job submit & Monitor

Fetch Result

Computing Grid

CE
Conclusion & Future Work

**Conclusion**
- The portal of the carbon flux data warehouse is enabled
- We approach a scientific workflow framework for analyses of all observation data.
- Made use of R as mathematical computation functions for statistical computing and deployed R service on grid
- To provide better environment for huge sensor data management and computation model.

**Future Work**
- the flexibility to integrate existing tools, the improvement of user interface, and the reusability of workflow components are the focus for advancement.
Questions & Contact information

HSU, Cheng-Hsin
Academia Sinica Grid Computing
Taiwan, R.O.C.
email: Jims.Hsu@twgrid.org

Thank you for your attentions.

http://gap.grid.sinica.edu.tw/
http://metacat.ndhu.edu.tw/clm_alpha/
http://metacat.ndhu.edu.tw/clm.html/instrumentation.html
Components

- **GAP (Grid Application Platform)**
  - a framework consist of high level API for access Grid Environment

- **AMGA**
  - a Metadata services background and possible uses on a grid environment

- **EML (Ecological Metadata Language)**
  - a metadata standard developed by the ecology discipline and for the ecology discipline.
  - Be implemented as a series of XML document types

- **Morpho**
  - a EML metadata editor, open source free software

- **MetaCat (Metadata Catalog)**
  - a flexible metadata database, open source free software

- **Grid Sphere**
  - a portal framework provides an open-source portlet based Web portal

- **Kepler**
  - a scientific workflow application, open source free software

- **R**
  - a language and environment for statistical computing and graphics, open source free software
<table>
<thead>
<tr>
<th>Number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Net radiometer</td>
</tr>
<tr>
<td>2</td>
<td>Net radiometer</td>
</tr>
<tr>
<td>3</td>
<td>Sunshine Sensor</td>
</tr>
<tr>
<td>4</td>
<td>Radiometer</td>
</tr>
<tr>
<td>5</td>
<td>Quantum Sensor</td>
</tr>
<tr>
<td>6</td>
<td>Aspired Temperature/Humidity Sensor</td>
</tr>
<tr>
<td>7</td>
<td>Unaspired Temperature/Humidity Sensor</td>
</tr>
<tr>
<td>8</td>
<td>Wind Monitor</td>
</tr>
<tr>
<td>9</td>
<td>Micro Response Anemometer</td>
</tr>
<tr>
<td>10</td>
<td>Barometer</td>
</tr>
<tr>
<td>11</td>
<td>Visibility Sensor</td>
</tr>
<tr>
<td>12</td>
<td>Leaf Wetness Sensor</td>
</tr>
<tr>
<td>13</td>
<td>Fog Collector</td>
</tr>
<tr>
<td>14</td>
<td>Soil Temperature Sensor</td>
</tr>
<tr>
<td>15</td>
<td>Soil Moisture Sensor</td>
</tr>
<tr>
<td>16</td>
<td>Soil Heat Flux Sensor</td>
</tr>
<tr>
<td>17</td>
<td>Eddy Covariance Measurement System</td>
</tr>
<tr>
<td>18</td>
<td>Data Logger</td>
</tr>
<tr>
<td>19</td>
<td>Computer</td>
</tr>
<tr>
<td>20</td>
<td>Temperature profile</td>
</tr>
<tr>
<td>21</td>
<td>$\text{CO}_2/\text{H}_2\text{O}$ profile</td>
</tr>
<tr>
<td>22</td>
<td>Sapflow sensor</td>
</tr>
</tbody>
</table>