The Cyberinfrastructure and Image Information Mining

Pushkar S. Pradhan, Roger L. King, Tomasz Haupt, and Valentine Anantharaj

GeoResources Institute, Mississippi State University

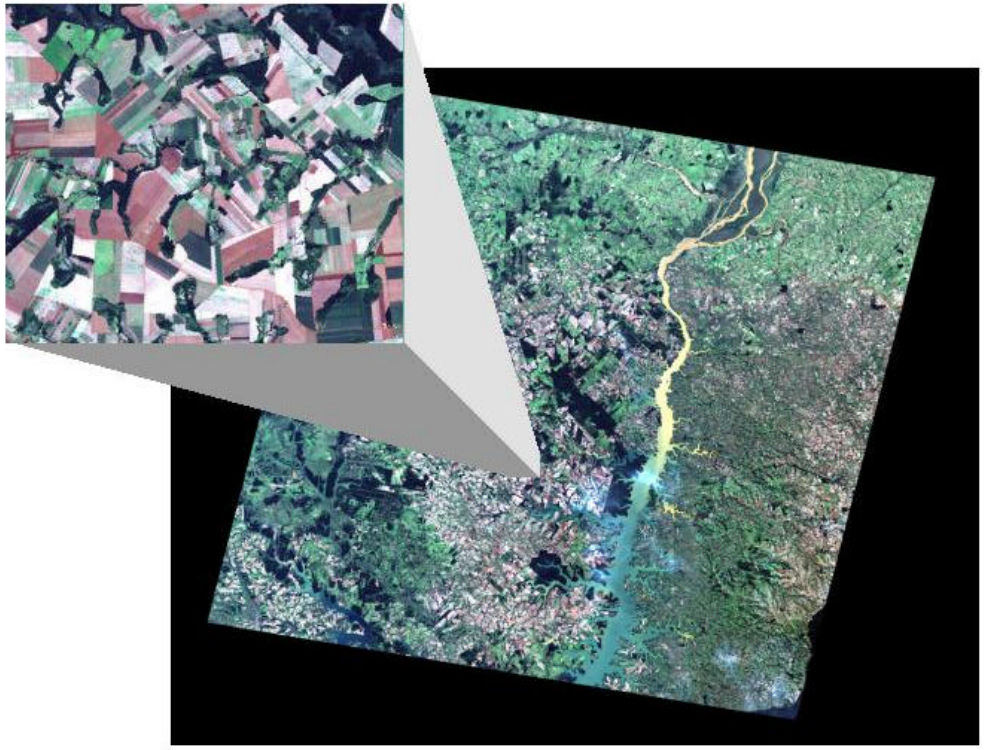
Problem

- Many realistic problems require expertise or data sources from globally distributed resources.
- These same problems may require processing of data into information from specific resources.
- How do I bring together geospatially diverse resources to facilitate sharing and knowledge discovery?
- How do I fuse data from a variety of sources (sensors, databases, images) into a useful information product?
- Constrained – compute horsepower, bandwidth, etc.

Crop Biosecurity Architecture

Ability to assimilate data from sensors and other data streams (e.g., human assets, laboratories, archives) and to use the information derived from these data sources to reduce the uncertainty in identifying agricultural threats.

This integrated observing web includes a variety of data types and vantage points (e.g., space, ports of entry, sentinel plots, production fields) and can stream from archived data and/or simultaneous, incremental or real time or near-real time sources.



Highway Watch® Program

- The *MSU EPE Center* has the important responsibility of *"establishing and refining a flexible, timely response mechanism that can be used to dramatically reduce response and recovery staging times and to provide a framework through which the trucking industry can be mobilized."*
- Need to provide an integrated systems framework that utilizes various technologies to detect critical threats and events and to communicate this information to a central site, where it can be quickly analyzed to determine the nature of the situation and the recommended response action, and then assist in identifying the locations of the appropriate emergency response organizations so that they can be quickly dispatched to the site(s) in question.

Comparison of Frameworks

- Crop Biosecurity
  - Building a reference scenario and framework
  - Anticipation
  - Prevention
  - Detection
  - Response
  - Recovery
- Highway Watch Program
  - Detection of threat or even
  - Communication of threat o event including location data
  - Analysis and impact assessment
  - Decision making and recommended response
  - Notification to first responders and impacted groups

Similarities

- Identification of threat
- Analysis of imagery (archives or near real time)
- Distributed databases
- Real time or near real time spatial data from sensors (e.g., ITS, in situ sensors, etc.)
- Domain specific software
- Decision making capability (knowledge representation)
- Recommendations

Cyberinfrastructure as a Solution

- Cyberinfrastructure is a NSF term used to refer to computational infrastructure that consists of:
  - computer hardware systems
  - application software and service
  - data and metadata management facilities and services
  - sociocultural elements of community building.
- The NSF recognizes that just as “infrastructure is required for an industrial economy, …, cyber infrastructure is required for a knowledge economy.”

Cyberinfrastructure Technologies

- Grid computing and Peer to Peer computing (p2p)
- Virtual Organization (VO) - An organization with its resources geographically distributed or when different organizations with different resources agree to share their resources in order to achieve a common goal.
- Resources - Data, CPU power, domain specific software, or human experts in their fields.

Grid Computing Middleware Features

- Delegation of user privileges.
- Integration w/ local security solutions.
- User must be able to specify resource usage policy (Who can share? What can be shared? How much can be shared?).
- Resource Information – The resource should publish its usage policy, state, availability, load, etc.
- Resource Discovery – Users must be able to discover resources. A directory based scheme can work well, which contains metadata about the resources.
- Resource allocation and reservation – Users must be able to schedule resources for future use.

Grid Computing Middleware Features

- The middleware must use standard protocols in order to be easily extendible, portable, and interoperable.
- The Open Grid Services Architecture (OGSA) defined by the Global Grid Forum has become the standard.
- The Globus toolkit is an implementation of the standard.

p2p Computing Middleware Features

- Peer is used to refer to a node participating in the computing environment.
- A peer can be either a client or a server or both.
- p2p computing is characterized by low bandwidth connections and intermittent presence by the peers. Thus failure will occur regularly and applications must be designed to handle them.
- Security - Since there are no *a priori* formal agreements between peers, it's difficult to identify a malicious user.
- p2p computing is also characterized by a large number of users.Thus scalability is an important issue.

Grid versus P2P Computing

- Pros:
  - Complex applications (grid computing evolved at universities, research labs).
  - Well accepted standards (OGSA) for middleware.
  - QoS – Nontrivial quality of service.
- Cons:
  - Scalability is not addressed well.

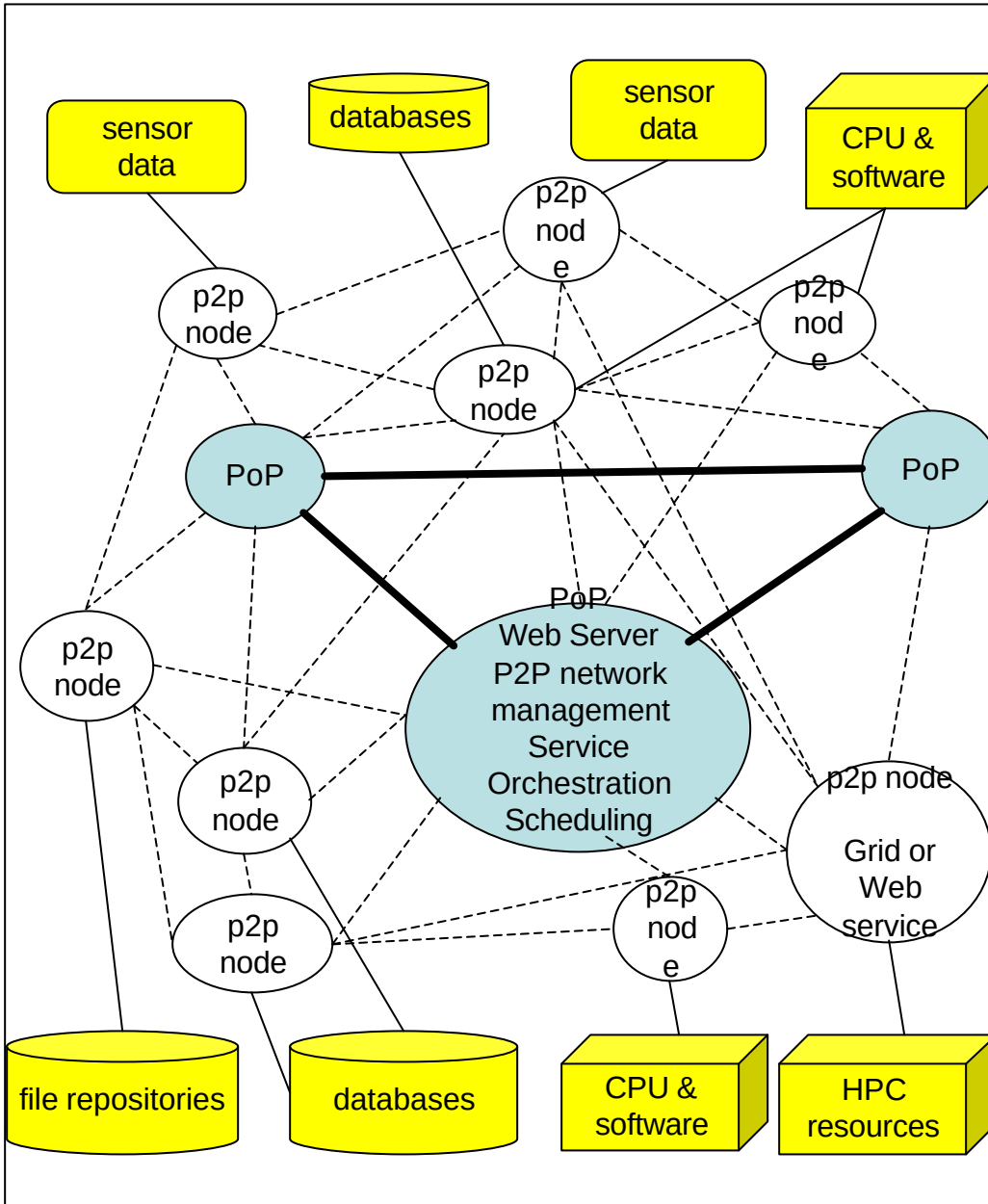
- Pros:
  - Dealing with a large number of peers and intermittent presence has led p2p to successfully address failure management.
- Cons:
  - No single middleware standard yet.
  - No QoS concept.
  - Applications are primarily file sharing and CPU sharing.

Cyberinfrastructure For Image Information Mining

- The cyberinfrastructure of choice often appears to be determined by the community of users.
- If the users are primarily sharing files they will opt for the less formal p2p approach.
- For example, a p2p based Java application to share satellite imagery with various researchers SATELLA (http://satella.geog.umd.edu/)
- DIGITAL PUGLIA pools together resources from University of Lecce, Italy, San Diego Supercomputing Center (USA) and California Institute of Technology (USA) in an active digital library of remote sensing data.

Cyberinfrastructure For Image Information Mining

- An I2M cyberinfrastructure must allow for on-demand aggregation of computational resources (such as computational servers, instruments and sensors, databases, and data repositories) across administrative domains at any time.



Conclusions

- Results from this web service can lead to discovery of new knowledge and understanding.
- Image information mining services will become critical middleware components of the cyberinfrastructure of the future.
- The task of this middleware will eventually be to operate not only on archived datasets, but also on data streams in near real time.
- The two vying approaches for a cyberinfrastructure application for image information mining both offer strengths.
- Research should be focused on using the best aspects from grid technologies and p2p computing for building an image information mining cyberinfrastructure.