

Experience With NASA's Grid Miner

Thomas H. Hinke

NASA Ames Research Center

Moffett Field, California, USA



Outline

- Why use the grid for data mining?
- Overview of Grid Miner
- Experience adapting existing stand-alone miner to grid
- A recent application of the Grid Miner



Grid Provides Computational Power

- Grid couples needed computational power to data
 - NASA has a large volume of data stored in its distributed archives
 - E.g., In the Earth Science area, the Earth Observing System Data and Information System (EOSDIS) holds large volume of data at multiple archives
 - Data archives are not designed to support user processing
 - Grids, coupled to archives, could provide such a computational capability for users



Grid Provides Re-Usable Functions

- Grid-provided functions do not have to be re-implemented for each new mining system
 - Single sign-on security
 - Ability to execute jobs at multiple remote sites
 - Ability to securely move data between sites
 - Broker to determine best place to execute mining job
 - Job manager to control mining jobs
- Mining system developers do not have to re-implement common grid services
- Mining system developers can focus on the mining applications and not the issues associated with distributed processing



Grid Will Provide Re-usable Services

- In the future, Grid/Web services will provide the ability to create reusable services that can facilitate the development of data mining systems
 - Builds on the web services work from the e-commerce area
 - Service interface is defined through WSDL (Web Services Description Language)
 - Standard access protocol is SOAP (Simple Object Access Protocol)



Grid Services: A Foundation for Grid Mining

- Global Grid Forum working groups on
 - Open Grid Services Architecture (OGSA) standard under development to specify a grid-enabled web services architecture. See “Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration”
 - Open Grid Services Infrastructure (OGSI) standard has been released. Specifies common interfaces that all grid services should support.



Grid Mining and OGSA/OGSI

- An OGSA/OGSI compliant mining service could be build
- Mining applications could be built by re-using capabilities provided by existing grid services.



Outline

- Why use the grid for data mining?
- Overview of Grid Miner
- Experience adapting existing stand-alone miner to grid
- A recent application of the Grid Miner



Grid Miner

- Developed as one of the early applications on the IPG
 - Helped debug the IPG
 - Provided basis for satisfying a major IPG milestones
- IPG is NASA implementation of Globus-based Grid
- Provides basis for what could be an on-going Grid Mining Service



Grid Miner Operations

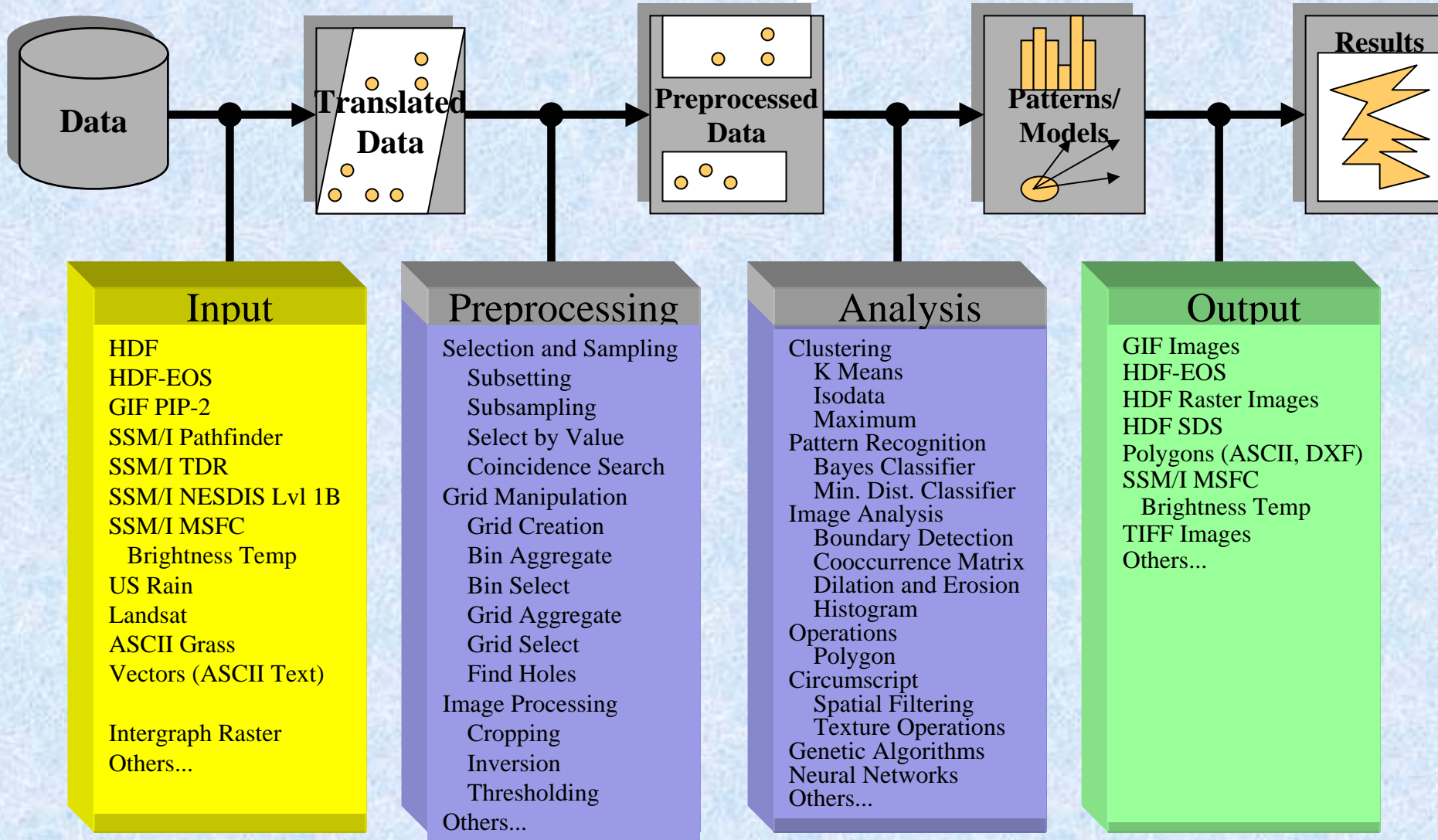
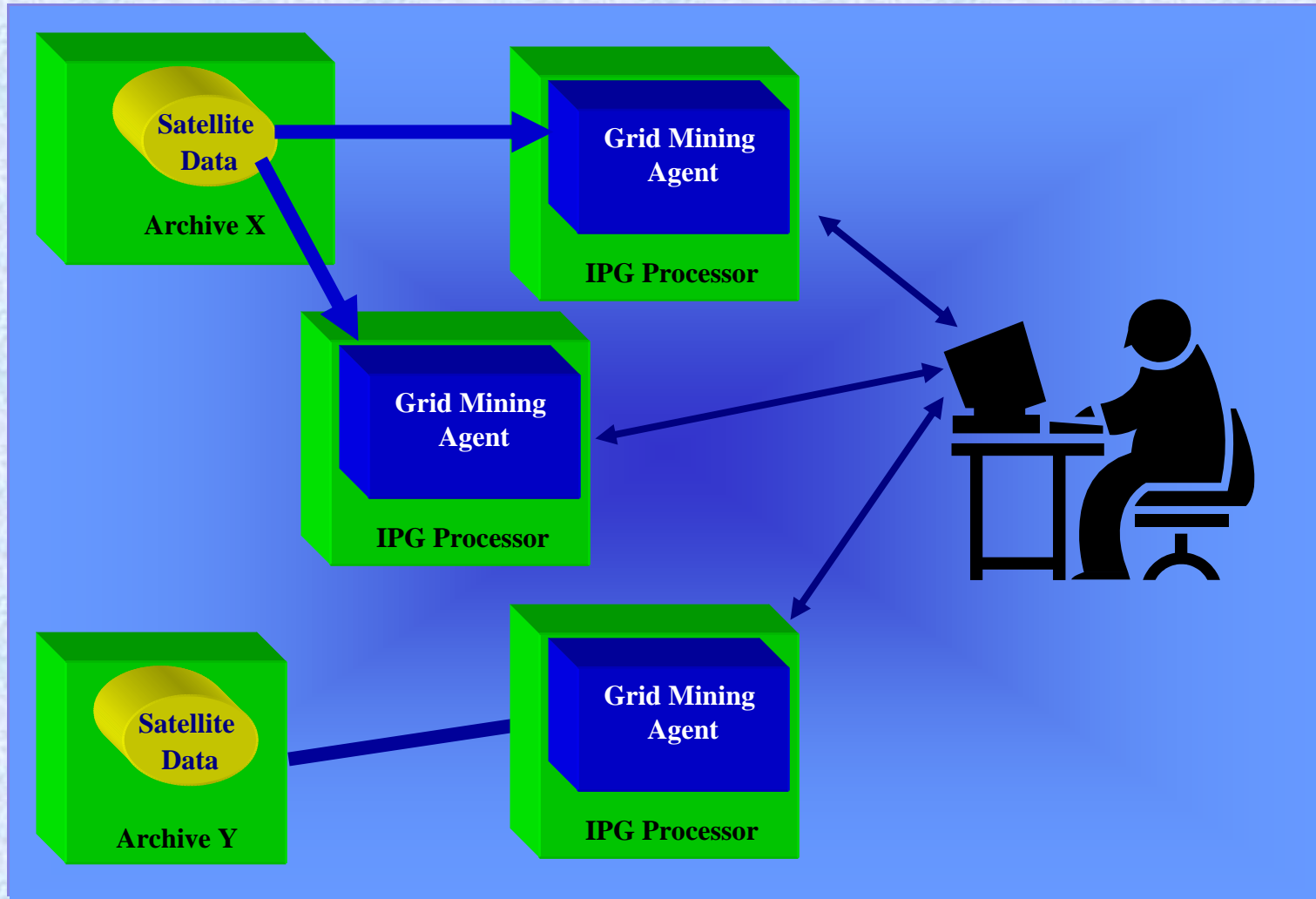
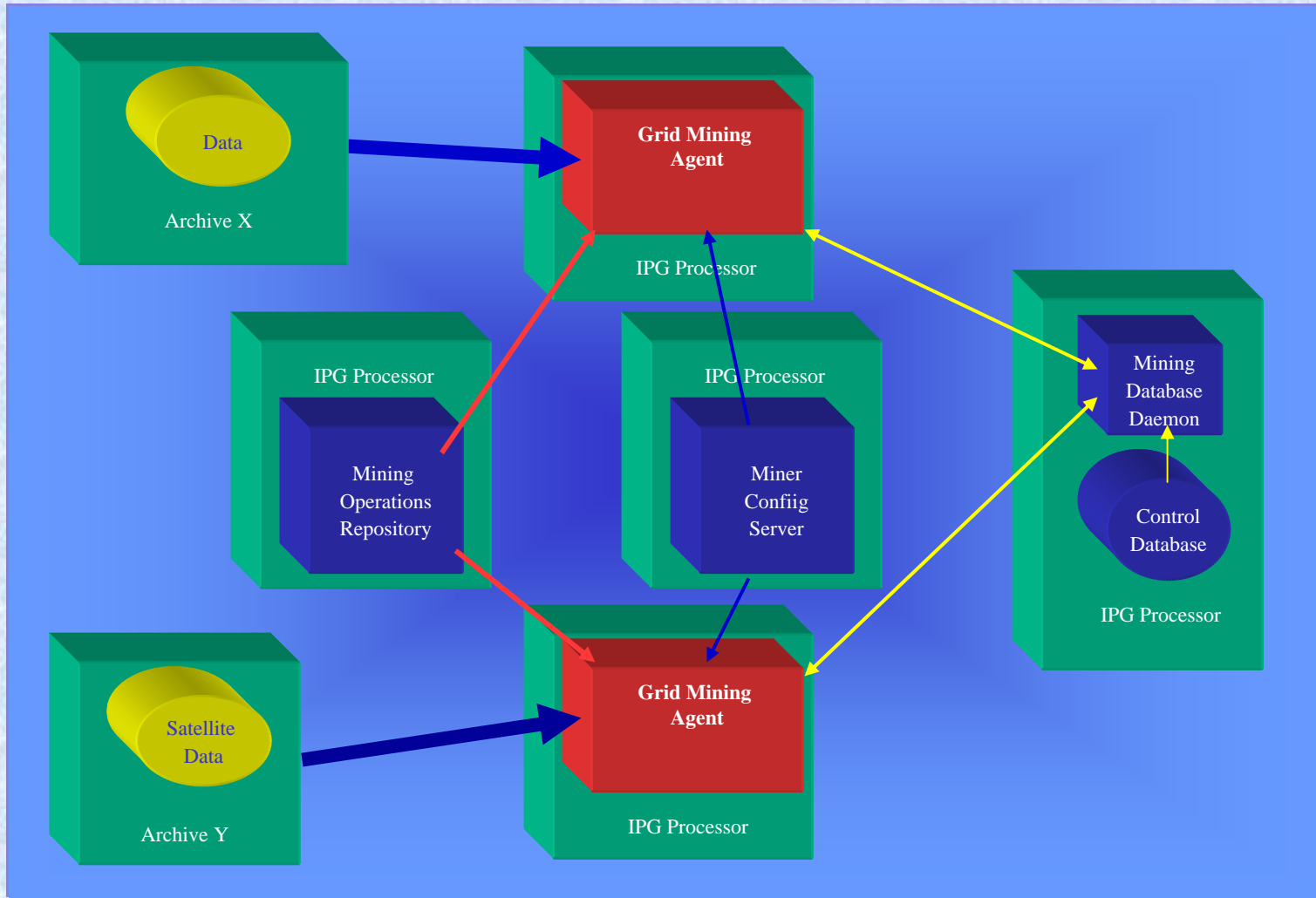


Figure thanks to Information and Technology Laboratory at the University of Alabama in Huntsville

Mining on the Grid



Grid Miner Architecture



Example: Mining for Mesoscale Convective Systems

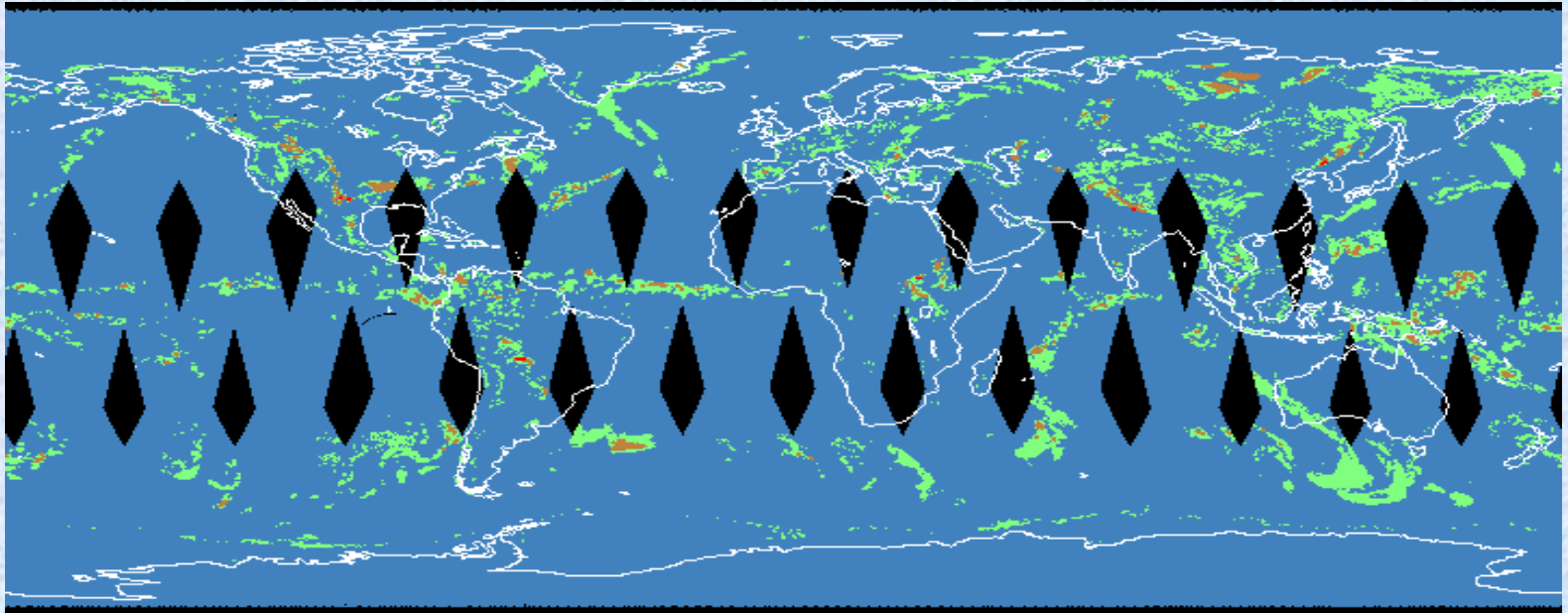


Image shows results from mining SSM/I data



Outline

- Why use the grid for data mining?
- Overview of Grid Miner
- Experience adapting existing stand-alone miner to grid
- A recent application of the Grid Miner



Starting Point for Grid Miner

- Grid Miner reused code from object-oriented ADaM data mining system
 - Developed under NASA grant at the University of Alabama in Huntsville, USA
 - Implemented in C++ as stand-alone, objected-oriented mining system
 - Runs on NT, IRIX, Linux
 - Has been used to support research personnel at the Global Hydrology and Climate Center and a few other sites.
- Object-oriented nature of ADaM provided excellent base for enhancements to transform ADaM into Grid Miner



Transforming Stand-Alone Data Miner into Grid Miner

- Original stand-alone miner had 459 C++ classes.
- Had to make small modifications to ADaM
 - Modified 5 existing classes
 - Added 3 new classes
- Grid commands added for
 - Staging miner agent to remote sites
 - Moving data to mining processor



Staging Data Mining Agent to Remote Processor

```
globusrun -w -r target_processor  
'&(executable=$(GLOBUSRUN_GASS_U  
RL)# path_to_agent)(arguments=arg1 arg2  
... argN)(minMemory=500)'
```



Moving Data to be Mined

```
gsincftpget remote_processor local_directory  
remote_file
```



Outline

- Why use the grid for data mining?
- Overview of Grid Miner
- Experience adapting existing stand-alone miner to grid
- A recent application of the Grid Miner



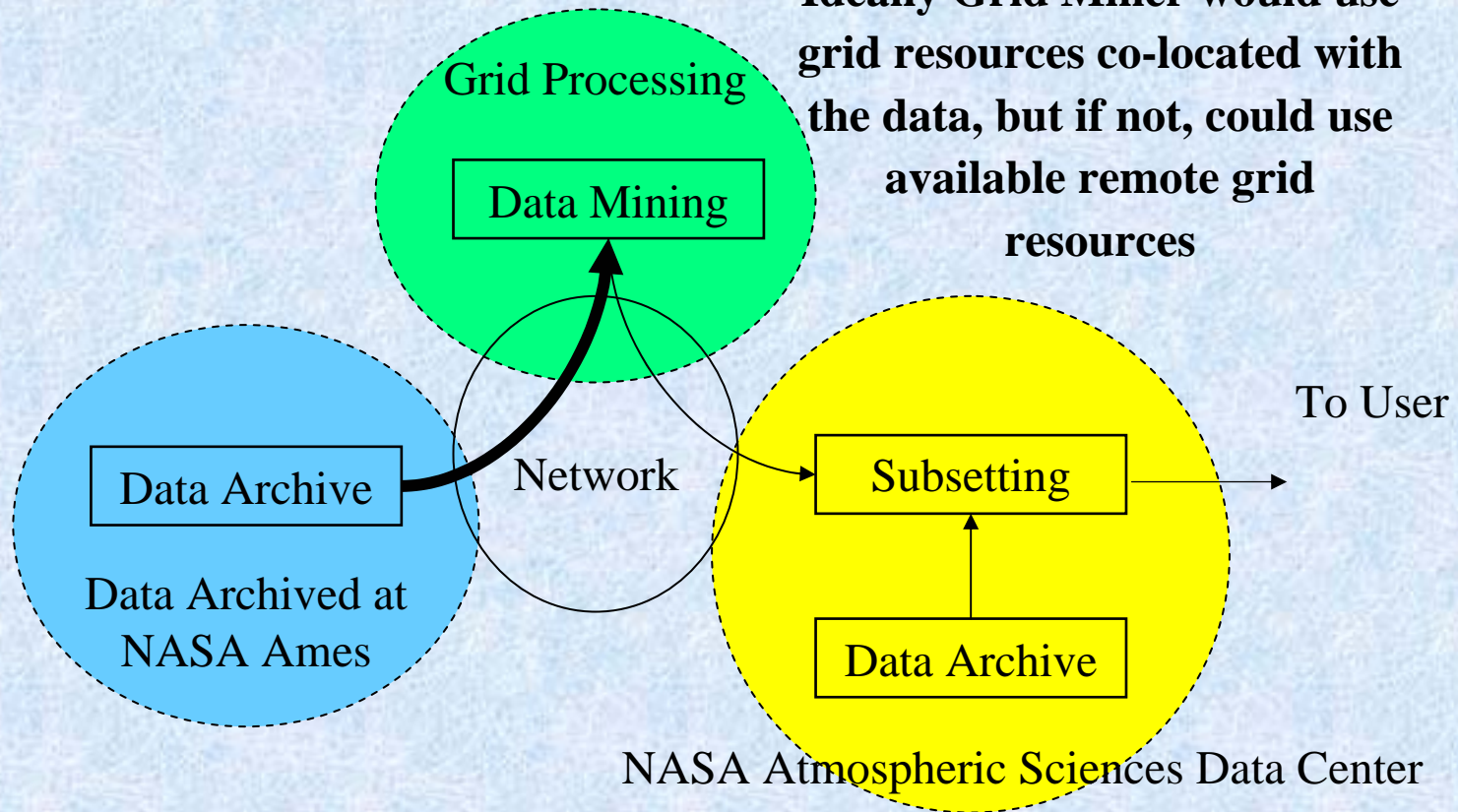
Demonstrate Grid Support for Interdisciplinary Earth Science Research

- Goal: Combine data from two distinctly different instruments (stored on two different grid-connected mass storage systems) to produce new insights by looking at data covering the same time and place across data from the two different instruments.
 - Approach:
 - Use Grid Miner to mine TMI data for mesoscale convective systems.
 - Generate feature index (convex hull polygon) for all mesoscale convective systems found.
 - Transmit polygons in form of XML document to subsetter.
 - Subset CERES SSF data that corresponds to mesoscale convective systems discovered by Grid Miner
-

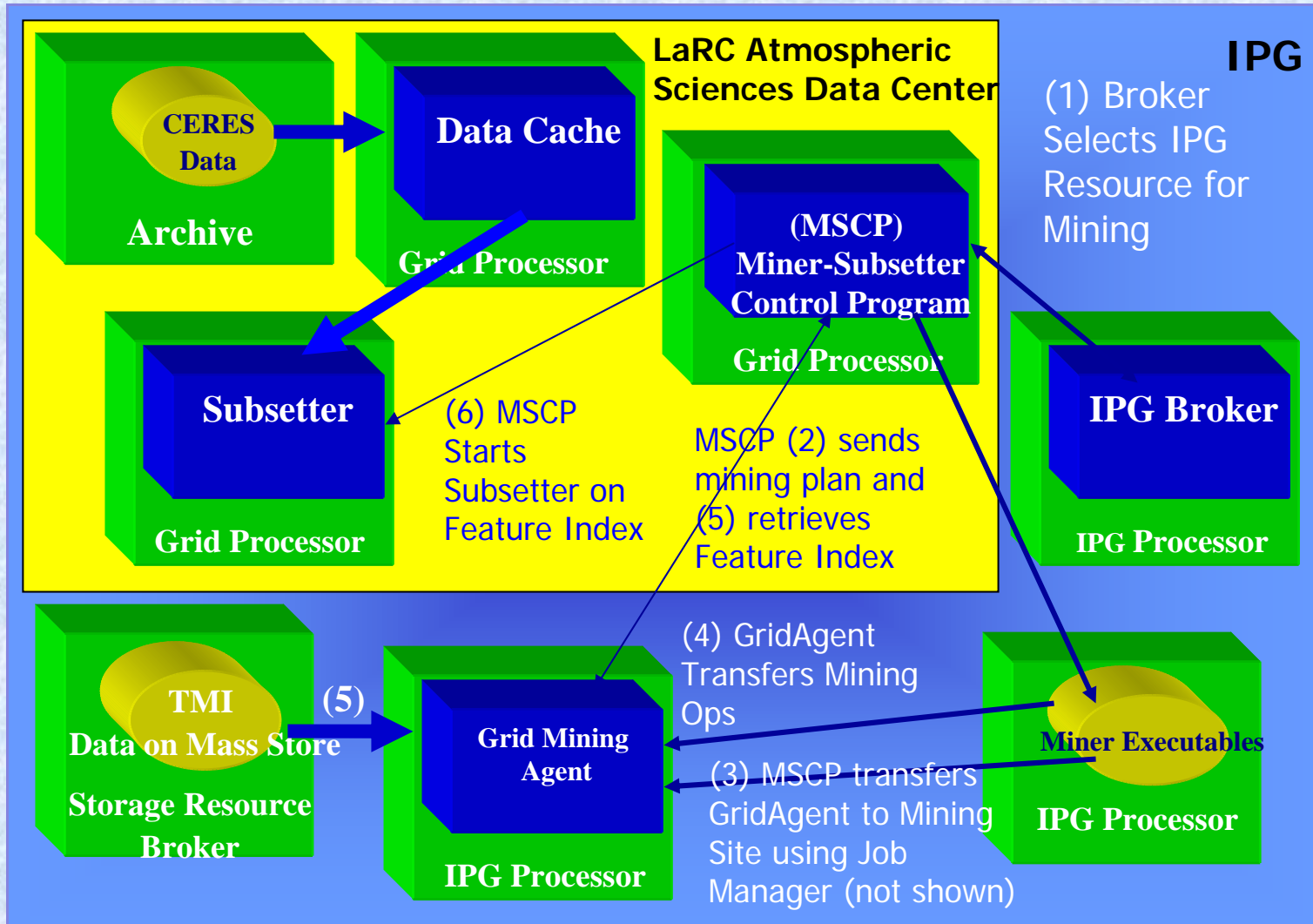


Desired Processing Pattern

Ideally Grid Miner would use grid resources co-located with the data, but if not, could use available remote grid resources



The Details



Example of Data Being Mined

- 230 MB contained in 15 orbit files for one day of TMI (TRMM [Tropical Rainfall Measuring Mission] Microwave Imager) data
- Much higher resolution data exists with significantly higher volume.



Mining and Subsetting Results

Grid Miner produced XML (Extensible Markup Language) document of polygons that circumscribe mesoscale convective systems. (MCSs) The following shows a portion of the XML description for two of the 64 vertices that comprise the convex hull polygon produced for the third MCS found by the miner in TMI data for April 1, 1998:

```
<polygon>
<julian_date_time> 2450904.754815 </julian_date_time>
<human_date_time> 1998-04-01 GMT 06:06:56 </human_date_time>
<size_in_square_km> 2083.126221 </size_in_square_km>
<region_type> 2 </region_type>
<vertices>
<number_of_vertices> 64 </number_of_vertices>
<vertex>
<latitude> -2.26 </latitude>
<longitude> -178.28 </longitude>
</vertex>
<vertex>
<latitude> -2.08 </latitude>
<longitude> -178.38 </longitude>
</vertex>
.
.
.
</polygon>
</polygon_list>
```

Grid Miner produced view of area mined using TMI data.



CERES SSF footprints for April 1, 1998 hour 6 corresponding to the third MCS found by Grid Miner

Convex Hull: 3 with 9		
Footprint:	1	15804
Footprint:	2	15805
Footprint:	3	16090
Footprint:	4	16091
Footprint:	5	16094
Footprint:	6	16376
Footprint:	7	16377
Footprint:	8	16381
Footprint:	9	16382

Current Status

- Currently works on the IPG as a prototype system
- User documentation underway
- Data archives need to be grid-enabled
 - Connected to the grid
 - Provide controlled access to data on tertiary storage
 - E.g., by using a system such as the Storage Resource Broker that was developed at the San Diego Super Computer Center
- Some earlier-adopter users need to be found to begin using the Grid Miner
 - Willing to code any new operations needed for their applications
 - Willing to work with system with prototype-level documentation



