Integrating and Delivering Massive Geospatial Datasets for Educational Use

Michael Haungs, Ph.D.
Rollin Strohman, Ph.D.
Our Mission

- Build a service to provide massive, remote geospatial datasets for educational use
“Satellite data is expensive, and using the data requires significant investments in software, hardware, and training. It has often been hard for university researchers to use or even access the data, particularly at smaller schools or research facilities. For three decades this has hindered applied research and made it difficult to train the workforce, both current and future.”
Related Work

- **TerraServer**
  - Microsoft/US Geological Survey (USGS)

- **Soil Data Mart**
  - Natural Resources Conservation Service
    - US Department of Agriculture

- **Sloan Digital Sky Survey**
  - NASA, Los Alamos National Laboratory, NSF, University of Chicago, Johns Hopkins, Japan Participation Group, many more…
Related Work

- The National Map (USGS)
- SkyServer (Sloan Digital Sky Survey)
- BioMOBY (NSF)
- NOTEBOOK Project (San Diego Supercomputing Center)
- GEON (SDSC, Penn State U, UC San Diego, Geological Survey of Canada)
Pedagogical use differs from scientific

- **Interactivity**
  - Limited lab time
  - Require quick, fluid responses
  - Treasure-hunt approach

- **Multiplicity**
  - Groups performing similar activities

- **Connectivity**
  - High speed networks
MapSurfer

- http://casil.ucdavis.edu/mapsurfer
Something like the following...
Movie showing dataset performance

- Receiving computer
  - IBM Thinkpad T23
  - 1 Ghz CPU, 768 MB RAM
  - 100 Base-T network connection
  - Running PhotoMapper and Camtasia Studio
  - Actual performance better than shown
Uses

- California missions (Virtual Field Trips)
- Monarch butterfly migration paths
- National wildlife reserves
- City layout and design
- Geographic Information Systems (GIS)
- Seems endless…
Benefits

- Gets students excited about learning:
  - Geography
  - History
  - Social Sciences
  - Architecture
  - City Planning
  - Forestry

- Exposure to GIS technology
Issues

- Which dataset should we use?
  - Proprietary? Build and maintain our own?
- How should we build the application?
  - Optimized for educational use
  - Thin client, Thick client, pure data export?
- How should we incorporate such a service into K-20 curriculum?
- What is the method of delivery?
Which dataset should we use?

- One meter resolution
- Proprietary Dataset
  - AirphotoUSA
    - Seamless of entire USA
- Exploring building own seamless orthophoto of California
  - CaliforniaView (http://gis.ca.gov/casil/usgs.gov/)
    - AmericaView (http://www.americaview.org/)
Why 1 meter resolution?

- Intended users have limited photo interpretation experience
- First task is to find your house
  - then investigate your neighborhood
  - learn to interpret top view in process
- Curiosity key to learning process
Using a proprietary dataset

- Airphoto USA
  - Seamless, one-meter resolution aerial orthophotography covering over 3 million square miles of the US
  - 20TB (2.5 TB compressed)
  - Generously made available by AirPhotoUSA for our study
  - Viewing software – PhotoMapper
    - Rapid pan and zoom for best user experience
Using a proprietary dataset
Using a proprietary dataset

- Issues:
  - Security
    - Licenses
    - Protecting the dataset
      - Shibboleth (http://shibboleth.internet2.edu)
  - Limited Customization
    - Do not have access to code
  - Cost
Using a proprietary dataset

- Advantages:
  - We have this now
  - It’s very good quality
  - Good compression/performance
  - Should be able to scale to large numbers of users
Building our own dataset

- Acquire Digital Orthophoto Quarter Quads (DOQQ) from free source
- Mosaicking (Leica Geosystems, ER Mapper)
- Compress to JPEG2000 (ERDAS)
  - Necessary over fast networks?
Building our own dataset

- **Issues**
  - Computational and Storage resources
  - Maintaining updates

- **Advantages**
  - Free use
How should we build the application?

- How should we build this application to best fit the needs of K-20 educators?
- Requirements
  - Scalable
  - Flexible
  - Available
  - Interactive
  - Intuitive
How should we build the application?

- Options
  - Web Service
  - Web Application
  - Distributed Application
Web Service

- Request tiles via a standard web services API
- Advantages
  - Flexible
    - Web page
    - GIS Applications
    - Custom Viewers
      - With or without layer support
Web Service

- Disadvantages
  - No client-side support
  - Duplicated effort
  - Requires high level of expertise from client
Web Application

- **Client-side**
  - Only requires a web browser (thin client)
    - Java plug-in
- **Server-side**
  - No need to conform to communication standard
Web Application

- **Advantages**
  - No client-side administration/maintenance required
  - Works well with proprietary data
  - Full control of application
    - Updates automatic

- **Disadvantages**
  - Extra “connection-time” load
  - Simple client-side application
Distributed Application

- **Client-side**
  - Provide full featured client to user (Thick Client)
  - Maintained/supported by us
    - Collaborate with a vendor?

- **Server-side**
  - General vector data repository
Distributed Application

- **Advantages**
  - Minimal user expertise
  - End-to-end optimizations
    - Load balancing
    - Caching
  - Quality Control

- **Disadvantages**
  - Resources
  - Cost
    - License software (3rd party vendor)
  - Inflexible
Current Work: Application

- Data Warehouse Approach, optimized for:
  - Proprietary dataset
  - Interactive, explorative
  - Internet2

- Calpoly Farmland GIS project
  - 6000 acres
  - 4000 students
  - ESRI products
K-20 Curriculum

- K-20 Curriculum
  - What do teachers want?

- Educator Collaboration
  - Curriculum development
  - Collaboration tools?

- Advertisement

- Training

- Delivery
K-20 Curriculum

- Pedagogical use differs from scientific
  - Interactivity
    - Limited lab time
    - Require quick, fluid responses
    - Treasure-hunt approach
  - Multiplicity
    - Groups performing similar activities
  - Connectivity
Internet2

- Project Focus
  - Internet2 Universities
  - Future Kindergarten through 12th grade
    - Network speed will continue to improve
Internet2 (Abilene Network)
Internet2

- CalPoly, San Luis Obispo, CA
  - 1 Gb/s backbone
  - 622 Mb/s pipe to ISP in L.A.
  - ISP connects to Abilene I2 network at 2.2 Gb/s
  - Future
    - Can grow up to 10Gb/s (Abilene Limit)
Internet2

- Changes basic assumption of client/server applications
  - Fast machines and slow communication
- Do old optimizations still make sense?
  - Caching? Compression? Remote desktops?
- Other optimizations?
  - Prefetching? Indirection?
Questions?

- Next: Some results of a study conducted for AirphotoUSA
Initial Study

- Understand existing bottlenecks of hardware and proprietary software for serving AirphotoUSA’s geospatial dataset.
Bottleneck Analysis
Bottleneck Analysis

- Measuring
  - CPU Utilization
  - Memory (Paged and non-paged)
  - I/O (Disk bytes and queue length)
  - Network Bandwidth (Bytes/sec)
Performance Metrics

- User-response time
  - What delays are noticeable to the user?
Experiment Setup

- Server setup duality:

```
  Internet2
  ↓
  Server
  ↓
  Server
  ↓
  Server
  ↓
  Server
  ↓
  Internet2
```

Storage Server

DATA
Experiment Setup

- **Machines**
  - Dell GX270
    - Intel Pentium 4 2.8 GHz CPU
    - 512KB L1 cache, 1MB L2 cache
    - 533MHz front-side bus
  - 512MB RAM
  - Gigabit Ethernet Card
  - 75GB ATA/100 IDE Hard drive
  - Windows XP SP1
Experiment Setup

- Maxtor 200GB External Hard Drive
  - 7200 RPM
  - 8MB Cache
  - USB 2.0 (up to 480Mbits/sec)
Experiment Setup

- Requests/Client
  - 40 random points in California
    - Maximum resolution
    - Image size: 1024x768, ~1/3 MB
  - Approximates 4 users/client
Response times (Server)
second response time (average) response time (max)
Results:: Server CPU
The graph shows the percentage of processor time (%) for different numbers of clients. The x-axis represents the number of clients, ranging from 10 to 100, while the y-axis represents the percentage of processor time, ranging from 0 to 120. There are two lines: the blue line represents the average processor time, and the red line represents the maximum processor time. As the number of clients increases, the percentage of processor time also increases, with the maximum processor time reaching its peak at 100 clients.
Server Network bytes
Server Packets
Server page faults
Server pool non-paged failures
Storage Server CPU
Storage Server Disk Bytes
Storage Server Disk Queue
Storage Server Network Bytes
Storage Server Packets
Storage Server Pool Non-paged Failure
Bottleneck Analysis Summary

- CPU not an issue
- Network and Disk load
  - Sufficient for 200 simultaneous, highly interactive users
- AirphotoUSA proprietary DLL for accessing data
  - BOTTLENECK for web services scenario
Recommendations

- AirphotoUSA
  - License server bottleneck
    - Possible race condition
    - Central and serial access
  - Port to Linux (or use Windows 2003 server)
    - Windows XP problems
  - Optimize DLL for web service use
    - Smaller working set, memory footprint
    - Asynchronous operation
      - Improve data retrieval
Recommendations

- **Client/Server organization**
  - (Storage) Server needs sufficient RAM (> 512MB)
  - Server-level OS
    - Network optimization, multiple processes
  - Standard 2-tier web service model works
  - Direct attached storage works
  - Currently investigating new ideas
    - Client-side centralized caching to capitalize locally on multiplicity
Recommendations

- Serving large, online datasets for educational use over Internet2
  - Storage Wide Area Network (SWAN)
    - Killer Application for I2?
Wrap up

- We have access to a seamless, orthophoto of entire US
- Conducted an initial feasibility study
  - Bottleneck analysis
  - Testing in CalPoly labs
- Have access to multiple data sources
  - Generate own
  - Pursue proprietary model
Wrap up

- End application
  - Pursuing multiple avenues
  - Investigating novel approaches
  - Need input from community

- Educator Collaboration
  - Pedagogical best practices
  - K-12 curriculum development
Future Work

- Prototype Implementation with set of I2 Universities
  - CalPoly, Stephen F. Austin, South Dakota, Oklahoma State
- Virtual Teaming
  - Educator Collaboration
- National Dissemination
  - All Internet2 Universities
  - K/20
  - Group data purchase
- Serving Other Datasets from Calpoly
- New Client/Server Architectures
  - Build them
Conclusion

- If you are interested in any aspect of this project, please contact us at:
  - Michael Haungs, mhaungs@calpoly.edu
  - Rollin Strohman, rstrohman@calpoly.edu

- Questions?
Client/Server Organization

- Direct access
  - NFS, Samba, TCP/IP session
  - LAN model
- Web services
  - HTTP, SOAP, XML, UDDI, WSDL, OpenGIS
- End-to-end Internet application
Approach 1

- Download data
  - Too much
  - Maintenance
    - Complicated System
  - Updates
    - Keeping data “fresh”
  - Duplicated Effort
Approach 2

- Standardize on an access protocol
  - SOAP, HTML, XML?
  - Samba, NFS?
  - Web Service API
    - OpenGIS?
  - Incentive to collaborate?
    - Open standards not optimized for specific use
    - Cost of handling access load
  - Availability
  - Load balancing
Approach 3

- Data Warehouse
  - Centralized Solution
    - Storage
    - Maintenance
    - Data Acquisition
  - Scalability
    - Sufficient Bandwidth and processing power
    - Who’s going to do this?
Speculative Client/Server Architecture

- How do these results translate into client/server architecture?
- Give examples