RAPID DELIVERY OF MASSIVE GEOSPATIAL DATA OVER INTERNET2

Rollin Strohman, Professor Emeritus
Michael Haungs, Assistant Professor
California Polytechnic State University
San Luis Obispo, CA 93407
rstrohma@calpoly.edu
mhaungs@calpoly.edu

ABSTRACT

We study the feasibility of the on-demand delivery of a massive geospatial dataset over Internet2 for educational use. The dataset (20TB uncompressed, 2.5TB compressed), generously made available for this study by AirphotoUSA, provides a seamless, one-meter resolution aerial orthophotograph covering over three million square miles of the continental United States. We identify factors that limit the scalability, availability and user-perceived performance of serving such a dataset. To do this, we conduct experiments that measure response times for various levels of network congestion, bandwidth, and load. We also provide a proof-of-concept experiment by serving the dataset over Internet2 to students at Oklahoma State University. Given this information, we determine the server-side architecture and resource requirements sufficient to serve this dataset from Cal Poly. We discuss the funding for wide distribution of high-resolution datasets to universities and the student response to use of this data for education.

INTRODUCTION

We are studying the feasibility of the on-demand delivery of a massive geospatial dataset over Internet2 for educational use. The dataset (20TB uncompressed, 2.5TB compressed), generously made available for this study by AirphotoUSA*, provides a seamless, one-meter resolution black and white orthophotograph covering over three million square miles of the continental United States. Without further funding, access to this data is limited to Cal Poly and Oklahoma State University. We hope to find funding to make the data widely available to educational institutions. We are identifying factors that limit the scalability, availability and user-perceived performance of serving such a dataset from a central location. To do this, we are conducting experiments that measure response times for various levels of network congestion, bandwidth, and load. We also provide a proof-of-concept experiment by serving the dataset over Internet2 to students at Oklahoma State University. Given this information, we determine the server-side architecture and resource requirements sufficient to serve this dataset from Cal Poly. To protect AirphotoUSA's commercial interests, the server architecture is constrained to be centralized and not rely on client-side caching.

Our findings have ramifications for future distributed applications that involve the processing and movement of very large datasets. Often, these types of services are constructed to minimize the bandwidth requirements between clients and servers. We believe that Internet2-level bandwidth allows us to simplify these architectures and reap the benefits of centralized administration and maintenance. With regards to AirphotoUSA, this has explicit relevance to customers considering purchasing their product and presents a new, subscription based business model to explore. There is also an indirect benefit to AirphotoUSA in that the next generation of GIS power users will have been exposed to their company's data and software. Last, it provides a cost effective solution for acquiring massive datasets for use in multiple departments across universities who often can't afford to replicate and maintain them individually.

---

* http://www.airphotousa.com/

HOW CAN THE PURCHASE OF LARGE COMMERCIAL DATASETS FOR EDUCATIONAL USE BE FUNDED?

To make the dataset used in our tests available to a larger number of universities, we need to adequately reimburse AirPhotoUSA for use of their development efforts. In the past years, ownership of data has shifted from the federal government to commercial companies. On July 23, 1972, Landsat 1 was launched and made it possible for educators to access imagery of large sections of the earth for education purposes. Programs such as AmericaView* now have made this imagery available for education at low or no cost because the data was federal owned. Once it had been purchased, it could be distributed to everyone with no additional reimbursement to the federal government. The Internet has made it easy to find and distribute this data. Unfortunately, the life of Landsat appears to be coming to an end.

Since early 1990's the government policy was changed to allow private companies to launch higher resolution satellites and market the data to the general public (Brown, 2004). Three companies, Space Imaging, DigitalGlobe, and ORBIMAGE offer black and white resolutions of less than one meter. ORBIMAGE and DigitalGlobe have each received 500 Million dollar NextView awards from the National Geospatial-Intelligence Agency (NGA) to insure the next generation of commercial satellites**. Other private companies such as AirPhotoUSA are using aircraft to capture images and produce orthophotos covering wide areas. For these companies to keep the cost down for each purchaser, they must be able to sell the same dataset to many customers and there must be a way to prevent persons who have not purchased the data from using it.

Commercial companies are repackaging federally owned data and selling value added services. The data for this test was produced by USGS, but this data has been commercially repackaged as a compressed seamless dataset covering forty-eight states and sells for $99,000 per copy. The PhotoMapper software they have developed allows one to easily pan and zoom to any location in the continental US. This ease of use capability adds to the usefulness of the data. To obtain this or other datasets for educational use by a large number of universities at an affordable price the data must be available only for educational use and not be made available in a way that will impact sales to commercial users (real estate, insurance, cities and counties) of the data. AirPhotoUSA provides control by a license server which controls which computers have access to the data, but there is no control as to which person has access to use the data. We are investigating the work of an Internet2 Shibboleth Working Group*** as a means of restricting use to only educational users. At the moment the emphasis seems to be on national defense and developing the commercial market, with little support for funding of national datasets at the one-meter resolution range.

WHAT IS INTERNET2?

"Internet2 is a consortium being led by 207 universities working in partnership with industry and government to develop and deploy advanced network applications and technologies, accelerating the creation of tomorrow's Internet. Internet2 is recreating the partnership among academia, industry and government that fostered today's Internet in its infancy. The primary goals of Internet2 are to:

- Create a leading edge network capability for the national research community
- Enable revolutionary Internet applications
- Ensure the rapid transfer of new network services and applications to the broader Internet community."****

The major pathway for Internet2 is the Abilene network that operates at 10Gbps*****. Top cable and DSL speeds are 3Mbs or 3000 times slower than Internet2. Whatever we determine about the benefits of such Internet2 speed today, it should be kept in mind that a few years ago only the standard dial up speed of 56Kbs was widely available. Thus in a short time there has been a 50 times improvement in speed. This speed improvement will continue.

---

* http://www.americaview.org/
** http://www.nga.mil/NGASiteContent/ StaticFiles/OCR/NextView20040930.pdf
*** http://shibboleth.internet2.edu/docs/internet2-mace-shibboleth-introduction-200404.pdf
**** http://www.internet2.edu/about/
***** http://abilene.internet2.edu/

Internet2 changes one of the fundamental assumptions made when developing distributed applications. This assumption is that the machines are fast and the network is slow. The reverse is now true and begs the question, "Do the old optimizations make sense?" We plan to explore that question. This educational environment lends itself to unique optimization opportunities, such as prefetching, while adding the additional constraints of interactivity and uniformity.

EDUCATIONAL MERIT

Massive scientific datasets would be of great educational benefit to many K-20 educators. Educators can use these datasets to augment student learning and expand the understanding of all regions of the continental US. For example, imagine forestry, agriculture, or city planning students using geographic datasets of the United States to view any national forest, farmland, or city instantaneously from the classroom and compare and contrast the different regions. The requirements and environment of pedagogical use of these datasets is quite different from that of scientific exploration. Students often have limited lab time and want quick, fluid responses. Students are grouped together in labs, exhibit similar access patterns, and are connected via high-speed network connections, such as Internet2.

There are three main differences between the average use of massive, online datasets and educational use:

1. Interactivity
   Students will not have a preconceived notion of the data they are looking for nor what they expect to find. Commonly, users of online datasets have a specific use in mind, e.g. looking at specific data to test a hypothesis or merely to get an aerial image to add as a backdrop to a map. Whereas students are more apt to take a treasure-hunt approach. This type of approach necessitates the interactive panning and zooming of the aggregated dataset.
2. Multiplicity
   Groups of students will be working simultaneously on a common assignment. While not working in lockstep, there will be significant overlap in their queries.
3. Connectivity
   Universities typically have a network connection speed 100 times faster than your average DSL connection.

PRELIMINARY RESULTS

Four Dell GX270 Intel Pentium 4 2.8 GHz blades with 512KB L1 cache, 1MB L2 cache, 533MHz front-side bus, 512 MB RAM, Gigabit Ethernet Card 75GB ATA/100 IDE Hard drive, and Windows XP SP1 were used. The California dataset was stored on a 200 GB Maxtor USB 2.0 (up to 480 Mbits/sec) drive. Software was written to simulate multiple users moving between 40 points in California (1024 by 768 pixel screens). The average response time with 100 simulated users was approximately 2 seconds. The results indicate the CPU speed is not an issue and 200 simultaneous highly interactive users can be supported. AirPhotoUSA's DLL may be a bottleneck in it's present implementation. RAM may be a limitation.

ACKNOWLEDGEMENTS

We acknowledge the support of the following persons in the development of this project:
Brent Cannon, Master's Candidate, Cal Poly
Tom Mastin, Instructor, Cal Poly
California Polytechnic State University (Cal Poly), San Luis Obispo, CA 93407
Paul Weckler, Assistant Professor, Oklahoma State University, Stillwater, Oklahoma

REFERENCES

Brown, Tim and John Pike. Watching the World's Hot Spots, Earth Imaging Journal Sept/October 2004 Vol. 1 No. 5 p12-19.