Processing Historic Wisconsin Aerial Photography with High Throughput Computing Resources

Throughput Computing 2024

Jim Lacy
State Cartographer's Office ~ Department of Geography ~ University of Wisconsin-Madison

We Have a Data Problem

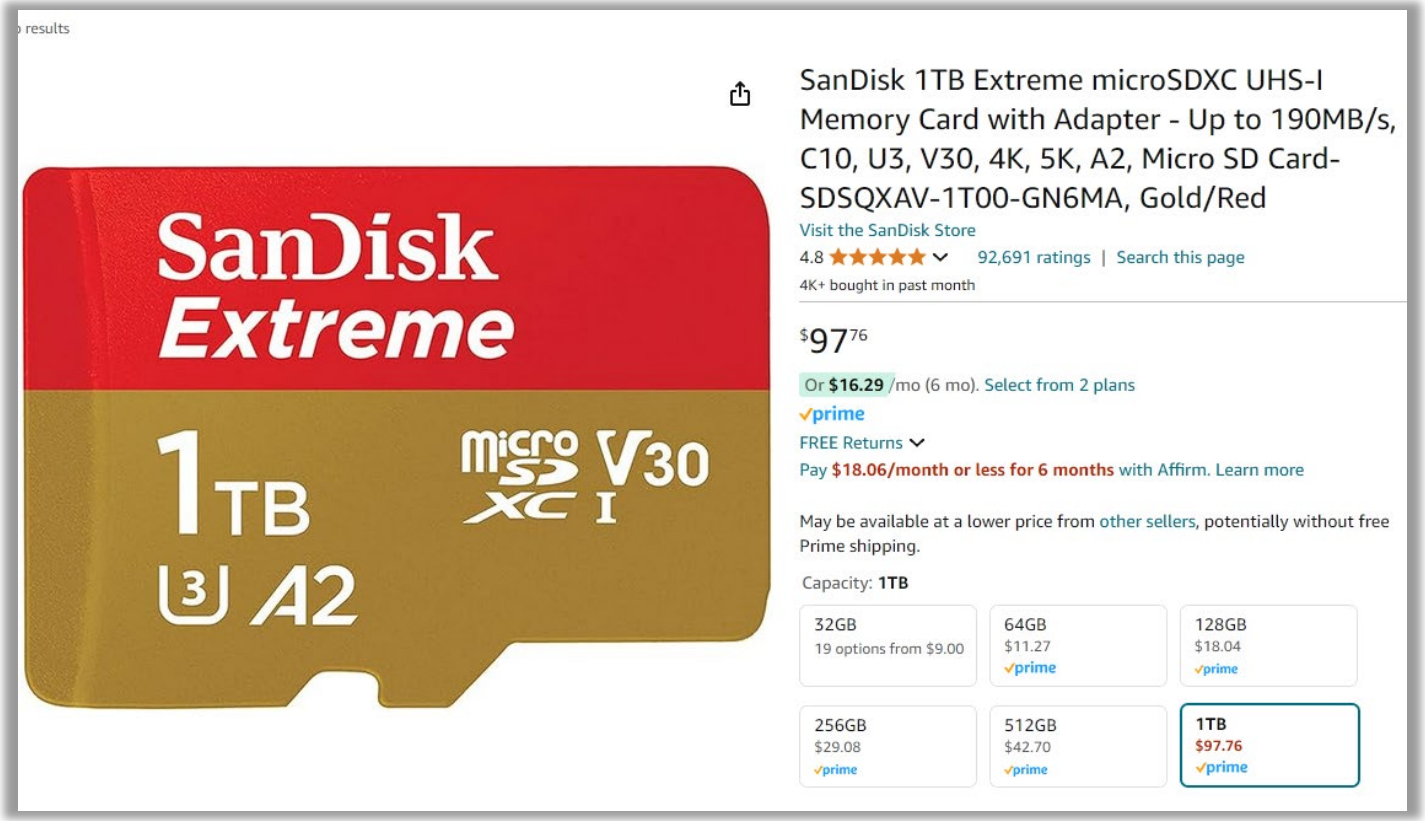
- ~100,000 images to convert to Cloud Optimized GeoTIFF ("COG") format for Web apps and public distribution
- Roughly 15TB
- This **could easily take an estimated 1.0 to 1.5+ months** of processing time running on a local PC using our typical methods
- Is there a better way???

Some Context


9-11-62
WU -3CC-261


USGS
NASA Earth Resources Aircraft Program
USGS
Storing the Nation's Data Management
Aerial Photography

Falling Storage Costs Makes This Work Possible!



Hmmm… What Are My Typical “Serial” Options?

- Batch processing in GIS software
- Python script(s)
- GDAL command line inside batch script(s)

Case Study: Dane County, WI

- 88 input files
  - 41.2 GB
  - 4-band GeoTIFFs with 0.60m spatial resolution
- Convert to COG and generate jpegs on local PC = 93 minutes
  - Total output data = 59.4 GB
  - Honestly not bad for a single county!
- But… it would take an estimated 79 hours to process 4,483 files for the statewide dataset

Enter: UW Center for High Throughput Computing

- Multi-disciplinary center with home in UW Computer Science
- Somewhere around 20,000 compute cores available
- Free (to UW fac/staff/students… big thank you to VCRGE/CS/WARF!)
- Lots of help docs, workshop, dedicated staff available to assist
- Copious amounts of (temporary) disk space if you act responsibly and clean up after yourself

Let’s Try This in Parallel

- Local PC Model = Sequentially loop through 4,483 files… process, then repeat

- CHTC Model = process 4,483 files in PARALLEL… all input images are processed at the same time(ish)*

* The number of jobs you can run at a given moment depends upon other users on system and what you request for resources (CPU, disk, memory.) During tests, I observed up to 500 of my files processing at the same time.

Case Study Revisited: Dane County, WI

- Local PC = 93 minutes
- **The CHTC Way = 5 minutes!**

Let's Scale Up to Wisconsin

- Data transfer to CHTC = 9.25 hours*
- Data processing = 1.5 hours
- Data transfer back to Science Hall = 12 hours*
- CHTC Total (compute + transfer overhead) = **22.75 hours**

- **Net savings of ~56.25 hours** vs local PC processing for this example

* Before I discovered Globus!

But Wait, There's More!

Challenges and Barriers to Entry

Challenge #1: Data Transfers

- How long does it take to get the input files TO the CHTC, and output files BACK to our server?
- Attempt #1 (estimate) of statewide data using the standard "scp" transfer method:

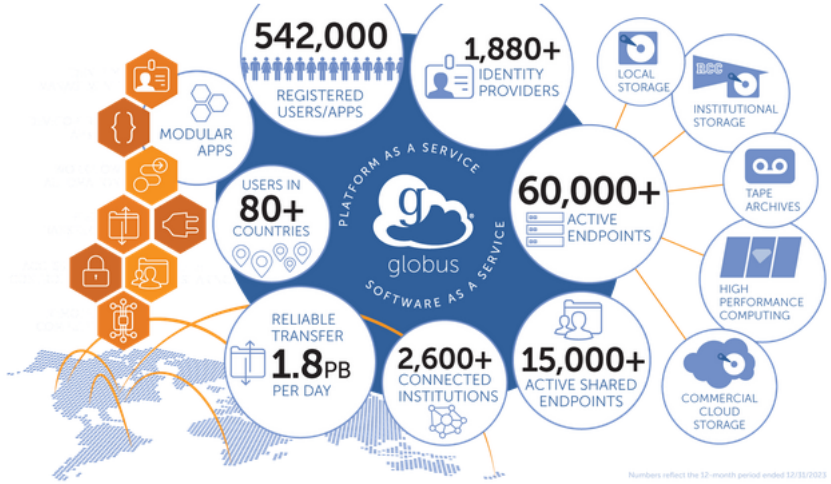
3 days there + 4 days back = was this a waste of time?

(Remember: local method was about 3.3 days of processing)

Globus is Your Friend!

What We Do

Globus is research cyberinfrastructure, developed and operated as a not-for-profit service by the University of Chicago.



With Globus, you can easily, reliably and securely **move, share, & discover data** no matter where it lives – from a supercomputer, lab cluster, tape archive, public cloud or laptop. Access and manage all your data, even protected data, from anywhere, using your existing identities, with just a web browser.

38635	Files
289	Directories
38635	Files Transferred
1.25 TB	Bytes Transferred
538.37 MB/s	Effective Speed
0	Skipped files on sync
0	Skipped files on error

Challenge #2: Command Line Can Be Scary

lacy@ap2002:~

```
_____________________________________________________________________
 #####  #     # #######  #####  Issues?  Email chtc@cs.wisc.edu
#     # #     #    #    #     # Unauthorized use prohibited by:
#       #     #    #    #       WI Statutes: s. 947.0125
#       #######    #    #       U.S. Code: 18 USC 1030
#       #     #    #    #       U.S. Code: 18 USC 2510-2522
#     # #     #    #    #     # U.S. Code: 18 USC 2701-2712
 #####  #     #    #     #####  U.S. Code: 18 USC § 1831
For off campus ssh access use https://www.doit.wisc.edu/network/vpn/
_____________________________________________________________________

   Virtual office hours are available once a week over the summer:
            Thursdays, 3:00 - 4:30pm (Central time)
        Join via this link: go.wisc.edu/chtc-officehours
     Sign in via this link: go.wisc.edu/chtc-officehours-signin
Filesystem quota report
Storage                  Used (GB)    Limit (GB)    Files (#)    File Cap (#)    Quota (%)
--------------------  -----------  ------------  -----------  --------------  -----------
/home/lacy                        3            50       114050               0            6
/projects/SCO_Imagery       15288.9         25000       148967          500000        61.16

(base) [lacy@ap2002 ~]$
```

Challenge #3: Impatience

- You need to invest time to save time
- Requires knowledge of Unix, shell scripts, command line principles… all things I already knew
- Required about 16 hours of learning/experimenting/reaching out for help
- Start small, test, scale up from there

Mitigating the Challenges…

- ✓ Smart people dedicated to helping users like me
- Gentle conceptual introductions to command line and batch processing in general
- “Cookbooks” with lots of examples
- Work to create connections in the user community … help them learn from each other

Kudos… it takes a team!

- UW-Madison Center for High Throughput Computing (CHTC)
- CHTC Facilitators (Christina K., Rachel L., Andrew O.)
- UW-Madison Geography Project Team (Jaime M., Hayden E.)
- UW-Madison Division of Information Technology Storage Team (Kevin K., Phillip D., Mark K.)
- UW-Madison Research Data Services (Michael L.)
- Morgridge Institute (Justin H., Brian B.)

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON