Processing Historic Wisconsin Aerial Photography with High Throughput Computing Resources

Throughput Computing 2024
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
Falling Storage Costs Makes This Work Possible!
Hmm… 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 Effective Speed
0      Skipped files on sync
0      Skipped files on error
Challenge #2: Command Line Can Be Scary

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

<table>
<thead>
<tr>
<th>Storage</th>
<th>Used (GB)</th>
<th>Limit (GB)</th>
<th>Files (#)</th>
<th>File Cap (#)</th>
<th>Quota (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>/home/lacy</td>
<td>3</td>
<td>50</td>
<td>114050</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>/projects/SC0_Imagery</td>
<td>15288.9</td>
<td>25000</td>
<td>148967</td>
<td>500000</td>
<td>61.16</td>
</tr>
</tbody>
</table>

(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.)