UC Berkeley
UC Berkeley Previously Published Works

Title

Community Access to MODIS Satellite Reprojection and Reduction Pipeline and Data Sets

Permalink

https://escholarship.org/uc/item/1rk2j48v

Authors

Hendrix, V

Li, J

Jackson, K

et al.

Publication Date

2012-04-12

Peer reviewed

Community Access to MODIS Satellite Reprojection and Reduction Pipeline and Data Sets

Val Hendrix[1], Jie Li[2], Keith Jackson[1], Lavanya Ramakrishnan[1], Youngryel Ryu[3], Keith Beattie[1], Christine Morin[4], David Skinner[5], Catharine van Ingen[6], Deb Agarwal[1]

[1]*Advanced Computing for Science*, Lawrence Berkeley National Laboratory, Berkeley, CA, United States.
[2]*Department of Computer Science*, University of Virginia, Charlottesville, VA, United States.
[3]*Dept. Landscape Architecture and Rural Systems Engineering*, Seoul National University, Seoul, Korea, Republic of.
[4]*INRIA*, Rennes, France.
[5]*National Energy Research Scientific Computing Center*, Lawrence Berkeley National Laboratory, Berkeley, CA, United States.
[6]*Microsoft Research*, San Francisco, CA, United States.

MODIS Community Portal

Goal

❖ **Provide access** to MODIS land and atmosphere swath products over land reprojected into the sinusoidal tiling coordinate system.

❖ **Enable calculations** that use a combination of MODIS land and atmosphere products over land.

Significance

❖ Moderate Resolution Imaging Spectroradiometer (MODIS), the key instrument aboard NASA's Terra and Aqua satellites is important to many scientific analyses.

❖ MODIS continuously generates data as the satellites cover the entire surface of earth every one to two days.

❖ Scientists, who are performing calculations over land and atmosphere products, need MODIS products a single coordinate system.

❖ Data procurement and processing at this scale can be challenging and cumbersome for user communities.
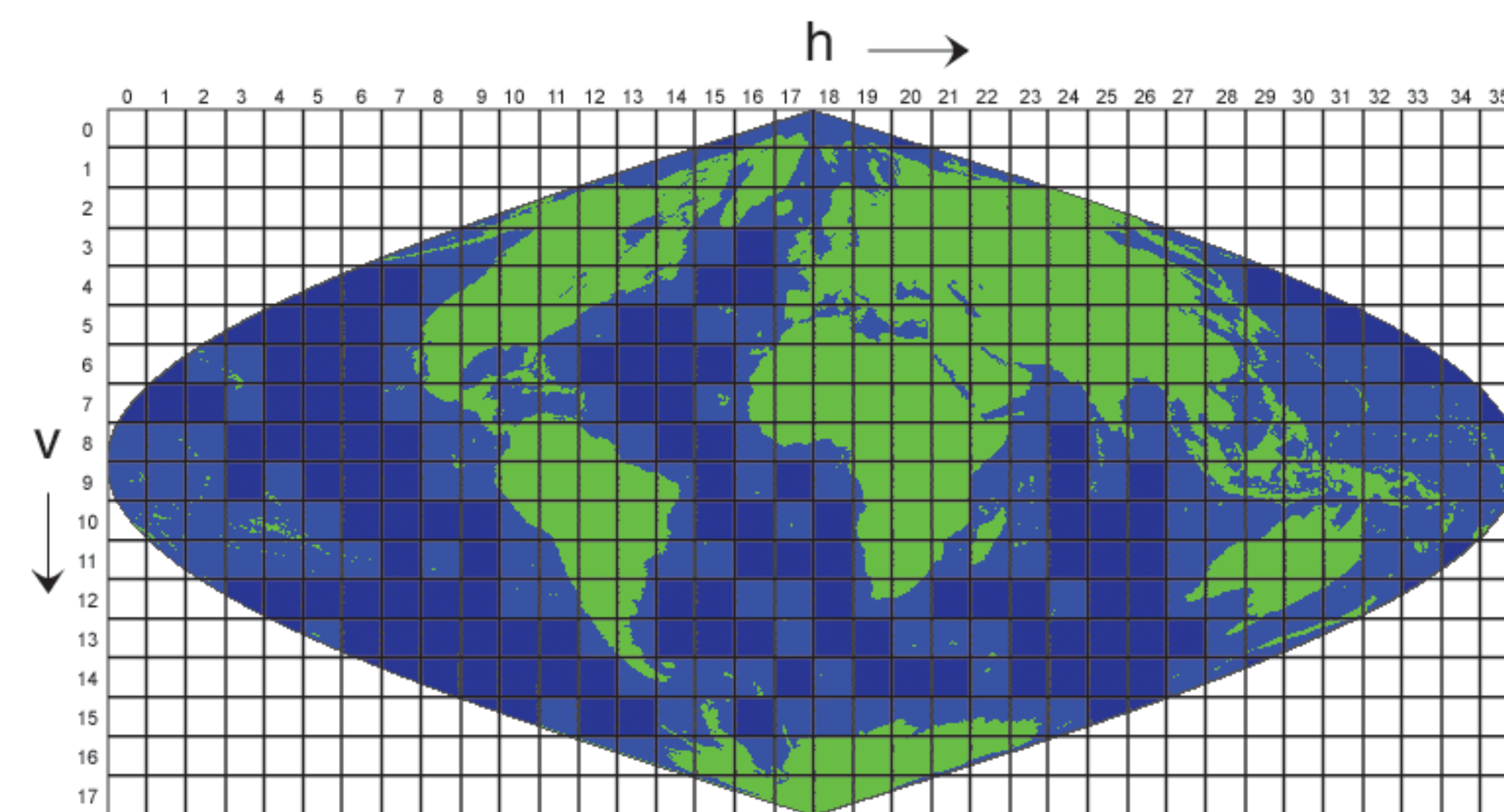
The Community Portal

Reprojected MODIS swath land and atmosphere products are available to the community through a front-end web portal. The portal allows users to download data (~1 TB/year) for the following land and atmosphere products over land:
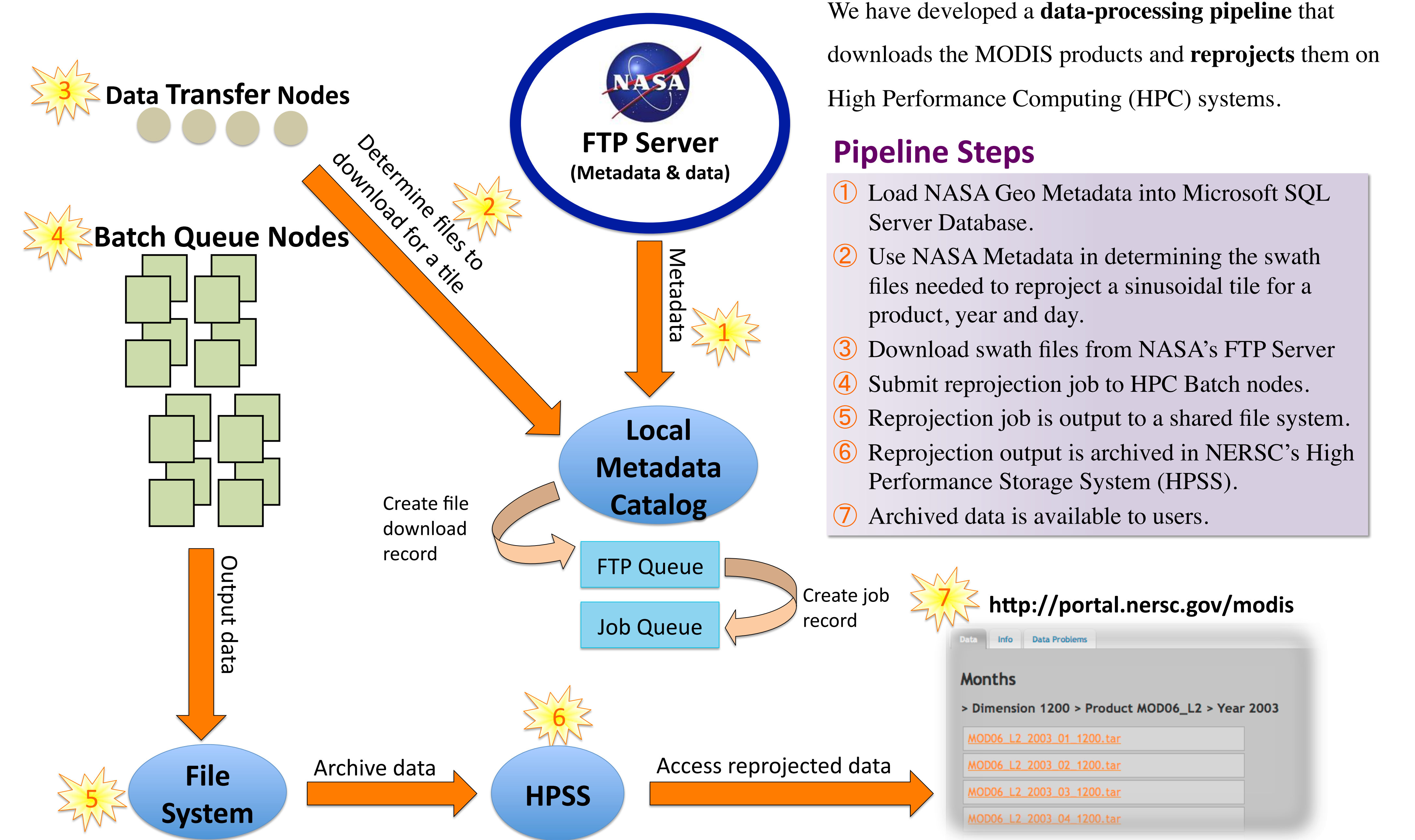
❖ MOD04_L2 (Aerosol)
❖ MOD05_L2 (Water Vapor)
❖ MOD06_L2 (Cloud)
❖ MOD07_L2 (Atmosphere Profile)
❖ MOD11_L2 (Land Surface Temperature Emissivity)

Years and resolutions of reprojected data available:

❖ 2000-2009 at 1km [North American tiles]
❖ 2000-2009 at 5km [Global tiles]

Data Processing Pipeline

We have developed a **data-processing pipeline** that downloads the MODIS products and **reprojects** them on High Performance Computing (HPC) systems.

Pipeline Steps

① Load NASA Geo Metadata into Microsoft SQL Server Database.
② Use NASA Metadata in determining the swath files needed to reproject a sinusoidal tile for a product, year and day.
③ Download swath files from NASA's FTP Server.
④ Submit reprojection job to HPC Batch nodes.
⑤ Reprojection job is output to a shared file system.
⑥ Reprojection output is archived in NERSC's High Performance Storage System (HPSS).
⑦ Archived data is available to users.

http://portal.nersc.gov/modis

Data Processing Challenges

Downloading

Downloading a year of the swath files for **one tile** for all products
=> ~3 files (typically) x 5 products x 365 days = **~5,475** files
Requires significant disk space and bandwidth to download and process data. Additionally, can experience performance and reliability variations on NASA FTP site which causes files to not get downloaded.
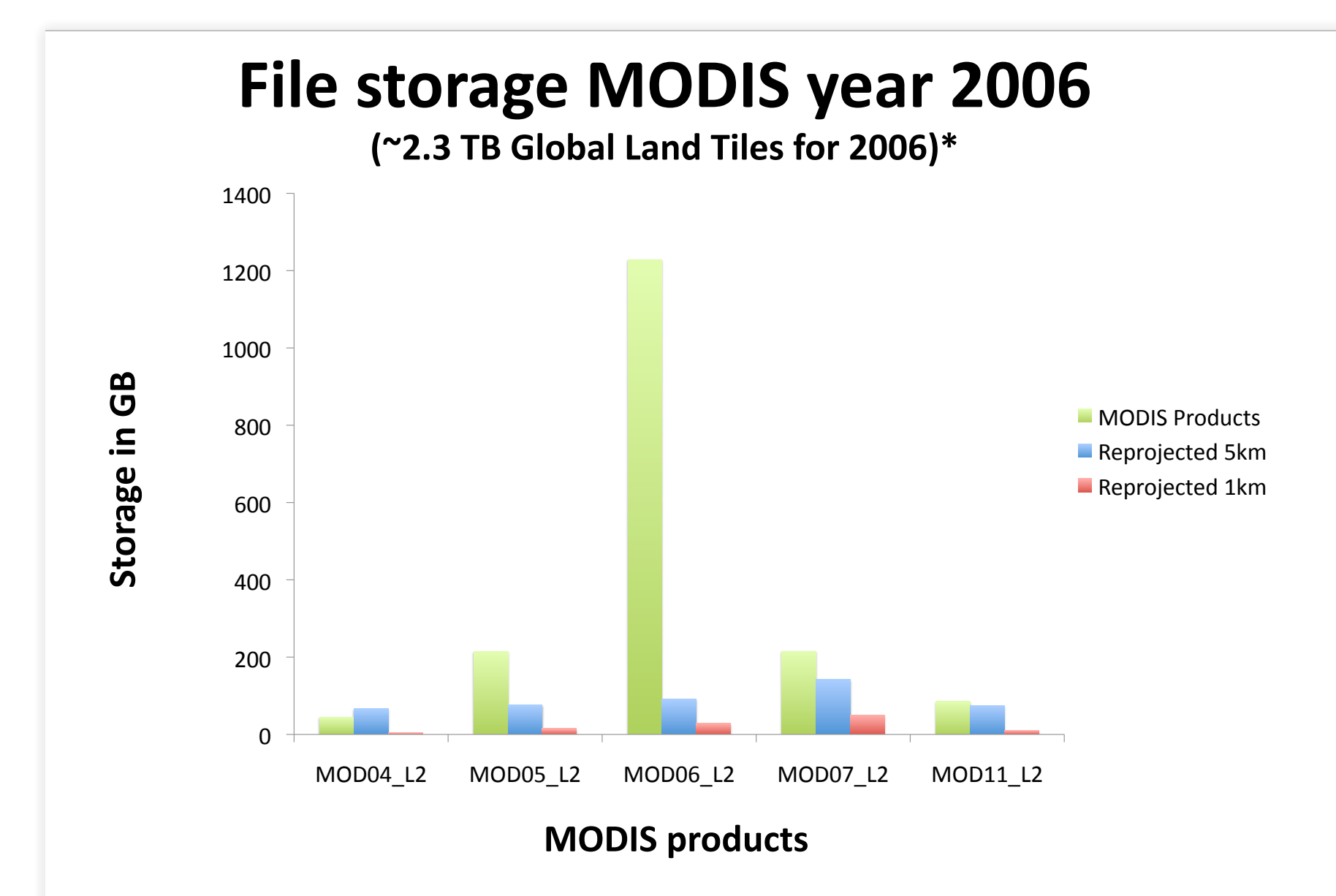
Storage

Swath files combined with the reprojected sinusoidal tiles use a considerable amount of storage.

File storage MODIS year 2006
(~2.3 TB Global Land Tiles for 2006)*
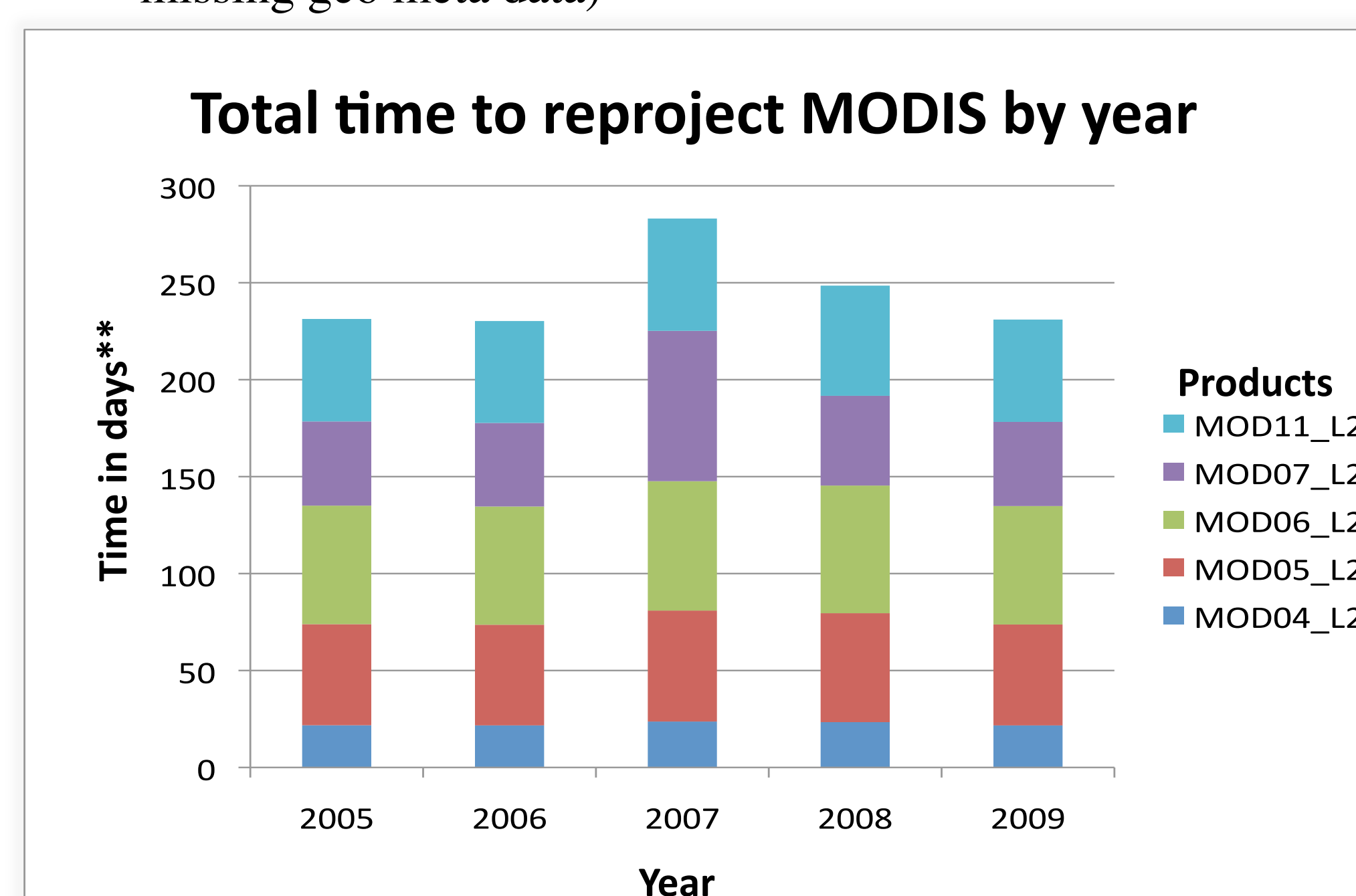
* 10 MODIS years of storage is approximately 23 TB

Other

Challenges addressed in the HPC data processing pipeline in the following three major areas:

❖ **Orchestration of data** included downloading in parallel across many nodes with decentralized storage before processing.

❖ **Validation** after each step of the pipeline - download, reprojection and data archival.

❖ **Failure recovery** involved repeating one or more steps in the pipeline to correct errors (e.g., missing swath product files, missing geo meta data)

Total time to reproject MODIS by year

Products: MOD11_L2, MOD07_L2, MOD06_L2, MOD05_L2, MOD04_L2

** Reprojection jobs ran in parallel on anywhere between 6 and 150 nodes at a time.

Next Steps

User Portal

We would like to continue our efforts to provide reprojected data in an easy to use manner.

❖ **User suggestions for ease of use:** feedback from users will inform future design changes of the web portal

❖ **Provide reprojected tiles beyond 2009**

❖ **User defined region:** provide data for a specified location and area on earth.

Reduction Pipeline

❖ **Provide ability to run calculations on MODIS data:** Allow users to run calculations, such as Evapotranspiration (ET), on 1km MODIS tiles that map to individual FLUXNET sites.

Reprojection Pipeline

❖ **On-demand reprojection pipeline:** provide user triggered reprojection tasks.

❖ **Versioning:** to inform users of what NASA data and what algorithm a reprojected tile was produced with.

❖ **Code Optimization:** increase performance of python reprojection algorithm

Download Data:

MODIS Community Portal: http://portal.nersc.gov/modis/

Contact:

Email: modis-help@lists.lbl.gov