Portable, scalable, high throughput geospatial analyses with Singularity containers on cloud and high performance computing.

Tyson L. Swetnam[1], Mats Rynge[2], Jon D. Pelletier[1], Viswanath Nandigam[3], and Yan Liu[4]

[1]University of Arizona, Tucson, AZ; [2]Information Sciences Institute, Los Angeles, California; San Diego Supercomputer Center, San Diego, California; [4]University of Illinois, Champaign, Illinois`

Abstract

Reproducible computational science requires portable, scalable, workflows. Here we present a method for parallel, distributed, geospatial analysis using free and open-source software developed by the Open Source Geospatial Foundation (OSGEO)[1] i.e. Geospatial Data Abstraction Library (GDAL), Geographic Resources Analysis Support System (GRASS), and System for Automated Geoscientific Analyses (SAGA); with a workflow management system, Makeflow[2], deployed via Singularity[3] containers. Our example involves the calculation of daily and monthly solar irradiation with an OpenMP version of the GRASS r.sun algorithm[4]. A virtual machine (VM) masters the workflow, while remote workers connect over Internet2. The workflow is in production on OpenTopography[5] where users can select any location on the terrestrial earth surface and calculate irradiation and hours of sun from global digital elevation models (DEMs). Our workflow links via the Opal2 toolkit[6] wrapping this particular scientific application as a Web service from a virtual machine on XSEDE Jetstream[7]. Presently, worker nodes are launched on demand on XSEDE Open Science Grid HTC[8]. Most importantly because the workflow is containerized with Singularity, it can be deployed on any combination of laptop, desktop, cloud, or HTC / HPC.

Objectives

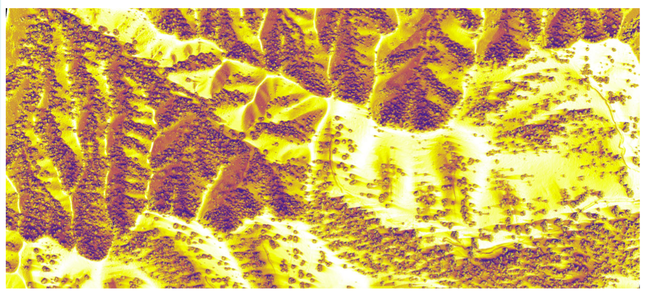
Goal: Model solar irradiation at very high spatial and temporal resolution, anywhere on the Earth.

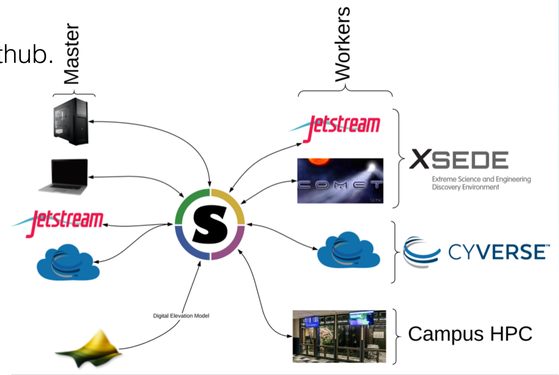
Requirements:
- Must use Free and Open Source Software
- Reproducible Workflows
- Scalable to any computational platform

Resources:
- XSEDE Extended Collaborative Support
- HPC via SDSC Comet
- HTC via Open Science Grid
- Virtual Machines on Jetstream

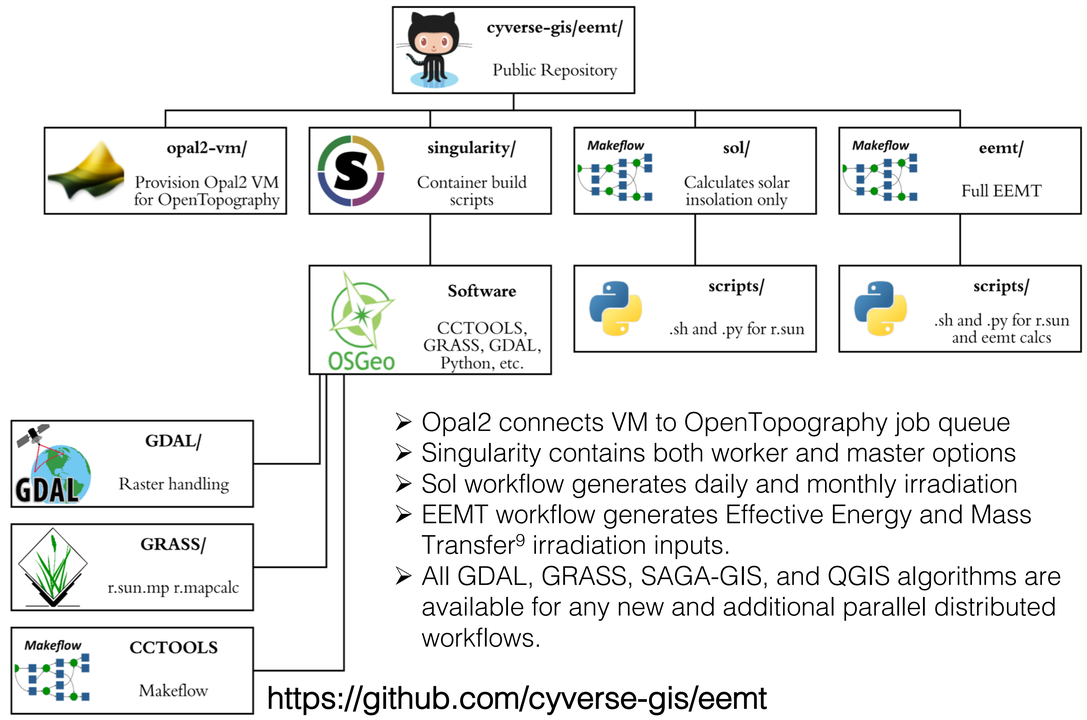
Outcomes

- Tool deployed on OpenTopography
- Workflow containerized with Singularity, hosted on Github.
- Scalable to any number of platforms or compute.





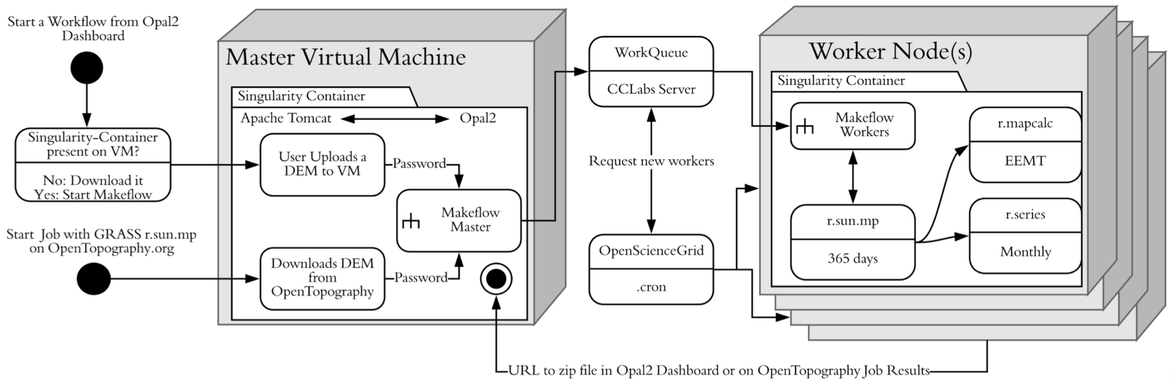
Workflow Hierarchy



- Opal2 connects VM to OpenTopography job queue
- Singularity contains both worker and master options
- Sol workflow generates daily and monthly irradiation
- EEMT workflow generates Effective Energy and Mass Transfer[9] irradiation inputs.
- All GDAL, GRASS, SAGA-GIS, and QGIS algorithms are available for any new and additional parallel distributed workflows.

https://github.com/cyverse-gis/eemt

Workflow Diagram



Jobs submitted on OpenTopography, Jetstream VM starts workers on HTC/HPC, VM hosts results data for one week before deleting. OpenTopography stores job metadata for reproducibility.

References

1. GDAL. 2018. GDAL - Geospatial Data Abstraction Library: Version 2.3.2, Open Source Geospatial Foundation, 2018. http://gdal.osgeo.org
GRASS Development Team, 2018. Geographic Resources Analysis Support System (GRASS) Software, Version 7.4. Open Source Geospatial Foundation. Electronic document:. http://grass.osgeo.org
QGIS Development Team, 2018. QGIS Geographic Information System. Open Source Geospatial Foundation. URL http://qgis.osgeo.org
Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., and Böhner, J. (2015): System for Automated Geoscientific Analyses (SAGA) v. 2.1.4, Geosci. Model Dev., 8, 1991-2007, doi:10.5194/gmd-8-1991-2015.
2. Zheng, C., Tovar, B., & Thain, D. (2017, May). Deploying high throughput scientific workflows on container schedulers with makeflow and mesos. In Proceedings of the 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (pp. 130-139). IEEE Press.
3. Kurtzer, G. M., Sochat, V.,& Bauer, M. W. (2017). Singularity: Scientific containers for mobility of compute. PloS one, 12(5), e0177459.
4. Hofierka, J., Lacko, M.,& Zubal, S. (2017). Parallelization of interpolation, solar radiation and water flow simulation modules in GRASS GIS using OpenMP. Computers & Geosciences, 107, 20-27.
5. Krishnan, S., Crosby, C., Nandigam, V., Phan, M., Cowart, et al. (2011, May). OpenTopography: a services oriented architecture for community access to LIDAR topography. In Proceedings of the 2nd International Conference on Computing for Geospatial Research & Applications (p. 7). ACM.
6. Krishnan, S., Clementi, L., Ren, J., Papadopoulos, P., & Li, W. (2009, July). Design and evaluation of opal2: A toolkit for scientific software as a service. In Services-I, 2009 World Conference on (pp. 709-716). IEEE.
7. J Stewart, C. A., Hancock, D. Y., Vaughn, M., Fischer, J., Cockerill, et al. (2016). Jetstream: performance, early experiences, and early results. In Proceedings of the XSEDE16 Conference on Diversity, Big Data, and Science at Scale (p. 22). ACM.
8. Pordes, R., Petravick, D., Kramer, B., Olson, D., Livny, M., et al. (2007). The open science grid. In Journal of Physics: Conference Series (Vol. 78, No. 1, p. 012057). IOP Publishing.
9. Swetnam, T. L., Pelletier, J. D., Rasmussen, C., Callahan, N. R., Merchant, et al. (2016). Scaling GIS analysis tasks from the desktop to the cloud utilizing contemporary distributed computing and data management approaches: A case study of project-based learning and cyberinfrastructure concepts. In Proceedings of the XSEDE16 Conference on Diversity, Big Data, and Science at Scale (p. 21). ACM.

Acknowledgements

CyVerse material is based upon work supported by the National Science Foundation (NSF) DBI-0735191 and DBI-1265383.

Jetstream is supported by NSF ACI-1445604.

XSEDE is supported by NSF OAC-1548562.

This material is based on services provided by the OpenTopography Facility with support under NSF Award Numbers 1226353&1225810