Geophysical Research Abstracts
Vol. 16, EGU2014-11259, 2014
EGU General Assembly 2014

Large-Scale, Multi-Sensor Atmospheric Data Fusion Using Hybrid Cloud Computing

Brian Wilson, Gerald Manipon, Hook Hua, and Eric Fetzer
JET PROPULSION LABORATORY, Atmospheric Remote Sensing, PASADENA, United States (Brian.Wilson@jpl.nasa.gov)

NASA's Earth Observing System (EOS) is an ambitious facility for studying global climate change. The mandate now is to combine measurements from the instruments on the "A-Train" platforms (AIRS, AMSR-E, MODIS, MISR, MLS, and CloudSat) and other Earth probes to enable large-scale studies of climate change over decades. Moving to multi-sensor, long-duration analyses of important climate variables presents serious challenges for large-scale data mining and fusion. For example, one might want to compare temperature and water vapor retrievals from one instrument (AIRS) to another (MODIS), and to a model (ECMWF), stratify the comparisons using a classification of the "cloud scenes" from CloudSat, and repeat the entire analysis over 10 years of data. To efficiently assemble such datasets, we are utilizing Elastic Computing in the Cloud and parallel map-reduce-based algorithms. However, these problems are Data Intensive computing so the data transfer times and storage costs (for caching) are key issues.

SciReduce is a Hadoop-like parallel analysis system, programmed in parallel python, that is designed from the ground up for Earth science. SciReduce executes inside VMWare images and scales to any number of nodes in a hybrid Cloud (private eucalyptus & public Amazon). Unlike Hadoop, SciReduce operates on bundles of named numeric arrays, which can be passed in memory or serialized to disk in netCDF4 or HDF5. Multi-year datasets are automatically "sharded" by time and space across a cluster of nodes so that years of data (millions of files) can be processed in a massively parallel way. Input variables (arrays) are pulled on-demand into the Cloud using OPeNDAP URLs or other subsetting services, thereby minimizing the size of the cached input and intermediate datasets.

We are using SciReduce to automate the production of multiple versions of a ten-year A-Train water vapor climatology under a NASA MEASURES grant. We will present the architecture of SciReduce, describe the achieved "clock time" speedups in fusing datasets on our own nodes and in the Cloud, and discuss the Cloud cost tradeoffs for storage, compute, and data transfer. We will also present a concept and prototype for staging NASA's A-Train Atmospheric datasets (Levels 2 & 3) in the Amazon Cloud so that any number of compute jobs can be executed "near" the multi-sensor data. Given such a system, multi-sensor climate studies over 10-20 years of data could be perform