What is ParCAT?

• Provides parallel (pre-)processing of large datasets
  – Focus is on datasets with large number of time steps and/or large number of variables rather than higher spatial resolution
  – Calculates point-wise temporal averages (seasonal, monthly, yearly, arbitrary), frequency distributions, and differencing of two datasets (plus difference of averages and average of differences)
  – Takes advantage of multi-core HPC offerings to improve parallelism and lessen IO issues

• Command-line utility
  – Written in C and utilizes MPI and pNetCDF
  – Provides scriptable and embeddable interface
  – Buildable and runnable on diverse architectures – laptops to supercomputers

• Reads netCDF input files, produces netCDF output files
Example 1

• CLM dataset, 15 year, monthly data. 349 variables, ½ degree resolution. ~71gb, 180 files (no CMIP5 postprocessing)
• Compute overall average for each variable across the time series using Titan
• Each node opens one or more files; each core processes one or more variables. Cores consolidate results and nodes create one or more output files.
Example 1 Timing Results

Map Average Time for 349 Variables, CLM Data, 1/2 Degree Resolution, 20 years

- Time went from ~475 seconds to ~15 seconds using only 256 processes on Titan
- The improvement should be even better for larger datasets
Example 1 – Loading into UV-CDAT
Average GPP over the entire data set
Example 2

• CAM dataset, 52 years, monthly data. 28 variables, 1 degree resolution. ~18gb, 28 files (postprocessed into CMIP5 standards)

• Compute overall average of each variable on Titan

• Each node opens one or more files, each core processes one or more time steps. Results are combined into one or more output files
Example 2 Timing Results

- Time went from ~150 seconds to ~50 seconds. More improvement should be possible with more variables or timesteps.