##### An integrative and parallel workflow system for climate data processing

Abstract: Climate scientists confront the complexity of large data processing with various software tools and computing environments. Traditionally the order of processing is controlled by programming scripts, which is difficult to maintain and reuse. In this paper, we build a specific scientific workflows platform based on Kepler for climate data processing, namely CDP-Kepler. The platform integrates the common existing tools in climate research to as workflow’s reusable components. We develop several parallel utilities based on MPI to accelerate the procedure and at the meantime simplify the trivial configuration to meet the need of large data processing in the high performance computing environment. We illustrate our solution with the real applications in climate research. The result shows that CDP-Kepler provides an easy-to-use, efficient, reusable and scalable platform for climate scientists to concentrate on information discovery and exploration.

1. Introduction

Due to the development of climate related instruments, the raw data grows big rapidly. As the climate modes evolve, some reanalysis datasets has exceeded several terabytes. Scientific discoveries in climate science are becoming increasingly data-intensive. How to efficiently analyze these huge datasets is a big challenge to the climate scientists.

All present climate model need numerical data as input files and produce numerical data as output files. Pre and post processing of climate data are common procedures in the climate research. Pre-processing is usually used to prepare input data for the model, which includes file format transformation, gridded data file producing, masking, etc. The output data of models are stored in binary files. In order to improve or present the model’s result understandable, researchers often have to deal with relevant information from the huge amount information stored in a multitude of output files. This procedure is called post-processing including noise reduction, features extraction, visualization, etc.

Pre and post processing covers a variety of data sets in different data formats. Common file formats are NetCDF, HDF, GRIB, ASCII, etc. For the file formats and requirements differs from processing tasks, no perfect processing tool or language exists. The tasks are often handled by a combination of software tools, such as NCO, CDO, NCL, GrADS, Feret, R, etc. Traditionally script (Shell, python …) is used to combine these complex tools together to finish a processing task. However, the problem is that the script is not easy understandable, revisable and reusable. Scientists have to spend lots of time dealing with script instead of exploring useful information in the data. Another problem is that the configuration of high performance computing(HPC) is complicated for scientists to run their big data processing task in parallel. So a more automated solution is needed.

Workflows today play an important role in the business communities. Workflows can also be used in scientific area, which can be called scientific workflows. Scientific workflows inherit many features from business workflows, but go beyond them. The big difference is that scientific workflows system has to meet the unique demands in scientific data processing or management. And these demands vary greatly in different scientific domains.

As scientific workflow sallow scientists to automate tasks through an efficient manner, many workflow systems emerge like the business domain. Kepler, which is dedicated to furthering and supporting the capabilities, use, and awareness of the free and open source, scientific workflow application[2], is one of the most popular.

Based on Kepler, we build a workflow system targeting at climate data processing, also known as CDP-Kepler. CDP-Kepler integrates frequently-used software tools(NCO,NCL,R …) in climate data processing. Moreover, parallel utilities(PNCO and PTrans)are developed as Kepler’s components to meet the needs of large scale data processing. Parallel NetCDF Operator (PNCO) is a parallel version of NCO, which takes advantage of MPI-IO to achieve high efficiency. Similarly, PTrans can transform different file formats(HDF5,GRIB,ASCII) to NetCDF4 in parallel. With CDP-Kepler, climate researcher can easily utilize our high-efficient tools or other related tools as they need in their climate data processing problem through a graphical user interface. Besides convenient composition of a workflow by mouse dragging, the architecture supports the reuse and sharing of climate data processing component quite well.

The paper is organized as follows. In Section 2, we describe Kepler workflow system in detail. Our work and related detail with kepler is covered in Section 3. Section 4 illustrates several scientific use cases with CDP-Kepler. Finally, we make the conclusion in Section 5.

1. Kepler Workflow System

Kepler is an open-source scientific workflow system building upon Ptolemy II system, which a mature platform supporting multiple models of computation suited to distinct types of analysis. Ptolemy[3] abstracts the model of computation to be reusable components that interact between each other. Ptolemy uses the directors to controls the execution model of a workflow. The actors are reusable software components that execute concurrently and communicate through messages between each other. Actors can represent data sources, format transformers, analytical tool, or arbitrary computational steps. Kepler inherits Ptolemy’s feature and adds lots of new features for scientific workflows.

With Kepler’s GUI, scientists can easily change, save, restore and execute workflows in an intuitive way. And its actor-oriented paradigm make the system be easily reusable and extendable. In order to meet various requirements in scientific domains, many customized workflow systems appeared. Kepler can be extended to cope with the challenges include rapid exploratory analysis, visualization, on-the-fly analysis ofstreaming data, recording, re-running, sharing procedures that utilize sensor data, etc. when accessing and using heterogeneous sensor data.[4]“Kepler + Hadoop”[5]integrates Hadoop which is an open source MapReduce frame work as an actor component to provide a parallel and scalable programming model for data-intensive processing. Chemistry researchers distribute the time-consuming parts of the calculation onto computational Grids using Kepler scientific workflow system.[6] Some researcher use Kepler workflow system to deal with the challenges in Ecological Niche Modeling.[7]Kepler/pPOD is developed as extension of the Kepler scientific workflow system for automating phylogenetic studies, orchestrating and routing data between invocations of local applications and remote services, and tracking the dependencies between input, intermediate, and final data objects associated with workflow runs.[8]

Because of the range of data typed and simulation programs, climate data processing usually combine multiple computing environments, software tools and application-specific code written in many programming languages. Some tools or codes are knowledge-intensive to use and maintain. And due to the rapid increase of data size, parallelized approaches shows great potentials but are not wildly used for lack of parallelized concrete tools and practical way to hide the complexity of accessing high performance computing resources. So our work focuses on these challenges in climate data processing. And we use real cases to examine the details of both the challenges and our solutions.

1. System Design

**System Framework**

The CDT-Kepler workflow system is an extension of the Kepler workflow system. We design dozens of actors and directors to facilitate the climate research in the High Performance Computing (HPC) environment. The system framework is shown in Figure 1.

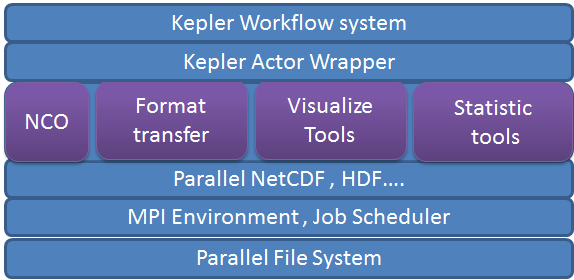


Figure 1 CDT-Kepler workflow system

**Workflow components**

We design ESM-director and actors to hide the detail information of the HPC environment and the usage of tools to the user. With the help of the ESM-director and actors the user can ignore the usage of the tools and the HPC environment and only focus on what to do rather than how to do.

**Director**

We design the ESM Director which could control the executing of a workflow to facilitate the using of Kepler in the high performance computing (HPC) environment. In the high performance computing environment, the works are submitted as jobs throw job scheduler. ESM Director support most job scheduler used in HPC environment such as LSF, SLURM, etc. The global configuration of the workflow is also in the director. Users can set the configurations by filling the parameters of ESM-Director including the how many machines and cores to use, and whether to write the script or not.

**Actors**

We designed dozen of actors to facilitate the Climate research, covering the whole data process chain of the Climate research. We also design parallel data process tools to accelerate the work flow of Climate research. We divided those actors into three classes: data process tools, data generate tools and data analysis tools. These actors take files as input and generate file to be used by other actors or as the final output to be used in the models. The actors can work together to finish the data process and analysis task.

**Functions and Implementation**

**Process Climate data in Kepler.**

Pre and post processing plays an important role in Climate research. We design dozens of actors to help climate research scientists to handle the data processing in Kepler. Pre and post processing covers a variety of data sets in different data formats. We design and integrate several data format transform tools, NetCDF operator tools and grid data generate tools into Kepler to solve this problem.

Most input data of the climate model is observational data collected by several of organizations, and those organizations are trend to use the format of their favor which is more efficient and convenient for them. The data used in the models is usually NetCDF. So the data should be transformed into NetCDF format before it can be used as the input data. In fact, there are tools to do this job. We design an actor for the every data format transform program to make it more convenient to use.

NetCDF Operator is consisting of a dozen standalone, command-line programs to manipulation the NetCDF files. They are used to edit the attributes, extract data, and view the content, contact files, do simple calculate both within one or several NetCDF file. They are widely used in the data process of the climate research. The command-line programs need many arguments, thus make them difficult to use. We designed an actor for each command-line program of NetCDF operator, the user only need to fill in the blanks of the actor parameters, then the actor will do the job for the programmer, in this way the scientist will not need to consider about the command-line program usage, they only need to concern on what they want to do rather than how to do it.

The scientists of climate research usually need to generate data to make the original data usable in their model, especially grid data. The data of climate model is usually grid data, which means all the data is meaningless without grid information which represents the latitude and the longitude of the data. There are various kind of grid all have different physical meaning.

**Analysis and diagnosis climate model data in Kepler.**

Data visualization and statistics is one of the most important ways to analysis and diagnosis the climate model systems.

The Grid Analysis and Display System (GrADS) is an interactive tool to visualization and manipulation climate science data. It is widely used in the post-data process and visualization of climate model. We use the GrADS as one of the post data process tool in the workflow system. User can use the basic function of GrADS or using the GrADS script in the GrADS actor.

R is an open source statistical system, and is used in Kepler workflow system as one of the components. In order to use R to help the climate research scientist, we modified the original R actor in Kepler and designed a NetCDF Adapter actor and NetCDF writer actor, with which the user can import data from NetCDF files into R and write a variable of R as a NetCDF file. The NetCDF adapter actor take a file name as the input, and change the file readable for R actor which has been modified to read the output of an adapter actor. The name of the R input port linked with adapter is the variable name to read and will be the variable refer to the data. A multi dimension variable in a NetCDF file will be treated as a multi-dimension array in R. To write a variable of R into a NetCDF file, the user only need to specify the dimension name of the variable.

**Parallel data process in Kepler.**

In the climate data process, there may be lots of file to deal or very large file. The NetCDF operator are all serial programs, there are no parallel counterpart as far as we know. It costs a long time to process in these situations. To solve this problem, we design a parallel NetCDF operator (PNCO) base on MPI which can do the main functions of NCO. We also design an actor for the parallel operator, which could be used in the same way as the other NCO operator actors. An operation will be transformed into an MPI job and submitted to the HPC system. We also designed parallel data transform programs for the formats which will benefits from parallel, such as binary, hdf, grib. With the development of the climate model, the resolution is increasing rapidly and the scale of the grid is exponential of precision. To meet the need of the grid data generate, we also design and implement parallel grid generate programs.

1. Scientific use cases

Climate data processing chain

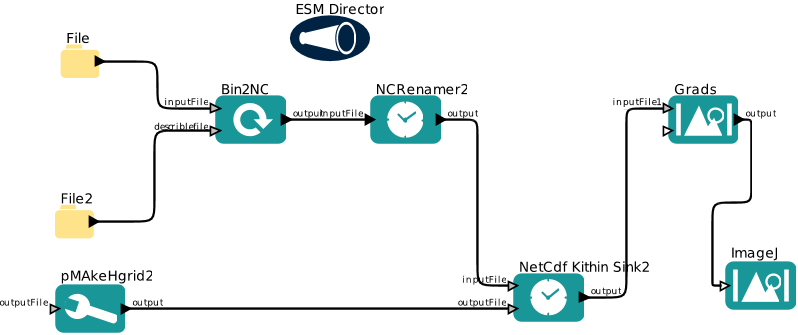
Demonstrations and example

**Parallel data process in Kepler.**

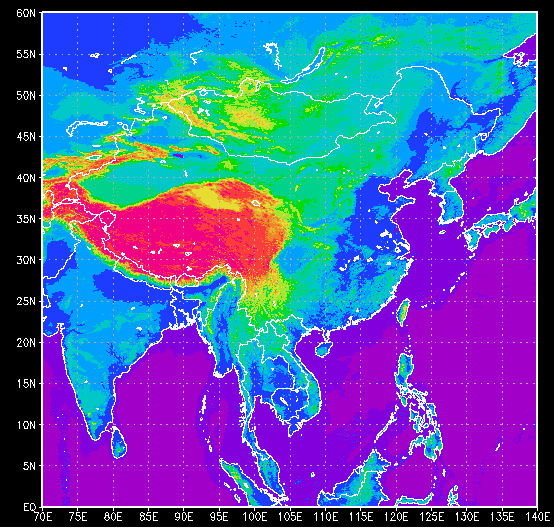
In this section, we will demonstrate the benefit of the parallel tools.

**Data format transforming and grid data generating in Kepler.**

The height of the land is commonly used in the climate research. The scientists need to get the data from other organizations because there is no need and impossible for them to collect the data themselves. The data they get can be in binary format or text format, which cannot be used in their model directly. The first job they need to do is to transform the data into NetCDF (which is commonly used in climate model). And the data they get may lack some information such as the latitude and longitude data, so they need to use related tools to generate these data. One of the most commonly used tools is grid maker, which can generate grids data used in climate model. In this example, we will demonstrate how we use our work flow system to process the data and make it usable in the model.

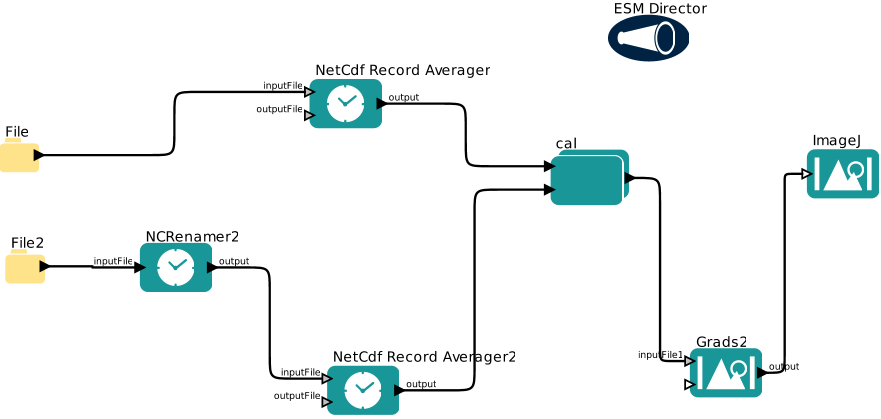


The data we get is the height information of China, the format is binary and there is no grid data in the file we get. What we know is the resolution of the data. We use a Bin2NC actor, which take the input file of binary text and an xml file to change to format of binary file into a NetCDF4 file. The xml file is used to tell the actor how to explanation the binary data. After the data in transformed, we append the grid data generated by the actor of grid maker into the data file. After doing these, we can get a file which is usable in the cliate model. We plot the data with grads demonstrate the correctness of result.

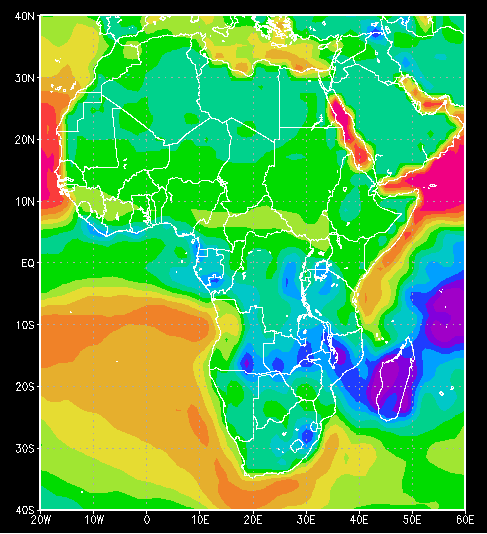
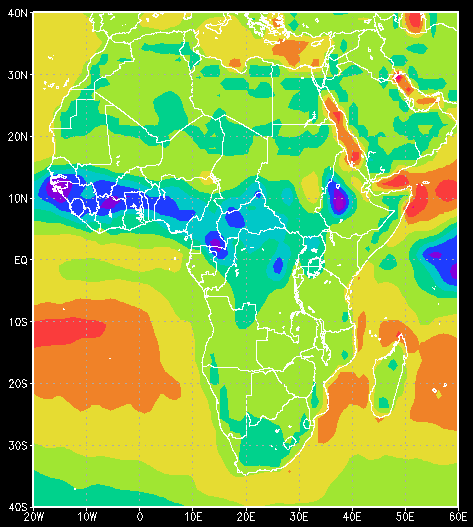


**Data processing and visualization in Kepler.**

This case is a real world use case from the climate research. We use the workflow system as the post-data process tool to help the climate model scientist to deal with the output data of the model. The input data are two files, one is the precipitation(降水) of 20 years, the other is the evaporation. The input data is the monthly precipitation(降水) and monthly evaporation (蒸发) of twenty years ,namely 240 monthly data of south Africa. We need to know the Africa is loss water or not in those years by month, so we need to calculate the average precipitation and evaporation of each month, and use grads to visualize the data.



If we want to calculate the average water loss of the January, the first thing we should do is to extract the precipitation and evaporation data from the input file, then calculate the average value, then get the water loss according to the unit of the variable, then we can visualize the data which we tell the status of each part of Africa intuitively. We use NetCDF Record Average actor to extract and average the data, the weighted average actor to do the weighted calculate of the two file, and then use grads to do the visualization. The left picture is the result for Feb and the right for July. We just change one parameter to get result for different mouth.

****

1. Conclusion

Difference with Kepler(just for climate, parallel)

Features(parallel->fast, integrative->completed)

1. References
2. <http://www.csc.ncsu.edu/faculty/mpsingh/papers/databases/workflows/sciworkflows.html>
3. <https://kepler-project.org/>
4. <http://ptolemy.eecs.berkeley.edu/ptolemyII/>
5. D. Barseghian, I. Altintas, M. B. Jones, D. Crawl, N. Potter, J. Gallagher, P. Cornillon, M. Schildhauer, E. T. Borer, E. W. Seabloom, and P. R. Hosseini, “Workflows and extensions to the Kepler scientific workflow system to support environmental sensor data access and analysis”, Ecological Informatics, vol. 5, no. 1, pp. 42–50, 2010.
6. J. Wang, D. Crawl, and I. Altintas, “Kepler + Hadoop: a general architecture facilitating data-intensive applications in scientific workflow systems”, in Proceedings of the 4th Workshop on Workflows in Support of Large-Scale Science, New York, NY, USA, 2009, pp. 12:1–12:8.
7. W. Sudholt, I. Altintas, and K. Baldridge, “Scientific workflow infrastructure for computational chemistry on the grid”, Computational Science–ICCS 2006, pp. 69–76, 2006.
8. D. D. Pennington, D. Higgins, A. T. Peterson, M. B. Jones, B. Ludäscher, and S. Bowers, “Ecological niche modeling using the Kepler workflow system”, Workflows for e-Science, pp. 91–108, 2007.
9. S. Bowers, T. Mcphillips, S. Riddle, M. Anand, and B. Ludäscher, “Kepler/pPOD: Scientific workflow and provenance support for assembling the tree of life”, Provenance and Annotation of Data and Processes, pp. 70–77, 2008.