Japan Geoscience Union Meeting 2014
(28 April - 02 May 2014 at Pacifico YOKOHAMA, Kanagawa, Japan)

MGI37-P01 Room:Poster Time:April 29 18:15-19:30

Construction of the Large-Scale Statistical Analysis Environment of the STP field data based on the NICT Science Cloud

YAMAMOTO, Kazunori[1*] ; NAGATSUMA, Tsutomu[1] ; KUBOTA, Yasubumi[1] ; MURATA, Ken T.[1] ; WATARI, Shinichi[1] ; TATEBE, Osamu[2] ; TANAKA, Masahiro[2] ; KIMURA, Eizen[3]

[1]National Institute of Information and Communications Technology, [2]University of Tsukuba, [3]Ehime University

There are two major research methods for geo-space science; one is computer simulation, and the other is satellite and/or ground-based observation. Both methods have their advantages and disadvantages: Computer simulations can provide data in whole time and space in the simulation domain, whereas satellite observation data are expected to provide more accurate information. Therefore it is effective for the improved reliability of data and the increased possibilities of explaining phenomena to utilize multi-satellite observation data in combination with sophisticated simulation data with high time resolution. It has a potential to lead to data assimilations in the future.

However, the amount of both multi-satellite observation data and simulation data with high time resolution is very large. We need computational techniques to analyze both data simultaneously. For statistical analysis and visualization, the typical data processing of both multi-satellite observation data and simulation data with high time resolution is called data intensive processing. In the data intensive processing, the same processing is applied to plenty of data files.

We have built a large-scale environment for the statistical analysis where the data obtained through satellites observations and computer simulation are used to construct a uniform, integrated dataset. In this environment, plenty of data are integrated in the following manner: (1)Archiving large quantities of data files, (2)Resampling time series and convert coordinates, (3)Extracting parameters from simulation data, and (4)Merging both data into one file.

(1)Archiving large quantities of data files: Using the STARS (Solar-Terrestrial data Analysis and Reference System) meta-database that provides meta-information of observation data files managed at distributed observation data sites over the internet, users download data files without knowing where the data files are managed. On the other hand, simulation data is saved from supercomputer to petabyte-scale distributed storage that is connected to 10GbE (JGN-X).

(2)Resampling time series and convert coordinates: We developed an original data class (SEDOC class) to support our reading of data files and converting them into a common data format. The data class defines schemata for several types of data. Since this class encapsulates data files, users easily read any data files without paying attention to their data formats. The SEDOC class supports a function of resampling time series through linear interpolations and converting them into major coordinates systems.

(3)Extracting parameter from simulation data: We have developed a 3-D visualization system that visualizes both of these data simultaneously and extract parameters from simulation data in the arbitrary coordinate value.

(4)Merging both data into a single file: Time scale and coordinate are regularized over data files.

We found a practical problem of the system, especially in case of long durational data analyses. It is the problem of the computational load on the processes two to four. It is necessary to solve this problem in order to achieve data-intensive processing for plenty of data files with non-negligible file I/O and CPU utilization.

To overcome this problem, we developed a parallel and distributed data analysis system using the Gfarm and Pwrake based on the NICT Science Cloud. The Gfarm shares both computational resources and perform parallel distributed processing. In addition, the Gfarm provides the Gfarm file system which behaves as a virtual directory tree among nodes. The Pwrake throws a job for each Gfarm node that has a target data file in the local disk. It utilizes local disk I/O to achieve effective load balance.

In today's presentation, we show latest results using archived long durational data and discuss the present Gfarm+Pwrake system extended to wide area.

Keywords: computer simulation data, satellite observation data, parallel distributed processing, Gfarm, Pwrake, NICT Science Cloud

MGI37-P01 Room:Poster Time:April 29 18:15-19:30