A High Level I/O Library For Numerical Simulation Data

Hong Chen
Institute of Applied Physics & Computational Mathematics
Fenghao East Road, Haidian District, Beijing, P. R. China, 100088
TEL: 0086 13910773233
chenhong@iapcm.ac.cn

Fang Xia
Institute of Applied Physics & Computational Mathematics
Fenghao East Road, Haidian District, Beijing, P. R. China, 100088
TEL: 0086 13910773233
xiafang@iapcm.ac.cn

ABSTRACT
In many research fields of numerical simulations, programs often produce a large amount of mesh data with complex structure. It is a fatal bottleneck for scientists to manage such large-scale simulation data. In this paper, we present a data model with high semantic for simulation data organization, which is based on the data characters such as multi-dimension, multivariable and time-space correlative. We use meta-data to efficiently organize and manage mesh data. proposes the way to recognize characteristics of many kinds of datasets, and give the definition and cataloguing principles for the metadata in scientific computation. Based on the high level data model and metadata, we develop an I/O library for various applications such as simulation and visualization applications or other data analysis tools to access data files with a unified method. Under the support of the I/O library, application programs store simulation data into file systems efficiently with the uniform data format including compressed format, on the other hand, they access datasets with metadata of high semantic, which are described by the data model in numerical simulation fields.

Keywords
numerical simulation; data model; I/O library; metadata.

1. INTRODUCTION
With the development of high-performance computer and parallel computing theory, numerical simulations of complex phenomena in full space with high precise and resolution are in sight. A large amount of scientific data with complex structure are produced by numerical simulations, especially by simulations for researching the problems of plasma dynamic and electromagnetism. These large scale of scientific data need both to be stored efficiently and to be exchanged and shared easily among applications. Nevertheless, so many formats of simulation data are formed as scientists employ an enormous variety of discrete approximations in modeling physical processes on computers. Without uniform method to store semantic information about data, problems occur when models based on different representations are required to exchange and share data with other. As the speeding-up of data scale and complexity in numerical simulations, the limitation of traditional data organization and management methods is more and more obviously, which becomes a very important bottleneck for scientific computing performance. With the development of high-performance computer and parallel computing theory, numerical simulations of complex phenomena in full space with high precise and resolution are in sight. A large amount of scientific data with complex structure are produced by numerical simulations. These large-scale simulation data need both to be stored efficiently and to be exchanged and shared easily among applications.

SIMUTools 2010 March 15–19, Torremolinos, Malaga, Spain.

2. DATA MODEL
The data model is a set of conceptional tools which describe data, data connections, data semanteme, and consistency constraints[1]. The aim of the data model is to offer a method to describe data contents[3], and to illuminate how data are described and used through application interface. The design of numerical simulation data model must be fit for structural characteristics of scientific computation data, and must be convenient for accessing, operating and interacting for experts in the field from the angle of physical analysis. We present a data model with high semantic for simulation data organization, which is based on the data characters such as multi-dimension, multivariable and time-space correlative. Researching on the mesh data in simulation computing, using the concepts and methods of building data model in database fields, the objects in the data model are recognized on numerical model semantics, and every object's properties, operations, and relationships between one and another are also presented. Under the support of data model, application programs store simulation data into file systems efficiently with the uniform data format including compressed format, on the other hand, they access datasets with metadata of high semantic, which are described by the data model in numerical simulation fields.

3. SIMULATION METADATA
Metadata technology is a common and efficient way to data management and access. It is fit to simulation data management too, which particularities are the intension and the extension of simulation data. According to the features of typical numerical simulation applications and their output data, we propose the way to recognize characteristics of many kinds of datasets, and give the definition and cataloguing principles for the metadata in scientific computation. Extracting metadata online is implemented, that means metadata are extracted from datasets which are writing into the file system in each output time-step, and then stored automatically into database. At the same while, the effect of metadata can be seen in helping scientists to analysis data. The key to set up a metadata management for scientific computation is creating and mining metadata. Metadata record the content and background information about data in archives, which is help users understand the logical structure of archive and provides

Digital Object Identifier: 10.4108/ICST.SIMUTOOLS2010.8628
http://dx.doi.org/10.4108/ICST.SIMUTOOLS2010.8628

information necessary for retrieve and interpret the data. After a series of tools generating or extracting metadata automatically to a database during the computational application running, scientists can analysis and viewing the metadata to help them find which data they want to get.

4. I/O Library

Based on the high level data model and metadata, we develop an I/O library supporting high-level simulation applications to save and read simulation data in applications. It consists of two main parts: the data-writing interfaces and the data-reading interfaces. In order to meet the need of saving and reading original data (physical variables, nodes coordinates, unit-connection information etc.) and their meta-data, the library is composed of the file interface, the original data interface and the meta-data interface. Because geometrical data and physical data have to be managed separately, it is also needed to define an interface for meta-data of these two types of data. The Fig.1 shows the output process of simulation data by a computing program.

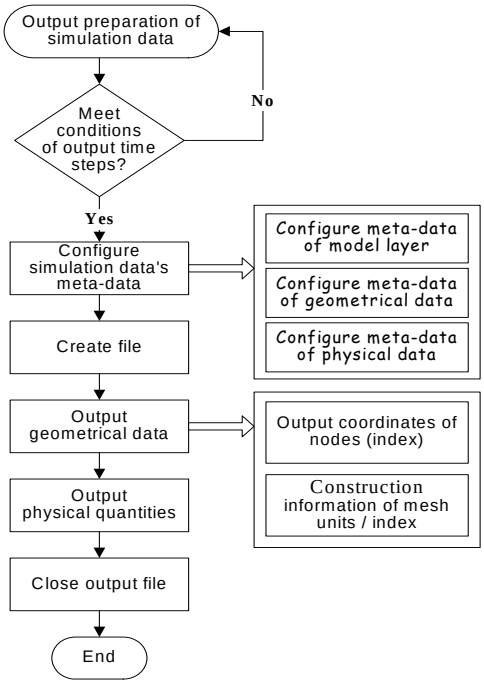

Figure 1. Output process of simulation data

First, the program determines output times by time steps controlling parameters. In each output time, the program creates a specified file after configuring meta-data of model layer, meta-data of geometrical data, and meta-data of physical data. Then, it is time to output geometrical information and physical-variable information of the corresponding mesh type. Finally, it will close the output file to end the output process.

Here the geometrical information is the construction information including coordinates of mesh nodes or mesh units. In this process one file is generated in one output time. For constant geometrical information only one output is needed, i.e. it is not necessary to output at each output time. For parallel programs, each process outputs its own file independently. To manage output files generated from a multi-process and multi-time program, it is possible to keep management information in an external abstract file. The information may be filenames, number of output process, starting and ending time step of the output, and the time step interval of the output.

We obtain data in an I/O command-driven mode and provide a simple and universal Application Programming Interface (API) so that the application program can use API to define data-field objects, and by means of operations of using data-field objects to run reading and writing operations, instead of directly assigning storage and using I/O sentences to save computing results.

API is easily accepted by users because it has reduced modifications to the original numerical simulation program, and it is not necessary to consider output by users. On the other hand, the data organization and storage are determined by API, being independent of computing researchers. Data management personnel take over the power of computing outputs, and they can directly obtain operations to save and manage data formats as well as their results. And meta-data, which are necessary for data management, are inserted in or drawn out from the data output. The unified data access interface is provided for computing programs and visualization programs. In addition, because the complexity of defining and saving data is obviated, because unified and self-describing binary data format can efficiently organize and save scientific data in the high-performance file system, the cross-platform data sharing can be realized with no deed to do format transforms among different numerical computing programs

5. APPLICATION AND CONCLUSION

Through applications of numerical simulating the interaction between plasma and laser, the molecular dynamics, and so on, we have found out many advantages of the I/O library . First, in applications users only need to use a few API interfaces to do output and read/write mesh data. Comparing with directly using HDF5[2] library, the code-writing time can be reduced. Second, for outputting different types of mesh data, it is only need to modify several interface parameters so that codes can be used repeatedly; it would be easy to realize the modularization or templates of the code. Third, there are few output interferes to computing programs because the data organization and storage are determined by scientific data access API. Other reasons for this are that data management personnel take over the power of computing output, and they can directly obtain operations and results needed in data format storage and management. And meta-data, which are necessary for data management, are inserted in or drawn out from the data output. Finally, the unified data access interface and the unified data format are more convenient for developing common data transformation tools and analysis tools.

6. REFERENCES

[1] Byung S. Lee. Modeling and Querying Scientific Simulation Mesh Data. University of Vermont, Department of Computer Science, Technical Report CS-02-7, February 2002.

[2] http://hdf.ncsa.uiuc.edu/HDF5

Digital Object Identifier: 10.4108/ICST.SIMUTOOLS2010.8628
http://dx.doi.org/10.4108/ICST.SIMUTOOLS2010.8628