Coupling Technique for Distributed High Performance Earth System Simulation Framework

Shanshan Li
Dept. of Disaster Information Engineering, Institute of Disaster Prevention Science and Technology
Email: lishanshan1981@126.com

***Abstract*—The Earth system includes a mass of interactive physical elements which interact to cause the constant changing of the Earth, even disaster. Disaster prediction of this system is crucial, but the complexity of this system brings a great challenge to researchers. Distributed high performance Earth system simulation framework is designed to solve this problem on our previous work. This paper proposes an efficient coupling interaction technique for this framework based on PRMI (Parallel Remote Method Invocation). It could automatically implement a series of complex underlying heterogeneous physics data remapping including grid remapping and data parallel distribution remapping, and ultimately realize the process to process parallel communication between physics-model components directly with no third-party conversion tools. Experiment results showed that this coupling interaction technique has achieved desirable performance, and successfully realized the numerical simulation of the coronal mass ejections.**

***Index Terms*—coupling interaction, framework, simulation, Earth system**

I. Introduction

The Earth consists of three physic domains: Sun-Earth, Earth Surface and Solid Earth. Each domain consists of many sub-domains. Different physical domains interact with each other through energy exchanging and material movement, which could cause the constant changing of the Earth, even disaster. Consequently, they work together to build a complex multi-physics coupling interactive system, Earth system. The research on this complex huge Earth system brings a great challenge to earth scientists.

The emergence of distributed high-performance computing technology brings a new way to the research on this complex system, which is high-performance distributed computing simulation [1]. In preliminary studies [2], we proposed a multi-physics coupled simulation framework for Earth system adopting CCA (Common Component Architecture) [3-5] which is a component specification for large-scale high-performance scientific computing proposed by America in 1998. This framework runs on a distributed high performance computing environment, which is divided into four layers: a basic framework layer for coupling interaction, a framework service component layer, a physical-model components layer and a GUI graphical user interface layer. The physics model is encapsulated into CCA physics-model component which is based on provides-uses service design pattern. CCA physics-model component has two type ports, provides port and uses port. A provides port represents the functionality that a physics component provides to other components. A uses port represents the functionality that a component needs. Two components are composed by connecting together with the two ports. This componentization mechanism makes the physical model pluggable and reusable. The framework [6] can make full use of a large number of distributed high-performance computing resources and geophysical models resources provided by different developers on the web to build a multi-physics loosely coupled simulation application of Earth system just by connecting the needed physic-model component service together on GUI. Ultimately, it could achieve realistic simulation of the Earth system, supporting the research and prediction of the earth disasters, As shown in figure 1.

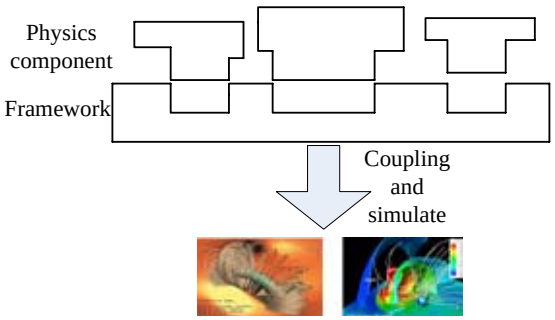

Figure 1. Blueprint of framework and physic-model component.(The two pictures on the bottom are quoted from T. I. Gombosi et al. [7] just for indication))

The core of this framework is the multi-physics coupling interaction. It is a very difficult problem to be solved. Sun-Earth system is composed by many different physic domain models which interact and couple with each other, including the corona, the hemisphere, the

Manuscript received January 1, 2013; revised June 1, 2013; accepted July 1, 2013.

doi:10.4304/jsw.8.10.2594-2599

global magnetosphere, the inner magnetosphere, the ionosphere and the upper atmosphere. Different physical domain models may use different grids, different spatial and temporal scales. In addition, due to the huge work load of the physical model computing, they are often parallelized on different high performance parallel computer, using different parallel mechanisms. Therefore, multi-physics coupling of the Earth system requires a coupling interaction technique which can achieve the grid remapping and data's parallel distribution re-mapping.

II. Related Studies

In recent years, the coupling interaction mechanism has been the research focus. Since the grid re-mapping problem relates to the practical application, researches often focus on parallel distribution re-mapping of complex data, such as PAWS [8], CUMULVS [9], and PRMI [10]. As shown in Table I, PRMI provides a framework-based data-parallel distribution re-mapping strategy and supports the distributed computing environment. It is a better choice for distributed multi-physics coupling interaction mechanism. However, it also has some shortcomings: (1) performance will be lost since it transfers too many messages; (2) it adopts the step-based data model to describe the data using PAWS. Although it is very flexible, it could lead to performance loss and occupy a lot of storage space; (3) scheduling algorithm for intersect index is inefficient, because of the need to calculate the intersection repeatedly between a tuple with other tuples; (4) it does not support the grid remapping.

TABLE I
EXISTING DATA'S PARALLEL DISTRIBUTION REMAPPING STRATEGY

Attribute	PAW	CUMULVS	PRMI
Method	component-based	component-based	framework-based
Supporting remote invocation	No	No	Yes
Description of Data distribution	Step-based	Block-based	Block-based
mode	M*N	M*1	M*N
Communication scheduling mode	Center mode	No	Distributed mode

This paper proposes a new coupling interaction technique for Earth System simulation framework based on PRMI. It not only brings the grid remapping solution, but also compensates the defect in the data's parallel distribution remapping of PRMI. The experimental results show that this coupling technique has better performance and could realize the multi-physic coupling simulation of Earth system.

III Coupling Interaction Strategy

While the physics components are making a remote method call, the coupling interaction policy could automatically implement a series of complex underlying operating including the grid remapping and data parallel distribution remapping on the data sender. It could ultimately realize the process to process parallel communication between physics-model components directly with no third-party conversion tools.

A. Coupling Interaction Process

The coupling interaction process could be divided into four process:

- Process 1: description and registration for data. When physics model developers release component of the physics model, they need to describe the information of its physical data and parallel distribution data. While the data is sending, the component of data sender and data receiver respectively registers its information of physical data and parallel distribution data to framework, making each process on the sender are informed of information of all the processes on the receiver, and finally make decisions. The collection of attributes of data in this process is $E\ (\Omega_{ij}, D_i(\Omega_{ij}), LO_i)$ on the sender and $E\ (\Omega_{ij}, D_j(\Omega_{ij}), LO_j)$ on the receiver;
- Process 2: grid remapping. Based on the physics data of both sides, the data grid of the sender is remapped through interpolation, so that the collection of attributes of the data reaches an intermediate status $E\ (\Omega_{ij}, D_j(\Omega_{ij}), LO_i)$;
- Process 3: parallel distribution remapping for data. Based on the parallel distribution data of both sides, the underlying framework computes to generate scheduling policy of parallel distribution remapping, ultimately makes the collection of attributes of data to reach the final state $E\ (\Omega_{ij}, D_j(\Omega_{ij}), LO_j)$;
- Process 4: data transmission. Based on the schedule policy, each process of the sender directly sent the matched data to the process of the other side.

B. Coupling Interaction Mechanism

This interaction mechanism improves PRMI on four points:

- Improves the parallel remote method invocation semantics. PRMI semantics define three remote method invocation semantics (M=N, M>N, M<N). But the definition of M>N case is inefficient. In this case, the N processes of the sender make one-to-one call to the N processes of the receiver, obtain the return value and store it in memory. The remained unmatched M-N-1 processes of sender make one-to-one call to the processes of the receiver. But it's not a real call, which just obtains the previously deposited return value directly from memory. The whole process cost much, especially in the M>>N case, a large number of unmatched processes need to package messages, transfer messages, unpack messages, and wait to obtain a return value on the receiver. Such process will greatly reduce system's performance. We propose

a new semantics: when M>N, the N processes of the sender make one-to-one call to the N processes of the receiver and obtain return value. Then the process which firstly receives return value broadcasts the value to the other M-1 processes. This semantic reduces the time consumed in method calls of the unmatched processes, which greatly improving the system's performance;

- Improves the data description and registration mechanism. According to the metadata required by defined grid remapping and data parallel distribution remapping, the system could automatically generate grid descriptor and DAD (descriptor of data parallel distribution) to achieve the description of data. Descriptor is not only very flexible, but also can be compressed stored in the memory of each process, which not taking up too much memory space;
- Improves schedule policy of data parallel distribution. The new schedule policy is very efficient based on the Hilbert space filling curve and the interval tree. Each process can directly send the overlapping blocks of data to corresponding process of the other end, which greatly improved operational efficiency.
- Adds the grid remapping mechanism. Based on the defined grid descriptor, we designed a grid converter adopting the Overture developed by Lawrence Livermore National Laboratory of California. It can automatically convert grid when the data is sending.

C. Coupling Interaction Synchronization Strategy

When the distributed components run concurrently in distributed high performance computing, it is likely to cause race conditions, deadlocks or other problems. Therefore, synchronization is needed to avoid the emergence of these problems. However, the full synchronization will block process, resulting in performance loss. Three levels synchronization strategy [11] were applied to different situations, which could avoid the above problems, and greatly reduce the performance loss. The three levels are:

- Synchronous: All the calling process must block, waiting the computing results brought by called component. This kind of way is very safe, but the performance loss is great;
- Non-blocking asynchronous: The caller component directly receives a token after the remote method calls and continues to run. When the call returns, it can use the token to get results.
- One-way asynchronous: It is a completely asynchronous mode which only applicable to one situation, that is when do not need to obtain the computing result of the called component.

The three levels synchronization strategy is depicted in figure 2.

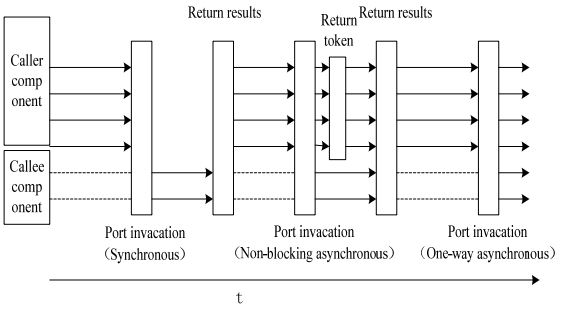

Figure 2. Synchronization strategy

IV DATA PARALLEL DISTRIBUTION REMAPPING STRATEGY

The data parallel distribution remapping strategy is mainly to resolve how the data is correctly sent to the matching process. From the mathematical point, it is to find the overlap region of the overlapping data blocks. Based on the Hilbert space filling curve [13] and interval tree, we design an efficient data parallel distribution remapping mechanism. Space filling curve can reduce the multidimensional data index space to one-dimensional index space. In the one-dimensional index space, interval tree is used to achieve the detection of intersecting data segment. The interval tree is an ordered binary tree which can improve the efficiency of intersecting detection. While the data is registering, the data parallel distribution remapping mechanism automatically and quickly detects overlapping data blocks and generates the scheduling policy of data parallel distributed remapping by the stubs and skeletons. Based on this policy, each process can directly send overlapping data blocks to the matched process. This greatly improves the operating efficiency. The whole implementation process includes four sub-processes:

(1) Build tree. It is actually a registration process of the caller component. The principle is that when the caller component is registering, firstly, the stub reduces the multidimensional data index space to one-dimensional data index space adopting the Hilbert space filling curve; then, generates the binary index range (start, end) of the one-dimensional index space according to data parallel distribution description, in which the start <=end; Finally, builds the interval tree based on the binary index range. Two main interfaces are encapsulated in this process of building tree:

Interface: array<Interval> Hilbert (in DAD DADname)

This interface maps the multidimensional data index to one dimensional index based on the DAD of the registered data. DAD is based on a template class named Templates which describes the layout of parallel data. The SIDL of Templates and DAD are as follows:

Templates (in array<int> Ptopology, in int dimension, in array<float>border, in array<Distribution> type).

DAD (in Data data, in Templates templatename, in array<int> pcoordiates, in Alignmentmap alignmentmap)

Interface: HItree(in Interval interval)

This interface will produce a tree by acquiring interval.

(2) Detect. The detection process is actually a registration process of the data parallel distributor of the called component. The principle is that, as the called component is registering its data parallel distributor, according to its DAD, generates one dimensional data indexing space based on Hilbert and detects the overlap space on the interval tree of the caller component. The detected overlap intervals indicate that these two processes overlap on this data block. Figure 3 describes the process that all of the processes on the caller component detect the overlap interval data block on the interval tree of the process P0 of the called component.

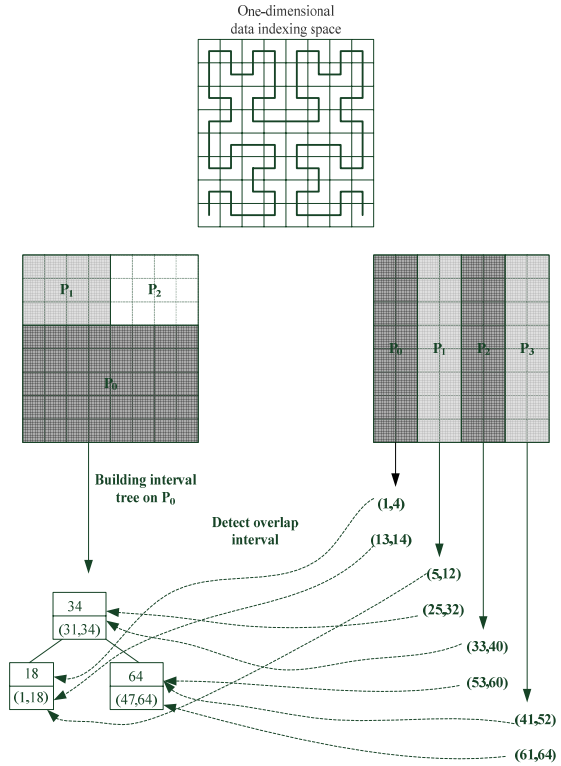

Figure 3. The process of detect overlap interval data block

A main interface is encapsulated:

Interface: Interval detectInterval (in Interval interval, in HItree T)

This interface will produce overlap intervals by acquiring interval and tree. The Algorithm of detection is as follow:

Algorithm of detection

a. *if (T≠NULL){ u← root[T]}*
else end;

b. *if (l≤val(u)≤r)*
{ R← (max(l, u.l_δ),min(r, u.r_δ));
T← u.leftchild, go to a;
T← u.rightchild, go to a;
}

c. *if (r < val(u)){*
if(r > u.l_δ){
R← (u.l_δ, r); }
T← u.leftchild, go to a; }

d. *if (r > val(u)){*
if(r < u.r_δ){
R← (l, u.r_δ); }
T← u.rightchild, go to a; }

(3) Generate scheduling policy of data parallel distributed remapping. According to the intervals detected out, data parallel distributed remapping scheduling policy is automatically generated and storied. This process involves two classes that are OverlapInterval and MXNScheduler. OverlapInterval class object is the overlap interval detected out, consists of three main attributes: Psource, Ptarget and Interval, which respectively represents the source process ID, the target process and the overlap interval class object. MXNScheduler class object is a data parallel distributed scheduler associated the template with object overlap interval. In subsequent interaction, as long as the data has the same templates, the same MXNScheduler will be used, without the need for rebuilding tree and redetection, resulting in improved efficiency. The main interface is as follow:

MXNScheduler (in Templates templatesname,in array< OverlapInterval > overlapInterval)

(4) Send data. According to MXNScheduler, the sender process directly sends the overlap interval data block detected out to the corresponding process on the other side, and record the InterMsgTable from the last checkpoint at the same time. InterMsgTable is a table which records the received and sent messages. It is a five tuple (ComponentIDsend ComponentIDreceive, Psend Preceive, num). ComponentIDsend represents the component ID of the sender. ComponentIDreceive represents the component ID of the receiver. Psend represents the process number of sender. Preceive represents the process number of the receiver. num represents the number of messages sent from Psend to Preceive after the last checkpoint. As shown in Figure 4, when the process P_0 on the called end detect overlap intervals on the interval tree of process P_0 and P_1 on the caller, they directly sent the overlap interval data block to process P_0 on the called end.

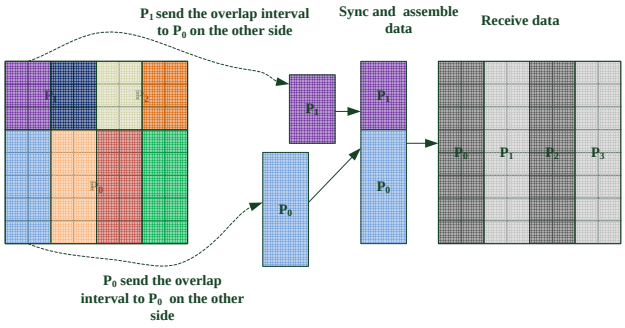

Figure 4. Data sending process

V EXPERIMENTS

We design three experiments. Experiment 1 is designed to test the performance of data parallel

distributed remapping policy. Experiment 2 is designed to test the performance of parallel invocation semantics. Experiment 3is designed to test the function of interaction mechanism supporting building simulation in Earth system simulation framework.

The tests run on two sites (A, B). Each site has a four node cluster. Each node has two processors. The peak performance of cluster is 8.973Gflops. Bandwidth is 100MB/s. The cluster of A site sends data to the cluster of B site, the grid size of which is 160*160. Parallel distribution mode of the two clusters is described as

TABLE II
PARALLEL DISTRIBUTION MODE ON THE TWO CLUSTERS

Site	Data distribution
A	(block, block)
B	(cyclic, cyclic)

shown in Table II. Process topology is shown in Figure 5.

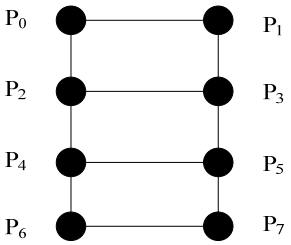

Figure 5. Process topology.

Experiment 1: is to compare the data parallel distribution remapping computing time Tp of this new interaction mechanism we designed and Te of SCIRUN PRMI through setting a different number of processors involved in the run (as the first column of M * N of Table

TABLE III
COMPUTING TIME OF REMAPPING IN DATA PARALLEL DISTRIBUTION

M*N	Te /s	Tp /s
4*2	0.14353	0.12491
4*5	0.16975	0.14432
4*8	0.21028	0.15147

III shown).

Figure 6 shows the data parallel remapping performance based on the data of Table III. The graph shows that when the number of target side processors N increases, the computing workload increases, which leads to Tp and Te increases. But Tp is always higher and increase quicker than that of Te. It suggests that the performance of new parallel remapping mechanism we designed is better than that of PRMI.

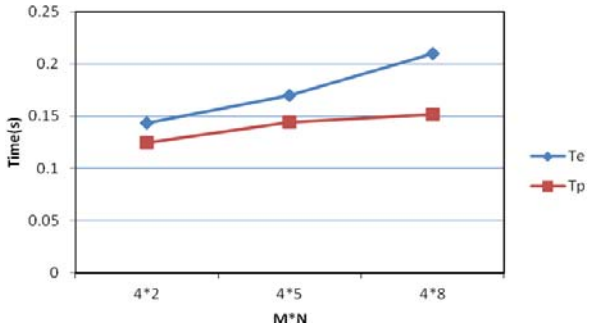

Figure 6. Data parallel distribution remapping performance comparison.

Experiment 2: Due to the invocation semantics mainly improved the M>N case of PRMI, we test the invocation time Ts and compare it to the time Tu which the PRMI requires.

TABLE IV
PERFORMANCE TEST OF PARALLEL INVOCATION SEMANTICS(M>N)

M*N	M-N	Tu/s	Ts/s
8*6	2	0.5376	0.4691
8*4	4	0.5489	0.4713
8*2	6	0.6511	0.4828
8*1	7	0.7094	0.4909

Based on Table IV, we get Figure 7. when in M>N case, Tu is always higher than Ts, and Tu increase significantly with increasing (M-N). In contrast with this, Ts make little change, and maintain a relatively stable state. Therefore the improved parallel invocation semantics could improve performance.

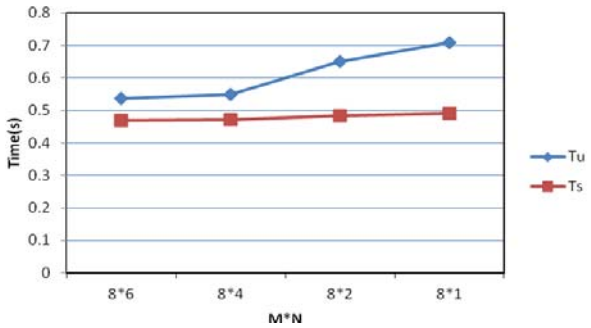

Figure 7. Data parallel distribution remapping performance comparison.

Experiment 3: Based on the new parallel and distributed interaction mechanism, we realize the numerical simulation of the CME (Coronal Mass Ejections) in the Earth system simulation framework, as shown in Figure 8.

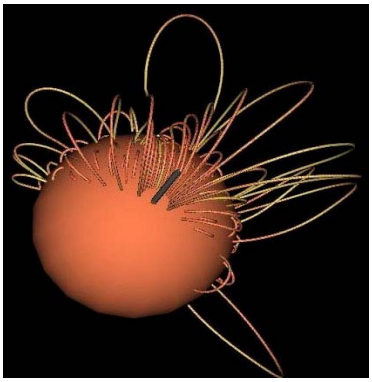

Figure 8. Numerical simulation of the CME in Earth system simulation framework.

VI. CONCLUSION

High performace distributed computing hardware environment and sofware environment, Earth system simulation framwork could support to relize a realistic physic simulation. This paper designs a new coupling interaction technique which is the core and key of the framework. It could automatically realize heterogeneous physical data coupling interaction in a transparent way, and ultimately efficiently achieve the multi-physics coupling simulation.

ACKNOWLEDGMENT

This work was supported by the China Earthquake Administration Teacher's Fund (Grant NO.20110112).

REFERENCES

[1] Shanshan Li, Wenqian Jiang, Qun Wang, "The research on integrated framework for large-scale high performance Earth system modeling and simulation, " *Earth Science Frontiers*, vol.14, no.6, pp. 54-62, 2007.

[2] The Common Component Architecture Forum, http://www.cca -forum.org/, 2012.

[3] Robert Armstrong, Dennis Gannon, Al Geist, "Toward a common component architecture for high-performance scientific computing", In *Proceedings of the Eighth IEEE International Symposium on High Performance Distributed Computing*, pp. 115-124, 2006.

[4] Benjamin A. Allan , Robert Armstrong , David E. Bernholdt , Felipe Bertrand , Kenneth Chiu , Tamara L. Dahlgren, et al., "A component architecture for high performance scientific computing, " *International Journal of High Performance Computing Applications*, vol.20, no.2, pp.163-202, 2004.

[5] Robert Armstrong, Gary Kumfert, Lois Curfman McInnes, Steven Parker, Ben Allan, Matt Sottile, et al., "The CCA component model for high-performance scientific computing, " Concurrency and Computation: Practice and Experience,vol.18, no.2, pp.215-229, 2006.

[6] Toth G, Volberg O, Ridley A J, et al., A physics-based software framework for Sun-Earth connection modeling. Multiscale Coupling of Sun-Earth Processes. Amsterdam: Elsevier, 2005, pp. 383-397 .

[7] T. I. Gombosi, G. Tóth, I. V. Sokolov, et al., "Halloween storm simulations with the space weather modeling framework," *in Proceedings of 44th AIAA. Aerospace Sciences Meeting*, AIAA , 2006.

[8] PAWS Homepage .http://acts.nersc.gov/paws/index.html, 2012.

[9] CUMULVS Homepage .http://www.csm.ornl.gov/cs/cumulvs.html, 2012.

[10] Benjamin A. Allan, Robert Armstrong, Felipe Bertrand, Kenneth Chiu, Tamara L. Dahlgren, Kostadin Damevski, et al., "Data redistribution and remote method invocation for coupled components," Journal of Parallel and Distributed Computing, vol.66, no.7, pp. 931-946, 2006.

[11] Damevski K, Zhang K, Parker S, "Practical parallel remote method invocation for the babel compiler," *Proceedings of the joint HPC-GECO/CompFrame Workshop*, Montreal, Canada, 2007.

[12] Bongki Moon, H.v. Jagadish, Christos Faloutsos, Joel H. Saltz, "Analysis of the clustering properties of the hilbert space-filling curve", *IEEE Trans on Knowledge and Data Engineering*, vol.13, no.1, pp.124-141, 2001.

Shanshan Li was born in Henan province of China, 1981. She received the doctoral degree of Cartography & Geographic Information Engineering from China University of Geosciences in Beijing of China, 2009.

She is an Assistant Professor at Institute of Disaster Prevention in Hebei province, and have worked almost for four years. She have published more than a dozen papers on international conferences/journals. Her main area of interest and research is computer simulation in Earth science.