Building a high performance ArcGIS cluster and its application to climate data processing

Dali Wang[1*], Ziliang Zhao[2], Shih-Lung Shaw[2], Gerald Ragghianti[3], Yaxing Wei[1]

[1] Climate Change Sciences Institute,
, Oak Ridge National Laboratory, Oak Ridge, TN, 37831, USA
Telephone: 18652418679
Email{wangd,weiy}@ornl.gov

[2]Department of Geography
The University of Tennessee, Knoxville, 37996, USA
Email:{sshaw,zzhao7}@utk.edu

[3]Office of Information Technology
The University of Tennessee, Knoxville, TN 37919
Email:gragghia@utk.edu

1. Introduction

ArcGIS Desktop provides an integrated GIS, combining object-oriented and traditional file-based data models with a set of tools to create and work with geographic data. Currently, ArcGIS Desktop software contains functions for mapping, data manipulation**,** data management, as well as a geoprocessing framework, which allows users to choose among the geoprocessing functions in a variety of ways to build extensions and toolboxes. From the early efforts by Openshaw and Abrahart 2000, high performance computing (HPC) has been adopted to address computational and data intensive geospatial problems. The integration of HPC and ArcGIS would create new capability and new possibilities of tackling challenging research questions. Herein, we report our effort of establishing a parallel ArcGIS environment on a computing cluster. First, we present the major components of our high performance computing environment. We then describe the configuration of ArcGIS in the cluster. Finally, we present a case study to demonstrate the parallel ArcGIS capability and performance. We hope those ideas can inspire further developments and deployment of ArcGIS on HPCs.

2. High Performance Computing Environment

Since ArcGIS is developed mainly for the Microsoft Windows operating system, Windows Server 2008 R2 is used as the operating system, and the MSHPC 2008 Pack R2 SP1 is our high performance computing platform. Windows HPC Server 2008 R2, Microsoft's third-generation HPC solution and successor to Windows HPC Server 2008, provides a comprehensive and cost-effective solution for harnessing the power of supercomputing. MSHPC helps developers build HPC applications more quickly, and enables end users to access HPC resources using familiar Windows-based desktops and applications.

2.1 MSHPC installation and configuration

Our MSHPC cluster contains a head node and several computing nodes. The deployment contains following two steps:

1) Deploy the Head Node

The deployment of head node consists of the following steps: 1) Install the Windows Server 2008 R2 on a computer that will act as the head node; install HPC Pack 2008 R2 on the head node; 2) Set the network topology of the HPC cluster network; 3) Add drivers for the operating system image; 4) Add two groups of users: administrator or user.

2) Add Nodes to the Cluster

Windows HPC Server 2008 R2 simplifies the process of deploying nodes by providing automatic node imaging, automatic naming of nodes, and other capabilities to streamline deployment tasks. It also provides tools to monitor the progress of the cluster deployment. Specifically, we use the *Add Node Wizard* and the *Deploy compute nodes from bare metal* method to add nodes to the HPC cluster.

2.2 Overview of the MSHPC environment

The hardware for our cluster is one head node with 8G memory, 2 duel-core computing nodes, each with 4G memory. The CPUs are Intel Xeon CPU X5670 @2.93GHz, and the interconnections are 1Gbps Ethernet links. All the jobs are scheduled in a queue via MSHPC scheduler. The MSHPC environment is illustrated in Figure 1.

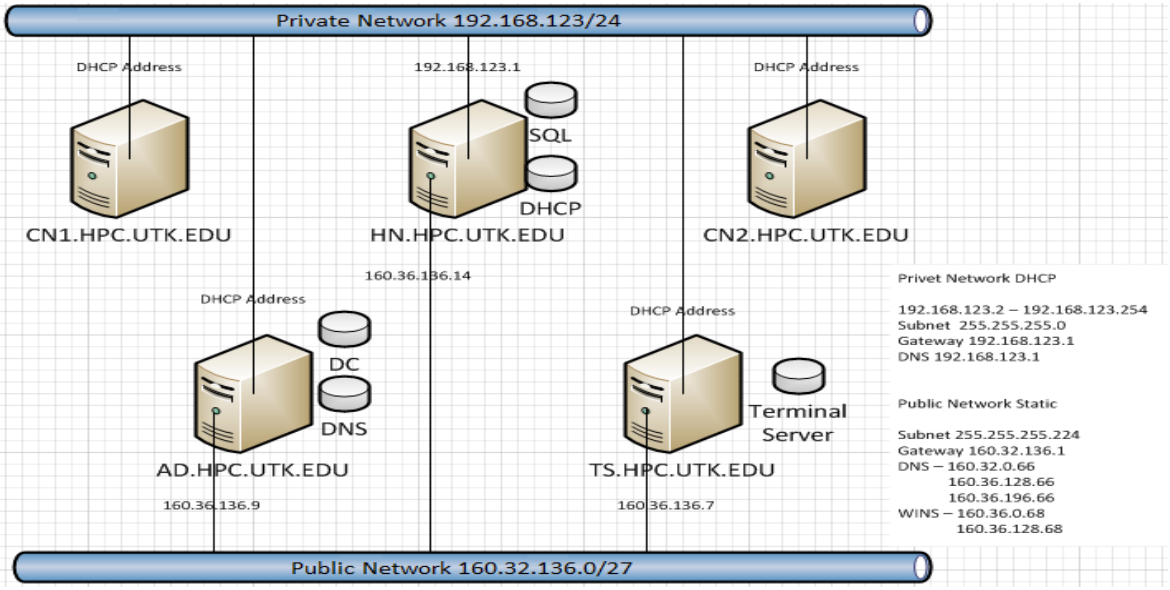

Figure 1. The configuration of MSHPC environment.

As shown in Figure 1, our cluster consists of five servers. They are: 1) Active Directory Domain Controller (AD.HPC.UTK.EDU), 2) Terminal Server Services (TS.HPC.UTK.EDU), 3) HPC Head Node (HN.HPC.UTK.EDU) and 4) two Computing Nodes (CN1/2.HPC.UTK.EDU)

3. ArcGIS Cluster Configuration

Based on the high performance computing environment settings, shown in section 2, our ArcGIS cluster environment consists of two components: software development stacks and communication middleware.

3.1 Software development stacks

The head node of our MSHPC cluster is used as a software development and analysis node to host the full suite of ArcGIS desktop 10 with several extensions and toolboxes. The software development environment for our cluster is the Microsoft Visual Studio 2010 Professional. The programming language we chose is the C# due to its easy integration with the ArcObjects. ArcGIS Engine Runtime is installed on all the computing nodes.

3.2 Parallel communication component: MPI.NET over .NET Framework 3.5

The parallel communication middleware for our cluster computing is MPI.NET http://www.osl.iu.edu/research/mpi.net/. MPI.NET is a high-performance, easy-to-use implementation of the Message Passing Interface (MPI) for Microsoft .NET environment. MPI is the *de facto* standard for writing parallel programs running on a distributed memory system, such as a compute cluster, and is widely implemented. MPI.NET provides support for all of the .NET languages (especially C#), and includes significant extensions (such as automatic serialization of objects) that make it far easier to build parallel programs that run on clusters.

3.3 License management

The ArcGIS cluster is enabled by the ESRI site license at The University of Tennessee, Knoxville. Under the agreement of ESRI site license, ArcGIS desktop license and software are installed on the login/head node to create ArcObject-enabled executable. On each computing node, one ESRI ArcGIS license and necessary libraries are installed.

4. Case Demonstration

4.1 Use Zonal Statistics to Upscale Land Cover Type Map for Carbon Cycle Modeling

Land cover type map is a key driver input data for carbon cycle models. Remote sensing-based global land cover type maps provide valuable Earth surface information at fine spatial resolutions, from 30 meters to several hundred kilometers. There is a need to upscale land cover type maps to match the spatial resolution of carbon cycle models (quarter degree to one degree), but still represent the key information on land cover type. In our study, the Zonal Statistics tool (with "MAJORITY" option) in the ArcGIS Spatial Analyst extension is used to identify the dominant land cover type map within each modeling grid cell used by carbon cycle models.

4.2 Data Sources

Two major data sources are used in this study: Synergetic land cover product (SYNMAP) and Global modeling grid at half degree (GMG05). SYNMAP is an improved global land cover type product with 48 classes at 0.00833 degree (about 1 km) spatial resolution. It fuses different global land cover products, including Global Land Cover Characterization (GLCC), Global Land Cover 2000 (GLC2000), and Moderate Resolution Imaging Spectroradiometer (MODIS) land cover products (Jung et.al, 2006). SYNMAP is in GeoTIFF format with a size of 721MB. SYNMAP is shown in Figure 2. GMG05 is an

ESRI Shapefile containing 720x360 polygons which define the boundary of all modeling grid cells.

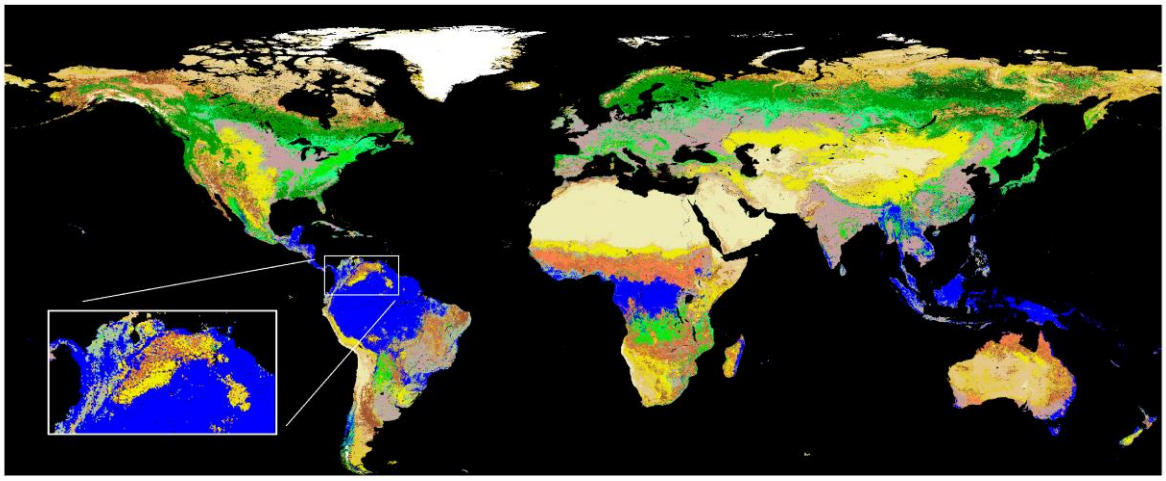

Figure 2: The overview of SYMMAP at original 1 km resolution.

4.3 Parallelization strategy and workflow

A simple data parallelization strategy is used in this study. The data was split into 4 parts with a similar size. An MPI communication is created with 4 processes. Each process is assigned to handle one part of data independently on compute nodes, which include invoking ArcGIS runtime, handling input datasets, upscaling land cover type map layer, and writing raster data output to local disks.

4.4 Result and performance

For a demonstration purpose, half degree (around 50 km) zonal results are shown in Figure 3. The colors represent the dominant land cover type in each cell. There are 48 land cover types shown in this map, illustrate very similar land cover type patterns with Figure 2.

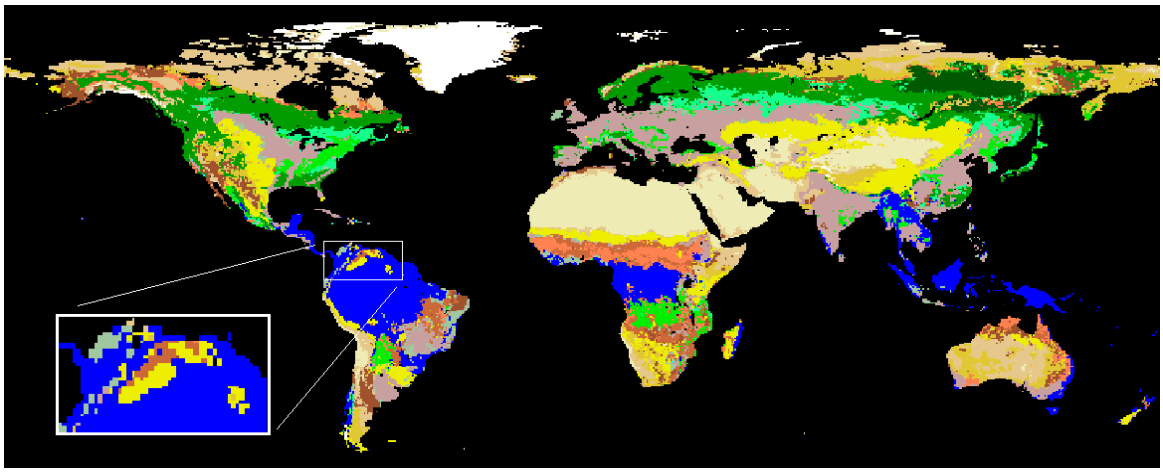

Figure 3: Upscaled SYNMAP (using “majority” zonal statistics) at half degree resolution.

It usually takes more than 40 minutes to finish the calculations using Zonal statistics function through Arc Toolbox on a desktop (with Intel Core2 Quad CPU Q6600 @2.4GHz and 3GB Memory). It also takes more than 20 minutes to process the 4 parts of data sequentially on the desktop. On the other hand, it takes only a total wall-time of 2 minutes and 8 seconds to finish those calculations with Zonal statistics function via C# API in our HPC environment and produce the same results, which demonstrates a major improvement over the desktop environment.

5. Conclusions

This article presents a roadmap to establish an ArcGIS cluster on a Microsoft HPC server and uses a zonal statistics case to demonstrate new potentials of using the current ArcGIS functionality to handle large datasets on a computing cluster. To our best knowledge, it is the first effort to establish parallel capability using ArcGIS and cluster computing. The proposed HPC approach and implementation methods can be useful to the broad ArcGIS community. Our future plans include further tests on HPC machines to better understand the performance improvement, and deployment of a larger ArcGIS cluster and development of a spatio-temporal data model in support of computing intensive research such as global climate change studies.

6. Acknowledgements

This research was funded by the University of Tennessee, Knoxville. Oak Ridge National Laboratory is managed by UT-Battelle LLC for the Department of Energy under contract DE-AC05-00OR22725.

7. References

Openshaw, S. and RJ Abrahart. R.J., 2000, *Geocomputing*, London: Taylor & Francis

Jung, M., Herold, M., Henkel, K., and Churkina, G., 2006, Exploiting synergies of land cover products for carbon cycle modelling, *Remote Sensing of Environment*, 101, 534-553.