Land-use Simulation at Large-scale using Big Data

Dan Li[*1]

[1] College of Environmental Science and Tourism, Nanyang Normal University, China

January 4, 2015

Summary

Land use/cover changes at larger scales have crucial impact on large-scale environmental problems. Cellular Automata (CA) has become a main tool to simulate and predict land use changes. But large-scale land-use change simulation with high-resolution data requires a large amount of data, complicated computing processes, and very long execution time. The data storage size can be hundreds of megabytes or even several gigabytes, we could consider these data as a kind of "Big Data", therefore these big data also lead to the problems of computational capability. A computing model called GPU-CA model is proposed to use the graphics processing unit (GPU) high-performance technique to execute and accelerate such simulations. The comparison indicates that the GPU-CA model is faster than traditional CA by 30 times. Such improvement is crucial for land-use change simulations at large-scales using big data.

KEYWORDS: Land-use Simulation, GPU, Cellular Automata, Large-Scale, Land Use Change

1. Introduction

The basic research object of Geospatial System is an open, complex, giant system that composes of natural, social and other elements with multi-scale and other characteristics of complex systems. In past study, the theory of complex systems was introduced to model and interpret the geospatial system. Cellular Automata (CA) is one of the most powerful tools in complex system theory, which has been used to simulate the geographical phenomena, such as urban development (Batty and Xie, 1994), disease expansion, fire spread, and etc. CA has become a main tool to simulate and predict the urban expansion, especially in the field of land use/cover change studies. In the previous study, Xia Li (2009, 2010a) have successfully developed Geographical Simulation and Optimization Systems (GeoSOS) via combining the theories of land-use change simulation and spatial optimization and extending the theory framework and range of application in related research fields.

2. Solution for land-use simulation at large-scale with big data

Land-use changes at larger scales (e.g. provincial, national, or even at global scale) have crucial impact on large-scale environmental problems, such as global climate change, food safety, carbon recycling, and so on. When simulating and predicting large-scale land-use changes, selecting proper data and experimental techniques are very important to generate reasonable results.

Raster data are usually used in spatial simulation experiments, thus, selecting proper data precision is considerable. For low-resolution data, a single data pixel represents multiple land-use types, and the composition of these types is expressed as percentages. This may affect the precision of model analysis results. However, for high-resolution data, each pixel can completely represent the dominant

[*] danl_163@163.com

land-use type within the involved space. Moreover, low-resolution data lose local information and non-linear features of geographic patterns, thus using high-resolution data in the simulations can ensure that micro-scale information will not be omitted. CA use a bottom-up approach to explore the emerging behavior of complex systems, it is preferable to use fine-grain data in simulations. Thus, it is essential to apply high-resolution data to large-scale land-use change simulations. However, large-scale land-use change simulation with high-resolution data requires a large amount of data, complicated computing processes, and very long execution time. For example, the data storage size can be hundreds of megabytes or even several gigabytes. We could also consider these data as a kind of "Big Data", which become more and more important and essential consideration in such large-scale simulations.

But Big Data also leads to the problem of computational efficiency. It is difficult for a PC to perform such simulation experiments because of its limited computing capability and CPU-based serial computing pattern. Parallel computing is an optimal method used to improve computational capability. There have been some studies on parallel computing-based land-use change simulation (add references) e.g., load balancing-based parallel CA simulation, grid computing-based CA simulation (Xia Li et al. 2010b), and others. However, the shortcomings of these methods lie in high computational cost, complex configuration, and lower performance relating to its computing acceleration. Thus, a new parallel computing pattern with low cost, simple configuration, and better acceleration performance is necessary, to provide better computational capability for large-scale land-use change simulations.

The graphics processing unit (GPU) is high-performance technique which can be used to CA simulation. GPUs make use of computer graphics card to execute general-purpose computations (Owens, 2008). It has characteristics of low cost, and a high degree of parallelization, programmability, and flexibility. The "CPU + GPU" computing pattern is a trend representing the future development of high-performance computing techniques. The TIANHE-1 supercomputer is also made by this hybrid computing architecture, which was the top 1 supercomputer in 2010 - 2011. NVIDIA proposed the CUDA (Compute Unified Device Architecture) computing platform in 2007, which provided a GPU-based general-purpose computing environment and software architecture that can be developed using C-like language. With this platform, general-purpose computing tasks can be performed with any CUDA-supported computer graphics cards, so the computational cost is relatively low. Presently, GPU based high-performance computing is extensively applied in fields such as physical simulation, image processing, three-dimensional terrain generation, signal processing, artificial intelligence, and others (Ferrando et al., 2011). There are greatly improves on the computational efficiency comparing the original CPU-based pattern. Geographic problems are generally complex, thus, applying the GPU computing technique to geographic simulations of large-scale and/or high-resolution land-use changes would be of great importance.

3. GPU-CA model

In this study, a computing model called GPU-CA model is proposed to map the CA computation procedure to the GPU programming and memory models. GPU general-purpose computing can be logically divided into three levels. The smallest computing unit is the Thread. Multiple threads compose a Block (one block has one shared memory), and one or more blocks compose a Grid. The GPU-CA computing model uses a data parallel computing pattern, which maps a thread to a cell in the cellular space, and this thread is responsible for the computation of the corresponding cell. The data with relating space variables are read by graphics cards from the computer mainframe (host) memory to global memory. These data are then copied to each shared memory of the blocks, where each thread reads the data according to the cell it maps. Since the CA model requires information about the neighborhood of each cell, each thread simultaneously reads data from the neighbors of the mapped cell. We adopted the 3×3 Moore neighborhoods. When the data reading is complete, a computing iteration is carried out using the multiple thread units, and resulting data are written back to global memory. The next iteration then begins, until the termination condition is reached. Figure 1 shows the GPU-CA computing model architecture.

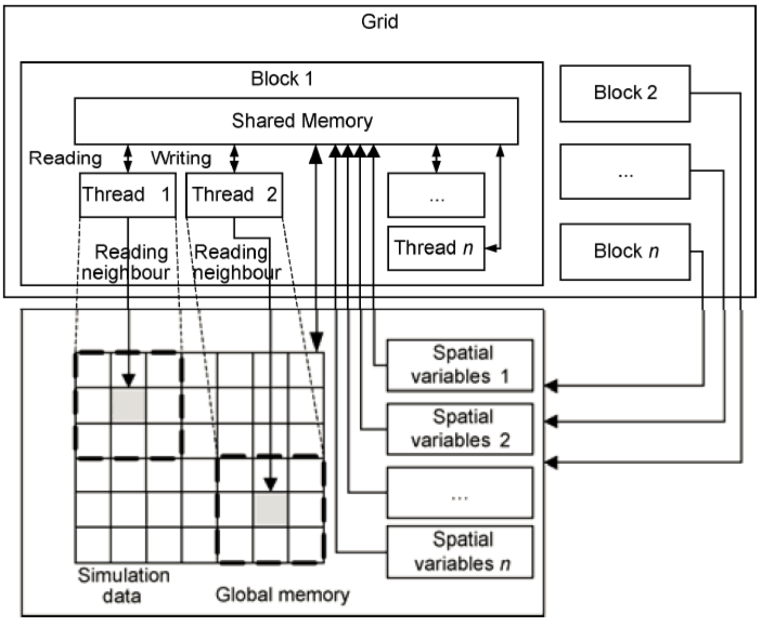

Figure 1 GPU-CA computing model architecture

4. Experiments and Results

We used the GPU-CA model to simulate the urban expansion in a rapid urbanization area as the study area, which is Guangdong Province, China, with a land area of 179800 km2. The land use classification data of Guangdong Province in 2000, 2005, and 2006 were acquired, as well as related space variable data. The land-use classification data are raster data files in ArcGIS ASCII format, with 80 m spatial resolution and the raster size is 9792×7376. The single data document has a storage size of approximately 500 MB, so the total amount of land-use classification data and space variable data exceeded 3GB, which is excluded the large amount of intermediate result data. There are six land-use types, namely, farmland, forest, grassland, water area, construction land, and unused land. The space variables include distance to city center for each raster cell, and distances to the railway, highway, and roadway. Based on the logistic regression CA model, the land-use classification and space variable data in 2000 and 2005 were used to determine the conversion rules, which define the conversion of non-construction to construction land in Guangdong Province. Subsequently, the land use data of 2005 was used as initial data to simulate the land conversion process in the period 2005–2006. After the simulation result was produced, it was compared with the 2006 land-use classification data using two accuracy assessment approaches to validate the simulation result, which are the point-to-point and naked-eye approaches. The general CA provided by GeoSOS software (GeoSOS, http://www.geosimulation.cn/) was used to simulate the same land-use change process, for analyzing the degree of improvement in computational efficiency by GPU-CA. Finally, the construction land conversion process in Guangdong Province during 2010 and 2015 was predicted. Figure 2(a)–(c) describes land use in Guangdong Province in 2000, 2005, and 2006, respectively. Figure 4(d) describes simulation results for 2006, using land use transition patterns extracted from the data of 2000 and 2005. Figure 4(e) and (f) shows predicted results for 2005–2010 and 2005–2015, respectively.

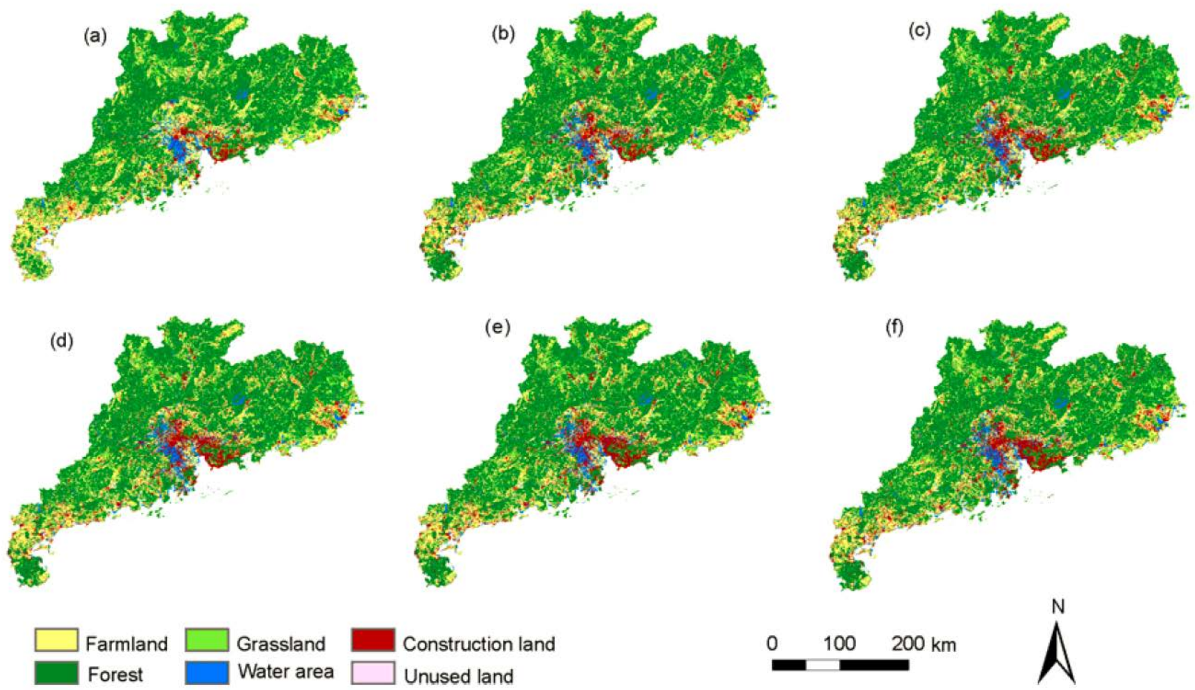

Figure 2 Diagram showing the simulated results of GPU-CA

To compare with the general CA using the CPU serial computing pattern, we adopted a powerful hardware platform configuration. This platform consists of 2× Intel Xeon E5620 2.5 GHz CPU, 24 GB mainframe memory, Windows 7 64-bit professional edition operating system, and Tesla 1060C graphics card manufactured by NVIDIA Inc., which has 4GB graphic memory and 1.30 GHz core frequency. The GPU computing environment is CUDA 3.2, and the development environment is Microsoft Visual C++ 2008.

By using the point-to-point comparison method which commonly used in image consistency evaluation, a confusion matrix is generated to test the simulation accuracy. The results indicate that the overall accuracy of GPU-CA simulation is 82.9%, indicating that the GPU-CA model is very effective and can be applied to large-scale land-use change simulations.

The concept of speedup is commonly used to evaluate the performance of parallel computation, which uses the ratio of parallel computation time to CPU serial computation time, to represent the acceleration performance of parallel computing. It shows that GPU-CA can remarkably accelerate the CA simulation, with the greatest speedup reaching 33.97. This is a much better improvement on the execution efficiency than the CPU computing pattern, indicating that GPU-CA is very efficient in computation. We also compare the acceleration performances of GPU-CA with the load balancing-based parallel CA. Parallel CA uses a 3×3 neighborhood to simulate land conversion in the Pearl River Delta region, with the computation time approximately 1200–1600s and a speedup around 2. The simulation time using GPU-CA is approximately 15s, and speedup around 18. Therefore, GPU-CA shows much improved computing performance over parallel CA. Moreover, parallel CA requires 8 PCs for parallel computing. GPU-CA only requires one computer with a graphics card supporting the CUDA environment, thus, the computational cost is much lower than parallel CA. This result fully demonstrates that GPU-CA has a low computational cost and high computing capacity.

5. Discuss and Conclusion

As described above, GPU-CA model is suitable, efficient and low-cost for large-scale land-use change simulations, which uses big data of geographical information and need powerful computation capability. Therefore, these are still some issues should be discussed in future. First, CA obtains conversion rules by statistical algorithms, usually the sampling data for statistics are 10% or even less of the whole spatial data. This is proper for small or medium spatial scale studies. But in large-scale land-use simulations, spatial heterogeneity will lead to the sampling data maybe only reflect the global law of the entire study area, but lose the local laws in different spatial areas. So the spatial-division algorithms should be applied to divide the whole study area to different zones according to the different spatial and humanity factors. Second, in present work we read all the computing data to GPU at once, for the amount of these data is not beyond the size of the GPU memory. But in practice, this condition will not be guaranteed all the time, so data-partition method as a more parallel approach should be used in later simulations. That will divide data to regular blocks, and then read some or all blocks to GPU memories at once according to the GPU memory size. This is a flexible framework which can be self-adaptive for different hardware environment. Also parallel computing protocols, such as OpenMP, can be used to construct a computing cluster that composed of multiple GPU graphics cards. It can improve computing efficiency and expand the data processing scale. These are all the subjects of future research.

6. Biography

Dan Li, got a doctoral degree of geographical information science at Sun Yat-sen university in 2011. Mainly interested in urban development simulation, high-performance geographical computing, GIS software development.

References

Batty M and Xie Y (1994). From cells to cities. *Environment and Planing B-Plan Design*, 21(7), 531–548.

Li X, Liu X P, He J Q, et al. (2009). A geographical simulation and optimization system based on coupling strategies (in Chinese). *Acta Geographical Science*, 64(8), 1009–1018.

Li X, Li D, Liu X P, et al. (2010a). The implementation and application of geographical simulation and optimization systems (GeoSOS) (in Chinese). *Acta Sinentiarum Natralium University Sunyatseni*, 49(4), 1–5.

Li X, Zhang X H, Yeh A G et al. (2010b). Parallel cellular automata for large-scale urban simulation using load-balancing techniques. *International Journal of Geographical Information Science*, 24(6), 803–820.

Owens J D, Houston M, Luebke D, et al. GPU Computing. *Proceedings of IEEE*, 2008, 879–899.

Ferrando N, Gosalvez M A, Cerda J (2011). Octree-based, GPU implementation of a continuous cellular automaton for the simulation of complex, evolving surfaces. *Computer Physical Communication*, 182(3), 628–640.