Scalable CyberGIS Analytics for Solving Complex Environmental, Geospatial, and Social Scientific Problems

Shaowen Wang[1, 2, 3, 5, 6], Wendy K. Tam Cho[1, 4, 6], and Yan Liu[1, 2, 6]

1. CyberGIS Center for Advanced Digital and Spatial Studies
2. CyberInfrastructure and Geospatial Information Laboratory
3. Departments of Geography & Geographic Information Science, Computer Science, Urban & Regional Planning
4. Departments of Political Science, Statistics
5. Graduate School of Library and Information Science
6. National Center for Supercomputing Applications

University of Illinois at Urbana-Champaign

CyberGIS—geographic information science and systems (GIS) based on advanced cyberinfrastructure (CI)—has emerged as a new-generation GIS with broad and significant societal impacts (Wang 2010; Wang *et al.* 2012). It emphasizes the integration of advanced CI, GIS, and spatial analysis and modeling capabilities to enable compute- and data-intensive research and education across a broad range of fields (Wang and Zhu 2008; Wright and Wang 2011). Scaling cyberGIS analytics to efficiently and effectively harness various high-spectrum computing resources is critical for cyberGIS to tame and integrate geospatial big data from numerous sources and to support computationally intensive spatial analysis and modeling for solving many challenging and important environmental, geospatial, and social scientific problems. For example, to assess the impact of global climate change and emergency management, we must analyze the interaction of geospatial patterns and their underlying processes, operating across a wide range of spatial, temporal, and social scales to understand dynamic human-natural systems that often exhibit complex, self-organized, emergent, and adaptive characteristics. Gaining a fundamental understanding of these characteristics will help answer challenging scientific questions such as *how climate and other environmental changes affect the vulnerability and resilience of coupled human-environment systems.*

The National Science Foundation (NSF) CyberGIS software initiative (http://cybergis.org) has contributed significantly to community-driven development of high-performance and scalable cyberGIS analytics (Wang *et al.* 2013). Computational performance analysis is an important step for creating scalable cyberGIS software and tools. In this step, each cyberGIS software component goes through a rigorous process of parallel algorithm development, performance profiling, identification of compute and data bottlenecks, and scaling tests on supercomputers (e.g., the Blue Waters and those provided by the NSF Extreme Science and Engineering Discovery Environment (XSEDE)). This community practice often leads to two desirable outcomes: 1) new parallel computing algorithms and software tailored for geospatial data and computing characteristics; and 2) improved scalability of existing codes. These new or improved components are then integrated into a cutting-edge cyberGIS software environment and deployed on advanced CI for community use.

Redistricting or the process of creating electoral maps is a specific example of scalable cyberGIS analytics. The problem amounts to arranging a finite number of indivisible geographic units into a smaller number of larger areas (i.e. districts). This problem can be defined as a multi-objective *NP*-hard discrete optimization problem with a configuration of objectives and constraints that represent political and geographical requirements and interests. The maps have significant societal impact and have interesting research roots in political science, geographic information science, and operations research. This fundamental computational challenge can be tackled through a scalable parallel genetic algorithm (PGA) library (Liu and Wang 2014). The PGA library,

implemented using the Message-Passing Interface (MPI), is designed to break the global synchronization barrier and provide desirable computation and communication overlap for achieving scalable use of supercomputers. Experiments on Blue Waters demonstrated desirable scalability up to multiple hundreds of thousands of cores. Significant research and development are needed to achieve major advances of computationally scalable algorithms, software, and tools for cyberGIS analytics to harness future high-spectrum computing in efficient and optimal ways.

In addition to computational scalability for cyberGIS analytics, scalable approaches to geospatial big data need to be tightly informed by future high-spectrum computing infrastructure. During the past few decades, geospatial data embedded with geographic references have been collected at an unprecedented pace and with significant complexity as location-based sensors and devices (e.g., environmental sensors, remote sensing satellites, and smart phones) have become commonplace, a trend that is clearly at a nascent stage. Geospatial big data along with innovative capabilities for enabling complex cyberGIS workflows are critical for knowledge discovery in numerous scientific domains (e.g., ecology, emergency management, geography and spatial sciences, geosciences, and social sciences, to name just a few).

The realization of cyberGIS's reach looms as the role of big data continues to enable and support collaborative, interactive, and scalable knowledge discovery through processing and visualizing complex and massive amounts of geospatial data and performing associated analysis, simulation, and visualization. Scientific problem solving based on scalable cyberGIS analytics requires high spectrum computing infrastructure to synergistically provide compute and data capabilities for enabling sophisticated and visual analytical workflows.

Acknowledgements

This material is based on work supported in part by NSF under grant numbers: 0846655, 1047916, and 1429699. Computational experiments used the Blue Waters and XSEDE.

References

Liu, Y. Y. and Wang, S. (2014) A Scalable Parallel Genetic Algorithm for the Generalized Assignment Problem. *Parallel Computing*, DOI: 10.1016/j.parco.2014.04.008.

Wang, S. (2010) A CyberGIS Framework for the Synthesis of Cyberinfrastructure, GIS, and Spatial Analysis. *Annals of the Association of American Geographers,* 100(3): 535-557, DOI:10.1080/00045601003791243.

Wang, S. (2013) CyberGIS: Blueprint for Integrated and Scalable Geospatial Software Ecosystems. *International Journal of Geographical Information Science (IJGIS),* 27(11): 2119-2121, DOI:10.1080/13658816.2013.841318.

Wang, S., Anselin, L., Bhaduri, B., Crosby, C., Goodchild, M. F., Liu, Y., and Nyerges, T. L. (2013) CyberGIS Software: A Synthetic Review and Integration Roadmap. *IJGIS*, 27(11): 2122-2145, DOI:10.1080/13658816.2013.776049.

Wang, S., Wilkins-Diehr, N., and Nyerges, T. L. (2012) CyberGIS – Toward Synergistic Advancement of Cyberinfrastructure and GIScience: A Workshop Summary. *Journal of Spatial Information Science,* 4: 125-148, DOI:10.5311/JOSIS.2012.4.83.

Wang, S., and Zhu, X.-G. (2008) Coupling Cyberinfrastructure and Geographic Information Systems to Empower Ecological and Environmental Research. *BioScience,* 58(2): 94-95, DOI:10.1641/B580202.

Wright, D. J., and Wang, S. (2011) The Emergence of Spatial Cyberinfrastructure. *Proceedings of the National Academy of Sciences,* 108(14): 5488-5491, DOI:10.1073/pnas.1103051108.