The Integration of Vario-scale Data Generalization with Heterogeneous Computing and Graph Database

Xuefeng Guan
LIESMARS, Wuhan University

Contents

- Background
- Past & Current Research
  - Parallelization of Geo-processing
  - Real-time GIS
  - Distributed ABM Simulation
- Future Research Plan
Background

PROFESSIONAL EXPERIENCE

- Lecturer, LIEMSARS, Wuhan University, 2011–Current

EDUCATION

- LIEMSARS, Wuhan University, Wuhan
  Ph.D. in Cartography & Geographic Information Engineering, 2011

- China University of Geosciences, Wuhan
  M.S. in Environmental Engineering, 2005
  B.S. in Environmental Engineering, 2002

Computing Facility

- 65 Computing Nodes;
- 500+ CPU Cores;
- 4 NVidia C2050 GPUs;
- 4TB+ Main Memory;
- 60TB+ Disk Storage;
- 10Gb Ethernet & 16Gb InfiniBand Network;
- Total cost: $0.5 million

Usage:
- Private Cloud;
- High Performance GeoComputation;
- Distributed Geospatial Database;
- High-Concurrency Geospatial Web Services
**Parallel Delaunay Triangulation**


“…..Experiments on a 2-Way-Quad-Core Intel Xeon platform show that ParaStream can triangulate approximately one billion LiDAR points (16.4 GB) in about 16 minutes with only 600 MB physical memory. The total speedup (including I/O time) is about 6.62 with 8 concurrent threads. “

**Concurrent DT Pipelines**

**The Hypergraph-based Scheduling strategy**

PI, “Parallel Task scheduling for Massive Spatial Data Processing with Hypergraph Partitioning”. National Science Foundation of China (41301411), 2014.1~2016.12

Scheduling Objective: Load balance and minimum I/O
The Hypergraph-based Scheduling strategy (cont.)

1. Hypergraph model with tasks & data

\[ T = \{ t_1, t_2, \ldots, t_n \} \text{ and files } F = \{ f_1, f_2, \ldots, f_m \} \]

\[ H = (V, N) \]

vertices and edges correspond to tasks and files

Weight determination of vertices and edges

2. Hypergraph Partitioning

\[ \Pi = \{ V_1, V_2, \ldots, V_K \} \text{ is a } K\text{-way partition of } H \]

\[ \text{cutsize}(\Pi) = \sum c_j(\lambda_j - 1) \]

\[ H' \leq H_{\text{opt}}(1+\epsilon) \quad 1 \leq k \leq K \]

3. Ordering of Task Exe & File Transmission

Heterogeneous Computing

- Heterogeneous computing refers to systems that use more than one kind of processor (typically CPUs and GPUs).

Heterogeneous Computing (cont.)

Contents

- Background
- Past & Current Research
  - Parallelization of Geo-processing
  - Real-time GIS
  - Distributed ABM Simulation
- Future Research Plan

Real-time GIS

The architecture of Real-time GIS

- Offline Time Series Data
- Static Data
- SNS
- Sensor Registry
- SOS
- Import Tool
  - Sensor Msg. Monitor
  - MSG Process
  - Select Sensor
  - Process Algorithm
  - DB Connection

Streaming Data Ingestion System

- Access API
- GI Services
- Integrated Spatiotemporal Index
- Spatiotemporal GIS Database

Time-critical Applications
Spatiotemporal Simulations
Emergency Decision Making
Real-time Visualization

SWE in Real-time GIS

- Sensor Web
- SensorML
- SOS Service
- Internal Database

OGC SOS Services in the real-time GIS platform act as:
- Sensor data encapsulation and sharing;
- Query and access interface
The proposed spatial-temporal data model

The prototype of Real-time GIS
The prototype of Real-time GIS

The ST-Hash indexing method

NoSQL database: Oracle NoSQL, MongoDB, Hbase, ....
Spatially supported by GeoHash;
Does not support direct spatiotemporal indexing.

GeoHash (2D) \( (x, y) \rightarrow \text{string} \)

ST-Hash (Augmented 3D) \( (x, y, t) \rightarrow \text{string} \)

Also supported by the 6th research project of WHU-Oracle Spatial Database Innovation Center, 2014.9~2015.9.
The ST-Hash indexing method (cont.)

- Longitude: 0001110001
- Latitude: 1001110001
- Time: 0110100111

ST-Hash Encoding

Value: 17 39 62 1 15
String: `R e + B P`

ST Range Query

\[ Q(E_x, E_y) \rightarrow \{ S_k = (p_1, p_2, \ldots, p_k) \} \]

The ST-Hash indexing method (cont.)

Range query performance comparison

Scalability over different data sizes
Distributed ABM Simulation

Agent-based modeling has been proved an important technology to understand these dynamic geospatial phenomena in a bottom-up manner.

- As the scale of research problems and the complexity of designed models continue to increase, agent-based modeling are often computationally demanding and inevitably faced with a series of computational issues.
- Integration of dynamic geospatial phenomena information can update the environment in real-time and enhance the accuracy of simulation outputs.
- Nearly all the parallel simulation frameworks focus on model description and performance improvement, efficient realistic visualization of parallel ABM simulations needs to be issued.

Distributed ABM Simulation (cont.)

System architecture

Contents

- Background
- Past & Current Research
  - Parallelization of Geo-processing
  - Real-time GIS
  - Distributed ABM Simulation
- Future Research Plan
Background

- Vario-scale Map Generalization
  - Polygon Split/Merge, Boundary simplification,........ Time-consuming
  - SSC generation, mixed-scale visualization ,........ Real-time

Objective

- The Integration of Vario-scale Data Generalization with Heterogeneous Computing and Graph Database

1) Use heterogeneous parallel computing to accelerate polygon collapse/merge and boundary simplification in map generalization.

2) Use mature graph databases to store and access massive tGAP dataset efficiently.
Method 1

- Decomposition considering data distribution: Field tree, quad-tree
- Task encapsulation with Intel TBB and NVidia CUDA
- Task scheduling and load balance

Method 2

- The database selection, storage schema, index building, and query interface will be designed here.
Research Plan

2015.09~2015.10---1 month
- Learn the details about the vario-scale map generalization (e.g. smooth tGAP structure, BLG tree, Fieldtree, SSC, map slicing);

2015.10~2016.02---5 months
- Algorithm analysis and parallelization design,
- Coding and algorithm implementation,
- Debug/experiment with massive real data;

2016.03~2016.07---5 months
- Database comparison and selection,
- Storage schema design and index building,
- Query interface design and implementation,
- Tune/experiment with massive real data;

2016.08~2016.09---1 month
- Work review and technical report writing

Thanks for your attention!

guanxuefeng@whu.edu.cn