GIS PROCESSING ON COMPUTE CLUSTERS

by

Nathan Thomas Kerr

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

ARIZONA STATE UNIVERSITY

August 2009

GIS PROCESSING ON COMPUTE CLUSTERS

by

Nathan Thomas Kerr

has been approved

August 2009

Graduate Supervisory Committee:

Dr. Daniel Stanzione, Chair
Dr. Robert Pahle
Dr. Yi Chen

ACCEPTED BY THE GRADUATE COLLEGE

ABSTRACT

Abstract goes here

Dedication goes here

ACKNOWLEDGEMENTS

Acknowledgments go here.

TABLE OF CONTENTS

Page

Chapter 1

Introduction

GIS processing is important

1.1 Current methods

Desktop GIS

Database GIS

1.2 Significance

Limitations of single processor implementations

1.3 Attempted solutions

Parallel DB

Parallel GRASS

Problems with desktop programs on clusters

1.4 Needed solution, cluster programming model

batch

mpi

scalable architecture

spmd

Thesis statement goes here

How this document goes about supporting the thesis statement.

Chapter 2

Related Work

2.1 Desktop GIS

ArcInfo, QuantumGIS, GRASS GIS

2.2 Database GIS

PostGIS, Oracle Spatial, Client-Server Paradise

2.3 GIS libraries

JTS, GEOS

2.4 Parallel DB

Paradise

2.5 Parallel GRASS

Other converted programs?

2.6 MRGIS MapReduce GIS

Chapter 3

Requirements

batch mode processing

data distribution options

standard geospatial operations

scalable, high performance

dataset centric

3.1 easy to use to form new operations

Sample program:

initClusterGIS

loadData (distributed/replicated)

process (user code here)

saveData (distributed/replicated)

finalizeClusterGIS

Chapter 4

Design/Implementation

How to fulfill requirements

Architecture of solution

Split dataset across tasks (MPI I/O)

Each task works on their own part

MPI to communicate if needed

Chapter 5

Experimental Setup/Design

How to show how well requirements were filled

5.1 Operation set

These operations are representative of problem space.

Data access based: parallelization is based off data decomp, so operations are too.

OGC standard requirements - survey of methods requiring 1, 2, or more data points.

5.2 Dataset Description

Full datasets; 34k employers, 1.2m parcels

sub datasets, how to generate from full

listing of datsets used

5.3 Hardware setup

Saguaro

5.4 implementations

5.4.1 PostGIS

5.4.2 Hadoop Prototype

5.4.3 ClusterGIS

Chapter 6

Results

6.1 Performance Analysis

vary procs, maintain data

vary data, maintain procs

vary data, vary procs

6.2 Comparisons

PostGIS as baseline

PostGIS to Hadoop

Hadoop to ClusterGIS

PostGIS to ClusterGIS

Chapter 7

Recommendations

Synthesis

7.1 Applicable Problem Spaces

Analysis of problem types that are good/bad for this approach

Chapter 8

Conclusion

8.1 Future Work

additional decomposition methods, combined with alternative MPI communicators

addition of preprocessing methods (indexing, etc)

chunking of replicated dataset