A TERAGRID-ENABLED DISTRIBUTED DISCRETE EVENT AGENT-BASED EPIDEMIOLOGICAL SIMULATION

Douglas J. Roberts
Diglio A. Simoni

Bioinformatics Program
RTI International
3040 Cornwallis Road
Research Triangle Park, N.C. 27709 U.S.A.

ABSTRACT

We discuss design issues related to the transformation of a mature Agent-Based Model (ABM) for computational epidemiology into a "grid-aware" version. EpiSims is a distributed discrete event ABM that has been in production for nearly a decade. Working under a grant from the National Science Foundation and the NIH (NIGMS) funded MIDAS project, we are reengineering EpiSims to run as a single job on multiple Linux clusters on the NSF TeraGrid.

1 INTRODUCTION

EpiSims is a distributed memory ABM that uses a discrete event update engine. EpiSims was designed to help us understand the spread of contagious diseases in urban populations. The system has been used to study the spread of various naturally occurring pathogens such as smallpox, bubonic plague, pneumonic plague, and various influenza strains, including H5N1 avian bird influenza. An example of the type of result produced by EpiSims is shown in Figure 1, which shows a hypothetical spread of the H5Ni virus in the Chicago metropolitan area. That particular study used an agent cardinality of [x million] individuals.

A typical EpiSims run will produce infection curves like those shown in Figure 2.

EpiSims was designed to run on distributed memory Linux clusters, and because it uses a discrete event update engine it has stringent synchronization requirements. Simulation time is advanced when each event queue running on every processor of the cluster pops its events off of the top of its own queue. We implemented a synchronization algorithm to maintain all the compute nodes in a cluster tightly coupled in time in order to prevent the possibility of events occurring in the past as individual person agents migrate between compute nodes The algorithm is at http://www.trnmag.com/Stories/2003/021203/Scheme_smooths_parallel_processing_021203.html.

2 GRID DESIGN ISSUES

With the introduction of grid computing environments, such as the NSF-funded TeraGrid (Figure 3), opportunities now exist for running much larger EpiSims simulations as a single job on multiple clusters on the TeraGrid. However, the computational heterogeneity of the TeraGrid resources introduces complexities for distributed discrete event-driven ABMs. Some of the most important issues that need to be addressed include global parallel I/O, problem partitioning, and synchronization methods. In particular, Grid computing introduces a new distributed discrete event (DES) synchronization requirement. We now identify two types of synchronization that must be performed in order to support efficient computation on the Grid. The on-cluster synchronization requirement remains tight, but we now have an inter-cluster synchronization requirement that is much looser (Figure 4).

Assuming that the social network input data has been appropriately partitioned such that regions of highly-connected network have been identified and assigned to the individual clusters in the grid configuration, we now encounter a loose inter-cluster synchronization requirement whose purpose is to ensure that person-agent migrations between network segments managed by different clusters do not create time order event errors. This inter-cluster synchronization requirement is less rigorous than the on-cluster requirement, if the network problem has been partitioned properly such that highly inter-connected regions are assigned to the separate clusters in a run configuration.

A key element of the Grid version of EpiSims is a tool called MPICH-G2 (http://www3.niu.edu/mpi/), which is a "grid-aware" version of MPI. With MPICH-G2 it is possible to send an MPI message from a compute node on a TeraGrid cluster to a compute node on any other TeraGrid cluster in the run configuration. MPICH-G2 will make it possible to extend the on-cluster synchronization algorithm used in EpiSims to also incorporate the new inter-cluster synchronization requirements.

3 SUMMARY AND FURTHER RESEARCH

This project is still a work in progress. Once the coding phase is complete, we will develop an experimental design that will allow us to perform parameter sweeps for identifying which systems parameters control system performance. Included in the parameter list are:

1. Inter-cluster synchronization time barrier value
2. Travel density between population segments running on different clusters
3. Global I/O bandwidth limitations
4. Initial population network partitioning between clusters

AUTHOR BIOGRAPHIES

DOUGLAS J. ROBERTS is a senior research scientist at RTI International where he works on HPC applications. Prior to RTI, he spent 20 years at Los Alamos National Laboratory where he worked on agent based models and HPC design implementations.

DIGLIO A. SIMONI is a senior computational scientist at RTI International. He began his career working on Hypercube MIMD machines with Geoffrey Fox at Caltech/JPL. For the past 20 years he has been involved in various modeling efforts in in computational seismology and geophysics, computational electrodynamics, computational magnetohydrodynamics, computational fluid dynamics, computational neuroscience, computational linguistics and more recently in computational epidemiology.

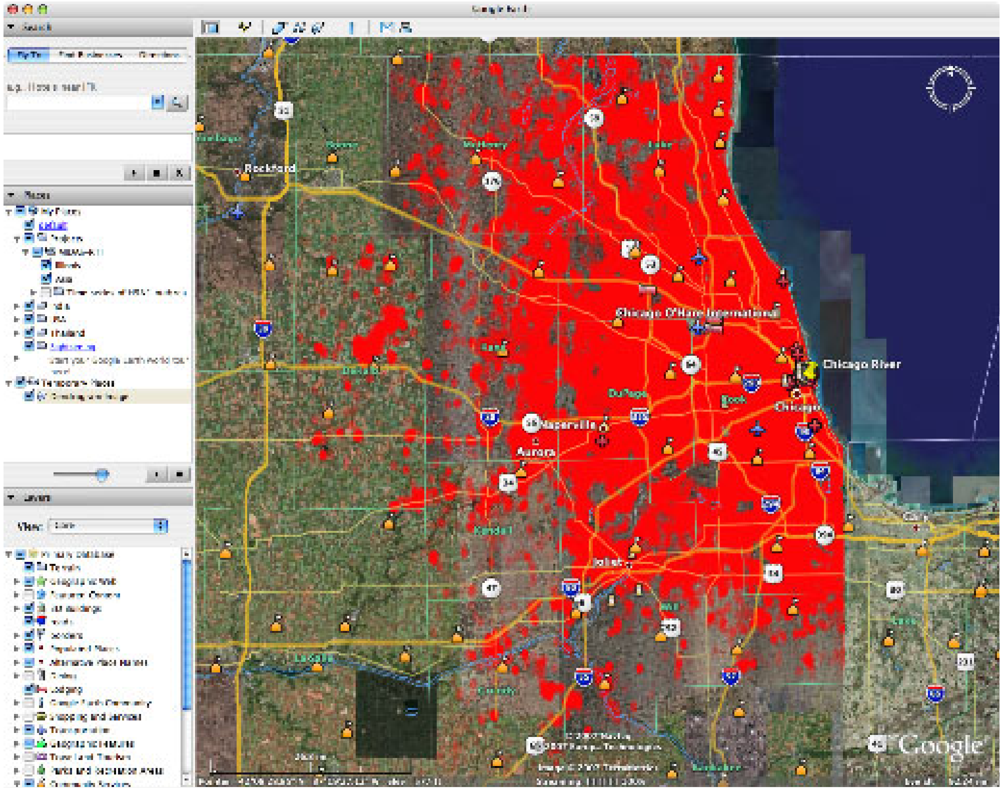

Figure 1. Chicago H5N1 Infections

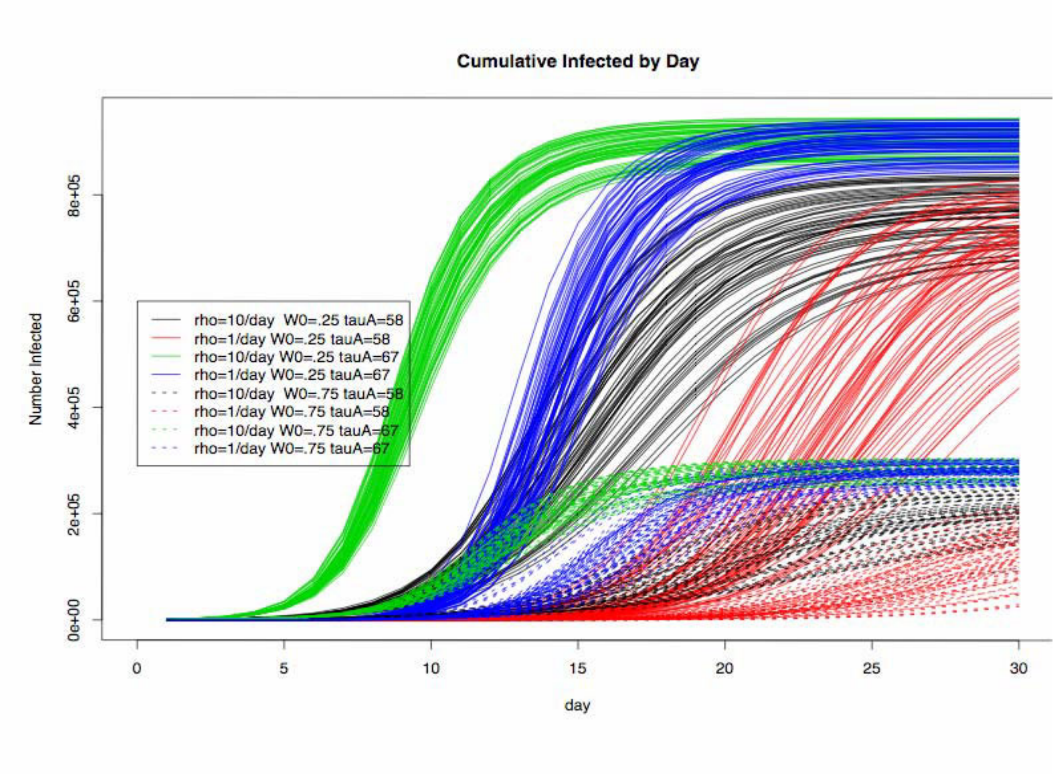

Figure 2. Counts of New Infections

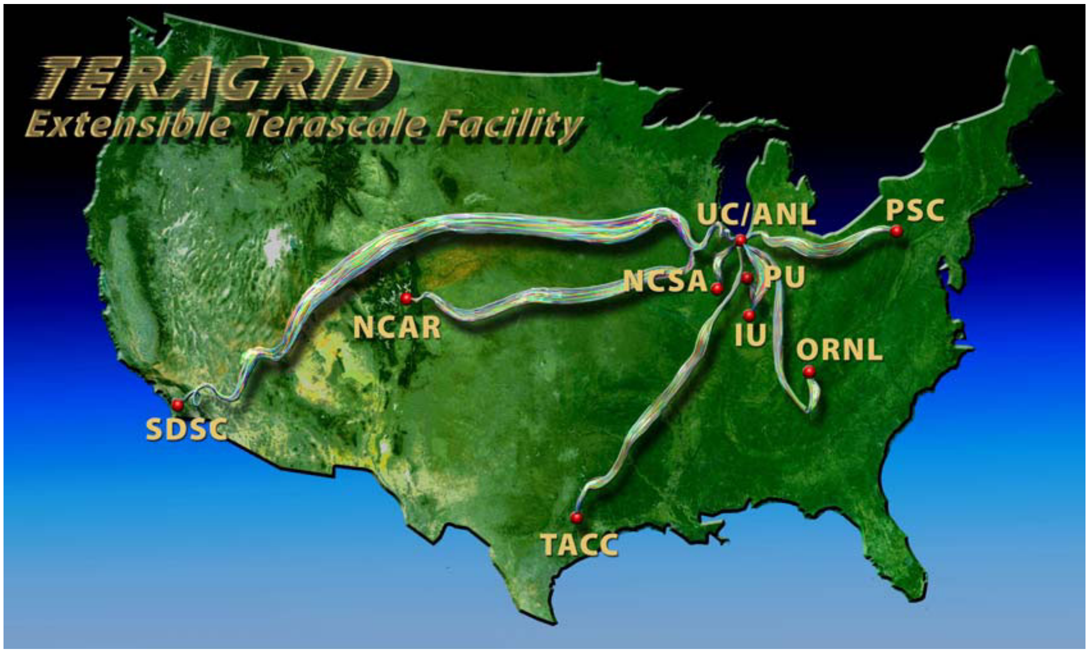

Figure 3. NSF TeraGrid

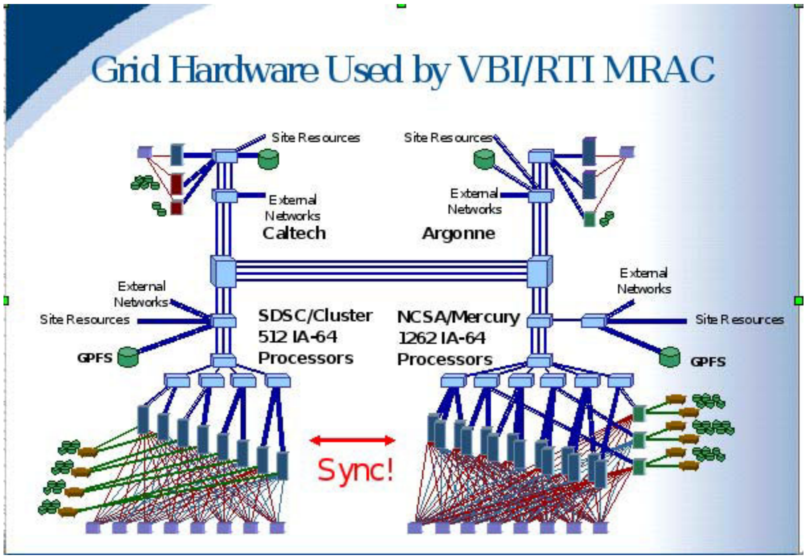
Grid Hardware Used by VBI/RTI MRAC

Figure 4. Inter and Intra Cluster Synchronization Messaging