Scalable Epidemiological Workflows to Support COVID-19 Planning and Response

Dustin Machi*, Parantapa Bhattacharya*, Stefan Hoops*, Jiangzhuo Chen*, Henning Mortveit*, Srinivasan Venkatramanan*, Bryan Lewis*, Mandy Wilson*, Arindam Fadikar‡, Tom Maiden†, Christopher L. Barrett* and Madhav V. Marathe*

* University of Virginia
‡ Argonne National Laboratory
† Pittsburgh Supercomputing Center

IPDPS 2021
The Covid-19 Pandemic

- The most significant epidemic event since the 1918 influenza epidemic
- Over 164 million confirmed cases worldwide
- Over 3.39 million confirmed deaths worldwide
- Estimated economic burden of over 9 trillion US dollars
Supporting Policy Makers

- We have been supporting decision makers March 25, 2020
- We provided weekly forecasts to:
  - Center for Disease Control and Prevention (CDC)
  - US Department of Defense (DoD)
  - Virginia Department of Health (VDH)
  - State Hospital Referral Regions (HRR)
  - University of Virginia
- Answering what if questions
  - Non-pharmaceutical Interventions (NPIs)
  - Mask mandate, school closures, local shutdowns, ...
Contributions and Significance

- A novel high performance computing (HPC) approach for executing epidemiological workflows
  - High resolution agent based models
  - Realistic representation of national scale social contact networks
    - 300 million nodes, 7.9 billion edges, 50 states + Washington DC
  - 5000–17,900 simulations per night
    - County level forecasts for 3140 counties
  - Split across two HPC clusters
We built very detailed Covid-19 disease models utilizing publicly available information from multiple sources.
EpiHiper: A distributed epidemic simulator

- An agent based discrete time simulation model
- Inputs: Disease model + Social contact network
- Provides custom domain specific language for programming NPI scenarios:
  - Voluntary home isolation, school closure, stay-at-home, pulsating shutdowns, partial reopening, custom vaccination schedules ...
- A distributed memory program written in C++/MPI
- Contact network is partitioned onto MPI ranks
- Shared inputs served via PostgreSQL database
Rivanna and Bridges: A multi-cluster setup

- **Rivanna**: Home cluster at University of Virginia
  - 50 nodes (40 CPU cores, 384 GB RAM per node)

- **Bridges**: Remote cluster at Pittsburgh Supercomputing Center
  - Limited access (10pm—8am every night)
  - 720 nodes (28 CPU cores, 128 GB RAM per node)

- Data transfer via Globus

- Both clusters used Slurm for scheduling
A typical timeline of tasks involving human efforts.

Orange boxes are automated.
Combined workflow: End to end timeline

Calibration Phase

Day 0

Data Transferred to Remote Supercluster

County-level Epidemic curves (Raw output: 5-110GB/cell)

Projection and Intervention Analysis

Day 3

Input Data (100MB-8.7GB)

Static Social Network Data (2 TB)

(One Time transfer)

UVA Rivanna Cluster Compute Resources

PSC Bridges Cluster Compute Resources

Summary output: 30-200MB/cell

Input Data (100MB-8.7GB)

Static Social Network Data (2 TB)

(Stored from Calibration)

UVA Rivanna Cluster Compute Resources

PSC Bridges Cluster Compute Resources

Summary output: 30-200MB/cell

UVA Rivanna Cluster Compute Resources

PSC Bridges Cluster Compute Resources

County-level Epidemic curves (Raw output: 5-110GB/cell)
Calibration Workflow

- Ranges for disease parameters
- Latin Hypercube Design
- $k$ model configurations
- Ground truth on day $t$
- State-specific county-level incidence curves
- Launch population trait databases
- Partition contact networks across distributed processes
- Execute System A model simulations
- Calibration using GP emulator
- Set of $k$ model configurations
- Compute Outcome using System A
- Final set of $k$ model configurations

Work on home cluster

Work on remote supercluster
Prediction Workflow

Final set of k model configurations

Construct various intervention scenarios (including no intervention scenario)

Launch population trait database

Partition contact network across distributed processes

Execute System A simulations

Detailed Transmission Trees from each simulation for all counties for each scenario

Summarization and organization

National County-level daily projections sent to home cluster for further analysis

- Work on home cluster
- Work on remote supercluster
Economic Workflow

1. **Ground truth on day $t$**
   - Data on disease incidence

2. **State-specific county-level incidence curves**
   - Calibration of disease model
   - NPIs with parameter sweep

3. **EpiHiper configurations of factorial design**
   - Launch 51 databases for 51 states

4. **Partition 51 contact networks across distributed processes**

5. **Execute EpiHiper based experiment with factorial design**

6. **Aggregate individual level simulation output to industry level**

7. **Join aggregate output with synthetic data**

8. **Compute medical costs due to hospitalizations, deaths**

9. **Compute GDP reduction with IO analysis**

10. **Combine to get total impact**

**Work on UVA Rivanna Custer**

**Work on PSC Bridges Cluster**
Case Study 1: Medical Costs of Covid-19

Medical costs of keeping the US economy open during COVID-19

Jiangzhuo Chen, Anil Vullikanti, Stefan Hoops, Henning Mortveit, Bryan Lewis, Srinivasan Venkatramanan, Wen You, Stephen Eubank, Madhav Marathe, Chris Barrett & Achla Marathe

Scientific Reports 10, Article number: 18422 (2020)  Cite this article

Abstract

We use an individual based model and national level epidemic simulations to estimate the medical costs of keeping the US economy open during COVID-19 pandemic under different counterfactual scenarios. We model an unmitigated scenario and 12 mitigation scenarios which differ in compliance behavior to social distancing strategies and in the duration of the
## Scale of Simulations

Large number of simulations are needed to explore the parameter space and to generate confident predictions for decision support purposes.
Workflow Orchestration

• Given:
  – A set of simulations to run
  – A set of compute nodes to run them on

• Objective:
  – Generate job ordering for Slurm
  – Minimize the total run time

• The problem can be mapped to
  – 2D Bin packing problem
  – a variant of the coloring problem (r-relaxed-coloring)

• Heuristics tested
  – First-fit decreasing time with database access constraints (FFDT-DC)
  – Next-fit decreasing time with database access constraints (NFDT-DC)

• Metric
  – System utilization

• Heuristic performance
  – FFDT-DC performs better (96.6% median utilization)
  – NFDT-DC performs worse (55.5% median utilization)
Conclusion

• We developed a novel HPC oriented workflow in order to support planning and response to pandemics such as Covid-19.
  – We used two geographically separated supercomputing facilities
  – Incorporated daily county-level surveillance data and policy data
  – National and high resolution agent-based simulations

• Real-time data driven high resolution epidemics science at national scale is indeed possible.
Thank You!