Design and Analysis of a Cloud-based Epidemic Simulation Framework

Polpat Sukcharoen¹, Sarunya Pumma¹, Onpreeya Mongkolsermporn¹, Tiranea Achalakul¹, Xiaorong Li²

1 Department of Computer Engineering, Faculty of Engineering
King Mongkut’s University of Technology Thonburi
2 Institute of High Performance Computing, A*STAR

golf.polpat@gmail.com, sarunya.pumma@gmail.com, iamonpreeya@gmail.com, tiraneae@cpe.kmutt.ac.th,
lixr@ihpc.a-star.edu.sg

Abstract—In this paper, we present a cloud-based epidemic simulation framework which is designed for simulating the spread of epidemic diseases globally. We enhance the SEIR, a popular global equation model for infection diseases study. The original SEIR model takes long to generate results. Hence, we modify SEIR by adding the level of severity into each of the epidemic state. Therefore, the time period is used instead of the rate of occurrences in the simulation which dramatically reduces the search space in parameter estimations while maintains the same level of accuracy. Moreover, we enable the modified SEIR model to be parallelizable using a multiple loop decomposition method. It partitions and distributes the workload of the epidemic simulation onto multiple compute resources concurrently. Experiments have been conducted using multiple virtual machines in a cloud environment and the results show that the performance can be improved substantially.

I. INTRODUCTION

The spread of infectious diseases has been one of the most important problems and a long-term issue to our society. Thus, the proper control of the spread should be made. The research on epidemic situation helps to provide a good understanding of the growth and spreading behaviors of those infectious diseases. The main purpose of epidemic simulation programs is to predict the risk of infection by simulating the epidemic behavior using the historical data. Mathematical models for epidemic forecast are formulated to perform the simulation.

There are a number of epidemic works that study the spread of epidemic around the world. Goedecke and team [1] are epidemiologists and developers who have developed the Global Equation Model (GEM) which applied SEIR model with dynamic population factors and air transportation. Kermack and McKendrick [2] have established the basic foundations of mathematical theory for epidemics. The basic fundamental epidemic model had been formulated and developed in order to understand the trend of infectious disease spread. Keeling and Danon studied the SIR model [3], which is the classic epidemic model that is widely used and currently. The proposed mathematical modeling of infectious disease described the relationship of individuals in a population. Individuals are divided into six basic classes: S, susceptible; E, exposed; I, infective; Q, quarantined; J, diagnosed; R, recovered. Different models may describe only the relationship of some important classes. For example, SIR model described the relationship between susceptible, infective, and recovered individuals.

SIR model is the fundamental epidemic model which categorizes the characteristic of individual into three different states; susceptible, infective, recovered. Susceptible state shows that an individual is neither being infected or able to transmit the disease.

SEIR model [4] was developed from SIR model by adding one more epidemic state: exposed. Similar to SIR model, the individual in SEIR moves from a state to another state. However, instead of moving from susceptible to infectious states, SEIR model also considers exposing of disease. Susceptible individual will be infected by infectious individual, but will not immediately become infected. When infected, the susceptible individual will move into exposed state and after a period of time, one will become infectious. The difference between exposed and infectious states is that exposed individual cannot transmit disease to susceptible individual.

In this paper, we adapt both GEM and SEIR models in order to create the epidemic simulation software that can be executed on a cloud-based environment. The framework design is described in the following section.

II. THE OVERVIEW OF THE SOFTWARE DESIGN

This section provides an overview of our software design. The software architecture is three-tier. User interface and visualization are separated from data repository and business logic. In other words, all components are independent of one another. They work together through services producing and consuming as depicted in a UML: component diagram in Figure 1.

From Figure 1, Epidemic Forecast software consists of 5 main components: 1) Web Portal: this web portal allows user to enter and configure the simulation period for data processing. Users can access the application from anywhere via the Internet. 2) Shared repository: this component stores all relevant files for simulation and visualization. 3) Epidemic simulation (modified-SEIR): the real-time data processing using the most up-to-date data in the repository will be performed immediately after user starts the execution from the portal. SEIR is, the mathematical model, used for prediction.
4) Could Scheduling: when parameter estimation is performed, the scheduler will divide the estimation process to be run on cloud [5] in order to increase the execution performance. 5) Visualization: this component renders the graphical result from the output text file obtained from data processing and returns the result to the epidemic portal for displaying. The simulation paths of the disease outbreak can then be visualized on a 2D world map as an animation with the color highlighted on each area to indicate the severity of the disease. In this paper, we emphasize the design of our epidemic simulation and cloud scheduling processes. Details are given below.

III. EPIDEMIC SIMULATION AND ITS ALGORITHMS

In our work, the simulation model consists of 2 main steps: parameters estimation and SEIR-based simulation. Parameters estimation is the process of calculating the reasonable input parameters with the least error for simulation. Parameters estimation determines the accuracy and reliability of the simulation. The SEIR-based simulation is a core module which will simulate the disease spreading behavior by using the estimated parameters as input. The predictive data will be written to file in a CVS format. The output file consists of the number of individuals in epidemic states for each city.

When implementing the process described above, we have slightly modified SEIR model to improve the performance of the simulation. The estimation algorithm, in particular, has been re-designed. The details of the algorithms are proposed in subsections A and B.

A. Parameters Estimation Algorithm

The estimation will iteratively simulate the spread of disease or the changes of epidemic states during simulated time according to the generated set of parameters: contact rate ($\lambda$), exposed period ($ep$), infected period ($ip$), mortality rate ($d$), and initial number of infectious ($iniE$). The different simulated outputs will be produced according to different set of parameters. Each parameter has the upper and lower bounds called max and min respectively. The max and min of each parameter will be entered in order to scope the size of a search space of each parameter. The focused output is the number of individuals in infected state that will be used to compare to the actual number of cases obtained from World Health Organization. The differences between the simulated number of infected individuals and actual numbers obtained from WHO [6] are considered as simulation error. The set of parameters that produce the least error is the optimized set of parameters for a disease of interest. Finally, the optimized set of parameters will be used as input data in the modified SEIR-based simulation program.

In addition, parameters estimation requires a high computational power in order to obtain a good accuracy within a reasonable amount of time. Therefore, cloud computing is adopted to improve the performance of the estimation.

B. SEIR-based Simulation Algorithm

The classic SEIR model is adjusted in order to improve the performance of simulating epidemic forecast. Originally, SEIR model is described by equations as shown in (1), (2), (3), and (4).

$$S(t+1) = S(t) - \lambda I(t)/N(t)$$  \hspace{1cm} (1)

$$E(t+1) = E(t) + \lambda R(t)/N(t) - \gamma E(t)$$  \hspace{1cm} (2)

$$I(t+1) = I(t) + \gamma E(t) - \delta I(t)$$  \hspace{1cm} (3)

$$R(t+1) = R(t) + \delta I(t)$$  \hspace{1cm} (4)

In our work, modifications are made to SEIR model to improve the performance of the simulation as well as to make the model more realistic. We modified the model to elaborate on the exposed, infected, and recovered states. We extend the parameters to differentiate the severity of infection. In our modified model, exposed, infected, and recovered states will be further divided into levels depending on number of exposed and infected days respectively. By dividing the each group
into levels, the model can better emulate the actual events occurred during an epidemic. Moreover, during parameters estimation, number of days (period) is used instead of rate of occurrences to trigger the state transitions. In other words, integer is used instead of floating points during the estimation step. Based on the mentioned concept, two modifications are introduced in our framework as described below.

1) State Transition

Equation (2)-(4) demonstrates the state transitions using rates. For example, the rate of exposed individual that changes to being infected. In our modified version, the ‘time period’ or number of days is used instead of rates during simulation. The number of exposed, infected, or recovered individuals will be divided into sub-groups depending on number of days after entering a state. In order to reflect such change in the simulation, equation (2) has been replaced by (5) and (6), (3) by (7) and (8), and finally (4) is replaced by (9). For example, equation (5) describes the change of number of exposed individuals being exposed for zero days, which is the number of susceptible being infected at time $t+1$. Equation (6) describes the number of exposed individuals being exposed for $τ_1$ days, which generally is the number of exposed individuals being exposed for $τ_1-1$ days. Equation (7)-(9) can be explained in a similar fashion.

$$E(t+1,0) = \frac{λ(t)}{N(t)}$$  \hspace{1cm} (5)

$$E(t+1,τ_1) = E(t,τ_1-1)$$  \hspace{1cm} (6)

where $τ_1$ is less than exposed period $(ep)$

$$R(t+1,0) = R(t,ep-1)$$  \hspace{1cm} (7)

$$R(t+1,τ_2) = R(t,τ_2-1)$$  \hspace{1cm} (8)

where $τ_2$ is less than infected period $(ip)$

$$R(t+1) = R(t) + I(t,ip-1)$$  \hspace{1cm} (9)

2) Integrate Traveling Model to Handle Multiple Regions

The global traveling can be applied to use together with epidemic model. Therefore, the spread of disease across regions is possible. The pattern and equations of global traveling is adapted from “Stochastic Equation-Based Model of a Global Epidemic” [1]. The total number of people travels per day from city $i$ to city $j$ is described by $Ω_{ij}$. Equation (10) illustrates the number of susceptible individual who migrates from city $i$ to city $j$. Thus, the number of susceptible in city $i$ will be reduced as described in Equation (10). Vice versa, the number of susceptible in city $j$ will be increased proportionally. The traveling of exposed and infected individuals will also be calculated in sub-states indicated by exposed and infected period in a similar fashion. Finally, the traveling for each state of each city will be added as shown in (10) to (17).

$$S(t+1) = S(t) - Ω_{ij}S(t)N(j)$$  \hspace{1cm} (10)

$$S(t+1) = S(t) + Ω_{ij}S(t)N(j)$$  \hspace{1cm} (11)

$$E(t+1) = E(t,τ-1) - Ω_{ij}E(t, τ)N(j)$$  \hspace{1cm} (12)

$$E(t+1) = E(t,τ-1) + Ω_{ij}E(t, τ)N(j)$$  \hspace{1cm} (13)

$$I(t+1) = I(t,τ-1) - Ω_{ij}I(t, τ)N(j)$$  \hspace{1cm} (14)

$$I(t+1) = I(t,τ-1) + Ω_{ij}I(t, τ)N(j)$$  \hspace{1cm} (15)

$$R(t+1) = R(t) - Ω_{ij}R(t)N(j)$$  \hspace{1cm} (16)

$$R(t+1) = R(t) + Ω_{ij}R(t)N(j)$$  \hspace{1cm} (17)

In our work, we implement the simulation model utilizing our modified SEIR with global traveling. The input data is the set of parameters for a specific disease: $λ$, $ep$, $ip$, and $d$, and first infected city number ($iniCity$), $iniE$, simulation period ($simulatedTime$) which is started from 0. Population number consists of number of population for 155 cities and traveling data consists of the traveling number from each city to the other cities.

Firstly, the parameters from input data will be set. The initial population number in each epidemic state for every city will also be initialized according to population data (population number, city name, city number, latitude, and longitude of 155 cities around the world), $iniCity$, and $iniE$. The population is assumed to be initialized in susceptible state, except for the first infected city that a certain number of people will initially be in exposed state. For each time $t$, the number of individuals in each epidemic state of every city will be calculated and updated according to our modified-SEIR model. After the numbers of individuals in epidemic states in each city are updated, the traveling between each city will be computed. The simulation will finish when time $t$ reaches the $simulatedTime$ and output will be generated for later display. The output will be the numbers of individuals in each epidemic state at time $t$ of every city.

IV. Scheduling Method On The Cloud

Since parameter estimation takes extremely long time to execute, we consider further optimizing the data processing by partitioning and distributing the workload in a cloud environment. With the ability to provide a large amount of computer resources, cloud allows the adoption of a large number of virtual machines, which enables the parallelization of a program executing on virtual machines. Furthermore, cloud allows the number of virtual machines used to expand or shrink on demand instantaneously.

We use Xen Cloud Platform (XCP) [7], an open-source cloud computing platform, to build a private cloud particularly for running the simulation. All virtual machines (VMs) have the same computational power and execution cost. All VMs are mounted on the shared repository. Therefore, files transmission between machines will be done by reading and writing files on the shared directory.

$$R = \left\lfloor \frac{λ_{min} - λ_{step}}{λ_{step}} \right\rfloor + 1 \% N$$  \hspace{1cm} (18)

$$T_{m} = \begin{cases} \left\lfloor \frac{λ_{min} - λ_{step}}{λ_{step}} \right\rfloor + 1 \% N & ; M < R \\ \left\lfloor \frac{λ_{min} - λ_{step}}{λ_{step}} \right\rfloor + 1 \% N & ; M ≥ R \end{cases}$$  \hspace{1cm} (19)

where $R$ is number of remaining jobs from equally dividing

$N$ is number of VMs

$M$ is machine number (0 to N-1)

$T_{m}$ is number of tasks on machine M (consider only $λ$)

$λ_{step}$ is the incremental step of $λ$
We parallelize parameter estimation by multiple loop decomposition and distribute the workload onto multiple VMs. All VMs are given the machine number starts from 0 to N-1 where N is a number of machines. We split only the outermost loop which is the \( \lambda \)-loop. Each VM will receive a similar number of tasks. Equation (18) and (19) are used for task decomposition. Decomposed tasks are then assigned to processors using a simple indexing method.

In summary, once the parameter estimation is performed, the relevant data and input data will be fetched from the shared repository. The scheduler will calculate a number of tasks to be run on each VM using multiple loop decomposition (Eq 18 and 19) and create the schedule. All VMs will execute the assigned piece of iteration simultaneously based on the schedule. After all tasks are executed, the output file of estimated parameter will be retrieved.

V. RESULTS AND DISCUSSION

The result shows the performance and accuracy of our estimation method by comparing our estimated basic reproduction number (contact rate/exposed rate or contact rate*exposed period) with the original SEIR model. Influenza (H1N1-2009) was estimated to find the optimized basic reproduction number through mathematical methods.

![Figure 2. Estimated number of influenza (H1N1-2009) cases.](image)

Figure 2 shows the result from our modified SEIR which is compared to the result from original SEIR and actual data. The data used in the simulation is Japan’s H1N1 that were collected during the 1st week of 2009 to the 28th week of 2010. The circle dots, dash line and bold line indicate actual number of influenza cases, simulated number of cases from an original SEIR and simulated number of cases from the modified-SEIR, respectively.

The parameters used in the model for H1N1 is estimated by our estimation program. The errors from both original and modified models, measured by mean square error, are also similar with less than 1% difference. However, the significant difference is in the performance of the estimation processes. Our estimation method uses the time period (exposed and infected periods) instead of rate of occurrences (exposed rate and recovery rate), which reduces the parameter search space for estimation substantially. Although the search space is reduced, the accuracy of modified SEIR model is still reliable (See Figure 2). The estimated basic reproduction number from modified SEIR and original SEIR are 1.15 and 1.21. However, the parameter estimation time is 20 minutes using modified SEIR model and almost 18 hours using the original SEIR model in the same computational environment.

We note that our estimation method involves the vast number of iterations. Hence, we attempt to further reduce the simulation time by using cloud resources as described in Section V. From the result, the estimation program normally runs for about 40 minutes. However, running the estimation program with two virtual machines, the time is reduced to 20 minutes. This shows that using two virtual machines can reduce the computation time by 50%. Moreover, the more number of virtual machines, the faster the estimation process. Using our cloud-based SEIR, the parameter estimation and simulation time can be reduced greatly.

VI. CONCLUSION

In this paper, we designed and developed a cloud-based epidemic simulation framework which enables and further optimizes the epidemic simulation in a highly scalable manner. We presented the software architecture of the epidemic forecast program and proposed a new algorithm which enhances the execution performance of the SEIR model. The software implementation provides user interfaces that allow the simulation of a variety of infectious diseases by extracting the estimation parameters of different diseases and applying to the simulation models. Moreover, our proposed framework can further improve the performance of the epidemic simulation by allowing further partitioning and distributing of the workload to multiple virtual machines in a cloud environment. We studied the performance of the algorithm in the cloud and the results show that our proposed framework can reduce the simulation time substantially.

REFERENCES