HemeWeb: Blood flow simulation in the cloud using Docker

Steven Steven

Master of Science

School of Informatics

University of Edinburgh

2016

Abstract

In this dissertation, we developed HemeWeb, an alternative interface to HemeLB, a complex computational fluid dynamic tools. This web application is developed to enable domain experts to use HemeLB simulation software without worrying about configurations of the software and the simulation. To measure the system, we run a usability evaluation and a performance benchmark. On the usability evaluation, respondents have to run a HemeLB simulation and reproduce past simulation, in which after both scenario, they will be asked to answer some questions. Respondents generally find the system subjectively satisfying to use and usable in doing the tasks given. On the performance benchmark, we compare the performance of HemeLB simulation on HemeWeb using cloud platforms compared to dedicated HPC infrastructure. We found out that HemeWeb perform worst and costs more while offering more flexibility in running the simulation.

Acknowledgements

I would like to thank both of my supervisors, Dr. Miguel O. Bernabeu and Dr. Rupert Nash, for allowing me to work on this interesting project. I would also thank them for helping me throughout the dissertation period with their numerous advice, guidances, and supports. Without them, this work would not have been possible.

Next, I would like to extend thanks to my dearest friends in Holyrood Residence Hall - Holyrood North. Alex, Henrique, Linn, Susanne, Tami, Eveline, Aisyah, Gil, Paskasius, Faisal, Caroline, and the others that are too many to be listed for providing a conducive environment for the dissertation. I am thankful to have met you guys and work together on our dissertations during our late night sessions.

Finally, I would like to thank Indonesian Government, specifically Indonesian Endowment Fund for Education (*Lembaga Pengelola Dana Pendidikan / LPDP*). With their support, I can study at the University of Edinburgh from start to end without any problems whatsoever.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Steven Steven)

Table of Contents

List of Figures

Listings

ix

Chapter 1

Introduction

1.1 Motivation

To study how blood flow in a given vessel, Mazzeo and Coveney [2] developed a fluid
dynamic simulation software named HemeLB. Currently, it is actively developed and
used by researchers to help their study. For example, Itani et al. [3] used HemeLB
for automated ensemble simulation of blood flow for a range of exercise intensities,
Bernabeu et al. [4] used it for detecting the difference of retinal hemodynamics with
regards to diabetic retinopathy, and recently Franco et al.[5, 6] used it to understand
branching pattern of blood vessel networks.

As I have written in the proposal for this dissertation [1], HemeLB works by cal-
culating fluid flow in parallel by using Lattice-Boltzmann method [2]. This calculation
allows HemeLB to simulate blood flow within a given blood vessel structure. In-
evitably, the calculation part is only a small part of the workflow to run the simulation.
There are multiple pre-processing and post-processing steps needed to run the simula-
tion from start to the end. These includes of preparing the input so HemeLB can work
on it, and also processing the output so it is ready to view or further analysis.

These long pipelines of steps needed to run the simulation, coupled with the com-
plexity of the configuration of HemeLB created a high barrier of entry for scientists
and doctors to use it. Furthermore, as observed previously [1], an interesting simu-
lation will require parallel computing resources like ARCHER supercomputer which
might be difficult to get access to by interested parties. While smaller simulation in-
stances can run on a typical laptop, most of the problems will require more powerful
machines.These facts might prevent usage of the software by domain experts. More
importantly, it shows there are still improvements that can be done to lower the barrier

of entry for users to use HemeLB. This is important for HemeLB, especially when it is envisioned to be part of future medical decision [7].

Another aspect that HemeLB workflow can be improved is with regards to its reproducibility and replicability aspect. Chris Drummond [24] make a distinction between these two terms by arguing that reproducibility require changes to the original experiment, while reproducibility avoid changes to it. Research projects that used HemeLB embrace these concepts as their priority. There are steps that are in place to make sure HemeLB and its simulation result are reproducible. First, the entire code base is publicly available on Github. Second, in running a simulation with HemeLB, version of the software is automatically recorded. Lastly, in addition to the version used, input files and configurations are also recorded automatically. These facts can be seen from the publications mentioned above [4, 3, 5, 6] that include all these information. This information allow researchers interested to replicate the simulation or tweak the simulation parameters to run their own simulation. Automation of these steps could further improve HemeLB's reproducibility and allow peers to replicate, duplicate, and audit its simulation results quickly and easily.

1.2 Objectives

Based on the needs to improve the usability and reproducibility aspect of HemeLB project, I will develop a prototype web interface for HemeLB. This prototype web interface will lower barrier of entry in using HemeLB software compared to the current approach of using command line interface. Currently, the approach consists of manual installation of HemeLB onto the HPC infrastructure, copying data, running simulation, and extracting outputs. These steps require users proficient with command line interface and is hard for people without the appropriate skills. By using web interface, it will allow domain experts to do all these steps in a more-familiar setting of web browser. In addition, it will also allow features to be added to the simulation workflow which might not be essential to the HemeLB core itself. For example, automating packaging, sharing, and reproducing simulation result. These features are not essential but useful for helping the overall workflow of blood flow simulation.

Using the dynamic capabilities of cloud computing vendor, the web backend should be able to dynamically scale effortlessly. On top of that, these infrastructures are available to everyone with a cost, can be started on demand, and user will only pay what was used. These benefits allow user to get access to computing resources to run HemeLB

simulation without having to get access to dedicated HPC infrastructure. The web backend will allow users to run a blood flow simulation without having to deal with the complexity of running each step of the workflow manually.

In addition to the web interface, I will also develop deployment script so that peer could deploy its own instance of the web interface. This will ease up deployment process for individuals or organizations intending to use HemeLB for its own purposes. This script will be developed as part of the project.

1.3 Outline

I provide a brief introduction to the topic of this dissertation in this chapter. The rest of the chapters will be organized as follow:

- **Chapter 2, Background**. I will provide background information that is necessary for readers to understand the concepts, technology, and implementation that are done in this dissertation. HemeLB, containerization technology, cloud computing, High-Performance computing infrastructure, and other topics will be discussed in details in this chapter.

- **Chapter 3, Design and Implementation**. I will discuss the bulk of the work in this chapter. Implementation details and design of the proposed solutions will be provided and discussed in details.

- **Chapter 4, Evaluation**. In this chapter, I will discuss how the success of this project will be measured. I detailed how I conduct an interview and performance benchmark to measure how the proposed solution perform.

- **Chapter 5, Result and Analysis**. I will discuss my findings from the evaluation of the project.

- **Chapter 6, Conclusion**. I will conclude and summarize the whole project.

- **Chapter 7, Future work**. Some recommendations on the future of the project in the context of HemeLB simulation workflow.

Chapter 2

Background

In this chapter, I will discuss about various background information that are used as a basis for the work presented in this dissertation. I will discuss about how HemeLB currently works, High-Performance Computing (HPC) infrastructure, and Docker.

2.1   Usability of complex applications

2.1.1   Current HemeLB workflow

Currently, running a blood flow simulation with HemeLB consists of multiple steps that needs to be run in sequence. These steps are done in a variety of interface, from command line to graphical user interface. Additionally, these steps also require various level of computing resources to work efficiently. In order to understand how the proposed work will improve the current conditions, I will elaborate on how HemeLB currently works.
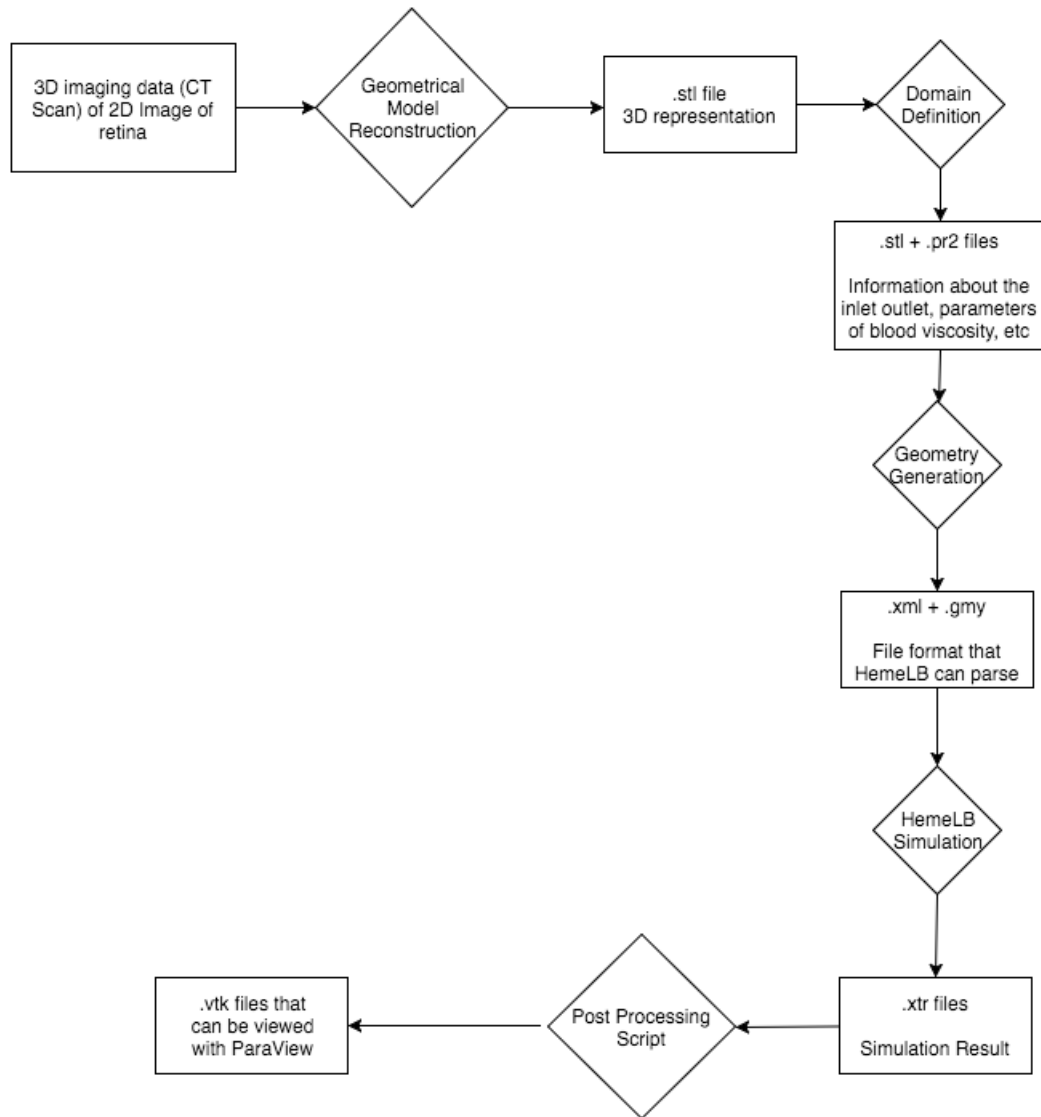
Figure 2.1: Current HemeLB workflow taken from HemeWeb proposal [1]

Figure 2.1 illustrates steps involved in running HemeLB workflow. These steps will be discussed in details below:

1. **Geometrical model reconstruction**

   In this step, a 3D model of vascular system is constructed from the raw microscopy image of it. Alternatively, the 3D model can also be constructed from CT scan with its 3D imaging data. From this step, a 3D geometry file are generated in the form of .STL(STereoLithography) file, a standard file format to describe 3D model object geometry. This process can run in a regular workstation. However, it is highly problem dependent as the tools needed to parse and generate the 3D model are dependent to the problem experts try to simulate.

2. **Domain definition**

   3D geometry model generated from the previous step is now used as an input for the domain definition step. In this step, a graphical user interface is used to add domain information to the 3D model. Information like blood viscosity, inlet outlet placement, and blood pressure will determine how the simulation will run. The HemeLB setup tool was developed for this particular needs. The setup tool provides a graphical user interface for domain experts to add these parameters. All parameters are then saved in a profile file with .pr2 format. This step can run on standalone commodity hardware and should not require a highly parallel computing resources.

3. **Geometry generation**

   This step will take the encoded information from domain definition step and the 3D model of the vascular system to generate files that can be understood by the main HemeLB program. These files contain similar information with the previous 3D model and the profile file. However, both of them are now formatted in a HemeLB parseable format, an XML configuration file and a GMY geometry file. This geometry generation step can also run on a commodity hardware. However, it requires users to use command line interface to operate with the files. The process is done with piping the input files to a python script which is part of HemeLB setup tools. In scientific computing this step is also known as "domain discretisation"

4. **HemeLB simulation**

   The main heavy computations of the workflow are done in this step. Configuration and geometry files that are generated in the previous step are feed into the HemeLB binary as input files. HemeLB will then run calculations that govern how blood will flow inside the provided vascular system for a number of iterative steps. The number of steps is defined in the configuration file that is generated in the domain definition steps.

   As observed in the proposal of this work [1], HemeLB can scale up from 1 up to 32,000 cores in running the simulation [8]. This means that a typical problem could run in a commodity hardware with a small number of cores. However, bigger and scientifically challenging problems will require a higher number of cores that requires high-performance computing resources as portrayed in studies by

Franco et al. [6, 5] and Bernabeu et al. [4]. In addition to that, users of HemeLB have to use command line interface to configure, run HemeLB simulation, and interact with the output files.

Output files generated by this step are written in parallel into output directory which is set when running the simulation. These output files represent the state of blood flow in the vascular system at a given step count. The interval in which HemeLB writes an output is also set from the domain definition step.

5. **Post processing**

The output files generated by the HemeLB simulations are not easily viewed by domain experts. The files are generated in a fashion that is efficient to write in parallel, however, it will need further conversion to make it easy to be interpreted. This is where the post-processing step comes in.

In this step, the output files are piped into two python scripts that are included in HemeLB tools to convert them into a .VTU files. These .VTU files are viewable in a separate software called ParaView [1]. With it, domain experts could visualize the result of the simulation in a graphical user interface which ParaView provided. This step can be run on commodity hardware without problems. However, to do this step, users will require to interact with command line interface.

All of these steps requires users to configure and install the tools required for each step by themselves. However, the target audience of HemeLB are biologists, clinician, and researchers, which might not have the capabilities and the technical know-how to do it. This is one of the motivating reasons for the proposed work in this thesis.

2.1.2 High performance computing infrastructure

Researchers increasingly use complex mathematical and computational approaches in doing their research. In understanding complex phenomenon, researchers use interdisciplinary approaches in providing insight into the problem [9]. Bioinformatics and computational biology are examples of this interdisciplinary fields.

The problems that these disciplines tries to understand require computational approaches that are costly. The smaller size of these problems would probably run on a a powerful workstation, more complex one will often require institutional or regional

---

[1] http://www.paraview.org

HPC facility. These resources are often found in the form of a supercomputer like ARCHER supercomputer. HemeLB software package is one of many examples of these computational approach that requires highly parallel computing resources.

To tackle problems that require large computing resources, two paradigms of computation is developed. These are High Performance Computing (HPC) and High Throughput Computing (HTC). While both of these disciplines are developed to solve problems that require large computing resources, they are different in the nature of the problems they are trying to solve.

High performance computing uses uniform computing nodes to perform a tightly coupled computation. They are placed in one physical location and connected to high bandwidth network amongst them. This network connection allows these nodes to communicate with each other efficiently, thus, allowing them to coordinate computation across the nodes[10]. A message passing interface (MPI) library is often used to perform this type of computation and it allows every process to communicate with each other in parallel. As observed in the proposal, computer clusters, GPUs, and supercomputers are the prime example of computing resources to run this type of computation.

On the other hand, High throughput computing tries to treat computing resources like a utility line. Users should not have to worry about where the computing resources come from, they can just request it and it will be given. This type of paradigm uses a middleware that allows non-uniform computing resources to communicate and cooperate in order to solve common problems[10]. These differing resources will then do different works that are scheduled independently.

HemeLB uses MPI library to communicate between processes and run tightly coupled computations on each of these processes. Based on the definitions outlined above, we can categorize HemeLB as an HPC application that requires an HPC infrastructure to run its computation efficiently.

Running HPC application like HemeLB requires access to HPC infrastructure that is often managed by a university or a research facility. These institutions give access to HPC projects by computing hours on the basis of the merit of their proposal. For example, this is how the Partnership for Advanced Computing in Europe (PRACE)[2] and Engineering and Physical Science Research Council (EPSRC)[3] operate. They conduct a peer-review of proposals that indicates the need for their infrastructure and give access selectively.

---

[2]http://www.prace-project.eu
[3]https://www.epsrc.ac.uk

The model of operation of this institute often does not prioritize reproducibility of a research or an exploratory one. Experts wishing to reproduce previous simulation have to compete for the limited amount of available computing hours with other projects. Most likely, reproducibility of past studies is not the top priority of this institutes, causing a barrier for reproducing computational research that relies on this type of infrastructure.

In addition to the above problem, most research that tackles biological and biomedical discipline often fall outside the scope of these institutions. Also at the time of writing, the equivalent funding bodies for these disciplines do not have access to HPC resources. These problems limit the capabilities of researchers to reproduce past studies easily.

2.1.3 Cloud computing

To answer the huge demand for computation powers by researchers and academics, a concept called grid computing was born in the 1990s [11, 12]. This concept treats computing resources like utility grid. Computing resources should be able to be acquired without users knowing how it was provided to them. This model caters mostly to the interest of researchers and academia that usually give CPU hours based on projects proposal [13]. An example of this is TeraGrid[4] project which ended its operation at 2011 .

Cloud computing paradigm was then developed by commercial operators based on the similar idea that computing resources should be available to the users without the user knowing where it came from. However, this is where the similarity ends. Cloud computing caters more towards the business aspect of these computing utilities. While grid computing prioritizes features and functionalities that researchers and academia would like, cloud computing vendors focus on features that business will pay. Cloud computing vendors are driven by economies of scale and it will not survive if businesses do not use their service [13].

Conditions outlined above have created a tight feedback loop between users and the cloud vendors. This led to the development of features that users need and will pay for. As observed in the proposal, Cloud vendor now is massively scalable, allows computing resources abstraction, configurable dynamically, and provisioned on-demand. This has led cloud computing vendors to be more relevant to common use-cases compared

---

[4]https://www.xsede.org/tg-archives

to grid computing.

Cloud vendors continue to grow significantly in the recent years. In 2013, it was reported that some cloud vendors reached had more than 90% growth per annum [14]. This growth enables them to keep more incentives for businesses and individuals to buy their services. In few instances [15, 16, 17], cloud vendors have cut their pricing for their service and fueled more demands. Renting computing resources is getting cheaper every year and could make more sense both economically and technically than building your own infrastructure.

On top of that, cloud vendors also do not vet projects based on their proposals. Projects could easily start on-demand computing units for their purpose. The business mindset of this cloud vendors allows reproducibility to be a priority in research, unlike requesting resources from research institutes. This scenario is perfect for researchers and institutions that do not own their own HPC infrastructure. Instead of building their own, they can rent from the cloud vendors and does not have to worry about maintenance.

Conditions above are also capitalized by cloud vendors like Amazon by attracting customers that need computing resources for HPC applications [10]. While it is reported that running HPC applications on cloud platform will incur performance overhead, it is a viable alternative to having your own dedicated infrastructure. This is shown in various studies in the past, for example, Nekkloud project [18], NASA HPC applications [19], and an HPC application benchmark in public cloud [20]. In addition to that, the capabilities to massively scale your computing resources, limited by one's purchasing power, is an attractive feat for HPC application that needs scalability. This is also why HemeLB software package can rely on the cloud platform to scale up or down depending on the problems.

2.1.4   How other HPC applications tackle usability issues

In this section, I will discuss how similar HPC applications solve problems like HemeLB faces. Recently, some complex HPC software packages have been deployed to the cloud. Their experience in deploying the software and their approach to solving similar problems will be indispensable for the work that I proposed.

The first project that tackles a similar problem like HemeLB is Nekkloud. In this project, Nektar++ faces usability issue just like HemeLB. It is a complex high-order finite element code which mainly is operated using command line interface. The original

workflow is complex and only limited number of people can operate it, barring people without computer expertise into actually taking advantage of the software package.

This usability problem coupled with the fact that users have to get access to HPC infrastructure, might not be a viable option for everyone. In order to answer these problems, Nekkloud project was hatched. Nekkloud was developed to make the software configuration invisible to the users. It uses a web application to provide an easy to use and learn interface for the domain experts. This allows people without computing expertise to actually run Nektar++ without having to be troubled with configurations. In addition to solving the usability problems, Nekkloud also uses public cloud vendors, alleviating concerns with regards to having dedicated HPC infrastructures from users.

Galaxy [21] is another project that is trying to tackle similar problems. It described itself as a web-based reproducible research platform and ran on public cloud infrastructure. With Galaxy, researchers can run various HPC applications that are compatible with an easy to use web interface. Users do not have to worry about the configurations and working in command lines, only worry about the software executions.

Developers of Galaxy developed a super-resolution spark (SRS) model to illustrate its use case. This model is an example of HPC application that requires a highly parallel computing resources like a supercomputer to run efficiently. However, running the SRS model in the cloud via Galaxy is observed to be a viable alternative. In addition to that, Galaxy provides an easy to use interface to run this model for the users. Making it easy for them to run their experiments and share it with their peers.

Nekkloud and Galaxy projects illustrate that web interface is a viable alternative interface for complex HPC applications. With the correct application and design, it will allow domain experts to operate the underlying HPC application with relative ease. However, the change of the underlying infrastructure from dedicated HPC infrastructure to the public cloud have some negative impacts.

One notable drawback is the lower performance of the HPC applications. It has been studied that running HPC applications on the public cloud means performance degradation compared to the dedicated HPC infrastructure. Mehrotra et al. [19] conducted a performance evaluation of the NASA HPC applications on Amazon EC2 and found out that in the worst case, using Amazon EC2 as an HPC infrastructure increase runtime up to a factor of five. The worst case is consistently found in the larger number of cores used, compared to the smaller number of cores. This severe penalty is largely caused by the inherent difference in the networking capabilities between cloud vendors like amazon with supercomputers, the more nodes the more severe the penalty is.

Amazon web service offers 10GigE (10Gb/s) compared to FDR InfiniBand (54.5Gb/s) on NASA supercomputer, Pleiades. In addition, Mehrotra et al. also find out that there was a small performance penalty on Amazon due to its virtualization layer. This penalty should be carefully considered when running HemeLB in the cloud.

2.2 Reproducibility of software

In this section, I will discuss the problem of reproducibility and how the proposed components try to solve it

2.2.1 Reproducibility problem in computational research

American Physical Society [22] describes science as "the systematic enterprise of gathering knowledge about the universe and organizing and condensing that knowledge into testable laws and theories". For theories and experiments to be testable, it has to be independently reproducible or replicable by peers. However, a recent study [23] highlights that some published psychology studies are not replicable to the same significance as reported. In machine learning conferences [24], a similar sentiment is being shared. Being not replicable does not necessarily means that the result of those studies is wrong. However, it suggests that the academia may not prioritize verification of results. The novelty of idea may be seen as better than verifying what we already know.

With the recent study highlighting a reproducibility problem of a research, previous pushes [25, 26] for reproducibility in studies become more relevant than ever. This is especially crucial in a discipline like computational research. Computational research discipline like bioinformatics and computational physics involves complex computation that requires huge computational power depending on the problem size. The complexity of configuration, algorithm, and execution process actually become a barrier for peers to replicate and reproduce works of studies of this discipline. Making results produced have less weight than it could be

This complexity is also one of the reasons why Galaxy project came into development [21]. While science values rigorous testing, replicable, and reproducible results, the complexity of HPC software package hinders that. Galaxy tries to solve that by producing automatic metadata that record the execution of an analysis done by tools on Galaxy platform. It records all the input files, configurations, and outputs of an

analysis and makes it available for the users to view and copy. This metadata information enables users to share the analysis with their peers. Allowing peers, to replicate or reproduce the analysis results independent of the original researcher's. Galaxy allows the ease of reproducibility and replicability for the domain experts.

2.2.2 Containerization Technology and HPC application

One of the main problems with replicating or even reproducing results of studies with complex software package is configurations. A complex software package like HPC application often requires hands-on configuration by the users for it to run correctly. This complex configuration coupled with complex usage become barriers for an independent party to verify results of a study.

Containerization technology, particularly Docker[5], can help with the issue above. Docker is a technology that is developed on top of kernel-level technologies like Linux Containers (LXC)[6] which allow containers to be a sandbox between processes [27]. Containers, allow applications to be securely placed in its own environment that shares the kernel with its host. Unlike full virtualization, containerization does not require a full operating system installed on the virtual environment so the application could run. Docker is arguably the most popular containerization technology currently. It is built on LXC and added features that allow an application to be deployed easily. It handles container image creation, versioning, sharing and archiving in addition to running the image like LXC does. On top of that, there is also a web application called Docker hub[7] that allows Dockerfile, a text file that contains the commands to build a container, to be inspected publicly if set correctly. It allows interested peers to audit a container and make changes to it for their own container.

Using containerization technology like Docker can help with reproducing results of studies. Complex configuration process can be recorded and encapsulated in the form of Docker container; hiding complexity to the casual users. Studies can package the complete software package in a Docker container to be shared with peers for independent replication and reproduction of an analysis. In addition to that, the versioning feature of Docker also allows the software package to continuously developed and for an analysis tied down to an older version of the package. An analysis can be replicated even if the software is currently way ahead compared to the time an analysis is

---

[5]https://www.docker.com/
[6]https://linuxcontainers.org
[7]https://hub.docker.com

made. It also creates a good opportunity for the study to be reproduced, an analysis could have different results if it is running with the new version of the package. Hence, containerization technology enabling reproducibility of results.

This is also the reason why Galaxy project supports Docker in its toolsets. Galaxy ecosystem uses Docker to create a secure, isolated environment for the tools and its dependencies [28]. Tools are versioned, archived, and shared with Docker containers so peers can download and audit all the tools. This openness is also important for the tools to be trusted.

HemeLB as a complex HPC application is also actively developed. While development is ongoing, it is often used as part of a research. Docker can help with the recording of the version used for a study by its versioning feature. In addition to that, configuration work is also taken out of the hands of the peer because it is done initially during the packaging of the application. Containerization technology sounds appropriate to be used for HemeWeb to enable easy reproduction, replication, and audit for HemeLB simulation.

Chapter 3

Design & Implementation

In this chapter, I will discuss HemeWeb's development and implementation. This will consist on how the HemeLB core Docker container is developed, how it is deployed, and how the web application is developed.

3.1 Overall architecture

The development process is divided into 5 phases. The planned phases are as follow:

1. Separate HemeLB core into its own container

2. Orchestrate the deployment of HemeLB cluster / infrastructure

3. Develop HemeWeb to accept user input

4. Extends HemeWeb to handle geometry generation workflow

5. Extends HemeWeb to handle domain definition step or Viewing of HemeLB simulation result

The development process loosely follows the agile method in which I regularly meet with the stakeholders every week to give an update and gather feedback on the project. The phases are designed in such a way to minimize the risk of having nothing at all during the end of the project phase. This is due to that HemeWeb can work on its own after finishing step 3. The HemeLB simulation can be done on its own. The rest of the steps are there to extend the functionalities of the HemeWeb to cover more functionalities.

During the first week of the development, I focused more on stripping the HemeLB core container into its own. I researched on how Docker and Dockerfile work, and finding out what are the issues with the current container. After identifying the issues, which are ssh service and full of functionalities which are not essential, I stripped down the image and changed the base image so that the container could avoid the mentioned problems. I end up with smaller container size and it is available online on https://hub.docker.com.

However, one particular issue with what I have currently is that the Dockerfile is published as a part of the HemeWeb source code. It should be tied down to the HemeLB development instead of HemeWeb. Currently, the development model of HemeLB is that there is an internal private repository where the less than stable build is pushed to it, and there are public repositories where only stable builds are pushed into. The release process should include adding the Dockerfile towards the core HemeLB source code repository and tagging the release correctly. HemeLB core containers should then be built automatically on Docker hub with regards to additional tags being pushed to the public repository.

After the development of HemeLB core container, I then focused on how the architecture could be primed for the HemeLB simulation. It involves on configuring the servers and all supporting infrastructures on cloud vendors to be ready for HemeLB simulation. Network configuration, security configuration, Docker configuration, and other should be handled automatically. I elect to choose ansible orchestration software because it is closely related to python language that I used.

In this phase, I successfully achieve the provision and deployment process that with the correct credentials and authorization, the script could provision and configure the architecture correctly so it is ready for HemeLB simulation. In addition to that, I successfully created the script so that it will be cloud vendors agnostic. I can deploy the architecture to google cloud vendors, amazon web service, and digital ocean.

However, it is to be noted that the deployment process that I achieve can only run HemeLB simulation from the command line. I have not considered the web application installation and configuration at this point of the deployment. I only considered the infrastructure being built and configured for HemeLB simulation.

Next, I started developing HemeWeb web application. I choose Django web application framework due to my experience with it. I created a basic interface, where job simulation is listed on the index page on the home interface. After that, I added a basic interface to add new job with a geometry file and HemeLB configuration file. The

web application will then add those input files into a newly created job instance and configure the job instance in the job configuration step. The job will then be submitted.

Here I developed a separate ansible script that will be called when a queued job is being worked at. The ansible script is responsible for starting up compute nodes needed by the job and executing HemeLB simulation. After the simulation is done, the script is also responsible for correctly terminating the compute node.

After the basic HemeWeb web application is achieved, I extend it to handle pre-processing. I added an extra form in the adding new job form to handle a profile file and geometry file. These two files will be converted into a geometry file and HemeLb configuration file that the job configuration steps expect. In addition to adding the interface, I also added a new function on the job instance that will run this pre-processing step on the background.

Lastly, due to the limited time, I can only manage to run a small post-processing step on HemeWeb. What I did was adding post-processing step that converts the Extracted results from HemeLB output into a format that can be viewed by third party software, ParaView. The results are piped into two scripts that will output a .vtu that is compatible with ParaView. This is done during the background activity after simulation result is outputted. However, currently, this is done on the master node.

In addition to that, I also managed to add persistence capability to HemeWeb application. What I did was to package the job simulation folder into a compressed archive and upload it to persistent storage service of cloud vendors. These archives can be queried by another instance of HemeWeb to get previous job IDs available for the particular cloud vendor account used to deploy HemeWeb. Also, the job simulation file URL is also showed on the web interface. Making it easy for the users to share the simulation file with their peers.

3.2 HemeLB core Docker container

HemeLB core is a Docker container that consists of essential software and services needed to run HemeLB simulation. It is the main component that runs the HemeLB simulation on the cloud. Calculations will be done on the container which will be started on the compute nodes of the HemeWeb architecture.

Previously, there was an effort to make HemeLB software package portable by creating its own container package[1]. It has all the necessary tools for HemeLB simulation

---

[1] https://github.com/mobernabeu/docker-hemelb

workflows such as setup tools, HemeLB binary, and post-processing scripts. Moreover, it also includes a Virtual Network Computing(VNC) capability that allows users to access the container's headless graphical user interface via HTML5 capable browser. All of these are necessary to make HemeLB more portable. Independent peers can replicate and reproduce simulation without having to configure all above tools. However, users will still need to configure Docker on their workstation to be able to run the container. HemeWeb will take this effort a step further. With HemeWeb, users will not be required to install any tools to run simulation workflows. Only web browser, which is normally installed by default, is needed.

In this phase, I took the previous container and modify the Dockerfile, the instructions file to build the container. The original Dockerfile uses a base image of Ubuntu[2], a popular open source operating system, that is provided by Docker. This image by Docker incorrectly handle Ubuntu's init system, Upstart[3], that results in some system services cannot be started correctly. One of this service is Secure Shell daemon(sshd)[4]. This daemon is responsible for allowing remote users to get secure encrypted access to the host, in this case, a Docker container. Without sshd, HemeLB, which is a Message Passing Interface(MPI) application, cannot run a multi-host simulation because communication cannot be done between hosts. To solve this issue, I switched the base image to base ubuntu provided by Phusion[5], a company in Netherland that released their version of Ubuntu container that solves above problem.

After switching the base image of the Docker container with Phusion's Ubuntu (See Appendix A), I added commands to correctly start the ssh service, and configure the access keys. It is currently configured to use a shared insecure key that is committed to the repository. Ideally, this should be automatically generated at each container build, however, due to the time limit of this project, this is not done yet. The next step is to strip out tools which will not be needed by the compute nodes to run HemeLB simulation. I added commands to purge the packages that are not essentials to the compute nodes. The resulting container is a minimal container that contains only the HemeLB binary and SSH service running. With these changes, the size of the container is also halved from originally 948MB[6] to 430 MB[7].

The modification of the Dockerfile of the container is the initial part to make

---

[2]`http://www.ubuntu.com`
[3]`http://upstart.ubuntu.com`
[4]`http://www.openssh.comm`
[5]`https://github.com/phusion/baseimage-docker#`
[6]`https://hub.docker.com/r/mobernabeu/hemelb/tags/`
[7]`https://hub.docker.com/r/seiryuz/hemelb-core/tags/`

HemeLB core container. To create the container correctly, this Dockerfile needs to be integrated into the development workflow of HemeLB. Currently, the modified Dockerfile still lives under the HemeWeb codebase[8]. HemeLB development should trigger an automated build of HemeLB core containers with each version of the software it pushes to the public GitHub repositories. In addition, the development team also needs to create a consistent tag naming in order for the HemeLB core containers to be created correctly.

3.3 Deployment scripts

The next step of the implementation is to develop deployment scripts to configure the overall software architecture. Due to the nature of the architecture that consists of many moving parts, manual configuration is not considered a manageable solution. I decided to create deployment scripts that facilitates the pain of deployment by provisioning and configuring the architecture with minimal manual intervention. The scripts are created using a configuration management and orchestration tool called Ansible[9].

Ansible is arguably the most popular tools in IT infrastructure automation space[29]. It is an open source tool that is used for provisioning, configuring, and orchestration of IT infrastructure. Moreover, it is an agentless software that allows it to work without installing client-side software on the hosts it manages. I decided to use Ansible for deploying HemeWeb's infrastructure because it is developed in Python[29]. In addition, Ansible being an agentless software is really useful because there is one less moving part that HemeWeb have to worry about. Another reason is because it uses YAML(YAML Ain't Markup Language[10]) syntax for the configuration management language which is described by the team as "easier for humans to read and write than other common data formats like XML or JSON"[11]. I find YAML is easy to work with in developing the scripts and also used extensively in HemeLB software package. Finally, there are a lot of community contributed components that are included with Ansible that makes developing the deployment scripts much faster. For example, there are modules that enable a script to provision instances from cloud vendors like Amazon Web Service, Google Cloud Platform, and Digital Ocean.

At the current time of writing, the deployment process consists of multiple files,

---

[8]https://github.com/SeiryuZ/HemeWeb/tree/master/hemelb_docker
[9]https://www.ansible.com
[10]http://yaml.org
[11]http://docs.ansible.com/ansible/YAMLSyntax.html

templates, and configuration scripts that are modularly assembled to be run as a set of instructions to build the infrastructure. These deployment scripts can be divided into two separate distinct functions, which are:

1. Provision instances and configure cloud vendor specific settings

2. Configure the provisioned instances to produce correctly set architecture

I will discuss the deployment script in details in the following sections.

3.3.1 Provision instances and configure cloud vendor specific settings

In this phase, the scripts are mainly responsible for provisioning master and compute instances for the infrastructure. The script will request instances from the cloud vendors based on the number and type of instances set in the script. Because cloud vendors have their own interface to interact with their platform, I decided to separate the scripts into their own directories for each cloud vendors. This is due to the fact that in Ansible, to interact with each vendor, it needs to use a different module.

At the time of writing, I developed the provisioning script for three different cloud vendors (See Appendix B, section B.1). They are Digital Ocean, Amazon Web Service, and Google Cloud Platform. Each script has slightly different YAML syntax due to the difference in the module that is used. In addition, there is a different model of provisioning that needs to be adhered to by the scripts. The most notable one is the Amazon Web Service one. The script has to be able to create a correct security settings for each instances or it will not allow the instances to be connected remotely. Without this ability, the deployment script will not be able to configure the instances correctly for the infrastructure.

To run the provisioning script, each cloud providers require authentication. Because this authentication is in the form of keys that need to be provided to the script, I decided that it should not be hard-coded in the script. The authentication credentials are to be set in the environment variable of the workstation that the script will be run from. The script will then look up the credentials from the environment and use it to authenticate with the cloud providers. This is done to prevent the credentials being publicly shown because the repository for the source code is public.

3.3.2 Configure the provisioned instances to produce correctly set architecture

After developing the script for instances provisioning, I developed the deployment scripts (See Appendix B, section B.2 and B.3). This deployment scripts will require addresses of the hosts to configure. However, because provisioning the instances will create new instances with new addresses every time it is run, it is not convenient to write a static list of addresses every time deployment script is run. Therefore, I find a community-contributed script[12] that allows addresses to be dynamically queried to the respective cloud providers. Each of this script lives in the same folder as the provisioning script and has a different configuration that needs to be set once. After entering the credentials in the configuration file for this dynamic addresses script, deployment script can now target the correct hosts every time.

For each cloud providers, there are different deployment scripts. However, these different deployment scripts are setting up the variable that will be used in the main deployment script that lives outside this cloud vendor specific folder. After setting variables like instance tag that varies between the vendors, it calls the main deployment script. In the main deployment script, it handles 3 separate tasks. First, to configure every host with common configurations. Second, configure the master instance. And finally, to configure the worker instances.

These configuration tasks are separated into its own Ansible "roles", modular component of tasks that live in its own directory. In each role, I defined tasks that are relevant to that specific roles. For example, in the "common" roles, I defined tasks to configure SSH service, update the instance's packages, and update the `/etc/hosts` file on each instance. These tasks are common to every type of instances so I put it into its own role. Another example of the role is the database role that I called "postgresql" that is only applied to the master instance. In it, I defined tasks that download and configure the database correctly so it is ready for HemeWeb uses. With these roles defined, I can include the roles in the main deployment script so it is included in the execution of the deployment script.

---

[12]`https://github.com/ansible/ansible/tree/devel/contrib/inventory`

3.4   HemeWeb web application

In this section, I will discuss the bulk of the work of this project which is developing the web application component. The web application component will be the interface for users to interface with HemeLB simulation workflow and is an essential part of this project. It is developed using Python 2.7 and Django web framework[13]. I chose Django web framework due to my previous experience with the web framework and also the existing HemeLB codebase has other tools written in python. Using Django web framework is to make sure that the codebase in HemeLB software package is done mostly consistent with Python.

All the source code for HemeWeb web application is available online at my GitHub's public repository[14].

3.4.1   Architecture
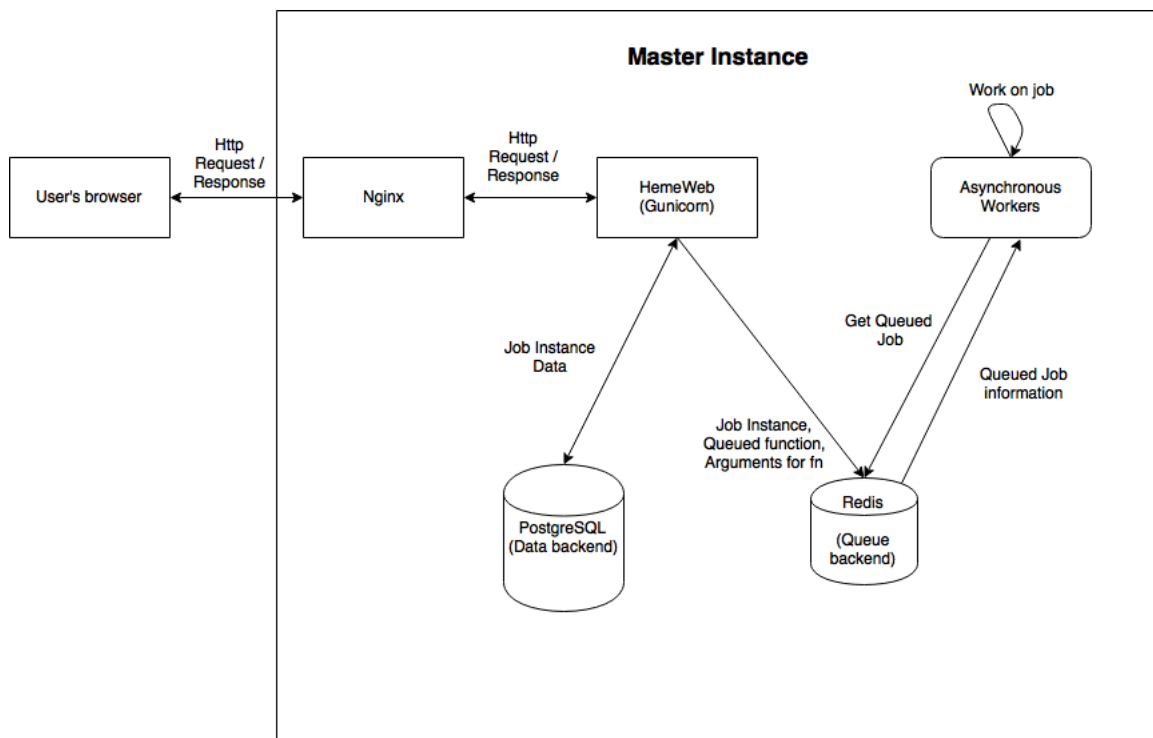
3.4.1.1   Web application components

Figure 3.1: HemeWeb architecture

---

[13]https://www.djangoproject.com
[14]https://github.com/SeiryuZ/HemeWeb/tree/master/src

Figure 3.1 above illustrate how HemeWeb application process interacts with the other components inside the master instance. It also illustrates how user's HTTP request start a chain of events inside the master instance that will eventually return a response to the user's browser.

To start, user's browser will send an HTTP request to the master instance. This request will be captured by Nginx process, a HTTP processing engine, that will act as a reverse proxy. Nginx[15] will proxy the HTTP request towards the correct web application server process or serve static files depending on the requested URL. If the request is routed to the web application, HemeWeb web application which is handled by Green Unicorn[16] HTTP server will accept the request. This library will run the HemeWeb python code to process the HTTP request by the user. Depending on the type and path of the request, the web application will serve a static HTML as a response, or handle job-related logic that might interact with another part of the system. One of the components the web application might interact with is the PostgreSQL database[17]. The database will persist job information locally on the instance to provide persistent information between HTTP request. However, it will be wiped out when the master instance is terminated and is not shared between HemeWeb instances.

Another part of the master instance the HemeWeb application can interact is with the queuing system. HemeWeb can submit a job into the queue which uses Redis datastore[18] as the queue backend. HemeWeb uses third party library called django-rq[19] that handles asynchronous background tasks handling using Redis backend. It uses Redis to store job information on a queue that will be processed by lightweight background job workers. HemeWeb process will store the function to be executed, the job instance and parameters to used by the function into the Redis backend. A background worker will look at the queue at an interval and work on a job if there's any in the queue. The worker will execute the function and update the instance with relevant job execution result. Finally, the worker will go back to being idle waiting for next job to be executed.

---

[15]https://www.nginx.com
[16]http://gunicorn.org
[17]https://www.postgresql.org
[18]http://redis.io
[19]https://github.com/ui/django-rq

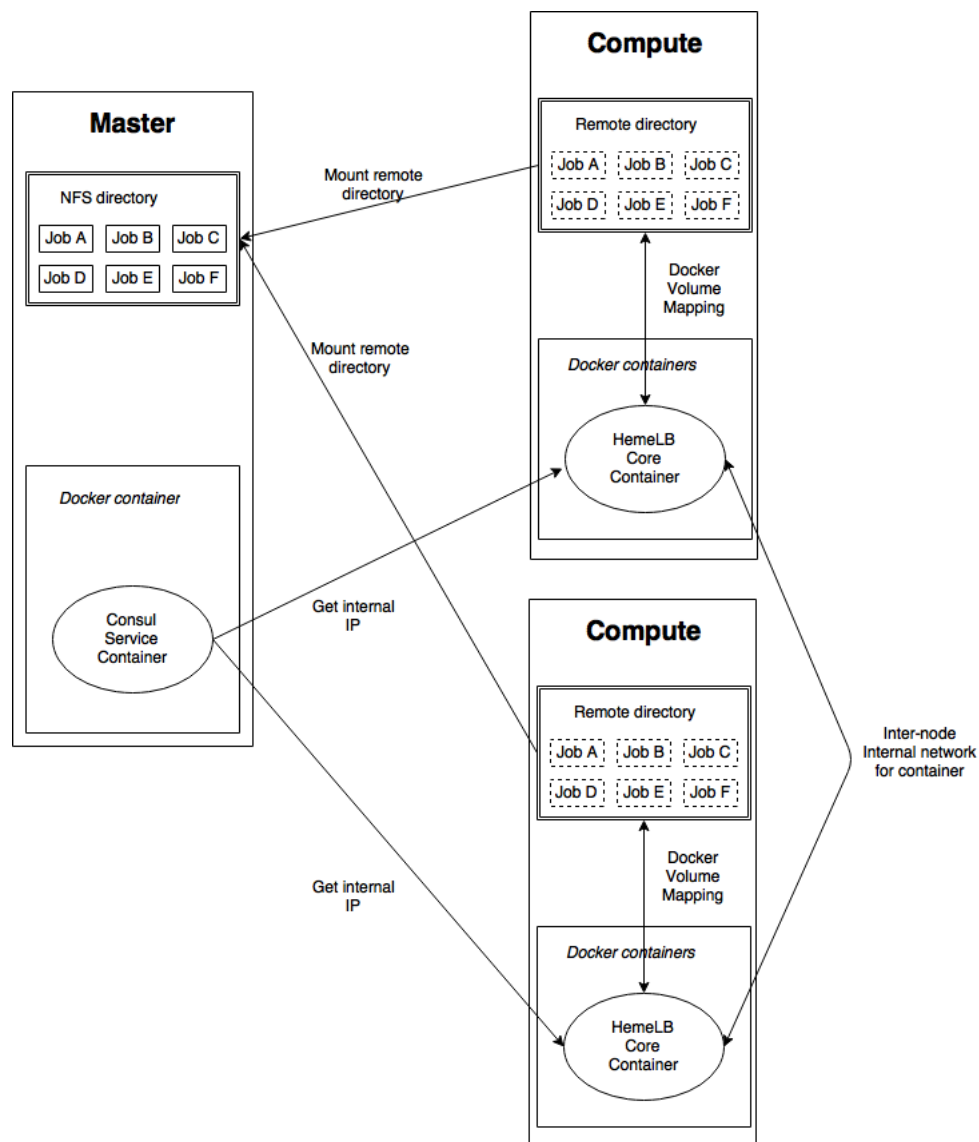### 3.4.1.2 Docker components

Figure 3.2: HemeWeb Docker component

In this section, I will discuss the interaction between Docker components and the hosts. As illustrated on figure 3.2, each host will run Docker service to run specific Docker container depending on their purpose.

In the master instance, a consul container is constantly running to provide inter-container communication mechanism. Consul[20] is used by the Docker containers to coordinate the internal networking communication between them. In a host, multiple

---

[20]https://www.consul.io

containers can be started without conflicting IP because they can communicate via the host. However, in HemeWeb the communication between containers will go beyond single hosts. Containers will need to communicate with other containers living on other hosts. Consul service is needed to coordinate this communication.

On the other hand, on the compute nodes, only HemeLB core container is run. The compute nodes will be started by a job submission to HemeWeb. After the nodes are started, HemeWeb process will instruct the compute nodes to pull the specified HemeLB core container version from Docker hub. If the specified version is locally cached on the compute nodes, no network activity will be made. HemeLB core container is then started to accept simulation command.

To start the simulation, the HemeLB core containers requires the job directories to be exported via Networked File System interface. HemeWeb process in the master node will prepare the job directories with the correct folder and files locally in the master node. Every job submission to HemeWeb will create a specific folder for storing submitted job-related files. The exported job directories will then be mounted by the compute nodes provisioned for the simulation. Finally, HemeLB simulation can use the mounted job directories to read input from and write output to.

3.4.1.3 Job instance structure

As mentioned above, HemeWeb uses Django web framework to handle web functionalities. Django framework follows the object-oriented principle where everything is modeled as an object. Job information that is handled by HemeWeb is also modeled as an object that is derived from object class of Django. Literally named **Job**, is a class that represents simulation information. Each instance of this class will contains information specific to a simulation instance. **Job** class inherits from *django.db.models.Model* class that is included within Django framework. HemeWeb **Job** class extends this basic class and add functionalities specifically related to job information.

HemeWeb's Job class has the following attributes:

- id: This is a Uniquely Universal Identifier (UUID) field that represents the Job ID. UUID field was chosen because it is appropriate for the possibility of sharing the job simulation files between different deployment of HemeWeb. UUID can prevent clashes of job ID between these instances.

- input_file, stl_file, profile_file, output_file, configuration_file: These attributes keep track of the files that are used by the job. It is stored as the path to the

file in the local filesystem, but with Django functionalities, HemeWeb can work with the file as an object.

- container_image: This attribute determines which container of HemeLB core will be used in the simulation. Currently, the choice of the field's value is set manually in the codebase.

- instance_type: This determines which compute node type will be started for the simulation.

- instance_count: This attribute determines how many compute node will be started for the simulation. Currently, this is also set manually in the source code.

- status: The attribute to determine Job's status, whether it is *queued, added, done, failed, etc*

- created: Attribute to keep track when the job is created

- updated: Attribute to keep track when the job is updated

3.4.1.4 Job directory structure

Each simulation done with HemeWeb have its own job directory. These directories are located in the master instance and are shared with the compute node. To provide a clearer picture how the application package and work with the job's files, I will discuss how HemeWeb structure each job's files.

```
<UPLOAD_FOLDER_DIR>/<JOB_ID>
<UPLOAD_FOLDER_DIR>/<JOB_ID>/inputs/*
<UPLOAD_FOLDER_DIR>/<JOB_ID>/logs/*
<UPLOAD_FOLDER_DIR>/<JOB_ID>/outputs/*
<UPLOAD_FOLDER_DIR>/<JOB_ID>/metadata
```

Listing above illustrates the structure of a job instance's directory structure. The first component of all job instance's directory is the UPLOAD_FOLDER_DIR. This is the path to a directory in which HemeWeb will upload all files. To change this parameter, one should change the path in the HemeWeb settings file and restart the web application so the changes took place.

The next part of the directory structure is the job folder named with its ID. The ID will be generated by HemeWeb using Universally Unique Identifier (UUID) scheme.

More specifically, UUID Version 4 that depends on the random number. Using UUID is important for HemeWeb because it is accurately approximated to have a really low chance of producing duplicate ID. Hence, HemeWeb does not have to take care of the possibility of jobs having duplicate ID.

Next, we have the inputs folder inside the job folder. This is where all the inputs and configurations are stored by the web application. There's also logs folder, where the job stdout, stderr, and HemeLB logs are stored. The web application will read from this folder and make it available on the web interface. Outputs folder will be used by the HemeLB simulation to write output files in this folder. One final file is the metadata file. This file is used by the web application to store the state of the job. The job is serialized into this metadata file using python pickle protocol so when it is downloaded, the web application can unserialize the state of the job instance and it is preserved, ready to be used for another simulation.

A Job instance is preserved with this job directory structure. This allows job information to be preserved in an external storage that can be used for backup. HemeWeb currently supports uploading job directory to Amazon simple storage service (S3). However, backup is not the only purpose for this functionalities. By making job directory files available to the public, independent peers can download these files and run the simulation with correct files and configurations. It allows peers to replicate a simulation exactly or with modifications.

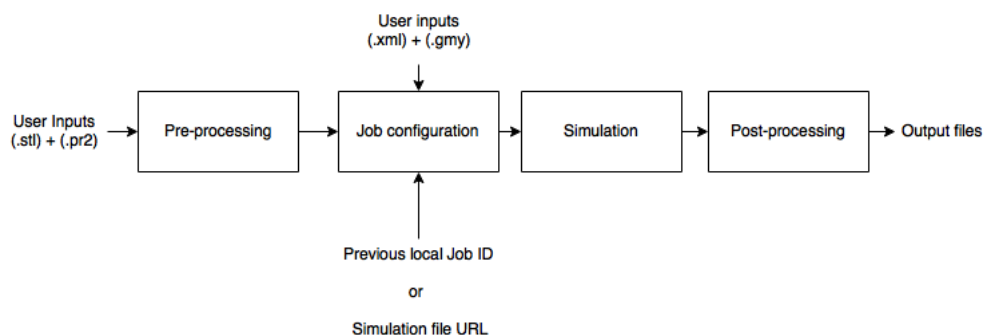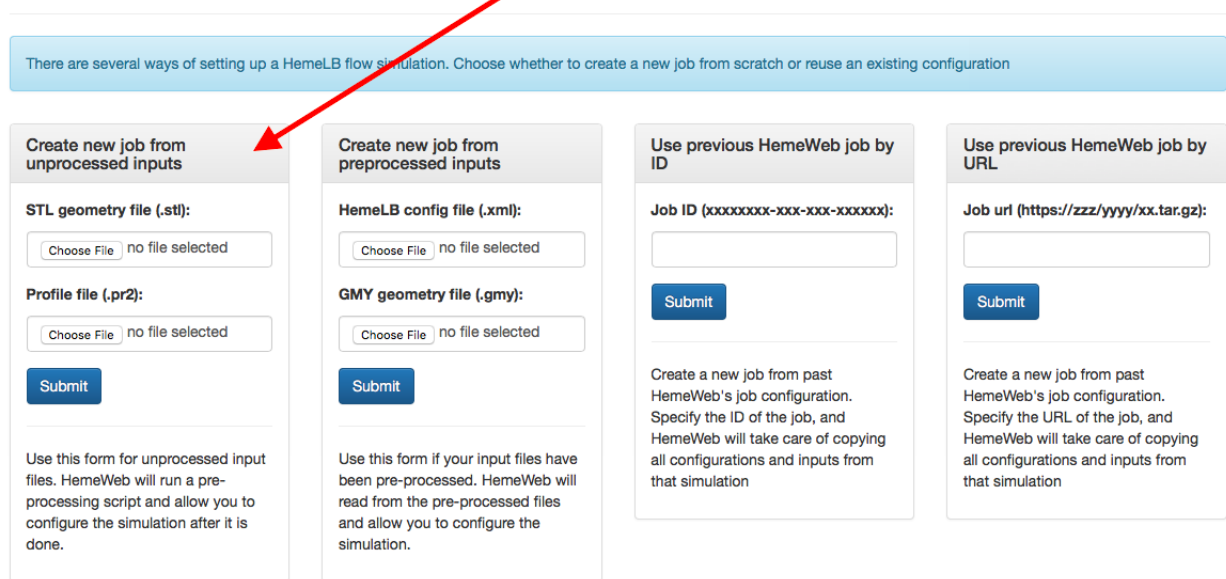### 3.4.2 Simulation workflow

Figure 3.3: HemeWeb flow

Figure 3.3 illustrates how the HemeWeb web application works. HemeWeb currently consists of 4 core activities that will be discussed in details in the following

section.

3.4.2.1 Pre-processing

In this step, HemeWeb handles pre-processing of inputs that are needed so that HemeLB simulation can parse the files. The user provides a geometry file (.stl) and a profile file (.pr2) to the web application for processing. HemeWeb will then create a job instance with these two files, save them locally on master instances and queue the pre-processing job.

Figure 3.4: HemeWeb pre-processing form

Figure 3.4 shows HemeWeb user interface to upload the unprocessed files. After submission of both files and job being queued, the asynchronous worker on the master instance will work on the job whenever they are free. It will run the pre-processing python script to generate the geometry files and HemeLB configuration file. These files will then be saved to the master instance, and HemeWeb will track these files by recording the path to these files on the job instance. Now the job instance is ready for the next step. of the workflow.

3.4.2.2 Job configuration

In this step, HemeWeb application will take a job instance with correctly set geometry file (.gmy) and HemeLB configuration (.xml). However, there are multiple ways that HemeWeb can get this correctly set job instance. As illustrated on Figure 3.3 , there are 4 possible entry points for this step. The web interface for these 4 entry points are also shown on Figure 3.4 They are:

- **From the post-processing step.** These files are generated from the previous pre-processing step. The job instance is directly used in this step

- **User's provided geometry and configuration file.** User have pre-processed their own file locally, or have their own geometry and configuration files available. HemeWeb will create a new job instance, save both files and keep them tracked with the job instance.

- **User's provided previous job ID.** There are two possible case when user specify previous job ID. First, the previous job is available locally on the HemeWeb instance. Second, the previous job is cached on the persistent storage on the cloud vendor and are not available locally. HemeWeb will download the previous file from the persistent storage if it is not available locally. It will then create a new job instance that copy the previous job's geometry file and configuration file to be used for further configuration.

- **User's provided simulation file URL.** The last alternative is for user to provide the simulation file URL. Simulation files are uploaded to a persistent storage at the end of the workflow. These files, if made public, can be used by other instance of HemeWeb to download the simulation files and use it as a basis to create a new job instance. The way the system work is the same as using previous job ID, but its source is not its own persistent storage, but other people's simulation files.

After the job instance is created from one of the four way possible discussed above, HemeWeb will then ask users for the job configuration.

Figure 3.5: HemeWeb HemeLB configuration form

Figure 3.5 shows the interface where users are asked to configure the HemeLB simulation parameter. In this page, an online XML editor will be provided for the user to directly edit the .xml file that is provided by the users or from the pre-processing step. This XML editor allow users to edit many simulation parameters on the web interface without changing the .GMY file. Users can directly edit values that affect simulation execution like inlet pressure, outlet pressure, blood viscosity, and etc. After configuring the simulation parameter it will then be redirected to job configuration page.

Figure 3.6: HemeWeb Job configuration form

Figure 3.6 shows the job configuration page. This page asks users about the pa-

rameter in which the simulation will be run with. These parameters are instance count, instance type, and HemeLB core container version. Instance count will determine how many compute node will be provisioned for this simulation by HemeWeb. Instance type will determine what type of compute node will be started, and the HemeLB core container version will determine what version of the container the compute node will use to run HemeLB simulation. After all of these parameters are set, users will be asked to confirm the job execution in the overview page. In the page, the user can then finally queue the job into the queue system.

3.4.2.3   HemeLB simulation

Once job instance is queued into the simulation queue, a free asynchronous worker will pop the queue and run the job. The worker will start up the configured amount and type of server instance from the cloud provider. These instances will then be further reconfigured by an ansible script so that it points to the correct master instance address. Next, input files are shared via Networked File System(NFS), the compute units will mount the input folders to their instance.

The correct HemeLB core container version will be pulled from Docker hub in the next step. This step will skip the download if the container asked are already cached in the image for compute units which are prepared on the deployment part. After all of these are done, then the simulation can finally begin. Master node will issue an MPI command to be run by the leader of compute nodes. The leader of compute node then will run this MPI command in the Docker container. This command will be run on multiple compute node if it is configured as such in the previous steps.

The HemeLB simulation will run until outputs are produced. The output will be written back to the correct output folder in the shared folder. This means that the master instance will have access to the outputs file and can do further processing. This step ends with the termination of the instances.

Figure 3.7: HemeWeb Job logs

Figure 3.7 shows how the Job execution will produce logs that can be viewed on HemeWeb web application. During simulation, logs are written to the respective job folder and are served to the browser by the HemeWeb app. Showing the logs allows users to easily view the progress of the job or even debug failed job.

3.4.2.4 Post-processing

After HemeLB simulation is finished, HemeWeb web app will do some post-processing steps to make sure the output files can be viewed easily. The outputs from HemeLB simulation are structured in such a way that makes it efficient to write in parallel. However, these outputs cannot be viewed by visualization system like Paraview. What HemeWeb will do is to pipe the output files to two python scripts that will format the output into a format that can be understood by ParaView.

However, the post-processing steps are not done yet. There are further steps that HemeWeb took to make sure that the simulation files, configurations, and results are preserved externally. HemeWeb will package the job directory, compressed it, and upload it into persistent storage that cloud vendors provide. As the time of writing, HemeWeb only supports amazon simple storage service. The simulation files are uploaded to this storage and made accessible to the public so other HemeWeb instance can use them. Also, with the job files persisted on persistent storage, the next HemeWeb instance deployed can take advantage of these files that it can use it as pre-

vious jobs to be used on current deployment

3.4.3 Implementation Challenges

In this section, I will try to outline and discuss the challenges in implementing this project, and if any, the solution that I choose.

3.4.3.1 Cloud vendors features and API difference

The challenge in developing the deployment script is the difference of cloud vendors' API and features. This has led to some problems when trying to create a common API to do a certain task. One notable problem is that the absence of image creation from running instance feature from one of the cloud vendors. Image creation feature is not an essential requirement of the project. However, with an image creation, the compute nodes that will be requested by the web application can be configured much quicker because all the pre-configuration that are done during the deployment phase. However, one of the cloud vendors, Digital Ocean, does not have this feature. This creates a situation where there is no elegant way to create image with the deployment script and users are asked to manually created the image on the web interface
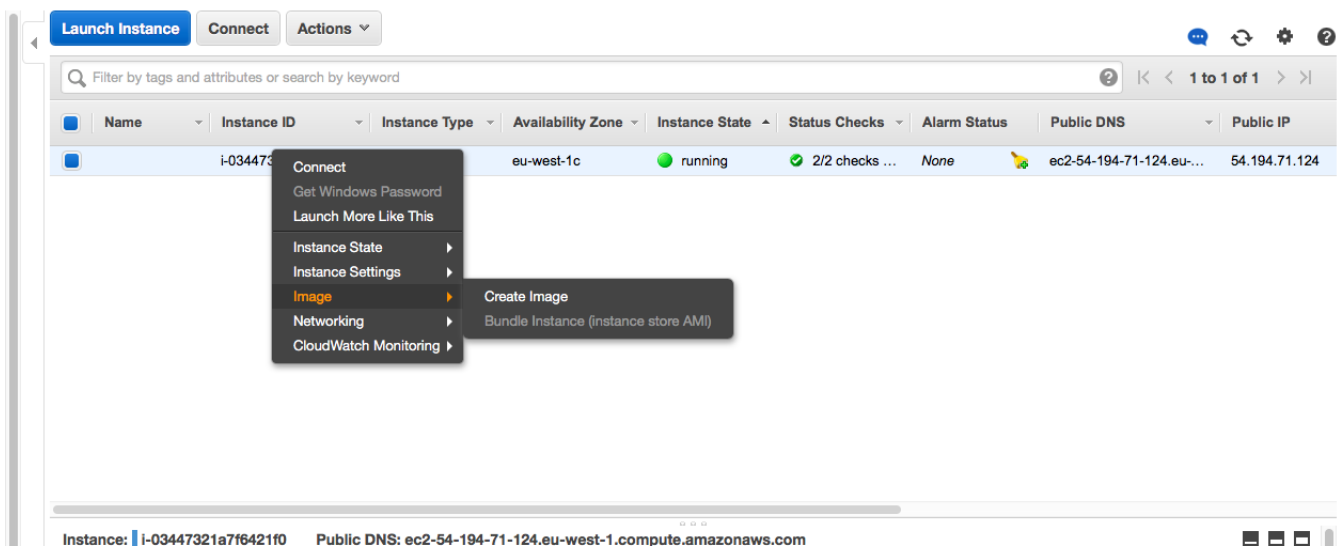
Figure 3.8: Manual image creation instead of automatic

Figure 3.8 shows how users are instructed to manually create an image from running instance. Users have to go the web interface of specific cloud vendors, right-click

on the running instance, and create the image from it. This is a simple workaround which is less complicated compared to accommodating different or missing features and API from different cloud vendors.

Another problem is the different storage API across cloud vendors. Due to the time constraint, I can only prioritize one vendor, Amazon Web Service. However, this means that the codebase is currently tied to one cloud vendors. Features like automatically reading past simulation files from cloud storage and uploading simulation files are tied to Amazon's infrastructure. It is possible to refactor these functionalities out to become more generic, however due to the time limit, I decided not to.

3.4.3.2 Security

Another challenge that I face during the development of HemeWeb is to handle the security of the application. However, security is not the main focus of this work and is apparent in the development of HemeWeb. I will still discuss the security issues so that I can give an objective assessment of the application.

The first security issues that I found is with regards to the compute node security in some cloud vendors. Digital Ocean, for example, does not provide private networking options that blocks other people's node to communicate with our compute node with the private network. Theoretically, this allow other people access to your private compute node if they have the credentials. In this case, I made sure that all the compute nodes have a sensible access policy to deter unauthorized access to the nodes. I only allow SSH access with a public key and disabled password access to SSH. Also, it is also much better to choose cloud vendors that have sensible setting for private networking like Amazon Web Service. On AWS, compute nodes are not accessible to other nodes that are not part of your own private network. This is much more secure and sensible.

"real" private networking option within compute nodes. They have a "shared" private networking options that allow other compute nodes, which are not even on your account have network connectivity to your node.

Another security issue is on how compute and master node share simulation job's files. It currently uses Network File System without any security measures towards the nodes that try to mount it via the private network. There is an opportunity to secure this communication by encrypting the job files, but it is not currently done.

Chapter 4

Evaluation

In this chapter, I will discuss on the system's evaluation. Success will be measured by answering these questions:

- Can users run a simulation using our system ?

- Can users reproduce past simulations using our system ?

- Are users satisfied in using our system ?

- Are users more likely to run a simulation using our system compared to the command line ?

- How does it perform compared to the existing infrastructure ?

In answering questions above, I conduct two sets of evaluation. First, an online questionnaire was conducted to measure user's experience and system's usability. And secondly, a performance analysis comparing the performance of HemeLB between dedicated hardware against cloud vendors. The questionnaire will answer most of the questions related to user experience and usability above, while the performance benchmark will be the basis for performance evaluation and justification in using HemeWeb.

4.1 Questionnaire

In measuring user's experience, I created a questionnaire for HemeWeb using google form. The questionnaire was live for 10 days, from 3rd of August 2016 until 12th of August 2016.

At the start of the questionnaire, respondents are asked about their background information. This information will be used to provide demographic insight on the respondents and how it affect the type of answers the respondents will most likely give. Respondents will be asked to fill out their age, gender, job, discipline, and level of familiarity with various software tools. From their responses, I can determine whether the sample population is representative of the target audience.

The questionnaire is set to make tester run specific scenarios that HemeWeb is developed for. These scenarios are running a simulation using the web interface and reproducing past simulation with it. Testers running the scenario are given an option to skip the scenario and go straight to the questions below it to measure their experience if they find it too difficult.

The first scenario asks testers to run a simulation with given inputs. The questionnaires list two input files, a geometry file and a HemeLB configuration file that testers will need to download to their computers. After downloading the input files, testers are asked to open their browser and go to a specific URL where HemeWeb was deployed for the evaluation purposes. In that URL, testers will then add a new simulation job with the downloaded files, configure the job, and submit the job to the queue. The scenario ends when the job is submitted.

The second scenario asks testers to reproduce past simulation with a given URL that contains past simulation files. It asks testers to create a new job from the given URL instead of using files that are downloaded in the past scenario. After creating a new job, testers then will change some configuration and parameters from the past scenario and submit the job. The scenario also ends when the job is submitted.

Following each scenario are questions to measure whether users skip the scenario and an after-scenario questionnaires that users have to answer. The after-scenario questionnaire is based on the usability measurement at IBM developed by James R Lewis [30]. It measures users' usability satisfaction with the system with regards to given scenario. The questionnaire give 3 statements which testers should agree or disagree, they are:

- Overall, I am satisfied with the ease of completing the tasks in this scenario

- Overall, I am satisfied with the amount of time it took to complete the tasks in this scenario

- Overall, I am satisfied with the support information (online-line help, messages, documentation) when completing the tasks

In addition to above questions, there are some questions about the tester's willingness to do exactly the same tasks as the scenario, but with the command line. This question will measure user's willingness in using the command line interface compared to the web interface.

After running both scenarios, testers are then redirected to the final questionnaire. The Post Study System Usability Questionnaire which is based on the same work by James R Lewis [30]. In this questionnaire, testers are given 19 statements where they should agree or disagree, they are:

- Overall, I am satisfied with how easy it is to use this system

- It was simple to use this system

- I can effectively complete my work using this system

- I am able to complete my work quickly using this system

- I am able to efficiently complete my work using this system

- I feel comfortable using this system

- It was easy to learn to use this system

- I believe I became productive quickly using this system

- The system gives error messages that clearly tell me how to fix problems

- Whenever I make a mistake using the system, I recover easily and quickly

- The information (such as online help, on-screen messages, and other documentation) provided with this system is clear

- It is easy to find the information I needed

- The information provided for the system is easy to understand

- The information is effective in helping me complete the tasks and scenarios

- The organization of information on the system screens is clear

- The interface of this system is pleasant

- I like using the interface of this system

- This system has all the functions and capabilities I expect it to have

- Overall, I am satisfied with this system

In addition to the above questions, testers also will be asked to list the most negative aspects and positive aspects of the system if they have any. These questions will measure users satisfaction with the overall system, whether it is useful, whether the information given by the system is any good, and the interface quality.

4.2 Performance benchmarks

The second part of the evaluation is the performance benchmark. HemeWeb is running HemeLB simulation outside its original scope of being used on a highly parallel computing resources like a supercomputer. This could have an interesting impact on the performance of the system because cloud vendors, while being easily accessible and provisioned, have an underlying infrastructure difference with the stand-alone infrastructure. This performance benchmarks will then measure whether the impact on the performance justify the claimed usability benefit that we measure on the first half of the evaluation.

The performance benchmark will be done by the internal tooling that is available inside HemeLB. Every HemeLB simulation will produce a report file that measures the performance of said simulation. On this evaluation, I will compare the performance of HemeLB simulation on three different infrastructure to measure the performance impact on having HemeLB simulation on the cloud with Docker containers.

They are:

- ARCHER supercomputer

- INDY2 HPC Cluster

- Cloud computing infrastructure with AWS EC2

On these three infrastructure, ARCHER supercomputer has the oldest processor architecture. It use a three-year old 2.7 GHz, 12-core E5-2697 v2 Ivy Bridge processor as the basis of its compute node. Each compute node consist of this two processors. INDY2 is a newly installed HPC infrastructure on EPCC that is currently in early-access testing. It has the newer Broadwell-based 18-core Intel Xeon CPU E5-2695 v4 @ 2.10GHz. Lastly, for AWS-EC2, we use Amazon's c4.8xlarge EC2 instance

which has Haswell-based E5-2666 v3 processor that has 18 virtual cores[1] or 36 vCPU with hyperthreading[2]. Essentially, we will be using 18 core on AWS-EC2 instance for the simulation, because HemeLB is a compute bound task that do not benefit from hyperthreading.

Next, the three different architecture also have different networking interface that could contribute to the performance difference. While AWS-EC2 offer 10 Gigabit per second interface, ARCHER use a Cray Aries router interconnectivity that provides higher network throughput (ranging from 12.5 Gigabit per second to 14 Gigabit per second) throughout the infrastructure. Finally, INDY2 has FDR InfiniBand network interface connected to each compute node that provides 54.5 Gigabit per second.

For the simulation, we used an input file which has 4,520,681 fluid sites[3] [4]. With these files, we are going to do a strong scaling analysis, where we keep the problem size the same while we increase the core count. According to the performance analysis by Groen et al. [8], HemeLB scales up near-linearly up to 32,768 cores. It also performs near its maximum efficiency when using 5,000 to 500,000 sites per core. This means that for the problem size we use, HemeLB should perform near maximum efficiency when using 9 cores up to 900 cores. Above that core counts, HemeLB simulation will incur a performance penalty.

The benchmark on ARCHER supercomputer will be the gold standard of the performance. It's a gold standard because the infrastructure has a clear purpose of being used for the HPC application. It has the necessary resources and components that are tailored for it. It should be the ideal performance scenario. Next, the INDY2 as the locally installed and managed HPC infrastructure. INDY2's performance will paints a picture if how HemeLB will perform on a private local infrastructure which is built to run highly parallel jobs. Based on all these infrastructure difference, our hypothesis is that AWS-EC2 will provide slower performance compared to INDY2 and ARCHER.

In running the simulation, I had help from Dr. Rupert Nash to run HemeLB simulation on ARCHER and INDY2. I need help because I did not have access to both of these infrastructures. On AWS-EC2 infrastructure, however, I run the HemeLB scenario myself. HemeLB simulation was run using the HemeWeb interface on AWS.

---

[1]https://aws.amazon.com/ec2/virtualcores/

[2]https://aws.amazon.com/ec2/instance-types/

[3]Available online at https://github.com/SeiryuZ/HemeWeb/blob/master/deployment/roles/hemeweb_master/files/990_Example2-skeleton_corrected_tubed_smoothed.gmy

[4]Config for the simulation can be found online at https://github.com/SeiryuZ/HemeWeb/blob/master/documents/resources/evaluation/performance/config.xml

Chapter 5

Analysis

5.1 Usability result and analysis

In evaluating HemeWeb's usability and capability to run and reproduce simulation, I conducted a usability evaluation via an online survey hosted by Google Form[1]. Respondents are given 2 tasks to complete, which are to run a simulation and reproduce a past simulation. After the tasks, they are given statements to respond to. Based on the answers, I can make an analysis about HemeWeb's usability.

5.1.1 Demography
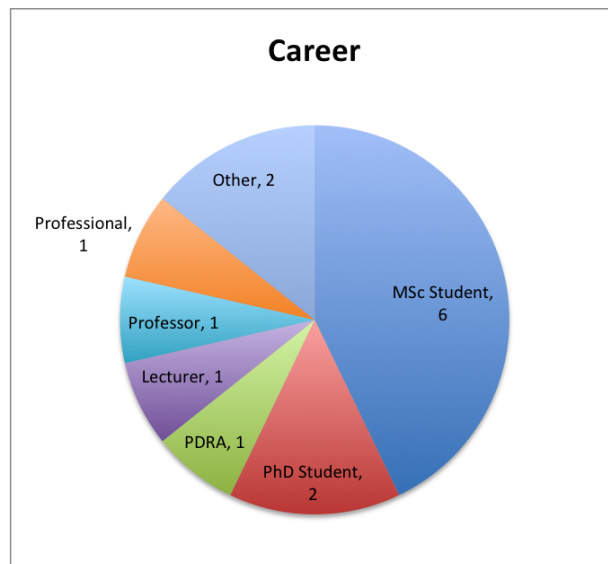
Figure 5.1: Career stage distribution

---

[1]https://goo.gl/forms/toYsRwnGIGumMBUD2

Figure 5.2: Career discipline

The survey was filled by 16 respondents over the period of 10 days (3rd August 2016 - 12th August 2016)[2]See Appendix D) [3]. From all the responses, one of the respondent was unable to run both of the tasks, citing errors preventing him to run the scenarios which we cannot reproduce. Another respondent also skip the second scenario without specifying reasons or contacting the author. Due to this problem, we have to remove these two responses from our analysis because it will not add meaningful information about the usability of the system when the scenarios are not run. In total, we got 14 valid responses out of the questionnaire period.

Figure 5.1 and 5.2 illustrates the distribution of career stage and discipline of our participants. Based on this distribution, we can further analyze the response we get on the survey questions based on their background. One meaningful comparison we can make is when we classify respondents based on their informatics-related discipline. 10 of our respondents are related to informatics background, while 4 of the respondents can be considered as domain experts which consists of Biologist, Clinician, and Biophysicist.

---

[2](

[3]Raw and compiled responses can be found at `https://github.com/SeiryuZ/HemeWeb/tree/master/documents/resources/evaluation/usability`
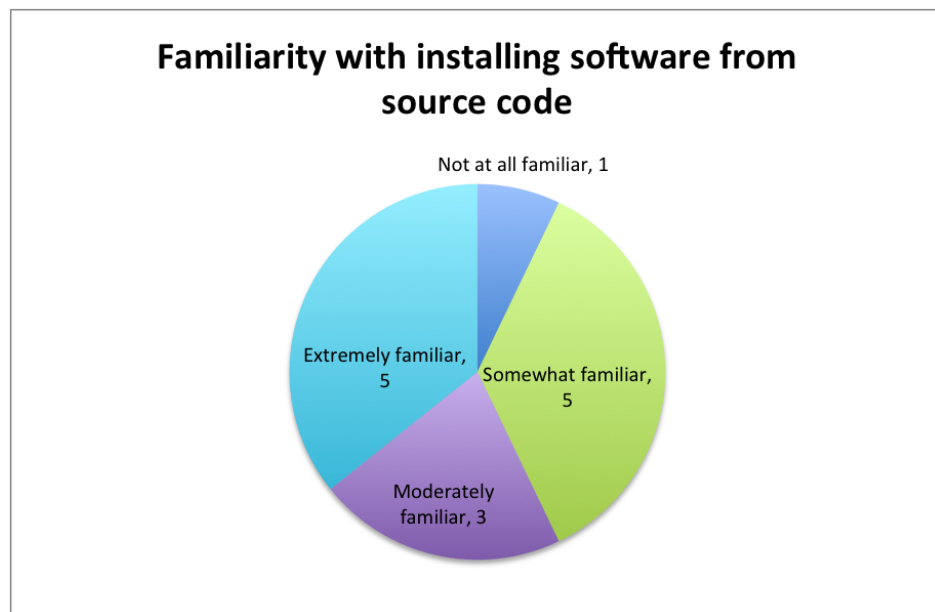
Figure 5.3: Familiarity with installing software from source code
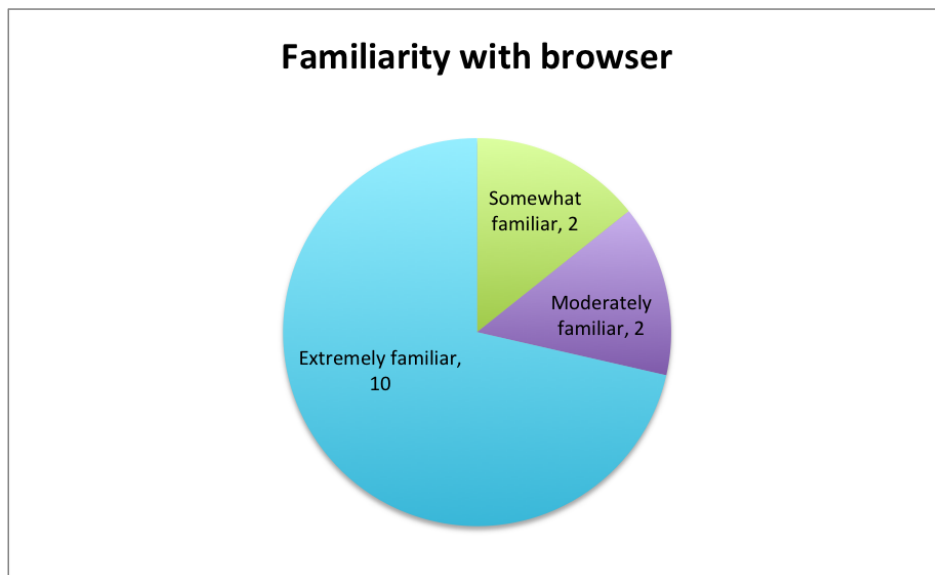
Figure 5.4: Familiarity with web browser
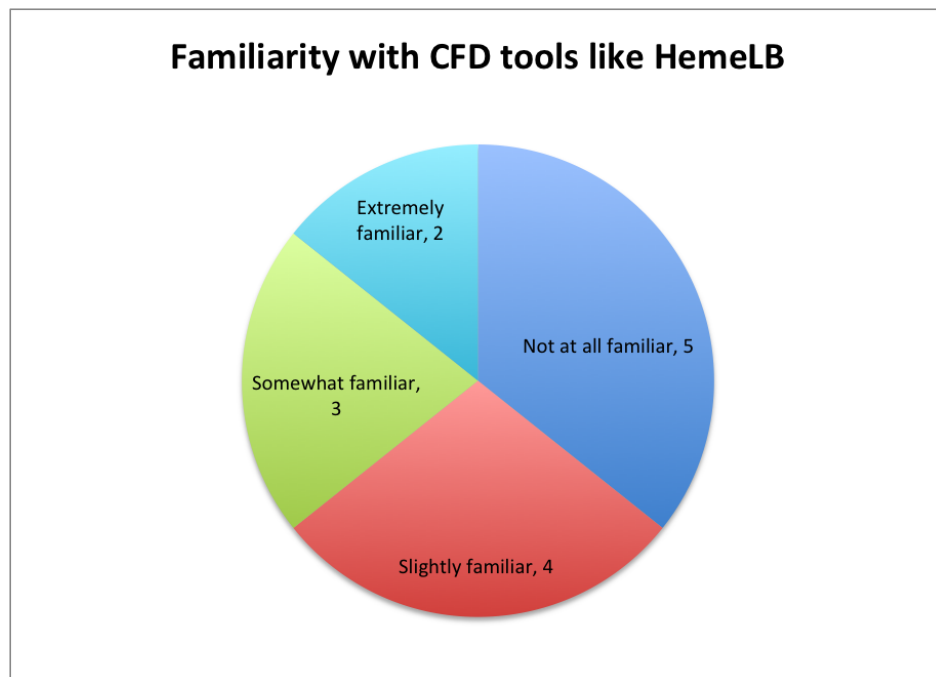
Figure 5.5: Familiarity with CFD tools like HemeLB

We can also further classify analyze the respondents' response based on the familiarity with browsers, computational fluid dynamic tools, and installing software from source code. All of this are shown in Figure 5.3, 5.4, and 5.5.

5.1.2 Scenario 1: Run a HemeLB simulation

In this scenario, respondents are provided with two input files necessary for running a simulation. Respondents are asked to download the files beforehand and follow the instructions provided in the online questionnaire to run a HemeLB simulation using HemeWeb. After running the simulation, Respondents are then asked to state agreement with three positive statements about HemeWeb that will measure their satisfaction with HemeWeb in running a HemeLB simulation scenario.
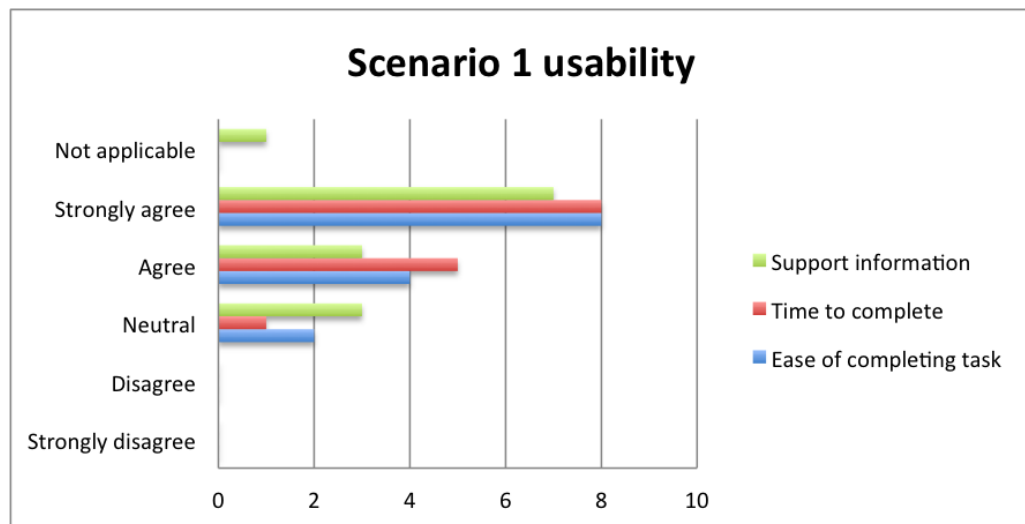
Figure 5.6: Scenario 1 usability

Figure 5.6 show the responses for the three positive statements about HemeWeb. From the 14 valid responses we get, all of them did not skip the instructions to run the simulation. From the responses, respondents tend to agree that they are satisfied with HemeWeb in running a HemeLB simulation. Respondents tend to equally agree on the three positive statements about the ease of completing a task, time to complete, and supporting information available to help complete the task. However, one respondent fills out "Not Applicable" towards the statement about HemeWeb giving them enough support information.

In addition to the general sentiment of the respondents, we can also put a number value to the responses to further measure the satisfaction objectively. We can calculate the After Scenario Questionnaire(ASQ) score. To do this, we assign an integer value for each response; 1 for "Strongly disagree", 2 for "Disagree", 3 for "Neutral", 4 for "Agree", and 5 for "Strongly agree". With this value, we can take the mean of the response for each question as a single ASQ score for the respondent. If a respondent skips a question, we can take the average of the remaining responses as the score. With this mechanism, we can calculate the respondent's average ASQ score, which is 4.4 when rounded. This score falls between "Agree" and "Strongly agree", therefore, we can conclude that in general respondents are satisfied with HemeWeb with regards to running a HemeLB simulation.
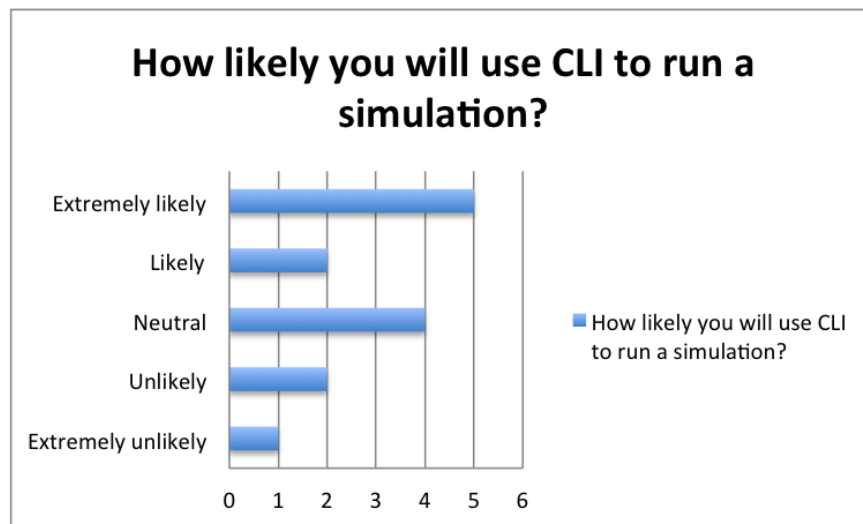
Figure 5.7: Scenario 1 command line preference
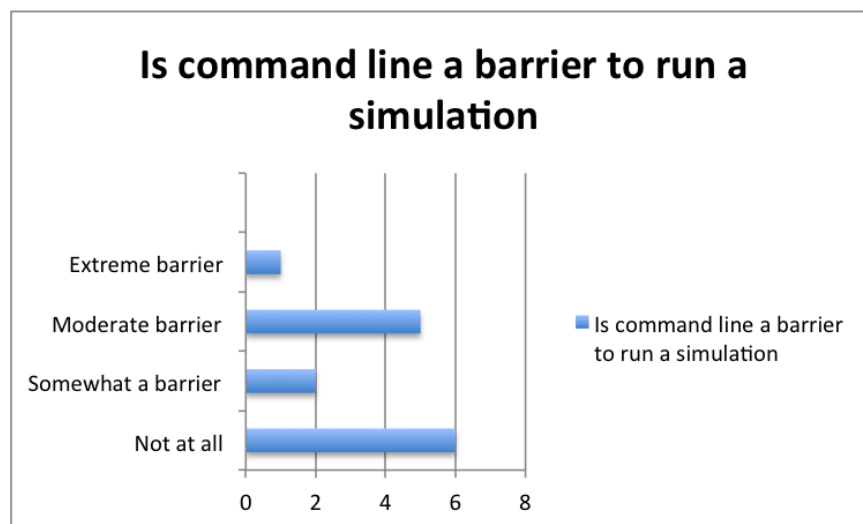
Figure 5.8: Scenario 1 barrier in using command line

After the above sentiments about using HemeWeb to run a HemeLB simulation, respondents were asked about their sentiment about using Command Line Interface(CLI) to do the same activity. Figure 5.7 and 5.8 shows the respondent's responses. Respondents are generally open to the likelihood of them running a simulation using CLI. If we quantify the results like we did with the ASQ score, we get 3.6, which is between neutral and likely.

The distribution of the responses also is quite spread out, that respondents fill out all possible responses with the highest frequency being "Extremely Likely" with 5 responses. However, we have to keep it mind the general background of the respondents that may explain the highest frequency answer being "Extremely Likely", which is

about familiarity with installing software with source code. To build software with source code, one will interact with the command line interface quite often. Thirteen of our respondents answered at least somewhat familiar with building software from source code, that can explain that our respondents are mostly quite competent in operating command line interface and would not shy away in using the command line interface. In addition to that, two out of three respondents has backgrounds in informatics and computational science. This could in effect explains why the respondents feel they are likely and extremely likely to do the same in CLI.

However, not all respondents who are at least somewhat familiar with building software from source code skew towards to the likely side of using CLI. There are respondents that, while familiar with the interface, running a HemeLB simulation using CLI is unlikely to be done by them. These responses might be explained by respondents' sentiment in using command line interface. This is further supported by the response of CLI being a barrier for the respondents. While six of the respondents think it is not a barrier at all to use CLI, eight of the respondents answered at least it is somewhat a barrier. With five of the nine respondents, feel it is a moderate barrier, and one of the nine feel it as an Extreme barrier. From these results, we can safely say that using command line interface is a form of a barrier to run HemeLB for almost two-third of the respondents.

5.1.3   Scenario 2: Reproduce past simulation

In the second scenario, respondents are asked to reproduce past simulation with HemeWeb. They are given instructions in the online questionnaire to create a HemeLB simulation job from past simulation. They have to enter a URL that contains past simulation job and modifies the simulation parameters to avoid only replicating the past simulation without changes. After reproducing past simulation, respondents are then asked to state agreement with the same questions like they had in the first scenario. These questions will also measure their satisfaction with HemeLB in reproducing past simulation.
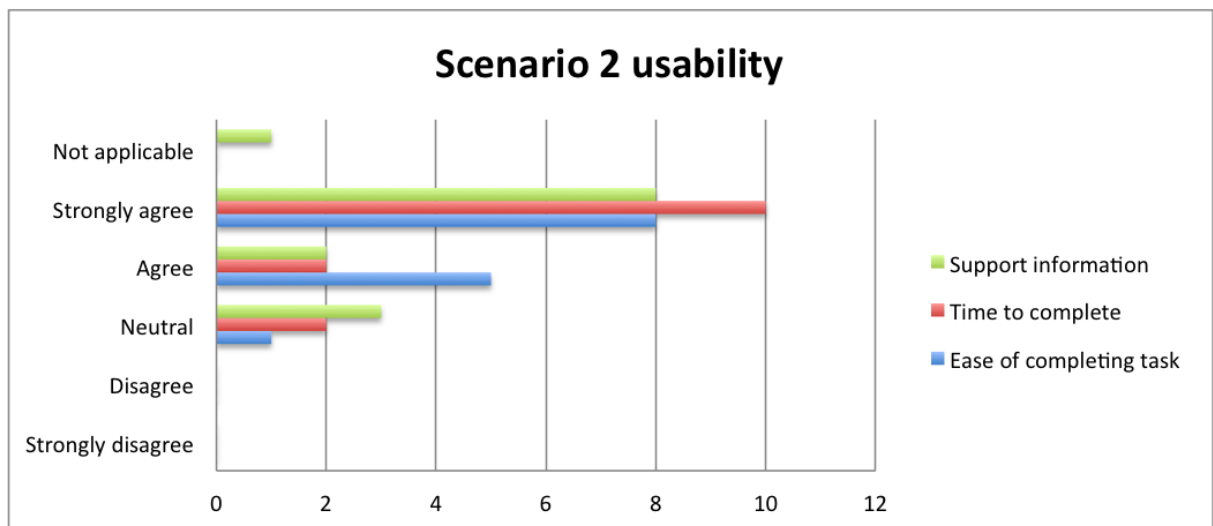
Figure 5.9: Scenario 2 usability

Figure 5.9 shows the general sentiment to the three statements that we provided. Generally, the response tends to skew towards agreeing that respondents are satisfied with HemeWeb with regards to reproducing past simulation. However, it has to be noted that 1 of the respondent skipped the instructions. This particular respondent did not give out any comments about encountering any problems, so we cannot deduce whether his skipping the instruction is due to usability problems or he just wants to skip it. Without further information, we cannot deduce why he skips the instructions and his response is one of the two taken out from the overall responses.

Also, similar to the first scenario, one respondent fills out "Not Applicable" to support information statement. In a nutshell, respondents tend to agree that HemeWeb is satisfying to use for the purpose of reproducing past simulation. If we convert the responses to a numerical value, we will get 4.5 of ASQ score. Which is in line with the sentiment that I described. The ASQ score falls between agreeing and strongly agree towards the positive statements we provided.
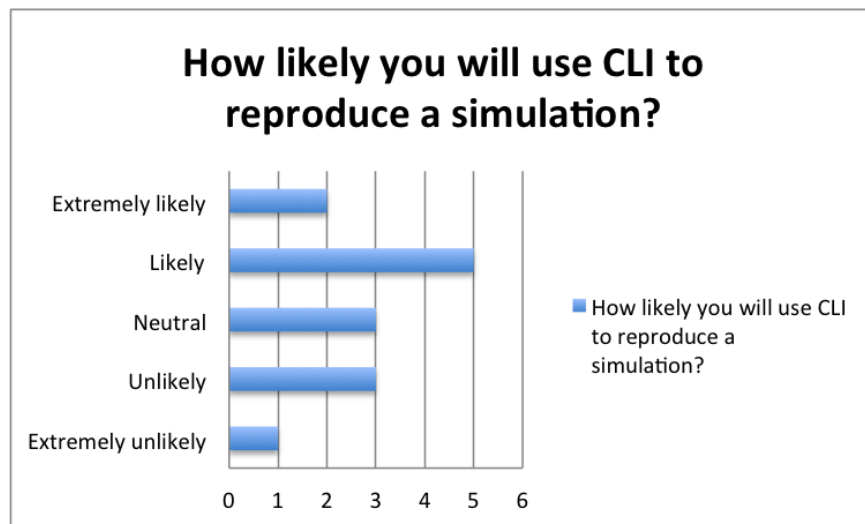
Figure 5.10: Scenario 2 command line preference
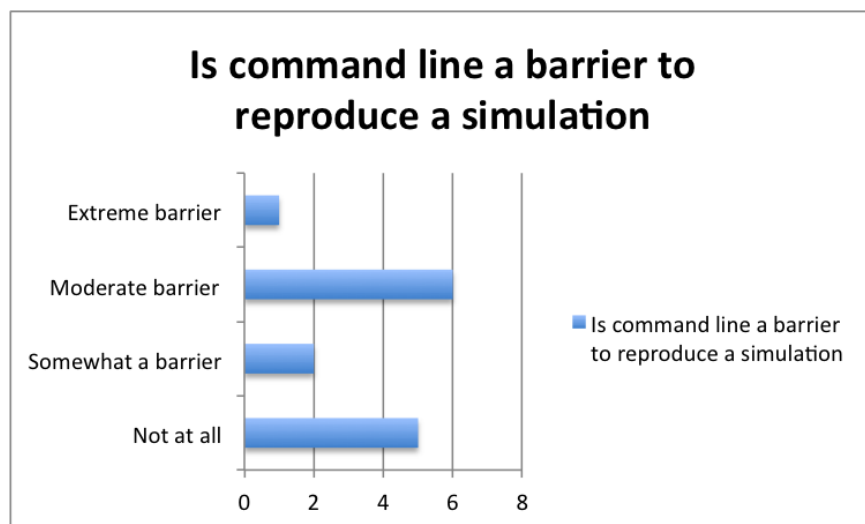
Figure 5.11: Scenario 2 barrier in using command line

Respondents are then given high-level steps to reproduce past simulation using the command line interface. Figure 5.10 and 5.11 shows the respondent's response. Similar to the first scenario, the respondents are generally open to the idea of operating command line interface to do their tasks. However, the scenario to reproduce past simulation provides a bit difference in the distribution of the answers. While in the first scenario the highest frequency of the answer is "Extremely likely", the highest frequency response in this scenario is "Likely".

This change of the highest frequency might be because there are extra step which might complicate the scenario to reproduce past simulation. respondents are more hesitant to answer "Extremely likely". However, the general nuance of the answer is

still the same, respondents are generally not afraid of using command line interface to do this task. If we convert the answers into a numerical form like we did before, we got 3.3, which is between neutral and likely.

In addition, same observation as the first scenario can be made. While some respondents are likely to reproduce past simulations, there are those who are not likely to reproduce past simulations even with their background that have dealt with building software from source code. This can be explained by Figure 5.11 where only five of the respondents did not feel that using CLI is a barrier at all. 2 of the respondents feel using CLI is somewhat a barrier, six feel it as a moderate barrier, and one feels it as an extreme barrier. This means that this nine respondents agree that CLI is a form of barrier to them, however small it is.

5.1.4 Overall usability

Here, I will discuss the overall usability of the system using the Post Study System Usability Questionnaire(PSSUQ) as the basis. This part of the questionnaire consists of 19 questions that can be divided into measuring three different component of the systems. These are system usefulness, information quality, and the interface quality. I will discuss each of them in details.

Table 5.1: System usefulness

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree	Not applicable
Overall, I am satisfied with how easy it is to use this system	0	0	0	7	7	0
It was simple to use this system	0	0	0	1	13	0
I can effectively complete my work using this system	1	1	1	3	4	4
I am able to complete my work quickly using this system	0	1	2	3	6	2
I am able to efficiently complete my work using this system	0	1	2	4	5	2
I feel comfortable using this system	1	0	0	3	10	0
It was easy to learn to use this system	0	0	1	0	13	0
I believe I became productive quickly using this system	0	0	2	3	5	4

Table 5.1 shows the respondents sentiment towards positive statements about the system usefulness. Generally, we can see the distribution of respondents mostly agreeing with the statements presented. Ignoring the respondents that answered with "Not

applicable", we can find the distribution is skewed to the agreeing side of the statements. There are respondents that disagree or even strongly disagree with some questions. However, the frequency is much lower compared towards the frequency of people agreeing to the statements.

Converting the response into a numerical value, we can get the score of 4.4 of system usefulness, which is quite high. However, there are still some improvements that can be made, and it is apparent in the feedbacks[4] of the system we got. For example, one respondent find the job configuration inflexible, another respondent find Firefox does not render the interface correctly. All of these can be addressed in the next iteration of HemeWeb development. In addition to these sentiments, there are respondents that respond to the statement by answering Not applicable. It is mostly on the framing of the simulation as a 'work'. Respondents might have no point of reference whether using the HemeLb via HemeWeb is quicker or efficiently because they are new to the system. This is apparent from their background that they don't have familiarity with HemeLB.

Table 5.2: Information quality

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree	Not applicable
The system gives error messages that clearly tell me how to fix problems	0	1	1	4	2	6
Whenever I make a mistake using the system, I recover easily and quickly	0	0	4	1	3	6
The information (such as online help, on-screen messages, and other documentation) provided with this system is clear	0	0	3	3	5	3
It is easy to find the information I needed	0	0	2	3	6	3
The information (such as online help, on-screen messages, and other documentation) provided for the system is easy to understand	0	0	2	5	6	1
The information is effective in helping me complete the tasks and scenarios	0	1	3	2	7	1
The organization of information on the system screens is clear	0	0	2	8	4	0

Table 5.2 gives more insight on the respondents sentiment on HemeWeb's information quality. This includes information like error messages, documentation, the organization of this information, and etc. From seven positive statements that we presented.

---

[4]See Appendix D - Figure D.4

most respondents tend to agree that the information quality is good. The overall sentiment is more towards agreeing compared to disagreeing. If we convert the sentiment into a score, it will get 4.1 average score, Which falls under "Agree" and "Strongly agree" sentiment.

However, there are more respondents answering "Not applicable" in this part of the questionnaire. This led to the remaining answer having more contribution towards the average sentiment because many of the answers are not counted. For example, there are 6 ignored responses on the first statement about error messages. This is likely because the scenario given to the respondents will not give out an error message if they are correct the first time. In the end, however, respondents tend to agree that information quality of HemeWeb is good.

Table 5.3: Interface quality

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree	Not applicable
The interface of this system is pleasant	1	0	1	7	5	0
I like using the interface of this system	0	1	1	8	4	0
This system has all the functions and capabilities I expect it to have	1	1	1	4	4	3

Table 5.3 shows the respondents sentiment with regards to HemeWeb's interface quality. In it, we see respondents strongly disagree and disagree with positive statements about the interface. While in general, most respondents tend to agree that the interface is good, some disagree. When we see this particular response, it seemed that the respondent had trouble with the interface not rendering correctly in firefox. This feedback means that the browser compatibility of HemeWeb should be improved. There are parts of the interface which are broken if it is viewed on firefox. There are other respondents who dislike the XML editor based on their feedbacks. However, all this are subjective in nature and further testing of the interface is needed.

However, if we convert the respondents answer to a numerical value, we got 4.0, which is "agree". It shows that while some respondents find the interface quality is not up to their standards, some find it good enough although a lot can be improved. Browser compatibility, web form choice, user experience, and etc should be improved on the next iteration of HemeWeb.

Table 5.4: Overall user satisfaction quality

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree	Not applicable
Overall, I am satisfied with this system	0	0	1	7	6	0

The last part of the question is the overall perceived satisfaction that the user has in using the system. The distribution of the response can be found in Table 5.4. Most of the respondents gravitate towards agreeing and strongly agreeing that they are satisfied with the system. When we convert it into a numerical mean, it is 4.4, which is between "Agree" and "Strongly agree". To get the overall user satisfaction score of the study, we average the numerical score between the 19 questions and we got 4.2. This means that the respondents mostly agree that they are satisfied with how the system performs, inform, and looks like.

5.2 Performance result and analysis

In this section, I will discuss the performance benchmark of HemeLB simulation done in three infrastructure, ARCHER supercomputer, INDY2 HPC cluster, and AWS EC2. Understanding the difference in performance is essential for HemeWeb project to justify the flexibility that cloud vendors provide.

In this benchmark, we are going to rely on the report file that HemeLB produced. It is produced after each simulation is done and contains performance metric of each simulation like how many seconds are spent on communication, initializations, reading input files, and most importantly total simulation time. We are going to take a look at the total simulation time because it is the metric that captures the difference in the infrastructure better. We ran the exact same simulation scenario on all three infrastructure . On each infrastructure, the simulations were done with an increasing number of compute node to measure the scalability of HemeLB. From the produced output, then we can measure the performance difference between architecture.

Table 5.5: HemeLB performance on INDY2 vs ARCHER vs AWS EC2

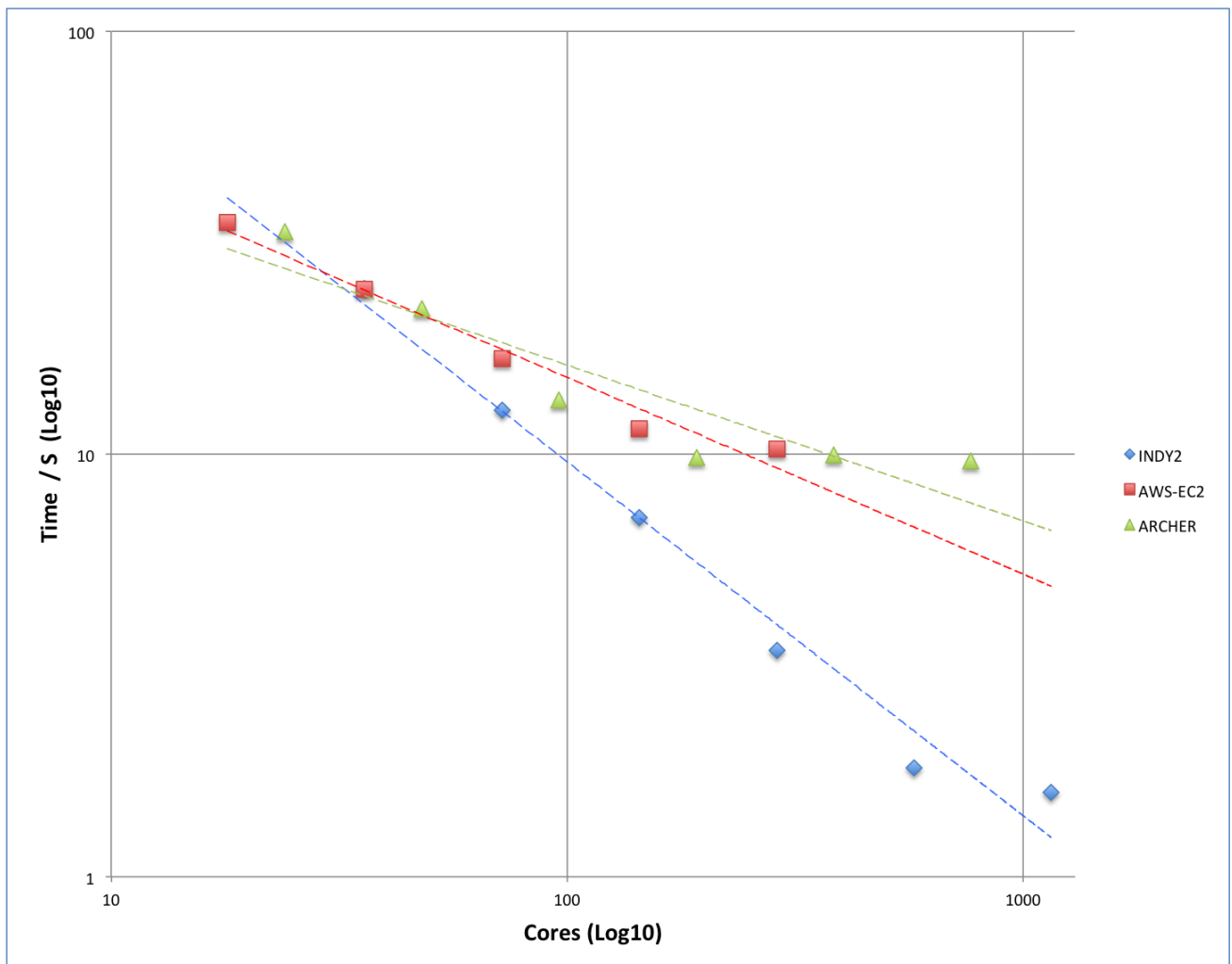| # Compute node(s)	INDY2		ARCHER		AWS-EC2	
	Core	Time / s	Core	Time / s	Core	Time / s
1	36	24.7	24	33.6	18	35.3
2	72	12.7	48	22.1	36	24.6
4	144	7.08	96	13.4	72	16.8
8	288	3.44	192	9.81	144	11.5
16	576	1.81	384	9.94	288	10.3
32	1152	1.58	768	9.64	N/A	N/A

Figure 5.12: HemeLB performance comparison

Table 5.5 shows the performance comparison between HemeLB total simulation time between AWS-EC2, INDY2, and ARCHER. This shows that AWS-EC2 performance is much slower compared to the dedicated infrastructure likes INDY2 and ARCHER. AWS-EC2 is slower 3 25% compared to ARCHER, and 42 469% com-

pared to INDY2 on various compute node size. However, it has to be noted that on each infrastructure, one compute node have different core count. This could affect the performance difference we got from this result.

Figure 5.12 also shows the difference in performance between the three infrastructure. INDY2 performed much faster than the other infrastructures. In addition, its performance also scales almost linearly up to 576 cores and had a performance degradation on 1152 cores. This closely follows the performance analysis by Groen et al. [8].

Next, ARCHER perform better than AWS-EC2. While it is still slower than INDY2, ARCHER provide more performance per compute nodes when we compare it to AWS-EC2. However, on 192 cores up to 768 cores, the performance became stagnant. It didn't achieve a better improvement as we scale. The performance difference between INDY2 and ARCHER can be explained by the different load on each system at the time of benchmark. INDY2 is on the early-access testing phase where at the time of execution, Dr. Rupert Nash was the only person using the infrastructure. On the other hand, when Dr. Rupert Nash executed HemeLB on ARCHER, the utilization was very high, which means that the HemeLB job potentially has to share network interconnectivity with other jobs.

Finally, AWS-EC2 perform slower on each compute nodes count we tested. The performance improved on increased the core count. However, this improvement is not nearly linear as modeled by Groen et al. [8]. This result could have many causes, but generally still in line with the benchmark that Mehrotra et al. did on NASA's HPC application [19]. The performance degradation comes from the network and the virtualization overhead that this cloud platform has. In addition, we do not have control over how the compute node are arranged on the cloud platform. We could have nodes with varying degree of interconnectivity on each simulation. A simulation might have started compute nodes near each other in the Amazon's data center, the next simulation might start compute nodes which need longer network interconnectivity.

Table 5.6: HemeLB parallel efficiency table of INDY2 vs ARCHER vs AWS EC2

INDY2			ARCHER			AWS-EC2		
Core	Speedup	Efficiency	Core	Speedup	Efficiency	Core	Speedup	Efficiency
36	1	1	24	1	1	18	1	1
72	1.94488189	0.972440945	48	1.520361991	0.760180995	36	1.43495935	0.717479675
144	3.488700565	0.872175141	96	2.507462687	0.626865672	72	2.101190476	0.525297619
288	7.180232558	0.89752907	192	3.425076453	0.428134557	144	3.069565217	0.383695652
576	13.64640884	0.852900552	384	3.38028169	0.211267606	288	3.427184466	0.214199029
1152	15.63291139	0.488528481	768	3.485477178	0.108921162	N/A	N/A	N/A
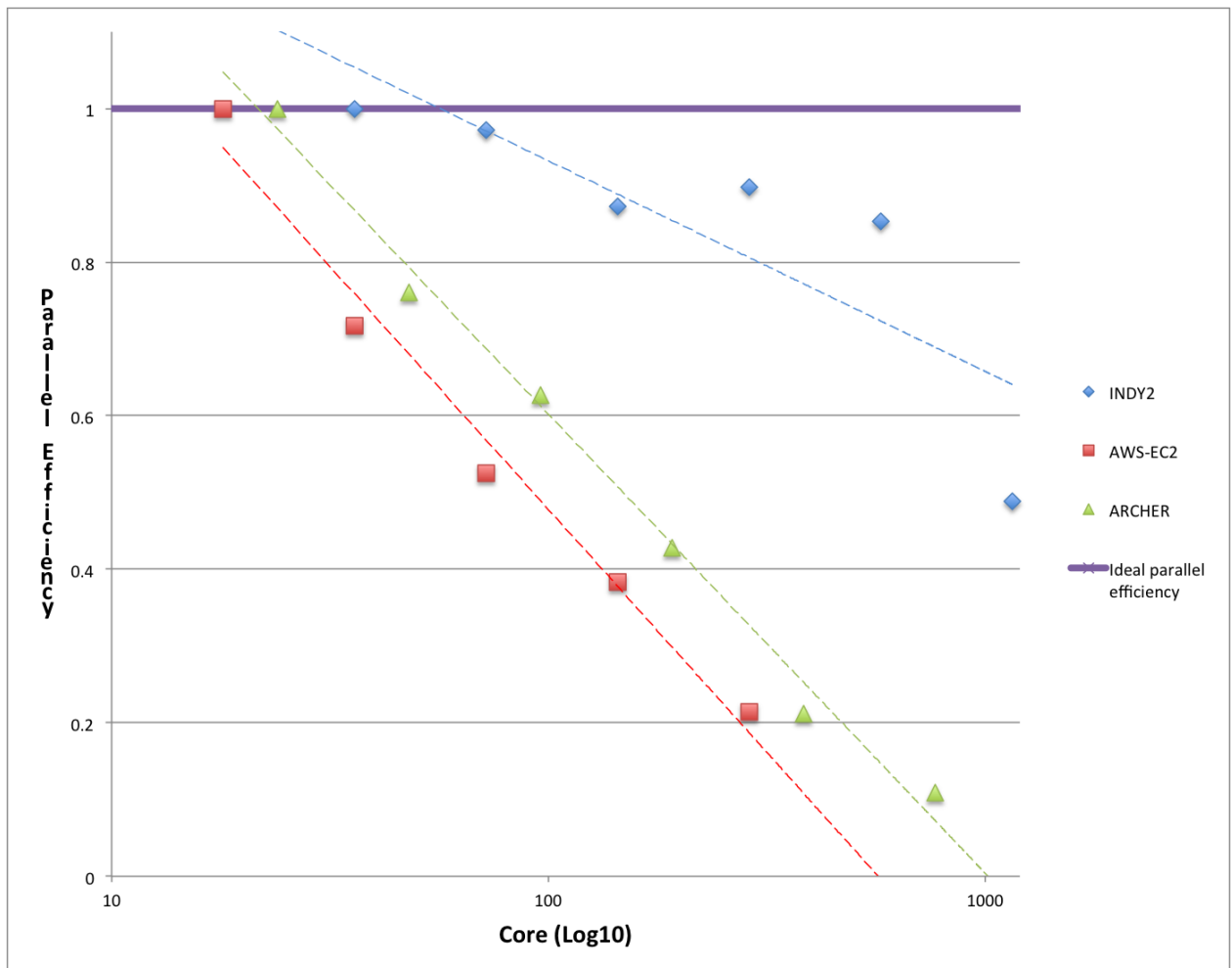
Figure 5.13: HemeLB parallel efficiency plot

From the raw performance information above, we can calculate the parallel scalability of HemeLB on all three infrastructure. Table 5.6 and Figure 5.13 shows the difference of the parallel scalability on the three infrastructures. INDY2 has the best parallel scalability that it has the efficiency nearest to the ideal parallel scalability. However, the last data point produces a much slower speedup compared to other points that reduce its efficiency at 1,152 cores.

ARCHER also perform better than AWS-EC2 in this evaluation. However, the parallel efficiency is still far below the level of INDY2 because of the utilization of the system. AWS-EC2 has the lowest parallel scalability as expected. This shows that running HemeLB on Amazon's infrastructure might not scale as well as using dedicated HPC infrastructure. However, on lower compute node count, AWS-EC2 provide a comparable parallel scalability (5% difference) and performance compared to ARCHER under very high utilization. This result concludes that our hypothesis is correct, that AWS-EC2 would provide slower performance in running HemeLB simulation.

5.2.1 Simulation cost analysis

In addition to performance analysis, we can further analyze the notional cost of a simulation job. For this, we need to come up with the way to calculate the simulation cost. This can be done by multiplying simulation time with cost per unit of time on that infrastructure.

Simulation cost = simulation time X cost per unit of time x number of nodes

In running the simulation for the benchmark, we used "c4.8xlarge" Linux instances on EU Ireland region which at the time of writing costs GBP 1.51[5] per hour[6]. Phase 2 XC30 ARCHER supercomputer give access to screened projects by compute hour unit which has the price of GBP 0.20 per compute node for research councils which are partnered and GBP 0.48 per compute node for non-partnered research council[7]. INDY2, however, have no public pricing released by the EPCC yet and cannot be included in this analysis.

Calculating simulation time can be done by taking the performance of compute nodes on each infrastructure divided by the base time. The base time will be chosen from the fastest performance of the smallest compute node count on each infrastructure. In this evaluation, we ignore INDY2 because of missing pricing, so the base time is ARCHER with one compute node that runs for 33.6s. This will make sure that the simulation time across infrastructure shrinks relative to the performance of that time. With this information, we can finally calculate the simulation cost of using various compute nodes on each infrastructure.

---

[5]USD 1.96 converted to GBP on https://www.oanda.com/currency/converter/ at 13th August 2016

[6]https://aws.amazon.com/ec2/pricing/

[7]Calculated on http://archer.ac.uk/access/au-calculator/

$$Simulation\ time\ ratio = \ simulation\ time\ /\ base\ time$$

Table 5.7: HemeLB simulation cost on AWS-EC2 vs ARCHER

Compute node	ARCHER time difference ratio	ARCHER-partner simulation cost	ARCHER-nonpartner simulation cost	AWS-EC2 time difference ratio	AWS-EC2 simulation cost
1	1	0.2	0.48	1.050595238	1.58639881
2	0.657738095	0.263095238	0.631428571	0.732142857	2.211071429
4	0.398809524	0.319047619	0.765714286	0.5	3.02
8	0.291964286	0.467142857	1.121142857	0.342261905	4.13452381
16	0.295833333	0.946666667	2.272	0.306547619	7.406190476
32	0.286904762	1.836190476	4.406857143	N/A	N/A
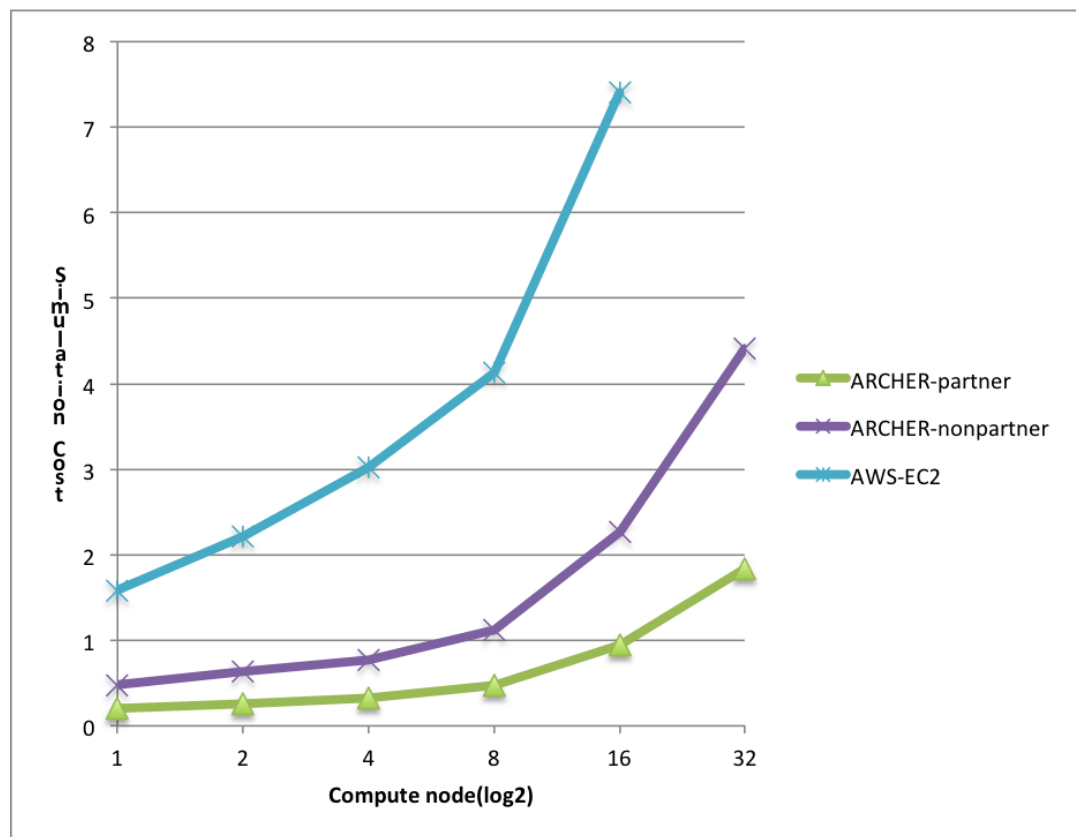
Figure 5.14: HemeLB simulation cost difference

Table 5.7 shows the breakdown of simulation cost between AWS-EC2 with ARCHER with partner institutions and non-partner. From the calculation, AWS-EC2 is 682% up to 846% more costly compared to ARCHER with partner pricing. When we compare to the price of a non-research partner, we get 230% to 290% more cost to run a simulation.

This result is as expected because, with slower performance, we are going to use

compute nodes for longer. Longer usage means more cost to be paid by the users. In addition, cloud platform on AWS-EC2 rent out their services with a premium that allows a certain flexibility that dedicated HPC infrastructure like ARCHER cannot afford. This flexibility like the amount of time needed to first start running a simulation or running an exploratory experiments.

While being slower in running simulation, cloud vendors like Amazon web service allow domain experts to run simulation on-demand. On the contrary, running a simulation, even a small exploratory one, on dedicated HPC infrastructure like ARCHER, require users to apply for a spot, write a proposal, and wait for the next application decision. This is the time of doing simulation that cannot be measured by this evaluation and should be used as one of the basis of using HemeWeb on cloud vendors compared to dedicated HPC infrastructure.

This result means that if you are an existing scientists with access to HPC infrastructure like ARCHER, using HemeWeb might not be the best way to run your HemeLB simulation because of the costs. However, if you are a scientists with no access to HPC infrastructure like ARCHER or INDY2 and does not have a constant demand for compute hour for your simulation, HemeWeb may worth the cost compared to applying for a spot on these HPC infrastructures. In addition, the on-demand nature of HemeWeb fits perfectly with exploratory studies type of job which usually needs a one-off job with relatively small problem size. With HemeWeb, users can start running the simulation on-demand.

5.3   Limitation of the evaluations

5.3.1   Usability evaluation

When observing the evaluation, one might argue that 5 point scale used in the questionnaire is not enough as pointed out by Kraig Finstad [32]. In his research, he argued that 5 point scale for a questionnaire allows more room for respondents to interpolate their response. He compared the same version of usability study but with five-point and seven-point scale and found out that 3% of the respondents answer to five point scale actually are interpolation, while the seven point scale have no interpolation. He concluded that administering usability study using seven point scale will achieve greater accuracy. However, I decided to use the five point scale of the questionnaire because of the limitation of google survey platform.

Figure 5.15: Google form hiding some options on 7 point scale

Figure 5.16: Google form render 5 point scale rating better

Figure 5.15 show the limitation of the platform. The response options are not rendered in complete, as in some are hidden in a scrolling element of the page. This leads to respondents have to scroll left and right to see the options and answers. I find this is a detriment to the respondents to do more work in answering the question so I decided to use 5 point scale to make sure everything is rendered better like shown in Figure 5.16.

Another limitation of the study is the way the study compare the experience of doing the scenarios in the questionnaire between using web browser and using command line. When measuring the usability of HemeWeb, respondents are asked to do the task

in the web browser, compared to the task in command line where it is just described in an overview. This decision is made because we cannot make sure the respondents have the necessary tools to run or reproduce a simulation in their computers. In addition, in order to run the scenario in command line interface, respondents have to install tools on their computer which are timely and prone to errors. I believe this will affect their sentiment in answering the survey questionnaire and decided against it.

5.3.2   Performance evaluation

On running the HemeLB simulation on the three systems compared, we cannot fully isolate the infrastructure from various system loads the infrastructure is handling. The performance might be affected by other jobs running in the system, which is apparent in the case of ARCHER supercomputer where many jobs are executing at the same moment. The same case with INDY2. The performance benchmark might be not as accurate as a completely isolated environment, however, it paints a pretty good picture of performance you get when using cloud vendors compared to the dedicated infrastructure.

Also, another limitation on the performance evaluation is that we cannot get more than 20 instances of c4.8xlarge instances on AWS. This causes the evaluation for AWS-EC2 cannot be done for the 6 compute nodes to compare it directly with INDY2 and ARCHER. Another limitation is that each infrastructure has different core counts in their compute nodes. ARCHER has 24, INDY2 has 36, and AWS-EC2 has 18. Ideally, we want exact core count so we can make a direct comparison between each infrastructure, but this is not the case. However, it could help in painting the general trend of the scaling capabilities and performance measure based on the compute node count.

Chapter 6

Conclusion

Computational bioinformatic applications like HemeLB is inherently complex. Domain experts, whose knowledge in the subject is invaluable, often face difficulties in using the software package. This complexity creates a barrier for them to run a study or even an exploratory use of the software. In addition to that, this barrier is also detrimental towards the replicable and reproducible aspect of the discipline. In this project, I alleviate this pain by providing an alternative interface for HemeLB. With it, I provide another example to the likes of Nekkloud and Galaxy projects on how a web application can be used as an alternative interface to complex application. In addition, this web application coupled with Docker enables it to provide a reproducible research platform for complex applications which is valuable to science in general. I believe using web application and cloud platform is the trend forward for complex applications like HemeLB. For example, at 10th August 2016, The VMTK Lab team announced that they will release a "Computational Fluid Dynamic on the Cloud" tools for their cardiovascular modeling application, VMTKLab, at 20th September 2016. The tools they add to the application allow users to run on-demand simulation using cloud infrastructure[33]. This is in line with what this project is about and shows that there are people trying to solve similar problem.

In this project, I designed and implemented HemeWeb, a web-based interface to interact with HemeLB simulation workflow. It is developed to answer the problems above. Originally, users are required to use command line interface, but HemeWeb allows them to use a web browser to configure and run HemeLB simulation.

In developing HemeWeb, I managed to produce 3 outputs. These are HemeLB core Docker container, HemeWeb deployment scripts, and HemeWeb web application. In the spirit of being open, all of these components are open-sourced and publicly

available on the public repository of this project[1]. HemeLB core Docker container is also published on the Docker hub. All of these enables an interested party to get their hands on the project to experiment or work with it.

The deployment scripts can configure HemeLB ready infrastructure on three cloud vendors. They are Digital Ocean, Amazon Web Service, and Google Cloud Platform. With it, users can configure cloud platform's instances to handle multi-host HemeLB simulation via command line interface. However, currently, the HemeWeb web application can only be deployed with Amazon Web Service. This is caused by an abstraction layer that currently only supports Amazon Web Service for a certain feature like simulation files persistence.

The HemeWeb web application itself allows domain experts to use the familiar interface of clicking on a web form to configure and run HemeLB simulation. Users can provide pre-processed input files or even unprocessed input files to the web app, and it will allow users to configure the job and run it within few clicks. In addition to that, HemeWeb also allows users to use past simulation information, both owned locally or remotely, to create a new simulation. This capability allows users to either replicate or reproduce a simulation.

From the survey, we found that the respondents are considerably satisfied with the system. HemeWeb allows the respondents to run a simulation and reproduce a simulation with the exception of one respondent. In doing both of these tasks, the respondents are considerably satisfied with the ease of doing it, the time it takes, and the support HemeWeb give them. Overall, HemeWeb is not perfect. There are things that could be better, especially the interface. However, considering that the user satisfactions are subjective, the respondent's sentiment for HemeWeb are generally good. In addition, some of the respondents find using command line interface is a barrier, so this provides a nice alternatives for them to run a HemeLB simulation.

Performance wise, HemeWeb perform worse up to 11.27 slower at its worst when compared against dedicated infrastructure for HPC application like INDY2 and ARCHER supercomputer. In addition, costs more than renting those HPC infrastructure resources and this will be worse when considering the job will finish much slower. However, this performance penalty should not be the only consideration in decision to use HemeWeb. HemeWeb offers flexibility in the way domain experts use computing infrastructures. One should only have a credit card and they can start running experiments. In addition to that, domain experts or institutions does not have to worry about maintaining

---

[1] https://github.com/SeiryuZ/HemeWeb

said infrastructure if they are building it themselves. Another reason is that while currently, cloud platforms cost more, companies have cut prices of their services in the pasts[15, 16, 17], making it potentially more enticing and easy to use their service compared to dedicated infrastructure in the near future.

Chapter 7

Future work

To better support HemeLB simulation workflow, HemeWeb can be further improved. I suggest the following areas to be further developed in the future:

1. Handle geometry generation step of HemeLB simulation workflow

   There are some steps which are not included in this project due to time constraints. One of them is the profile generation step. In this step, the domain experts should generate a profile file by pointing out how the simulation will run. They need to point where the blood will flow into the 3D model of the vascular system, where it will flow out, the blood viscosity, and other various parameters that will affect the simulation result. This process will most likely require a graphical user interface for the domain experts to interact with.

2. Viewing simulation result on the browser.

   HemeLB simulation outputs are in a format that is viewable by a third party tool, ParaView. It will be ideal if HemeWeb can be one stop solution for HemeLB simulation that domain experts do not have to bother with all other tools to view the output of it. A ParaView integration can be done in the next step of the development so that simulation result can be directly viewed on the browser so users do not have to bother with an extra tool to configure and install.

3. HemeLB simulation security

   As outlined in the implementation challenge of the project, security was not the main focus of this project. However, if this project is to be an essential part of the future medical decision, security will need to be addressed seriously. After

all, the patient's private health information will be used for the simulation. A system using such highly private information should be better secured.

Also, in line with the simulation security. HemeWeb instance should be better protected. Currently, HemeWeb does not have protected user area. Individuals without correct credentials can just run the simulation by entering the address of the application on their browser. This is not ideal because this simulation costs money. It should be better protected by having appropriate security measures against unauthorized access.

4. HemeWeb notification

Currently, users are asked to wait until simulation is done to get the output. The user had to keep the page open or checking back in an interval to see whether it is finished. It would be a better experience for users to be notified when a simulation is finished. The user can leave the web application and get back to it when notification is sent.

5. Better interface to configure HemeLB parameters.

Echoing the suggestions that are collected during the evaluation of HemeWeb, there should be a better interface to configure HemeLB parameters. Currently, HemeWeb provides an online XML editor for users to update the HemeLB parameter directly. However, this process is prone to error and might not be the best interface for new users to configure HemeLB simulation. A better solution can be in the form of an automatic web form building from the XML file. With it, users can use the easier web form to select and specify the parameters for HemeLB simulation. This is currently not done because the active development cycle of HemeLB means that the format of XML is also actively developed and could create complications in the building in the form if the format is changed.

6. Cloud vendor abstraction on web application

One challenge of the project was the difference between cloud vendors. Due to the time constraint, the developed web application is tied down to amazon web services only. It would be ideal if HemeWeb could work on any cloud vendors with minimal changes. This is going to be more of a reconciling the difference between cloud vendors and make an abstraction layer that HemeWeb will need to call whenever it needs to interact with the cloud vendors' feature.

The project did achieve cloud vendor abstraction for the deployment scripts. The infrastructure can be deployed to three different cloud vendors easily. They are google cloud platform, amazon web service, and digital ocean. However, the web application needs more work to achieve the similar feat. Infrastructure can be deployed on these infrastructures, but HemeWeb still cannot work on those infrastructure.

To further improve the deployment, the abstraction layer should be able to handle deployment on dedicated HPC clusters. This will improve the flexibility of deployment even further that HemeWeb can be run not only on various cloud platforms, but also dedicated infrastructure that an institution built.

Bibliography

[1] S. Steven, "Hemeweb: Container based high performance computing scenario in cloud infrastructure for hemelb," informatics research proposal, University of Edinburgh, April 2016.

[2] M. D. Mazzeo and P. V. Coveney, "Hemelb: A high performance parallel lattice-boltzmann code for large scale fluid flow in complex geometries," *Computer Physics Communications*, vol. 178, no. 12, pp. 894–914, 2008.

[3] M. A. Itani, U. D. Schiller, S. Schmieschek, J. Hetherington, M. O. Bernabeu, H. Chandrashekar, F. Robertson, P. V. Coveney, and D. Groen, "An automated multiscale ensemble simulation approach for vascular blood flow," *Journal of Computational Science*, vol. 9, pp. 150–155, 2015.

[4] M. O. Bernabeu, Y. Lu, J. Lammer, L. P. Aiello, P. V. Coveney, and J. K. Sun, "Characterization of parafoveal hemodynamics associated with diabetic retinopathy with adaptive optics scanning laser ophthalmoscopy and computational fluid dynamics," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 8070–8073, IEEE, 2015.

[5] C. A. Franco, M. L. Jones, M. O. Bernabeu, I. Geudens, T. Mathivet, A. Rosa, F. M. Lopes, A. P. Lima, A. Ragab, R. T. Collins, *et al.*, "Dynamic endothelial cell rearrangements drive developmental vessel regression," *PLoS Biol*, vol. 13, no. 4, p. e1002125, 2015.

[6] C. A. Franco, M. L. Jones, M. O. Bernabeu, A.-C. Vion, P. Barbacena, J. Fan, T. Mathivet, C. G. Fonseca, A. Ragab, T. P. Yamaguchi, *et al.*, "Non-canonical wnt signalling modulates the endothelial shear stress flow sensor in vascular remodelling," *Elife*, vol. 5, p. e07727, 2016.

[7] C. Green, "Computer simulation could become 'integral' in the diagnosis of," 2014.

[8] D. Groen, J. Hetherington, H. B. Carver, R. W. Nash, M. O. Bernabeu, and P. V. Coveney, "Analysing and modelling the performance of the hemelb lattice-boltzmann simulation environment," *Journal of Computational Science*, vol. 4, no. 5, pp. 412–422, 2013.

[9] M. Huerta, G. Downing, F. Haseltine, B. Seto, and Y. Liu, "Nih working definition of bioinformatics and computational biology," *US National Institute of Health*, 2000.

[10] A. W. Service, "Whitepaper: Intro to hpc on aws." `https://d0.awsstatic.com/whitepapers/Intro_to_HPC_on_AWS.pdf`, August 2015. (Accessed on 08/05/2016).

[11] F. Berman, G. Fox, and A. J. Hey, *Grid computing: making the global infrastructure a reality*, vol. 2. John Wiley and sons, 2003.

[12] I. Foster and C. Kesselman, *The Grid 2: Blueprint for a new computing infrastructure*. Elsevier, 2003.

[13] I. Foster, Y. Zhao, I. Raicu, and S. Lu, "Cloud computing and grid computing 360-degree compared," in *2008 Grid Computing Environments Workshop*, pp. 1–10, Ieee, 2008.

[14] fsn.co.uk, "The economy is flat so why are financials cloud vendors growing at more than 90 percent per annum?." `http://www.fsn.co.uk/channel_outsourcing/the_economy_is_flat_so_why_are_financials_cloud_vendors_growing_at_more_than_90_percent_per_annum#.UbmtsPlJPGA/`, March 2013. (Accessed on 08/06/2016).

[15] J. Barr, "Aws price reduction 42 – ec2, s3, rds, elasticache, and elastic mapreduce — aws blog." `https://aws.amazon.com/blogs/aws/aws-price-reduction-42-ec2-s3-rds-elasticache-and-elastic-mapreduce/`, March 2014. (Accessed on 08/06/2016).

[16] S. Martin, "Announcing reduced pricing on storage — blog — microsoft azure." `https://azure.microsoft.com/en-us/blog/storage-price-match/`, January 2014. (Accessed on 08/06/2016).

[17] F. Lardinois, "Google announces massive price drops for its cloud computing services and storage, introduces sustained-use discounts." `https://techcrunch.com/2014/03/25/google-drops-prices-for-compute-and-app-engine-by-over-30-cloud-storage-by` March 2014. (Accessed on 08/06/2016).

[18] J. Cohen, D. Moxey, C. Cantwell, P. Burovskiy, J. Darlington, and S. J. Sherwin, "Nekkloud: A software environment for high-order finite element analysis on clusters and clouds," in *2013 IEEE International Conference on Cluster Computing (CLUSTER)*, pp. 1–5, IEEE, 2013.

[19] P. Mehrotra, J. Djomehri, S. Heistand, R. Hood, H. Jin, A. Lazanoff, S. Saini, and R. Biswas, "Performance evaluation of amazon ec2 for nasa hpc applications," in *Proceedings of the 3rd workshop on Scientific Cloud Computing Date*, pp. 41–50, ACM, 2012.

[20] Q. He, S. Zhou, B. Kobler, D. Duffy, and T. McGlynn, "Case study for running hpc applications in public clouds," in *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*, pp. 395–401, ACM, 2010.

[21] J. Goecks, A. Nekrutenko, and J. Taylor, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences," *Genome biology*, vol. 11, no. 8, p. 1, 2010.

[22] APS, "What is science?." `http://www.aps.org/policy/statements/99_6.cfm`, November 1999.

[23] O. S. Collaboration *et al.*, "Estimating the reproducibility of psychological science," *Science*, vol. 349, no. 6251, p. aac4716, 2015.

[24] C. Drummond, "Replicability is not reproducibility: nor is it good science," 2009.

[25] D. L. Donoho, "An invitation to reproducible computational research," *Biostatistics*, vol. 11, no. 3, pp. 385–388, 2010.

[26] G. K. Sandve, A. Nekrutenko, J. Taylor, and E. Hovig, "Ten simple rules for reproducible computational research," *PLoS Comput Biol*, vol. 9, no. 10, p. e1003285, 2013.

[27] D. Merkel, "Docker: lightweight linux containers for consistent development and deployment," *Linux Journal*, vol. 2014, no. 239, p. 2, 2014.

[28] F. Moreews, O. Sallou, Y. Bras, *et al.*, "A curated domain centric shared docker registry linked to the galaxy toolshed," in *Galaxy Community Conference 2015*, 2015.

[29] M. Mohaan and R. Raithatha, *Learning Ansible*. Packt Publishing Ltd, 2014.

[30] J. R. Lewis, "Ibm computer usability satisfaction questionnaires: psychometric evaluation and instructions for use," *International Journal of Human-Computer Interaction*, vol. 7, no. 1, pp. 57–78, 1995.

[31] C. Quan, "Network interconnect evaluation on archer," Master's thesis, The University of Edinburgh, August 2014.

[32] K. Finstad, "Response interpolation and scale sensitivity: Evidence against 5-point scales," *Journal of Usability Studies*, vol. 5, no. 3, pp. 104–110, 2010.

[33] OROBIX, "Vmtklab cfd on the cloud." `http://vmtklab.orobix.com/cfd`, August 2016. (Accessed on 08/14/2016).

Appendix A

Appendix: Dockerfile

A.1   Cloud vendor specific provisioning scripts

Listing A.1: Stripped down version of the Dockerfile

```
1  # Taken from Miguels docker file for the docker−hemelb
       vnc
2  # https :// github .com/ mobernabeu / docker −hemelb / blob / master
       / Dockerfile
3
4
5
6  # Use phusion / baseimage as base image . See
7  # http :// phusion . github . io / baseimage −docker /
8  FROM phusion / baseimage : 0.9.18
9
10 # Use baseimage −docker 's init system .
11 CMD ["/ sbin / my_init "]
12
13 # ... put your own build instructions here ...
14
15
16
17
18 ##
```

72

```
19  # Dependencies
20  ##
21  # Ubuntu's VMTK package is in the Multiverse repository,
       OpenMPI in Universe
22  RUN apt-get update && \
23      apt-get install -y software-properties-common && \
24      add-apt-repository "deb http://us.archive.ubuntu.com/
          ubuntu/ trusty universe" && \
25      apt-get update
26
27  # CppUnit fails to compile if downloaded by HemeLB's
       CMake, install it system-wide
28  RUN apt-get install -y \
29    git \
30    cmake \
31    libcppunit-dev \
32    build-essential \
33    libopenmpi-dev \
34    libctemplate-dev \
35    openmpi-bin
36
37  ##
38  # Download and install HemeLB
39  ##
40  WORKDIR /tmp
41  RUN git clone https://github.com/UCL/hemelb.git
42  RUN mkdir hemelb/build && \
43      cd hemelb/build && \
44      cmake .. -DHEMELB_STEERING_LIB=none -DHEMELB_USE_SSE3
          =ON && \
45      make
46
47  VOLUME /data
48
49
```

```
50  # Enable SSH
51  RUN rm −f /etc/service/sshd/down
52
53  # Regenerate SSH host keys. baseimage−docker does not
        contain any, so you
54  # have to do that yourself. You may also comment out this
        instruction; the
55  # init system will auto−generate one during boot.
56  RUN /etc/my_init.d/00_regen_ssh_host_keys.sh
57
58  # TODO: Change this to real ssh keys distribution stuff
59  # Enable insecure key
60  RUN /usr/sbin/enable_insecure_key
61
62  COPY container_key /root/.ssh/id_rsa
63  COPY ssh_config /root/.ssh/config
64
65  RUN chmod 400 /root/.ssh/id_rsa
66
67  # Set root password
68  RUN echo "root:Docker!" | chpasswd
69
70  # Clean up packages, except libctemplate−dev and openmpi−
        bin
71  RUN apt−get purge −y \
72      git \
73      build−essential \
74      cmake \
75      libcppunit−dev
76
77  # Clean up APT when done.
78  RUN apt−get clean && rm −rf /var/lib/apt/lists/* /tmp/* /
        var/tmp/*
```

Appendix B

Appendix: Deployment Script

B.1 Cloud vendor specific provisioning scripts

Listing B.1: Amazon web service specific provisioning script

```
1
2  - hosts: localhost
3    connection: local
4    gather_facts: false
5    vars:
6      - region: eu-west-1
7      - master_node_count: 1
8      - worker_node_count: 0
9      - master_instance_type: m3.large
10     - worker_instance_type: c4.xlarge
11
12   tasks:
13
14     - name: Create ssh security group
15       ec2_group:
16         name: ssh
17         description: HemeWeb - Security group that allow
             SSH from public IP
18         region: "{{ region }}"
19         rules:
```

```yaml
20            - proto: tcp
21              from_port: 22
22              to_port: 22
23              cidr_ip: 0.0.0.0/0
24
25      - name: Create web security group
26        ec2_group:
27          name: web
28          description: HemeWeb - Security group that allow
                web traffic
29          region: "{{ region }}"
30          rules:
31            - proto: tcp
32              from_port: 80
33              to_port: 80
34              cidr_ip: 0.0.0.0/0
35
36      - name: Provision master node
37        ec2:
38          region: "{{ region }}"
39          image: ami-f95ef58a
40          instance_type: "{{ master_instance_type }}"
41          key_name: hemeweb
42          groups:
43            - default
44            - ssh
45            - web
46          wait: true
47          count_tag:
48            name: master
49          exact_count: "{{ master_node_count }}"
50          instance_tags:
51            name: master
52
53      - name: Provision worker nodes
```

```
54        ec2:
55          region: "{{ region }}"
56          image: ami−f95ef58a
57          instance_type: "{{ worker_instance_type }}"
58          key_name: hemeweb
59          groups:
60            - default
61            - ssh
62          wait: true
63          count_tag:
64            name: worker
65          exact_count: "{{ worker_node_count }}"
66          instance_tags:
67            name: worker
```

Listing B.2: Digital ocean specific provisioning script

```
1
2  − hosts: localhost
3    connection: local
4    gather_facts: false
5    vars:
6      - image_slug: ubuntu−14−04−x64
7      - droplet_size: 8gb
8      - region_id: lon1
9      - ssh_key_id:  1732394 # Enter your ssh key id from
         digital ocean API
10
11
12   tasks:
13
14     - name: Provision master node
15       digital_ocean:
16         id: 9999
17         name: master
18         unique_name: true
```

```yaml
19          state: present
20          command: droplet
21          size_id: "{{ droplet_size }}"
22          region_id: "{{ region_id }}"
23          image_id: "{{ image_slug }}"
24          private_networking: yes
25          wait: false
26          ssh_key_ids: "{{ ssh_key_id }}"
27
28
29      # TODO: Loop these
30      - name: Provision worker nodes
31        digital_ocean:
32          id: 1
33          name: worker1
34          unique_name: true
35          state: present
36          command: droplet
37          region_id: "{{ region_id }}"
38          size_id: "{{ droplet_size }}"
39          image_id: "{{ image_slug }}"
40          private_networking: yes
41          wait: false
42          ssh_key_ids: "{{ ssh_key_id }}"
43
44      - name: Provision worker nodes
45        digital_ocean:
46          id: 2
47          name: worker2
48          unique_name: true
49          state: present
50          command: droplet
51          region_id: "{{ region_id }}"
52          size_id: "{{ droplet_size }}"
53          image_id: "{{ image_slug }}"
```

```
54        private_networking: yes
55        wait: false
56        ssh_key_ids: "{{ ssh_key_id }}"
```

Listing B.3: Google cloud platform specific provisioning script

```
1
2  - hosts: localhost
3    connection: local
4    gather_facts: false
5    vars:
6      - hemeweb_project_id: 127280593056
7      - image: "ubuntu-1404-trusty-v20160610"  # https://
         console.cloud.google.com/compute/images?_ga
         =1.221098584.1390118076.1465823354&project=hemeweb&
         filter=name:ubuntu*
8      - region: "europe-west1-b"
9      - master_node_count: 1
10     - worker_node_count: 2
11     # https://cloud.google.com/compute/docs/machine-types
         #standard_machine_types
12     - master_instance_type: n1-standard-1
13     - worker_instance_type: n1-standard-1
14
15   tasks:
16
17
18     - name: Provision master node
19       gce:
20         image: "{{ image }}"
21         machine_type: "{{ master_instance_type }}"
22         instance_names: master
23         project_id: "{{ hemeweb_project_id }}"
24         state: active
25         tags: master, http
26         zone: "{{ region }}"
```

```
27
28        - name: Provision worker nodes
29          gce:
30            image: "hemeweb-worker-node"
31            machine_type: "{{ worker_instance_type }}"
32            instance_names: worker1
33            project_id: "{{ hemeweb_project_id  }}"
34            state: active
35            tags: workers
36            zone: "{{ region }}"
```

B.2   Cloud vendor specific deployment scripts

Listing B.4: Amazon web service specific deployment script

```
1 ---
2
3 - include: ../deploy.yml
4    vars:
5      master_hosts_group: "tag_name_master"
6      worker_hosts_group: "tag_name_worker"
7      preferred_network_interface: "eth0"
8      preferred_ansible_interface: "ansible_default_ipv4"
```

Listing B.5: Digital ocean specific deployment script

```
1 ---
2
3 - include: ../deploy.yml
4    vars:
5      master_hosts_group: "master"
6      worker_hosts_group: "workers"
7      preferred_network_interface: "eth1"
8      preferred_ansible_interface: "ansible_default_ipv4"
```

Listing B.6: Google cloud platform specific deployment script

```
1 ---
```

```
 2
 3  - include: ../deploy.yml
 4    vars:
 5      master_hosts_group: "tag_master"
 6      worker_hosts_group: "tag_workers"
 7      preferred_network_interface: "eth0"
 8      preferred_ansible_interface: "ansible_default_ipv4"
```

B.3 Common deployment script

Listing B.7: deploy.yml

```
 1  ---
 2  # Configure all nodes
 3  - hosts: all
 4    become: true  # allow priviledge escalation
 5    roles:
 6      - common
 7
 8
 9  # Specific to master node
10  - hosts: "{{ master_hosts_group  }}"
11    become: true
12    roles:
13      - nginx
14      - redis
15      - postgresql
16      - hemeweb_master
17
18
19  # Specific to worker nodes
20  - hosts: "{{ worker_hosts_group }}"
21    become: true
22    vars:
23      master_ip: "{{ hostvars[groups[master_hosts_group
            ][0]][ preferred_ansible_interface ]['address'] }}
```

```
              "
24      roles:
25          - hemeweb_worker
```

B.3.1 Common role

Listing B.8: roles/common/tasks/main.yml

```
1   #
2   # Common tasks to be run on the entire architecture
3   # Configuration tasks
4   #
5
6   ---
7
8   - include: ssh.yml
9   - include: apt.yml
10  - include: hosts.yml
11
12  - name: Create shared Folder
13    file: path=/shared state=directory owner={{
          ansible_ssh_user }} group={{ ansible_ssh_user }}
```

B.3.1.1 SSH specific common role

Listing B.9: roles/common/tasks/ssh.yml

```
1   #
2   # Configure SSH config and keys on nodes
3   #
4
5   ---
6
7   - name: Copy hosts private key
8     copy: src=insecure_key dest=~/.ssh/id_rsa force=yes
          mode=0600
9     become: false
10
```

```
11   − name: Copy containers private key
12     copy: src=container_key dest=~/.ssh/container_key force
          =yes mode=0600
13     become: false
14
15   − name: Copy insecure key pub to authorized key
16     become: false
17     authorized_key: user={{ ansible_ssh_user }} key={{
          lookup('file', 'insecure_key.pub') }}
18
19   − name: Configure SSH hosts config
20     become: false
21     template: src=ssh_config.j2 dest=~/.ssh/config force=
          yes
```

B.3.1.2   Package manager specific common role

Listing B.10: roles/common/tasks/apt.yml

```
1  #####################################################
2  # Configure APT, install packages and upgrade all
     packages
3  #####################################################
4
5  ---
6
7  − name: Configure apt repository
8    apt_repository: repo={{ item.name }} state=present
9    with_items:
10     - name: "deb http://us.archive.ubuntu.com/ubuntu/
          trusty universe"
11     - name: "deb http://us.archive.ubuntu.com/ubuntu/
          trusty multiverse"
12
13  − name: Add docker keyserver to apt
14    command: apt−key adv −−keyserver hkp://p80.pool.sks−
          keyservers.net:80 −−recv−keys 58118
```

```
          E89F3A912897C070ADBF76221572C52609D
15
16  − name: Add docker repository to apt
17     shell: sh −c "echo ’deb https://apt.dockerproject.org/
              repo ubuntu-$(lsb_release -sc) main’ | cat > /etc/
              apt/sources.list.d/docker.list"
18
19  − name: Update apt cache
20     apt: update_cache=yes
21
22
23  #####################################################
24  # Update kernel and restart if necessary
25  #####################################################
26  − name: Update kernel
27     apt: name=linux−image−generic−lts−vivid
28     register: kernel_install
29
30  − name: Restart all nodes
31     shell: sleep 2 && shutdown −r now "Ansible updates
              triggered"
32     async: 1
33     poll: 0
34     ignore_errors: true
35     when: kernel_install.changed
36
37  − name: Waiting for all nodes to come back
38     local_action: wait_for host={{ inventory_hostname }}
              port=22 state=started delay=5 timeout=300
39     become: false
40     when: kernel_install.changed
41
42
43  #####################################################
44  # Install softwares
```

```
45  ####################################################
46  - name: Install needed software
47    apt: name={{ item.name }} state=latest
48    with_items:
49      - name: python
50      - name: openmpi-bin
51      - name: docker-engine
52      - name: libffi-dev
53
54  - name: Upgrade packages
55    apt: upgrade=yes
```

B.3.1.3 Host specific common role

Listing B.11: roles/common/tasks/hosts.yml

```
1  ####################################################
2  # Update /etc/hosts file
3  # We need separate task for this, because this task
4  # can be run independently of configuration tasks
5  ####################################################
6
7  ---
8  # From https://gist.github.com/rothgar/8793800
9  - name: Update /etc/hosts file
10   template: src=hosts.j2 dest=/etc/hosts
```

Listing B.12: roles/common/templates/hosts.j2

```
1  # {{ ansible_managed }}
2  127.0.0.1    localhost localhost.localdomain localhost4
       localhost4.localdomain4
3  ::1          localhost localhost.localdomain localhost6
       localhost6.localdomain6
4
5
6  {{ hostvars[groups[ master_hosts_group ][0]][
       preferred_ansible_interface ]['address'] }}   master
```

```
 7
 8
 9  {% if worker_hosts_group in groups %}
10    {% for host in groups[ worker_hosts_group ] %}
11      {{ hostvars[host][ preferred_ansible_interface ]['
          address'] }} hemelb-node-{{ loop.index }}
12    {% endfor %}
13  {% endif %}
```

B.3.2 Redis role

Listing B.13: roles/redis/tasks/main.yml

```
 1  #
 2  # Configuration specific for Redis
 3  #
 4
 5  ---
 6
 7  - name: Add redis ppa
 8    apt_repository: repo='ppa:chris-lea/redis-server'
 9
10  - name: Update apt cache
11    apt: update_cache=yes
12
13  - name: Install redis-server
14    apt: name=redis-server state=latest
15    register: install_result
16
17  - name: Restart redis
18    service: name=redis-server state=restarted
19    when: install_result.changed
```

B.3.3 Nginx role

Listing B.14: roles/nginx/tasks/main.yml

```
 1  #
```

```
 2  # Configuration specific for Nginx
 3  #
 4
 5  ---
 6
 7  - name: Add nginx ppa
 8    apt_repository: repo='ppa:nginx/stable'
 9
10  - name: Update apt cache
11    apt: update_cache=yes
12
13  - name: Install nginx
14    apt: name=nginx state=latest
15    register: install_result
16
17  - name: Configure nginx
18    template: src=nginx.j2 dest=/etc/nginx/nginx.conf force
          =True
19    register: nginx_configuration
20
21  - name: Configure reverse proxy for HemeWeb
22    template: src=default.j2 dest=/etc/nginx/sites-
          available/default force=True
23    register: sites_available
24
25  - name: Restart nginx
26    service: name=nginx state=restarted
27    when: install_result.changed or nginx_configuration.
          changed or sites_available.changed
```

Listing B.15: roles/nginx/templates/default.j2

```
1  server {
2          listen 80;
3          server_name default_server ;
4
```

```
 5            root  / var /www/ hemeweb / src ;
 6
 7            location  /  {
 8                    try_files  $uri  @proxy_to_app ;
 9            }
10
11   location  / shared  {
12      root  /;
13                    try_files  $uri  @proxy_to_app ;
14    }
15
16            location  @proxy_to_app  {
17                    proxy_set_header X−Forwarded−For
                            $proxy_add_x_forwarded_for ;
18                    proxy_set_header Host $http_host ;
19                    proxy_redirect  off ;
20                    proxy_pass  http ://127.0.0.1:8000;
21            }
22 }
```

Listing B.16: roles/nginx/templates/nginx.j2

```
 1 user www−data ;
 2 worker_processes  auto ;
 3 pid  / run / nginx . pid ;
 4
 5 events  {
 6        worker_connections  768;
 7        # multi_accept  on ;
 8 }
 9
10 http  {
11
12        ##
13        # Basic  Settings
14        ##
```

```
15
16          sendfile on;
17          tcp_nopush on;
18          tcp_nodelay on;
19          keepalive_timeout 65;
20          types_hash_max_size 2048;
21          # server_tokens off;
22
23   client_body_timeout 6000s;
24   client_max_body_size 5G;
25
26          # server_names_hash_bucket_size 64;
27          # server_name_in_redirect off;
28
29          include /etc/nginx/mime.types;
30          default_type application/octet-stream;
31
32          ##
33          # SSL Settings
34          ##
35
36          ssl_protocols TLSv1 TLSv1.1 TLSv1.2; # Dropping
                 SSLv3, ref: POODLE
37          ssl_prefer_server_ciphers on;
38
39          ##
40          # Logging Settings
41          ##
42
43   log_format upstreamlog '[$time_local] $remote_addr -
         $remote_user - $server_name  to: $upstream_addr:
         $request upstream_response_time
         $upstream_response_time msec $msec request_time
         $request_time ';
44
```

```
45              access_log  /var/log/nginx/access.log upstreamlog;
46              error_log  /var/log/nginx/error.log;
47
48          ##
49          # Gzip  Settings
50          ##
51
52          gzip  on;
53          gzip_disable  "msie6";
54
55          # gzip_vary  on;
56          # gzip_proxied  any;
57          # gzip_comp_level  6;
58          # gzip_buffers  16 8k;
59          # gzip_http_version  1.1;
60          # gzip_types  text/plain text/css application/json
                    application/javascript text/xml application/
                 xml application/xml+rss  text/javascript;
61
62          ##
63          # Virtual  Host  Configs
64          ##
65
66          include  /etc/nginx/conf.d/*.conf;
67          include  /etc/nginx/sites-enabled/*;
68 }
69
70
71 #mail {
72 #          # See sample  authentication  script  at:
73 #          # http://wiki.nginx.org/
     ImapAuthenticateWithApachePhpScript
74 #
75 #          # auth_http localhost/auth.php;
76 #          # pop3_capabilities  "TOP"  "USER";
```

```
77  #           # imap_capabilities "IMAP4rev1" "UIDPLUS";
78  #
79  #          server {
80  #                  listen      localhost:110;
81  #                  protocol    pop3;
82  #                  proxy       on;
83  #          }
84  #
85  #          server {
86  #                  listen      localhost:143;
87  #                  protocol    imap;
88  #                  proxy       on;
89  #          }
90  #}
```

B.3.4 Postgresql role

Listing B.17: roles/postgresql/tasks/main.yml

```
 1  #
 2  # Configuration specific for PostgreSQL
 3  #
 4
 5  ---
 6
 7  - name: Install postgresql
 8    apt: name={{ item.name }} state=latest
 9    with_items:
10      - name: postgresql
11      - name: postgresql-contrib
12      - name: postgresql-server-dev-all
13    register: install_result
14
15  - name: Allow postgresql to trust 127.0.0.1
16    template: src=pg_hba.j2 dest=/etc/postgresql/9.3/main/
         pg_hba.conf force=True
```

```
17    register: config_update
18
19 − name: Restart postgreql
20    service: name=postgresql state=restarted
21    when: install_result.changed or config_update.changed
22
23 # This is ugly because becoming an unpriviledged user '
       postgres' is problematic
24 # http://docs.ansible.com/ansible/become.html
25 − name: Create superuser
26    command: sudo −u postgres createuser −−superuser {{
          ansible_ssh_user }}
27    ignore_errors: true
28
29 − name: Create database
30    become: false
31    command: createdb hemeweb
32    ignore_errors: true
```

Listing B.18: roles/postgresql/templates/pg_hba.j2

```
 1 # PostgreSQL Client Authentication Configuration File
 2 # =================================================
 3 #
 4 # Refer to the "Client Authentication" section in the
       PostgreSQL
 5 # documentation for a complete description of this file.
        A short
 6 # synopsis follows.
 7 #
 8 # This file controls: which hosts are allowed to connect,
        how clients
 9 # are authenticated, which PostgreSQL user names they can
        use, which
10 # databases they can access.  Records take one of these
        forms:
```

```
11  #
12  #  local      DATABASE   USER   METHOD   [OPTIONS]
13  #  host       DATABASE   USER   ADDRESS  METHOD   [OPTIONS]
14  #  hostssl    DATABASE   USER   ADDRESS  METHOD   [OPTIONS]
15  #  hostnossl  DATABASE   USER   ADDRESS  METHOD   [OPTIONS]
16  #
17  #  (The  uppercase  items  must  be  replaced  by  actual  values
        .)
18  #
19  #  The  first  field  is  the  connection  type:  "local"  is  a
        Unix−domain
20  #  socket,  "host"  is  either  a  plain  or  SSL−encrypted  TCP/
        IP  socket,
21  #  "hostssl"  is  an  SSL−encrypted  TCP/IP  socket,  and  "
        hostnossl"  is  a
22  #  plain  TCP/IP  socket.
23  #
24  #  DATABASE  can  be  "all",  "sameuser",  "samerole",  "
        replication",  a
25  #  database  name,  or  a  comma−separated  list  thereof.  The  "
        all"
26  #  keyword  does  not  match  "replication".  Access  to
        replication
27  #  must  be  enabled  in  a  separate  record  (see  example  below
        ).
28  #
29  #  USER  can  be  "all",  a  user  name,  a  group  name  prefixed
        with  "+",  or  a
30  #  comma−separated  list  thereof.  In  both  the  DATABASE  and
         USER  fields
31  #  you  can  also  write  a  file  name  prefixed  with  "@"  to
        include  names
32  #  from  a  separate  file.
33  #
34  #  ADDRESS  specifies  the  set  of  hosts  the  record  matches.
```

```
            It can be a
35  # host name, or it is made up of an IP address and a CIDR
         mask that is
36  # an integer (between 0 and 32 (IPv4) or 128 (IPv6)
         inclusive) that
37  # specifies the number of significant bits in the mask.
         A host name
38  # that starts with a dot (.) matches a suffix of the
         actual host name.
39  # Alternatively, you can write an IP address and netmask
         in separate
40  # columns to specify the set of hosts.  Instead of a CIDR
         -address, you
41  # can write "samehost" to match any of the server's own
         IP addresses,
42  # or "samenet" to match any address in any subnet that
         the server is
43  # directly connected to.
44  #
45  # METHOD can be "trust", "reject", "md5", "password", "
         gss", "sspi",
46  # "krb5", "ident", "peer", "pam", "ldap", "radius" or "
         cert".  Note that
47  # "password" sends passwords in clear text; "md5" is
         preferred since
48  # it sends encrypted passwords.
49  #
50  # OPTIONS are a set of options for the authentication in
         the format
51  # NAME=VALUE.  The available options depend on the
         different
52  # authentication methods — refer to the "Client
         Authentication"
53  # section in the documentation for a list of which
         options are
```

```
54  # available for which authentication methods.
55  #
56  # Database and user names containing spaces, commas,
       quotes and other
57  # special characters must be quoted.  Quoting one of the
       keywords
58  # "all", "sameuser", "samerole" or "replication" makes
       the name lose
59  # its special character, and just match a database or
       username with
60  # that name.
61  #
62  # This file is read on server startup and when the
       postmaster receives
63  # a SIGHUP signal.  If you edit the file on a running
       system, you have
64  # to SIGHUP the postmaster for the changes to take effect
       .  You can
65  # use "pg_ctl reload" to do that.
66
67  # Put your actual configuration here
68  # ————————————————————————————————
69  #
70  # If you want to allow non−local connections, you need to
       add more
71  # "host" records.  In that case you will also need to
       make PostgreSQL
72  # listen on a non−local interface via the
       listen_addresses
73  # configuration parameter, or via the −i or −h command
       line switches.
74
75
76
77
```

```
78  # DO NOT DISABLE!
79  # If you change this first entry you will need to make
       sure that the
80  # database superuser can access the database using some
       other method.
81  # Noninteractive access to all databases is required
       during automatic
82  # maintenance (custom daily cronjobs, replication, and
       similar tasks).
83  #
84  # Database administrative login by Unix domain socket
85  local    all                postgres
                                                    peer
86
87  # TYPE   DATABASE          USER              ADDRESS
                            METHOD
88
89  # "local" is for Unix domain socket connections only
90  local    all                all
                                                    peer
91  # IPv4 local connections:
92  host      all               all               127.0.0.1/32
              trust
93  # IPv6 local connections:
94  host      all               all               ::1/128
                        md5
95  # Allow replication connections from localhost, by a user
       with the
96  # replication privilege.
97  #local     replication      postgres
                                                    peer
98  #host      replication      postgres          127.0.0.1/32
                  md5
99  #host      replication      postgres          ::1/128
                        md5
```

Appendix C

Appendix: Survey

HemeWeb survey

Hi there,

My name is Steven and I am currently writing my MSc dissertation titled "HemeWeb: Blood flow simulation on the cloud". Today, I will be conducting a usability testing for the project's evaluation with your help.

In this evaluation, I will ask you to run two scenarios in the simulation workflow. First, You are going to configure and run a simulation. Secondly, you are going to reproduce previous simulation with modification to simulation parameters. Following each scenario, I will ask few questions to measure your experience. At the end, there are final set of questions to measure the overall experience in using the system.

This session should last around 20 minutes and if you have any questions or problems, I can be reached via email at s1561690@sms.ed.ac.uk.

Once again, thank you for helping me with my work.

Steven

* This survey will be closed at 11:59PM Friday, 12th August 2016

*Required

About the tools

HemeWeb is an alternative interface to use a computational fluid dynamic software called HemeLB. Traditionally, to use HemeLB to run a simulation, command line interface is used. HemeWeb tries to make it easier for domain experts to run HemeLB and use it for their study.

In addition to being an alternative interface, HemeWeb also add features that might not fit in HemeLB's scope, one such example is convenience in reproducing past simulation.

Demographic
Some questions to know you better

1. **Age?** *

    ....

2. **Gender** *
    Mark only one oval.

    ◯ Male

    ◯ Female

3. **Career Stage** *
Mark only one oval.

( ) MSc student

( ) PhD student

( ) PDRA

( ) Lecturer

( ) Professor

( ) Researcher

( ) Professional

( ) Other: ....

4. **Career discipline** *
Mark only one oval.

( ) Informatics / Computer Science

( ) Computational Scientists

( ) Biologist

( ) Clinician

( ) Other: ....

5. **Level of familiarity with installing software from source code?** *
Mark only one oval.

	1	2	3	4	5	
Not at all familiar	( )	( )	( )	( )	( )	Extremely familiar

6. **Level of familiarity with operating web browser (Safari / Firefox / Chrome / IE / Opera) ?** *
Mark only one oval.

	1	2	3	4	5	
Not at all familiar	( )	( )	( )	( )	( )	Extremely familiar

7. **Your level of familiarity with computational fluid dynamic tools like HemeLB ?** *
Mark only one oval.

	1	2	3	4	5	
Not at all familiar	( )	( )	( )	( )	( )	Extremely familiar

Running a blood flow simulation
Scenario

Imagine that you are a computational researcher. You are trying to understand how haemodynamic and structural changes relate in the context of vascular remodelling. HemeLB, a computational fluid dynamic solver, can help you to simulate different scenarios and find interesting correlations.

Coincidentally, computing support have just deployed HemeWeb to the division's computing resource. You open up your browser and access HemeWeb, the web application that allows you to run HemeLB simulation from the web browser.

Instructions

Try to run a blood flow simulation from your browser. If you find it too difficult, you can skip these instructions and go straight to the questions below

Before we start, you need to download these input files first

https://s3-eu-west-1.amazonaws.com/hemeweb-evaluation/input.xml (2.5 KB)
https://s3-eu-west-1.amazonaws.com/hemeweb-evaluation/990_Example2-skeleton_corrected_tubed_smoothed.gmy (43 MB)

or if the links above does not work, here's an alternative link
https://drive.google.com/open?id=0B9-H3nR0aNncR3NqZnQ5ck5mQUU (2.5 KB)
https://drive.google.com/open?id=0B9-H3nR0aNncTkM1SFhfUE9BRlE (43 MB)

These input files were generated with the HemeLB Setup Tool based on a confocal microscopy image of a mouse retinal vascular plexus. The HemeLB Setup Tool can be run standalone on commodity hardware and provides a graphical interface for HemeLB simulation definition. See image below.
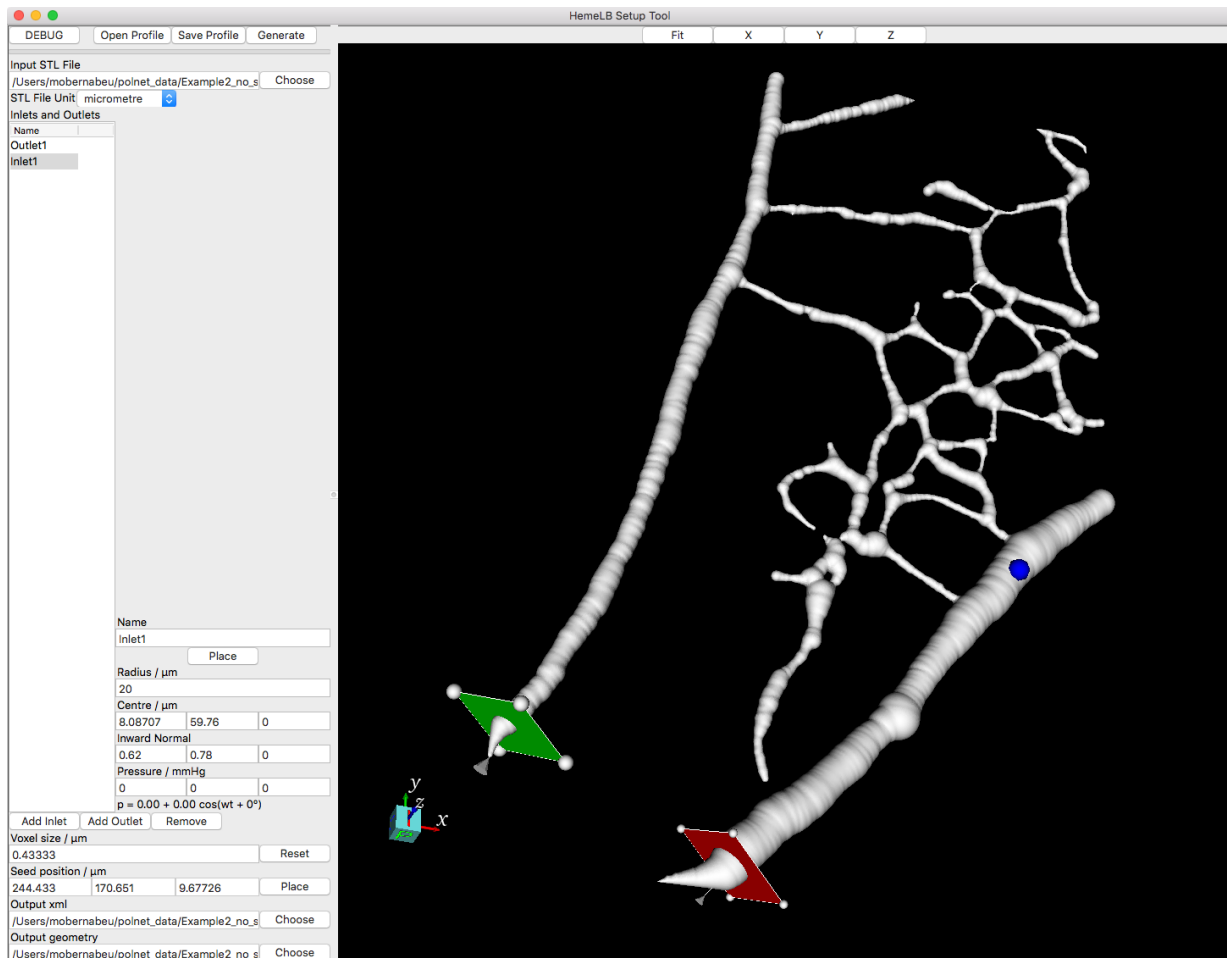
1. To start, open up your browser and access http://54.229.202.129
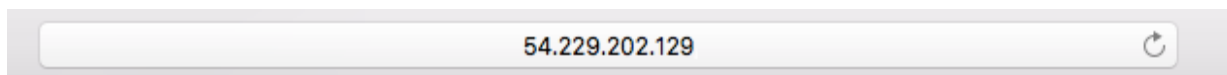
2. This is HemeWeb homepage, click on "Add new job" button

HemeWeb: HemeLB in the cloud

Welcome to HemeWeb homepage! HemeWeb aim to allow users to run blood flow simulation using HemeLB from web browser. To create a new simulation job, please press the Add new job button. To view past simulation's overview please click on one of the Job IDs below.

+ Add new job

HemeWeb Job Index

Job ID	Submission time	Status

3. Use the input file downloaded above to create a new job from preprocessed inputs

4. You will be redirected to HemeLB configuration page (Which is read from the .xml file you uploaded). You can change the configuration here before submitting a job, but we will not make any changes now

Job 234533e3-fa63-40fb-94cf-23eb5352a3b1

HemeLB Config   Job config   Overview

Configure HemeLB execution

The panel below contains content from the HemeLB configuration file you provided. You can modify the content of the file and save it back before proceeding to the next step.

```
36    <visualisation>
37      <centre units="m" value="(0.0,0.0,0.0)" />
38      <orientation>
39        <longitude units="deg" value="45.0" />
40        <latitude units="deg" value="45.0" />
41      </orientation>
42      <display brightness="0.03" zoom="1.0" />
43      <range>
44        <maxvelocity units="m/s" value="0.1" />
45        <maxstress units="Pa" value="0.1" />
46      </range>
47    </visualisation>
48    <initialconditions>
49      <pressure>
50        <uniform units="mmHg" value="0.0" />
51      </pressure>
52    </initialconditions>
53    <monitoring>
54      <incompressibility />
55      <steady_flow_convergence terminate="false" tolerance="1e-5">
56        <criterion type="velocity" units="m/s" value="0.001" />
57      </steady_flow_convergence>
58    </monitoring>
59    <properties>
60      <propertyoutput file="surface-pressure.xtr" period="5000">
61        <geometry type="surface" />
62        <field type="pressure" />
63      </propertyoutput>
64    </properties>
65  </hemelbsettings>
66
```

💾 Save (Ctrl+S)   Next step ❯

5. In the Job config page, configure the job execution with the parameters specified below (Number of Machine: 5, Type of Machine: 4 Core, HemeLB container Version: 0.0.3) and click save.

Job 234533e3-fa63-40fb-94cf-23eb5352a3b1

HemeLB Config   Job config   Overview

Configure HemeWeb job execution

Here, you can configure how the simulation will run. Configure the type of compute node and how many of it HemeWeb should start for the simulation. Also, specify the HemeLB container version to use.

How many machine:

5

Machine type:

4 Cores

HemeLB container Version:

seiryuz/hemelb-core:0.0.3

💾 Save

6. In the overview page, make sure that the parameters are correctly set

7. Click on "All good! queue job" If all parameters are set correctly. Your job are queued and will run when the workers are ready. (In the job execution page, there might be low level errors on stdout output that can be safely ignored)

(Optional) If you are willing to wait, the job will run and finish in around 45 minutes if there are no other jobs queued. A finished job will show a link to download output file. This file is a compressed folder of output directory that contains a .VTK file viewable with ParaView. In addition, a simulation file URL is also shown for sharing simulation files with other HemeWeb installation

Job status: Done

Output file

/shared/ea24198f97f24b2496841920abaf6eb5/result/ea24198f-97f2-4b24-9684-1920abaf6eb5.tar.gz (156.7MiB)

Simulation file URL

https://s3-eu-west-1.amazonaws.com/hemeweb-jobs/ea24198f-97f2-4b24-9684-1920abaf6eb5.tar.gz

Stdout

Questions

8. **Did you skip the above instructions (Step 1 -7)?** *
   Mark only one oval.

   ◯ Yes

   ◯ No

9. **Indicate whether you agree or disagree with the following statements?** *
   Mark only one oval per row.

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree	Not Applicable
Overall, I am satisfied with the ease of completing the tasks in this scenario	◯	◯	◯	◯	◯	◯
Overall, I am satisfied with the amount of time it took to complete the tasks in this scenario	◯	◯	◯	◯	◯	◯
Overall, I am satisfied with the support information (online-line help, messages, documentation) when completing the tasks	◯	◯	◯	◯	◯	◯

10. **Do you have suggestions to improve the experience in using the web interface to run the simulation? (Optional)**

Running a simulation using command line

Alternatively, one can run the above scenario with these high-level steps on command line (you don't have to do this scenario):

1. Install and configure docker tools
2. Download the HemeLB core docker container from hub.docker.com
3. Run docker container in your machine
4. Find out the internal IP of your running container
5. Run HemeLB on the running docker container

11. **How likely are you to run a simulation command line interface?** *
    Mark only one oval.

	1	2	3	4	5	
Extremely unlikely	◯	◯	◯	◯	◯	Extremely likely

12. **Do you think using command line interface to run the simulation is a barrier to run a simulation?** *
    Mark only one oval.

	1	2	3	4	
Not a barrier	◯	◯	◯	◯	Extreme Barrier

Reproducing past blood flow simulation
Scenario

You came across a paper that uses HemeLB to study how haemodynamic and structural changes relate in the context of vascular remodelling. In that paper, the authors include a link that can be used to rerun their simulations on their HemeWeb infrastructure. You want to reproduce those simulation with your own infrastructure and tweak the parameters a bit to see how it affects simulation.

Once again, You open up your browser and access your division's internally deployed HemeWeb and try to reproduce the result with some modifications.

Instructions

Try to run previous blood flow simulation that is ran by your peer in their infrastructure. If you find it too difficult, you can skip these instructions and go straight to the questions below

1. Open the browser and go to http://54.229.202.129

2. Click add new job

3. Enter your peer's simulation file URL https://s3-eu-west-1.amazonaws.com/hemeweb-jobs/234533e3-fa63-40fb-94cf-23eb5352a3b1.tar.gz

4. You noticed in the past simulation that the inlet pressure is set to 45. You decided to change inlet amplitude from 45 to 30 because you think it will run better. Click save, and then click next step

5. Change the job configuration with our own configuration. (Machine type: 4 core, How many machine: 5, and use HemeLB container version 0.0.3)

6. In the overview page, make sure that the parameters are correctly set

7. Click on "All good! queue job" If all parameters are set correctly. Your job are queued and will run when the workers are ready. (In the job execution page, there might be low level errors on stdout output that can be safely ignored)

(Optional) If you are willing to wait, the job will run and finish in around 45 minutes if there are no other jobs queued. A finished job will show a link to download output file. This file is a compressed folder of output directory that contains a .VTK file viewable with ParaView. In addition, a simulation file URL is also shown for sharing simulation files with other HemeWeb installation

Job status: Done

Output file

/shared/ea24198f97f24b2496841920abaf6eb5/result/ea24198f-97f2-4b24-9684-1920abaf6eb5.tar.gz (156.7MiB)

Simulation file URL

https://s3-eu-west-1.amazonaws.com/hemeweb-jobs/ea24198f-97f2-4b24-9684-1920abaf6eb5.tar.gz

Stdout

Questions

13. **Do you skip the instructions (Step 1 - 7)?** *
    Mark only one oval.

    ◯ Yes

    ◯ No

14. **Indicate whether you agree or disagree with the following statements?** *
    Mark only one oval per row.

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree	Not Applicable
Overall, I am satisfied with how easy it is to use this system	◯	◯	◯	◯	◯	◯
Overall, I am satisfied with the amount of time it took to complete the tasks in this scenario	◯	◯	◯	◯	◯	◯
Overall, I am satisfied with the support information (online-line help, messages, documentation) when completing the tasks	◯	◯	◯	◯	◯	◯

15. **Do you have suggestions to improve the experience in using the web interface to reproduce a simulation? (Optional)**

    ....

    ....

    ....

    ....

    ....

Reproducing past simulation using command line

Alternatively, one can run the above scenario with these high-level steps on command line (you don't have to do this scenario):

1. Install and configure docker tools
2. Find out the input and configuration of past experiments
3. Download the correct HemeLB core docker container from hub.docker.com
3. Run docker containers in your machine
4. Find out the internal IP of your running containers
5. Run HemeLB on the running docker container with the correct input and parameters

16. **How likely are you to reproduce simulations in command line interface?** *
   Mark only one oval.

	1	2	3	4	5	
Extremely unlikely	◯	◯	◯	◯	◯	Extremely likely

17. **Do you think using command line interface to reproduce the simulation is a barrier?** *
   Mark only one oval.

	1	2	3	4	
Not a barrier	◯	◯	◯	◯	Extreme Barrier

Overall evaluation

18. **Indicate wether you agree or disagree with the following statements ?** *
   Mark only one oval per row.

	Strongly disagree	Disagree	Neutral	Agree	Strongly Agree	Not Applicable
Overall, I am satisfied with how easy it is to use this system	◯	◯	◯	◯	◯	◯
It was simple to use this system	◯	◯	◯	◯	◯	◯
I can effectively complete my work using this system	◯	◯	◯	◯	◯	◯
I am able to complete my work quickly using this system	◯	◯	◯	◯	◯	◯
I am able to efficiently complete my work using this system	◯	◯	◯	◯	◯	◯
I feel comfortable using this system	◯	◯	◯	◯	◯	◯

It was easy to learn to use this system	◯	◯	◯	◯	◯	◯
I believe I became productive quickly using this system	◯	◯	◯	◯	◯	◯
The system gives error messages that clearly tell me how to fix problems	◯	◯	◯	◯	◯	◯
Whenever I make a mistake using the system, I recover easily and quickly	◯	◯	◯	◯	◯	◯
The information (such as online help, on-screen messages, and other documentation) provided with this system is clear	◯	◯	◯	◯	◯	◯
It is easy to find the information I needed	◯	◯	◯	◯	◯	◯
The information (such as online help, on-screen messages, and other documentation) provided for the system is easy to understand	◯	◯	◯	◯	◯	◯
The information is effective in helping me complete the tasks and scenarios	◯	◯	◯	◯	◯	◯
The organization of information on the system screens is clear	◯	◯	◯	◯	◯	◯
The interface of this system is pleasant	◯	◯	◯	◯	◯	◯
I like using the interface of this system	◯	◯	◯	◯	◯	◯
This system has all the functions and capabilities I expect it to have	◯	◯	◯	◯	◯	◯
Overall, I am satisfied with this system	◯	◯	◯	◯	◯	◯

19. **List most negative aspect(s) of the system** *

20. **List most positive aspect(s) of the system** *

Appendix D

Appendix: Survey Response

D1	Age?
D2	Gender
D3	Career Stage
D4	Career discipline
D5	Level of familiarity with installing software from source code?
D6	Level of familiarity with operating web browser (Safari / Firefox / Chrome / IE / Opera) ?
D7	Your level of familiarity with computational fluid dynamic tools like HemeLB ?
S1_1	Overall, I am satisfied with the ease of completing the tasks in this scenario
S1_2	Overall, I am satisfied with the amount of time it took to complete the tasks in this scenario
S1_3	Overall, I am satisfied with the support information (online-line help, messages, documentation) when completing the tasks
S1_4	How likely are you to run a simulation command line interface?
S1_5	Do you think using command line interface to run the simulation is a barrier to run a simulation?
S2_1	Overall, I am satisfied with the ease of completing the tasks in this scenario
S2_2	Overall, I am satisfied with the amount of time it took to complete the tasks in this scenario
S2_3	Overall, I am satisfied with the support information (online-line help, messages, documentation) when completing the tasks
S2_4	How likely are you to reproduce simulations in command line interface?
S2_5	Do you think using command line interface to reproduce the simulation is a barrier?
O1	Overall, I am satisfied with how easy it is to use this system
O2	It was simple to use this system
O3	I can effectively complete my work using this system
O4	I am able to complete my work quickly using this system
O5	I am able to efficiently complete my work using this system
O6	I feel comfortable using this system
O7	It was easy to learn to use this system
O8	I believe I became productive quickly using this system
O9	The system gives error messages that clearly tell me how to fix problems
O10	Whenever I make a mistake using the system, I recover easily and quickly
O11	The information (such as online help, on-screen messages, and other documentation) provided with this system is clear
O12	It is easy to find the information I needed
O13	The information (such as online help, on-screen messages, and other documentation) provided for the system is easy to underst
O14	The information is effective in helping me complete the tasks and scenarios
O15	The organization of information on the system screens is clear
O16	The interface of this system is pleasant
O17	I like using the interface of this system
O18	This system has all the functions and capabilities I expect it to have
O19	Overall, I am satisfied with this system

Figure D.1: Survey response Question code

	R1	R2	R3	R4	R5	R7	R9	R10
D1	39	34	23	39	34	26	45	28
D2	Male	Male	Male	Male	Female	Male	Male	Male
D3	Professor	Lecturer	PhD student	Project Manage	PDRA	Professional	Software Archit	PhD student
D4	Computational Sc	Informatics / Comp	Informatics / Cc	Computational	physics/Biophy	Clinician	Informatics / Cc	Biologist
D5	5	5	4	4	3	1	5	3
D6	5	4	3	5	5	3	5	5
D7	5	5	3	2	3	1	1	3
S1_1	Neutral	Strongly agree	Agree	Strongly agree	Strongly agree	Neutral	Strongly agree	Strongly agree
S1_2	Neutral	Strongly agree	Strongly agree	Strongly agree	Agree	Agree	Strongly agree	Strongly agree
S1_3	Neutral	Strongly agree	Neutral	Agree	Strongly agree	Strongly agree	Strongly agree	Strongly agree
S1_4	5	5	5	5	3	2	1	4
S1_5	1	3	1	3	2	3	3	1
S2_1	Agree	Agree	Agree	Strongly agree	Strongly agree	Agree	Strongly agree	Agree
S2_2	Neutral	Strongly agree	Strongly agree	Strongly agree	Strongly agree	Strongly agree	Strongly agree	Agree
S2_3	Neutral	Strongly agree	Neutral	Strongly agree	Strongly agree	Strongly agree	Strongly agree	Agree
S2_4	5	4	4	5	2	4	1	4
S2_5	1	3	2	3	1	3	3	1
O1	Agree	Strongly Agree	Agree	Agree	Agree	Strongly Agree	Strongly Agree	Agree
O2	Strongly Agree	Strongly Agree	Strongly Agree	Strongly Agree	Strongly Agree	Strongly Agree	Strongly Agree	Strongly Agree
O3	Neutral	Strongly disagree	Agree	Not Applicable	Disagree	Strongly Agree	Not Applicable	Strongly Agree
O4	Neutral	Disagree	Agree	Not Applicable	Neutral	Strongly Agree	Strongly Agree	Strongly Agree
O5	Neutral	Disagree	Agree	Not Applicable	Neutral	Strongly Agree	Strongly Agree	Strongly Agree
O6	Strongly disagree	Agree	Strongly Agree	Strongly Agree	Strongly Agree	Strongly Agree	Strongly Agree	Agree
O7	Strongly Agree	Strongly Agree	Strongly Agree	Strongly Agree	Strongly Agree	Strongly Agree	Strongly Agree	Strongly Agree
O8	Not Applicable	Neutral	Strongly Agree	Not Applicable	Not Applicable	Strongly Agree	Strongly Agree	Agree
O9	Neutral	Not Applicable	Agree	Agree	Disagree	Strongly Agree	Not Applicable	Agree
O10	Neutral	Not Applicable	Neutral	Agree	Neutral	Strongly Agree	Not Applicable	Neutral
O11	Neutral	Strongly Agree	Agree	Agree	Neutral	Strongly Agree	Strongly Agree	Agree
O12	Neutral	Strongly Agree	Neutral	Not Applicable	Agree	Strongly Agree	Strongly Agree	Agree
O13	Agree	Strongly Agree	Neutral	Agree	Agree	Strongly Agree	Strongly Agree	Strongly Agree
O14	Neutral	Strongly Agree	Neutral	Agree	Disagree	Strongly Agree	Strongly Agree	Strongly Agree
O15	Agree	Agree	Neutral	Agree	Strongly Agree	Strongly Agree	Agree	Agree
O16	Strongly disagree	Agree	Strongly Agree	Agree	Strongly Agree	Strongly Agree	Strongly Agree	Agree
O17	Disagree	Agree	Strongly Agree	Agree	Agree	Strongly Agree	Strongly Agree	Agree
O18	Strongly disagree	Disagree	Strongly Agree	Not Applicable	Agree	Strongly Agree	Not Applicable	Agree
O19	Neutral	Agree	Strongly Agree	Agree	Agree	Strongly Agree	Strongly Agree	Agree

Figure D.2: Survey response part 1

	R11	R12	R13	R14	R15	R16
D1	26	24	25	26	24	22
D2	Male	Male	Male	Male	Male	Male
D3	MSc student	MSc student	MSc student	MSc student	MSc student	MSc student
D4	Informatics / Cc	Informatics / Cc	Informatics / Cc	Informatics / Co	Informatics / Co	Biologist
D5	5	4	5	3	3	3
D6	5	5	5	4	5	5
D7	2	2	1	1	2	1
S1_1	Agree	Agree	Strongly agree	Strongly agree	Agree	Strongly agree
S1_2	Agree	Agree	Agree	Strongly agree	Strongly agree	Strongly agree
S1_3	Strongly agree	Agree	Strongly agree	Not Applicable	Neutral	Agree
S1_4	4	3	5	2	3	3
S1_5	1	1	2	3	1	4
S2_1	Strongly agree	Strongly agree	Strongly agree	Strongly agree	Neutral	Strongly agree
S2_2	Agree	Strongly agree	Strongly agree	Strongly agree	Neutral	Strongly agree
S2_3	Strongly agree	Strongly agree	Strongly agree	Not Applicable	Neutral	Agree
S2_4	4	3	3	2	2	3
S2_5	1	1	2	3	3	4
O1	Strongly Agree	Strongly Agree	Strongly Agree	Strongly Agree	Agree	Agree
O2	Strongly Agree	Strongly Agree	Strongly Agree	Strongly Agree	Agree	Strongly Agree
O3	Strongly Agree	Agree	Strongly Agree	Not Applicable	Not Applicable	Agree
O4	Agree	Strongly Agree	Strongly Agree	Strongly Agree	Not Applicable	Agree
O5	Agree	Agree	Strongly Agree	Strongly Agree	Not Applicable	Agree
O6	Strongly Agree	Agree	Strongly Agree	Strongly Agree	Strongly Agree	Strongly Agree
O7	Strongly Agree	Strongly Agree	Strongly Agree	Strongly Agree	Strongly Agree	Neutral
O8	Agree	Agree	Strongly Agree	Strongly Agree	Not Applicable	Neutral
O9	Not Applicable	Not Applicable	Strongly Agree	Agree	Not Applicable	Not Applicable
O10	Not Applicable	Not Applicable	Strongly Agree	Strongly Agree	Not Applicable	Not Applicable
O11	Strongly Agree	Not Applicable	Strongly Agree	Not Applicable	Not Applicable	Neutral
O12	Strongly Agree	Agree	Strongly Agree	Strongly Agree	Not Applicable	Not Applicable
O13	Strongly Agree	Agree	Strongly Agree	Not Applicable	Agree	Neutral
O14	Strongly Agree	Agree	Strongly Agree	Strongly Agree	Not Applicable	Neutral
O15	Agree	Agree	Strongly Agree	Strongly Agree	Agree	Neutral
O16	Agree	Neutral	Strongly Agree	Agree	Agree	Agree
O17	Agree	Neutral	Strongly Agree	Agree	Agree	Agree
O18	Agree	Agree	Strongly Agree	Strongly Agree	Not Applicable	Neutral
O19	Strongly Agree	Agree	Strongly Agree	Strongly Agree	Agree	Agree

Figure D.3: Survey response part 2

List most negative aspect(s) of the system	List most positive aspect(s) of the system
Firefox does not seem to render web interface as intended.	Helps to make simulations reproducible.
It's too inflexible in the configuration stage, with only one code version and only changes possible in the XML or GMY file (e.g., not in inlet files).	It is extremely easy to use, the interface is very well designed, and it enables people to quickly run pressure-based simulations with default code settings. Definitely a major step ahead compared to the command-line solutions for many users.
Not clear how I can now use this to create my own reproducible experiments to give to someone else (which is something that you might want to do). It would be nice to be able to view the VTK files without having to have a local ParaView install.	Good display of information all in one place (e.g the config file and geometry files etc are all displayed). Nice separation of stdout/stderr on the output (again avoiding two seperate log files as you potentially need on command line). Simplicity to use and requires no command line training (for example you didn't even need to know how to untar the files in the second example. This is very useful for non-expert users).
Not clear where output files end up	Flexible
The simulation failed and it did not provide a useful error message nor I could find instructions on how to fix the problem.	It is easy to use and great for researchers who don't have much computational power themselves. A reference to a paper which explains the method and the results and how to interpret them could help.
nothing	easy to use
The need to edit raw XML.	No need to install software or run at the command-line which makes it easy for novices.
potential (but needed) slow-learning curve, requires a detailed troubleshooting documents	Online-based, easier to use to low-computational skills scientists, efficient
Cannot find something to say as a negative part.	How easy it is to use.
just neutral interface, overall is good!	easy and fast
The result takes time to out	It is easy to use and the instruction guide is clear enough.
Maybe should avoid the users to see the code	If just download and upload sth then I can receive the result, it will be good
Necessity to edit XML directly on a webpage. I think it would be good to have links like "Help me choose machine type. Help me choose number of machines." in case the user is not sure what to do.	In general, HemeWeb is a much more user-friendly interface compared to a command line. It is very easy to re-run simulations.
The job overview page may be confusing for someone who's doing it the for first time	The interface and the possibility of easily reproduce someone else's simulation

Figure D.4: Survey response feedback