E2C: A Visual Simulator to Reinforce Education of Heterogeneous Computing Systems

Ali Mokhtari, Drake Rawls, Mohsen Amini Salehi
High Performance Cloud Computing (HPCC) Laboratory,
School of Computing and Informatics,
University of Louisiana at Lafayette, LA 70503, USA
E-mail: {ali.mokhtari1, drake.rawls1, amini}@louisiana.edu

Abstract—Heterogeneity has been an indispensable aspect of distributed computing throughout the history of these systems. In particular, with the increasing popularity of accelerator technologies (e.g., GPUs and TPUs) and the emergence of domain-specific computing via ASICs and FPGA, the matter of heterogeneity and understanding its ramifications on the system performance has become more critical than ever before. However, it is challenging to effectively educate students about the potential impacts of heterogeneity on: (a) the performance of distributed systems; and (b) the logic of resource allocation methods to efficiently utilize the resources. Making use of the real infrastructure (such as those offered by the public cloud providers) for benchmarking the performance of heterogeneous machines, for different applications, with respect to different objectives, and under various workload intensities is cost- and time-prohibitive. Moreover, not all students (globally and nationally) have access or can afford such real infrastructure. To reinforce the quality of learning about various dimensions of heterogeneity, and to decrease the widening gap in education, we develop an open-source simulation tool, called E2C , that can help students researchers and practitioners to study any type of heterogeneous (or homogeneous) computing system and measure its performance under various system configurations. To make the learning curve shallow, E2C is equipped with an intuitive graphical user interface (GUI) that enables its users to easily examine system-level solutions (scheduling, load balancing, scalability, etc.) in a controlled environment within a short time and at no cost. In particular, E2C is a discrete event simulator that offers the following features: (i) simulating a heterogeneous computing system; (ii) implementing a newly developed scheduling method and plugging it into the system, (iii) measuring energy consumption and other output-related metrics; and (iv) powerful visual aspects to ease the learning curve for students. We used E2C as an assignment in the Distributed and Cloud Computing course. Our survey study indicates that E2C can improve students' understanding of heterogeneous systems by around 80%.

1 Introduction

Heterogeneity has been an indispensable aspect of distributed computing throughout the history of these systems. In the modern era, as Moore's law is losing momentum due to the power density and heat dissipation limitations [8], [15], heterogeneous computing systems have attracted even more attention to overcome the slowdown in Moore's law and fulfilling the desire for higher performance in various types of distributed computing systems. In particular, with the increasing prevalence of accelerator technologies (*e.g.,* GPUs and TPUs) and the emergence of domain-specific computing via ASICs [16] and FPGA [5], the matter of heterogeneity and harnessing it has become a more critical challenge than ever before to deal with.

Examples of heterogeneity can be found in any type of distributed system. Public cloud providers offer and operate based on a wide variety of machine types. Hyperscalers such as AWS and Microsoft Azure provide computing services ranging from general-purpose X86-based and ARM-based machines to FPGAs and accelerators [1]. In the context of Edge computing, domain-specific accelerators (ASICs and FPGA) and general-purpose processors are commonly used together to perform near-data processing, thereby, unlocking various real-time use cases (*e.g.,* edge AI and AR/VR applications) [2], [3]. In the HPC context, deploying various machine types with different architectures on HPC boards to fulfill the power and performance requirements is becoming a trend [6].

Heterogeneity plays a key role in improving various performance objectives of distributed systems, such as cost, energy consumption, and QoS. That is why harnessing system heterogeneity has been a longstanding challenge in distributed systems (*e.g.,* [7], [9], [11]), and educating it to Computer Science/Engineering (and more broadly STEM) students, and researchers has become necessary. Making use of real infrastructure (such as those offered by the public cloud providers) for benchmarking the performance of heterogeneous systems, for different applications, with respect to different objectives, and under various workload intensities is cost- and time-prohibitive. As an example, consider an IoT-based system that offers multiple smart applications to its users (*e.g.,* object detection, face recognition, speech recognition, etc.); there exists a wide range of machine types with different architectures (such as x-86 or ARM-based multi-core CPUs, different types of GPUs, FPGAs, and ASICs) that can process these services. To find an optimal configuration, a student must examine all permutations of these configurations. Moreover, there can be multiple workload intensities and scheduling policies that can affect performance of the system and the student must examine them too. Last but not least, learning about the energy consumption of the heterogeneous computing system

in question adds another dimension to the evaluation process that needs to be conducted by the student.

To avoid the burden of examining all cases, we need simulation tools that can help the students and researchers to study the performance of various system configurations and effectively learn about impacts of heterogeneity in a distributed system. To that end, in this paper, we introduce E2C that is an open-source discrete event simulator that simulate any type of heterogeneous (and homogeneous) computing system. By using E2C , the students can easily examine their system-level solutions (scheduling, load balancing, scalability, etc.) in a controlled environment within a short time and at no cost. In particular, E2C offers the following features: (i) defining user-defined workload generation scenarios with various number of applications (a.k.a. task types) and arrival intensities; (ii) simulating a heterogeneous computing system; (iii) implementing a newly developed scheduling method and plugging it into the system, (iv) measuring power and other output-related things, and (v) visual aspects to ease the learning curve for students. These features help students who study resource allocation solutions in distributed systems to test and evaluate their solutions easier and faster. Moreover, the graphical user interface would help students to gain a deeper knowledge of resource allocation procedures in distributed computing systems.

We used E2C as an assignment in our Distributed and Cloud Computing class to examine various types of scheduling methods for heterogeneous (and homogeneous) systems under various workload intensities. We conducted a survey on the learning outcomes of the simulator and its usability aspects. Analysis of the survey results showed that E2C has been effective in improving the knowledge of students by around 80%. Moreover, it is easy to use and around 75% of students indicated that they are willing to recommend it to others to learn about heterogeneous systems.

In the rest of this paper, we first elaborate on the features of E2C in more detail. Then, in Section 3, we discuss the potential user base for the simulator and how it can help them. Next, in Section 4, we describe our experience of using E2C as a class assignment for Computer Science and Engineering students. In Section 5, the evaluation of E2C and the results we obtained are discussed. Availability of the simulator and the conclusion and future extensions of E2C are explained in Sections 6 and 7, respectively.

2 Simulating a Heterogeneous Computing System via E2C

Figure 1 shows an overview of the E2C simulator that includes the following major components: (i) workload, (ii) batch queue, (iii) scheduler, (iv) machine queue, and (v) a set of (homogeneous or heterogeneous) machines. In addition, there are two more components that contain canceled and dropped tasks. This is to support circumstances where tasks have hard deadlines and there is no value in executing them beyond their deadline.

A workload is defined as a large group of tasks where each task is a request for an application (task type). In the real world, a heterogeneous computing system can be configured to execute several task types. For instance, a heterogeneous system processing satellite images should support task types for object detection, noise removal, and image enhancements to be performed on the received images. E2C enables us to define the task types, arrival distribution for each task type, and their arrival duration. Each task in the generated workload of E2C has an arrival time and deadline as well.

The machines in the distributed system can be identical (homogeneous) or non-identical (heterogeneous). Note that the heterogeneity of the system is modeled by a matrix, called the Expected Execution Time (EET) matrix [4], [13], [14]. This matrix defines the expected execution time of each task type on each machine. This is to model a real world heterogeneous system, where any given task type (e.g., object detection, noise removal, etc.) is expected to have a differing execution time across heterogeneous machines. The opposite holds true for a homogeneous system where any given task type has identical execution time across all machines. As shown in Figure 2, the user has access to the EET matrix by selecting the workload component. Users can either modify the EET matrix manually or load the desired one as a CSV file.

As shown in Figure 2, the user can load the desired workload trace as a CSV file in this section. The user must keep in mind that the workload trace must conform to the EET matrix. That is, there can be no task type within the workload that is not defined within the EET. Upon the arrival of a task, the simulator transfers the task to the batch queue. The batch queue is where tasks are held before being scheduled. Next, based on the selected scheduling method, the scheduler selects a task from the arrival queue.

Figure 3 shows the scheduler options. The user can choose between immediate scheduling or batch scheduling [10]. Immediate scheduling is when incoming tasks are immediately scheduled to a machine upon arrival, whereas, with batch scheduling, tasks are buffered in the batch queue so the scheduler can make a more informed decision. Typically, immediate mode scheduling methods impose a lower overhead and generally load balancers use this type of scheduling [10]. The following immediate policies are currently implemented into E2C as options: FirstCome-FirstServe (FCFS), Min-Expected-Completion-Time (MECT), and Min-Expected-Execution-Time (MEET). For batch policies, E2C currently implements: ELARE, FELARE, MinCompletion-MinCompletion (MM), MinCompletion-MaxUrgency (MMU), and MinCompletion-SoonestDeadline (MSD). An explanation of these methods can be found in [11].

There exist two options for the scheduled tasks: (i) it might be canceled because of missing its deadline before assignment; or (ii) it might be mapped to one of the available machines. The status of a canceled task is set to “canceled” and no more process is needed. The canceled tasks component shows the number of tasks have been canceled so far. In the case of mapping decisions, the task is appended to the local queue of the assigned machine until the machine queue is saturated. Tasks are executed on the assigned machine in a sequential manner by default. If a task missed its deadline while executing on the machine, it is dropped from the machine. As shown in

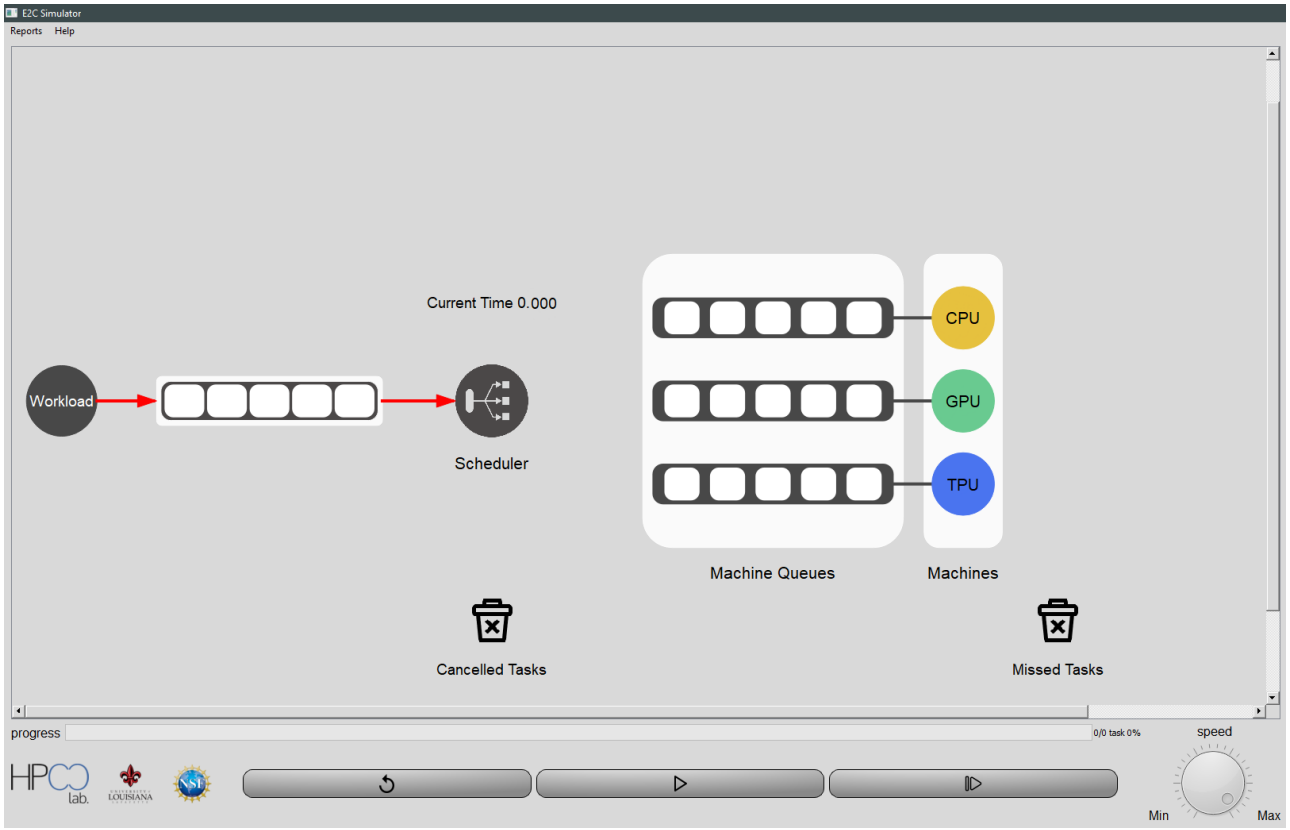


Fig. 1: Overview of the E2C Simulator that includes major components, being the source workload, a batch queue of arriving tasks, scheduler (a.k.a. load balancer), and a set of heterogeneous machines, represented with different colors. Each machine has a "machine queue" where the assigned tasks are queued for the execution.

Figure 4, the Missed Tasks component shows the tasks that missed their deadline.

Importantly, E2C is designed to be modular, hence, providing the ability for the user to modify the existing scheduling methods or adding their own custom-designed scheduling methods. This feature is particularly helpful for researchers to examine new methods under various conditions and configurations.

After the user selects and submits the EET and workload, they will press the "Play" button near the bottom-middle of the GUI. This will begin the animation of tasks flowing from the incoming workload to scheduler to machines, along with the "Current Time" which will update continuously during simulation. If you press the "Play" button again during the simulation run-time, the simulation will be paused. The button right of the play button is the "Increment" button, which when pressed while the simulation is paused will perform the next individual step that would performed (i.e. a task being submitted to a machine by the scheduler, or a task's execution being completed by a machine, etc.). This can be helpful if you wish to analyze each specific action of the simulation. To the left of the "Play" button is the "Reset" button, which can be used either during a pause or after completion of a simulation. This will allow you to begin a new simulation, also allowing you load in a new EET and/or workload should you choose. Along with these three options, during the simulation run-time, you can choose to alter the speed at which the simulation runs by using the speed dial located at the bottom right. This can be useful for either getting quicker results or for better visibility of the animated simulation.

Upon completion of a simulation within E2C, the user may view a report, and optionally, save the report as a CSV file. There is an option for a "Full Report," "Task Report," "Machine Report," and "Summary Report." The Full Report displays the majority of relevant information regarding the simulation - this is the option to view all data related to each task and and how each machine performed on it. The Task Report displays information that is more centric to the individual tasks of the workload, whereas the Machine Report displays data more relevant to the machines of the system. Lastly, the Summary Report displays a summary of the workload data without the specifics of each individual task.

3 Potential Users of E2C Simulator

The E2C simulator can be implemented as a learning tool for undergraduate and graduate students, and also serve practical solutions for researchers and practitioners. Through E2C, students can gain the ability to analyze, design, implement, and test distributed computer systems and components. They can deeply investigate scheduling methods, how they work, and gain insights into their advantages and disadvantages. Along with this, they can develop their own scheduling method(s) and use E2C as a means to implement it. Students can

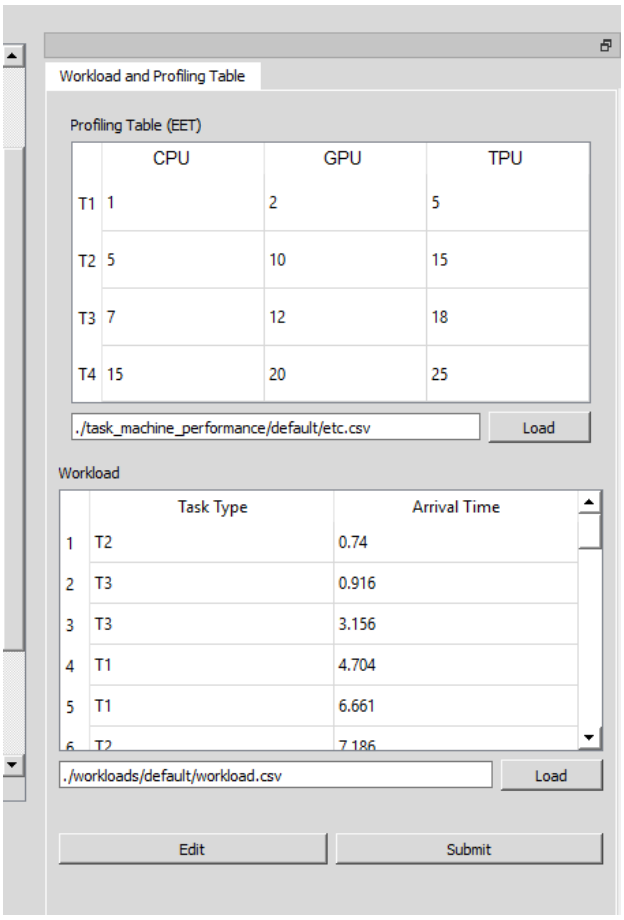


Fig. 2: Workload component. Here, the user can load EET and Workload CSV files. The user can also modify the EET matrix and arrival times of the task with the "Edit" button. Upon loading new CSV files or editing values, the user must press the "Submit" button. EET and Workload files must be compatible. `T1, T2, T3` represent different task types in this simulation.

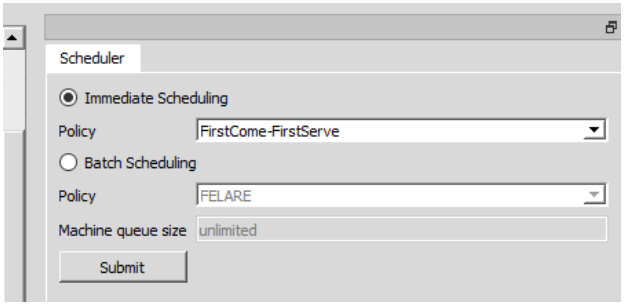


Fig. 3: Scheduler component. Here the user may select between the various immediate or batch scheduling policies, along with setting the machine queue size. The machine queue size is limited to infinite for immediate policies, but can be changed for batch policies.

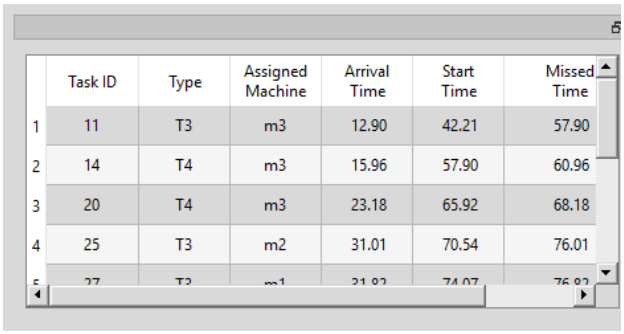


Fig. 4: Missed Tasks component shows the task ID that missed its deadline, along with its task type, assigned machine, arrival time, start time, and the time when it missed.

also learn how heterogeneity can improve the performance of the system through defining machines that have better performance for executing specific task types. Moreover, they can study the energy consumption of the system once a certain scheduling method is applied, allowing them to learn about resource management. So far, we have used the E2C simulator for students in "Distributed and Cloud Computing" courses to examine the impact of different scheduling policies on homogeneous and heterogeneous systems with various workload intensities. Similarly, the simulator can be used for the "Operating Systems" and "Computer Networks" courses at the undergraduate and graduate levels to teach students about the impact of scheduling at different levels.

Researchers in the resource allocation area and cloud solution architects can employ the E2C simulator to test their solution prior to implementation. Being highly customizable, they can configure E2C to represent its real world counterpart. Through this, they may test the outcome of a heterogeneous system with different scheduling methods without spending real resources, saving both money and time. The outcome to be tested can be things such as QoS through task completion percentage (versus missed and cancelled tasks), energy consumption of machines (resource management), and how different scheduling methods perform on any given system. This way, researchers can apply practical use of E2C in order to help design and compare their own real world distributed systems or clusters. As an example, in [12], we have used E2C to examine energy efficiency and fairness of scheduling methods on a heterogeneous edge. Also, in [17], we extended E2C to simulate the memory allocation policies of multi-tenant applications on a homogeneous edge computing system.

4 Class Assignment for Computer Science and Engineering Students

The E2C was used, and will continue to be used, by undergraduate and graduate students of the University of Lafayette's Distributed Cloud Computing course. The assignment involving E2C was used to teach students about the impact of various scheduling methods operating under various workload intensities. It also asked the students to develop and implement new scheduling policies. The installation and interface/GUI of E2C is user friendly, which makes it easy to pick up and

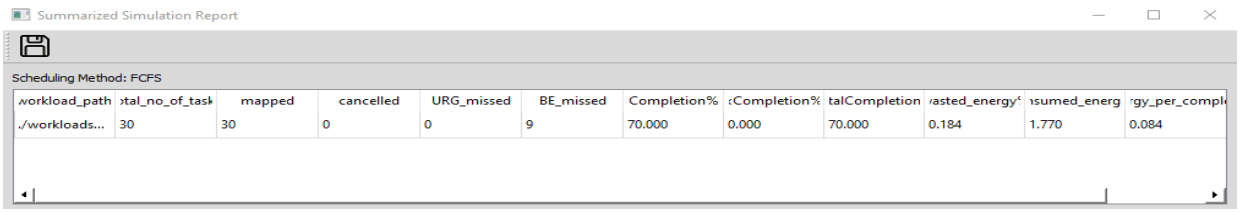


Fig. 5: The "Summary Report" menu which displays general information related to the simulation that was ran. At the top left, there is a button to locally save the report as a CSV (as goes for any other report type).

use for projects or assignments. We have created a web-based documentation[1] where all the features of the simulator—from installation to reporting—are explained.

In this assignment, students were to read and learn about the basic components of E2C, being task types, machines, EET matrix, workload trace, and task deadlines. The students would then use the simulator to evaluate the different scheduling methods currently implemented by E2C on both a homogeneous and a heterogeneous system. For the homogeneous system, students were to use three workload traces with arrival intensities ranging from `low`, `medium`, to `high` to stress the system at different levels. For each arrival intensity level, they ran the simulation and saved the CSV output files, provided by E2C , summarizing all the data related to the simulation for three different immediate scheduling methods, namely FirstCome-FirstServe (FCFS), Minimum-Expected-Completion-Time (MECT), and Minimum-Expected-Execution-Time (MEET). Students then created a bar graphs to depict the percentage of completed tasks that each scheduling method results under each intensity level. The expected results is that higher intensity workloads lead to a lower completion rate (i.e., more tasks missing their deadlines). In addition to observing this behavior, the students had to analyze and report the behavior of different scheduling methods.

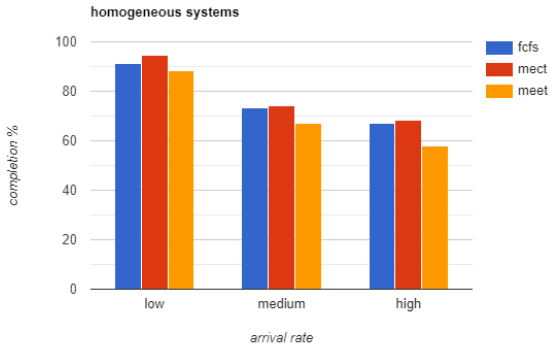


Fig. 6: A bar graph with completion % for immediate scheduling methods on a homogeneous system, showing results for varying intensities using FCFS, MECT, and MEET policies.

For the next part of the assignment, they would do similarly

1. E2C documentation can be accessed at: https://hpcclab.github.io/E2C-Sim-docs/

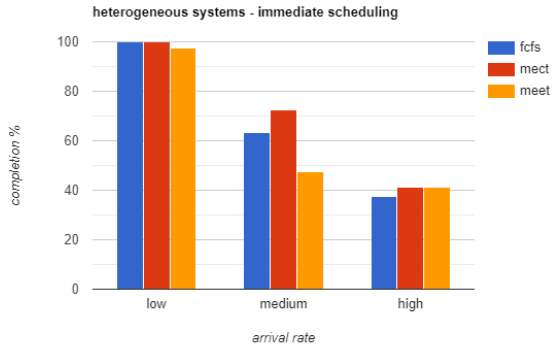


Fig. 7: A bar graph with completion % for immediate scheduling methods on a heterogeneous system, showing results for varying intensities using FCFS, MECT, and MEET policies.

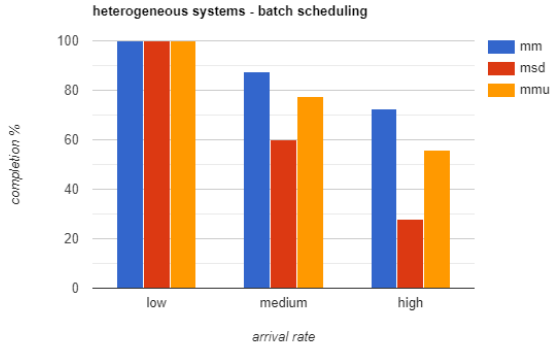
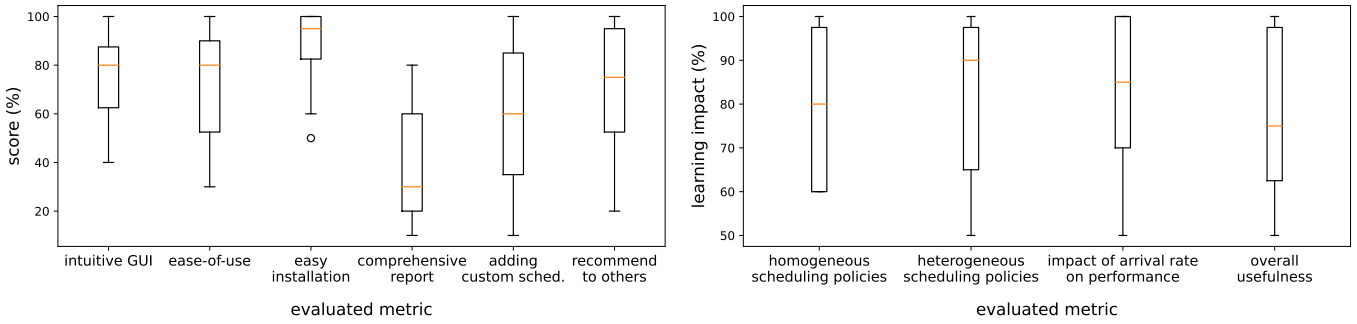


Fig. 8: A bar graph with completion % for batch scheduling methods on a heterogeneous system, showing results for varying intensities using MMU, MSD, and MMU policies.

but with a heterogeneous system instead. For this part, in addition to the immediate scheduling policies, they would also be testing the batch mode policies: MinCompletion-MinCompletion (MM), MinCompletion-MaxUrgency (MMU), and MinCompletion-SoonestDeadline (MSD). Required by the graduate students and optional to undergraduates as a bonus, the third part of this assignment was to create and implement their own scheduling method for the heterogeneous system that enabled fairness across various task types in the system. After these simulations and implementations were complete, students were to perform an analysis of their findings on both the homogeneous system and heterogeneous system, and



(a) Evaluation of user experience with E2C

(b) Evaluation of learning objectives

Fig. 9: Illustration of the survey on the students' experience in accomplishing their distributed systems assignment via E2C (a) This subfigure demonstrate the HCI experience of the students with E2C (b) This subfigure shows how much E2C simulator could help students in understanding the characteristics of task scheduling policies in the homogeneous and heterogeneous configurations.

answer questions that show what they have learned about scheduling and its related methods.

The creation of graphs to evaluate their findings is straightforward due to the way saving data from simulations is within E2C . Once a simulation is complete, all students needed to do is go to the reports menu and save the report as a CSV file (Figure 5).

For the bar graphs that the students create for both their findings on homogeneous and heterogeneous systems, they plot the completion percentage (completed tasks/total tasks in workload) for each scheduling method. Some examples of their findings show a bar graph depicting completion percentage for immediate scheduling policies on a homogeneous system (Figure 6), immediate scheduling policies on a heterogeneous system (Figure 7), and batch scheduling policies on a heterogeneous system (Figure 8).

The learning outcomes of this assignment was to understand the impact of different scheduling methods in face of homogeneous and heterogeneous systems, and to analyze the advantages and disadvantages of each. For instance, they analyzed why Minimum-Expected-Completion-Time (MECT) performs better than FirstCome-FirstServe (FCFS) method, and why the batch policies outperform immediate scheduling policies for heterogeneous systems.

5 EVALUATING LEARNING OUTCOMES OF E2C

As mentioned in Section 4, E2C have been examined as an assignment in the Distributed and Cloud computing course. After the assignment, we conducted an optional survey among the students to evaluate the impact of E2C on their learning. Ten students participated in this survey study. The questions of the survey were in two categories: (i) Those related to the user interactions (experience) with the E2C simulator that is shown in Figure 9a; and (ii) Those focuses on the specific learning outcomes, i.e., how much the knowledge of students was improved as a result of doing this assignment. The result of this category is shown in Figure 9b.

The user experience part studies the user-friendliness of the E2C interface and how it makes technical concepts intuitive. Installing E2C is the first experience of such nature. Figure 9a shows that students evaluated that the installation part is an easy and straightforward procedure with the median score of 95%. The intuitive Graphical User Interface (GUI) of E2C is another metric of the user experience. The median score of 80% for this metric shows that the students has had no difficulty in dealing with the E2C through its GUI. Moreover, the median score of 80% for ease-of-use metric demonstrate that students assess the overall technical part of E2C is intuitive and easy to understand. However, the students assessed the report section with the median score of 30%. Although the reports are comprehensive, we realized that the structure of the GUI for the report section is not intuitive, therefore, the students could not find their required reports easily. To address this issue, we are rearranging the report section in the GUI and make different reports and their fields more informative. In case of developing a custom scheduling in E2C, the graduate students responded that E2C was useful, with the median score of 60%, in implementing and evaluating their custom scheduling policy. In general, the students responded that they are willing to recommend E2C to others with the median score of 75%, as shown in Figure 9a.

Figure 9b summarizes the students' responses in terms of their learning outcomes. The results show that they found it helpful in understanding scheduling characteristics in heterogeneous and homogeneous systems with the median score of 90% and 80%, respectively. In addition, as explained in Section 4, they utilized three workload traces with varying intensities to learn about the impact of arrival rate on the system performance in terms of on-time completion rate. As shown in Figure 9b, they responded that E2C could help them in understanding the impact of arrival rate on the system performance with median score of 85%. Overall, based on the survey results shown in Figure 9b, students assessed E2C is useful in developing their knowledge in the distributed systems course with the median score of 75%.

At the end of the survey study, we asked them to write us their feelings and suggestions that they would like to see in the next version of the E2C simulator. Here, is the main suggestions we received from them: "The simulator clarified the working of different scheduling methods well with its

visual animation." "The application was intuitive when it comes to the context of this course and it was relatively easy to use." "I must commend the great work done by the everyone at the HPCC lab that contributed to the E2C simulator. This is a wonderful software." As for the suggestions, students reported several bugs that we already fixed. Some others had suggestions to make the GUI more intuitive, *e.g.,* by changing the mouse pointer when it is hovered on various components; also there were suggestions to enable drag and drop feature to the simulation scenario.

6 E2C Code and Resource Availability

E2C core is available for download at the following address: https://github.com/hpcclab/E2C-Sim

The manual document on how to run E2C and its options and full documentations are available here:

https://hpcclab.github.io/E2C-Sim-docs/

The video resources for E2C are in this YouTube page.

7 Conclusion and Future Works

E2C provides a free (open-source) learning tool for students enrolled in courses like Distributed Systems, Operating Systems, and Computer Networks as well as researchers by delivering an intuitive way to simulate heterogeneous and homogeneous systems. It particularly helps the students to gain insight into the performance of different scheduling methods upon various heterogeneous systems and under various workload intensities without the need to use and expend for real infrastructure. As such E2C is a step towards reducing the widening educational gap nationally, and even at the global scale. The users of this system can employ several existing scheduling methods built into the simulator, but also have the ability to develop and test their own custom method. As we experienced it in our Distributed and Cloud Computing class, it is an effective accompaniment that can remarkably improve the knowledge of students in the area of heterogeneous computing and scheduling. E2C comes with user friendly GUI for quick usage by beginners, but is also configurable enough to meet the needs of researchers and practitioners in the field. Based on the feedback we received from our students, we plan to extend E2C with several other features, including various communication paradigms and the ability to drag and drop components into the simulator.

Acknowledgement

Development of E2C was made possible by the funding support provided by National Science Foundation (NSF) under awards# CNS-2007209 and CNS-2047144 (NSF CAREER Award).

References

[1] Amazon sagemaker. https://aws.amazon.com/sagemaker/.

[2] Glass enterprise edition 2. https://www.google.com/glass/tech-specs/, note = Acessed: September 2023 .

[3] Qualcomm reveals the world's first dedicated xr platform. https://www.qualcomm.com/news/releases/2018/05/qualcomm-reveals-worlds-first-dedicated-xr-platform, note = Posted on: May 28, 2018.

[4] Shoukat Ali, Howard Jay Siegel, Muthucumaru Maheswaran, Debra Hensgen, Sahra Ali, et al. Representing task and machine heterogeneities for heterogeneous computing systems. *Journal of Applied Science and Engineering*, 3(3):195–207, 2000.

[5] Christophe Bobda, Joel Mandebi Mbongue, Paul Chow, Mohammad Ewais, Naif Tarafdar, Juan Camilo Vega, Ken Eguro, Dirk Koch, Suranga Handagala, Miriam Leeser, et al. The future of fpga acceleration in datacenters and the cloud. *ACM Transactions on Reconfigurable Technology and Systems (TRETS)*, 15(3):1–42, 2022.

[6] Suma George Cardwell, Craig Vineyard, Willam Severa, Frances S Chance, Frederick Rothganger, Felix Wang, Srideep Musuvathy, Corinne Teeter, and James B Aimone. Truly heterogeneous hpc: Co-design to achieve what science needs from hpc. In *Smoky Mountains Computational Sciences and Engineering Conference*, pages 349–365. Springer, 2020.

[7] Chavit Denninnart, James Gentry, Ali Mokhtari, and Mohsen Amini Salehi. Efficient task pruning mechanism to improve robustness of heterogeneous computing systems. *Journal of Parallel and Distributed Computing (JPDC)*, 142:46–61, 2020.

[8] Hadi Esmaeilzadeh, Emily Blem, Renee St. Amant, Karthikeyan Sankaralingam, and Doug Burger. Dark silicon and the end of multicore scaling. In *Proceedings of the 38th annual international symposium on Computer architecture*, pages 365–376, 2011.

[9] James Gentry, Chavit Denninnart, and Mohsen Amini Salehi. Robust dynamic resource allocation via probabilistic task pruning in heterogeneous computing systems. In *Proceedings of the 33rd IEEE International Parallel & Distributed Processing Symposium*, IPDPS '19, May 2019.

[10] Muthucumaru Maheswaran, Shoukat Ali, Howard Jay Siegel, Debra Hensgen, and Richard F Freund. Dynamic mapping of a class of independent tasks onto heterogeneous computing systems. *Journal of parallel and distributed computing*, 59(2):107–131, 1999.

[11] Ali Mokhtari, Chavit Denninnart, and Mohsen Amini Salehi. Autonomous task dropping mechanism to achieve robustness in heterogeneous computing systems. In *Proceedings of 34th IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 17–26, 2020.

[12] Ali Mokhtari, MD Abir Hossen, Pooyan Jamshidi, and Mohsen Amini Salehi. FELARE: Fair Scheduling of Machine Learning Applications on Heterogeneous Edge Systems. In *Proceedings of the 15th IEEE International Conference on Cloud Computing*, IEEE Cloud '22, 2022.

[13] Sanjaya K Panda and Prasanta K Jana. Efficient task scheduling algorithms for heterogeneous multi-cloud environment. *The Journal of Supercomputing*, 71(4):1505–1533, 2015.

[14] Sanjaya K Panda and Prasanta K Jana. An energy-efficient task scheduling algorithm for heterogeneous cloud computing systems. *Cluster Computing*, 22(2):509–527, 2019.

[15] Michael B Taylor. Is dark silicon useful? harnessing the four horsemen of the coming dark silicon apocalypse. In *DAC Design Automation Conference 2012*, pages 1131–1136. IEEE, 2012.

[16] Michael Bedford Taylor, Luis Vega, Moein Khazraee, Ikuo Magaki, Scott Davidson, and Dustin Richmond. Asic clouds: Specializing the datacenter for planet-scale applications. *Communications of the ACM*, 63(7):103–109, 2020.

[17] SM Zobaed, Ali Mokhtari, Jaya Prakash Champati, Mathieu Kourouma, and Mohsen Amini Salehi. Edge-MultiAI: Multi-Tenancy of Latency-Sensitive Deep Learning Applications on Edge. In *Proceedings of 15th IEEE/ACM International Conference on Utility and Cloud Computing*, UCC '22, Dec. 2022.