REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
06-06-2022	Final Report	7-Jan-2019 - 6-Mar-2022

4. TITLE AND SUBTITLE
Final Report: Resource Allocation in Massively Heterogeneous Computer Systems: A Distributed Approval

5a. CONTRACT NUMBER
W911NF-19-1-0036

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

6. AUTHORS

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Carnegie Mellon University
Associate Director, Sponsored Programs
5000 Forbes Avenue
Pittsburgh, PA 15213 -3890

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
72812-CS.9

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Carlee Joe-Wong
UU	UU	UU	UU		19b. TELEPHONE NUMBER 818-419-7508

Standard Form 298 (Rev 8/98)
Prescribed by ANSI Std. Z39.18

RPPR Final Report
as of 06-Jun-2022

Agency Code: 21XD

Proposal Number: 72812CS   **Agreement Number: W911NF-19-1-0036**

INVESTIGATOR(S):

Name: Carlee Joe-Wong
Email: cjoewong@andrew.cmu.edu
Phone Number: 8184197508
Principal: Y

Organization: **Carnegie Mellon University**
Address: Associate Director, Sponsored Programs, Pittsburgh, PA 152133890
Country: USA

DUNS Number: 052184116   EIN: 250969449
Report Date: 06-Jun-2022   Date Received: 06-Jun-2022
Final Report for Period Beginning 07-Jan-2019 and Ending 06-Mar-2022
Title: Resource Allocation in Massively Heterogeneous Computer Systems: A Distributed Approval
Begin Performance Period: 07-Jan-2019   **End Performance Period:** 06-Mar-2022
Report Term: 0-Other
Submitted By: Carlee Joe-Wong   Email: cjoewong@andrew.cmu.edu
Phone: (818) 419-7508
Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: 0   **STEM Participants:** 2

Major Goals: The goal of this project is to develop distributed placement and scheduling algorithms for heterogeneous computing jobs run across a network of heterogeneous computing devices. We characterize heterogeneity as follows. Computing jobs may arrive in the system at different times and are characterized by their resource requirements, which may encompass multiple types of resources, e.g., requirements for both compute power and memory. Devices, in turn, are characterized by their heterogeneous resource availability, e.g., providing different amounts of CPU or GPU resources, memory, etc. These devices may even have different types of computing paradigms, e.g., CPUs compared to GPUs, and will have various amounts of these resources available at different times. Our algorithms to match jobs to providers over time should consider heterogeneity of both devices and jobs, and are designed to scale to the potentially massive number of jobs and devices present.

While centralized matching algorithms allow users to easily coordinate their assignment of jobs to users, they may not scale well to massive numbers of jobs and devices. Thus, we focus on distributed algorithms that empower users and devices to find a mutually satisfying matching that meets job needs within device resource constraints. In particular, our framework is based on distributed pricing algorithms, in which devices announce virtual “prices” for their resources and users attempt to allocate their jobs to resources so as to incur the lowest cost. These prices indicate the capacity limitations of each device relative to users’ demands for them, and thus serve as a means for users to indirectly coordinate their job scheduling and placement. Thus, it requires little exchange of knowledge between devices and users; devices set the prices based on resource availability and users react based on their job requirements. This work will develop mechanisms for resource providers to set their prices, and users to react to them, that aim to maximize the overall system welfare. By adjusting the prices over time, we can also account for changes in the system, such as job arrivals and departures.

The main results of this project are anticipated to be (1) analytical results comparing the effectiveness of a distributed pricing approach to a centralized one, (2) a software suite that allows users and providers to utilize virtual pricing in order to allocate heterogeneous resources to heterogeneous jobs, and (3) a testbed demonstration of the software suite for Army-relevant information processing applications.

The planned effort under this project is divided into three research thrusts, each of which can be divided into three sub-tasks. First, we will establish performance benchmarks by formulating and solving (possibly in a centralized manner) the problem of allocating jobs to devices, under both deterministic and stochastic job arrivals. We will implement our solution algorithms to find performance benchmarks under both types of job arrivals, when different system performance metrics are optimized. Second, we will develop distributed, virtual pricing algorithms for users and providers to optimize their task allocations, considering cases where the system objective is decomposable into a sum of user-specific objectives or not and finally considering auction-based negotiation mechanisms. Third, we will conduct a performance analysis of our distributed algorithms compared to the centralized benchmarks. We will quantify the convergence of the negotiated pricing algorithms developed in Thrust 2 and compare their performance to the centralized benchmarks, and implement our distributed pricing algorithms. We will finally conduct a testbed evaluation that empirically assesses the effectiveness of our proposed algorithms on a testbed of heterogeneous computing devices.

Accomplishments: Major Activities: Our major activities were organized into sub-tasks as described in our “Major Goals”.

1) Tasks 1.1 and 1.2: Formulate and solve our job matching problem for deterministic (1.1) and stochastic (1.2) job arrivals.

2) Task 1.3: Implement our solution algorithms from Tasks 1.1 and 1.2.

3) Tasks 2.1 and 2.2: Develop distributed virtual pricing algorithms to match job components to compute devices for decomposable (2.1, i.e., ones that may be written as a sum of component-specific objectives) and non-decomposable (2.2) system objectives.

4) Task 2.3: Derive optimal user and provider strategies for resource auctions.

5) Tasks 3.1 and 3.2: Quantify the convergence of the algorithms from Tasks 2.1 and 2.2 and compare their optimality to the centralized benchmarks from Thrust 1.

6) Implement Thrust 2’s distributed algorithms on a testbed of networked computing devices.

We elaborate on the specific objectives, significant results, and key outcomes of these tasks below.

Specific Objectives, Significant Results, and Key Outcomes:

1) In Task 1.1, we extend the formulation in our preliminary work to incorporate jobs’ having preferences for being run on certain providers’ resources. For instance, the completion time is affected by the physical distance between a provider and the origin of a job. The main significant result is to extend our framework to account for streaming jobs that do not have a definite start and end time, e.g., analyzing a continuous stream of data.

In Task 1.2, we extend the framework from Task 1.1 to account for sources of uncertainty. First, we consider uncertain costs for running jobs, e.g., uncertain resource requirements. Second, we consider failures in resource availability, e.g., if providers suddenly lose power or network connectivity. We propose to replicate jobs across multiple providers so as to ensure that sufficient resources are always available, taking into account failure correlations, e.g., if two providers share physical infrastructure. Third, we consider stochastic job arrivals. To do so, we find an expression for the service rate of a job given the set of resources on which it runs and constrain it to exceed the job arrival rate. The resulting optimization problem is NP-hard, so we design an iterative algorithm that assigns resource to jobs as they arrive.

2) In Task 1.3, we implement a simulator in Python that can evaluate our proposed solution algorithms on realistic job traces and provider resource availability. The simulator allows for arbitrary job arrival patterns of arbitrary size, as well as arbitrary job- and resource-specific costs of processing these jobs on different resources. Upon determining an allocation for each individual job, it evaluates the resulting costs.

3) In Tasks 2.1 and 2.2, we solve the formulation from Thrust 1, producing significant results in the form of iterative online algorithms for distributedly allocating resources to jobs based on virtual prices. First, we suppose the objective of our formulation is to minimize the sum of the cost of processing each user’s job, subject to resource constraints at each provider. In this case, we can reduce the allocation problem to a shortest-path problem on a resource graph and design an algorithm that iteratively reserves provider resources for each user. We next account for stochasticity in the job costs and arrivals by adapting a time-windowed approach in which we need only predict job costs and arrivals for a short time into the future, and re-optimize after this time has passed. Our third significant result considers non-decomposable objectives, e.g., distributed machine learning applications where the goal is to maximize the accuracy of the trained model. We show that the training accuracy may be upper-bounded by a decomposable sum of accuracy terms. We then derive and characterize the optimal solution when the learning job

is distributed across multiple devices in the system.

4) In Task 2.3, we derive optimal user and provider strategies for resource auctions. Our first significant result is an auction-based algorithm that allows users to submit bids to providers reflecting their resource needs and willingness to pay for these resources. We show that the resulting multi-unit combinatorial auction is tractable if users' resource requirements do not change over time and propose an allocation algorithm that incentivizes truthful user bids in expectation. We show that the problem of finding the optimal bids over repeated auctions can be formulated in the reinforcement learning framework.

Our second significant result is an algorithm for providers to learn the right virtual prices to offer to users. We formulate this learning algorithm as follows: providers first send virtual prices to users. Users respond by sending a set of demands to the resource providers. We design an accelerated learning algorithm that takes advantage of the "side information" that results from users adjusting their demands on a much faster timescale than the virtual prices.

5) In Tasks 3.1 and 3.2, we analyze the convergence and optimality of the algorithms developed in Tasks 2.1 and 2.2. We show that the decomposable resource allocation problem is NP-hard with respect to the numbers of users and providers, but that our distributed allocation algorithm runs in polynomial time. Our second goal was to numerically evaluate the optimality of our algorithms on our simulator. We show that our distributed algorithm for the decomposable problem achieves less than a 20% increase in cost but runs 50% faster than a brute force solution to the NP-hard optimization problem. The non-decomposable problem for distributed machine learning jobs achieves comparable accuracy and a 50% reduction in processing costs compared to centralized learning algorithms. We further find that if there are no resource constraints, then our heuristic finds the optimal solution.

We then evaluate our algorithms in the stochastic setting, which raises a new challenge: it is difficult to know the communication overhead associated with placing each sub-task at a given device a priori. We then observe a tradeoff between exploration and exploitation: we may want to initiate communications between devices to collect a new communication sample, even if doing so is not optimal. We tackle this challenge with a multi-armed bandit approach to placing tasks on devices. However, these algorithms are generally designed to handle unknown parameters in the objective , while our placement problem, the unknown communication parameters often occur in the problem constraints. We exploit the fact that the system itself will translate any infeasible placement input into one that satisfies its physical limitations (e.g., the actual amount of bandwidth available on each link). We show that under certain conditions on the system translation function, we can still learn the optimal feasible placement over time.

Our final significant result for this task is to analyze the convergence and optimality of our proposed algorithms for non-decomposable objectives. We evaluate the optimality of approximating non-decomposable objectives with decomposable ones by comparing this approximation to the true objective. We show that a greedy algorithm that sequentially chooses the placement that maximizes the marginal increase in the objective value converges, but may not be optimal.

6) In Task 3.3, we used our Python simulator and a testbed of twelve Raspberry Pis connected to the Amazon Web Services cloud to empirically evaluate our proposed distributed optimization algorithms. We were able to reproduce the 20% gain in efficiency compared to prior heuristics with distributed placement algorithms and decomposable objectives from our previous simulations. For federated learning algorithms, for which the objectives are non-decomposable, we confirmed a 50% reduction in compute and communication costs compared to baselines.

Training Opportunities: The grant has partially supported the research of three Ph.D. students: Taejin Kim, Jinhang Zuo, and Yuhang Yao. The PI worked closely with these students, meeting with them once or twice a week, to define concrete research problems, review existing problem solutions from the relevant literature, and identify gaps in the current literature. The students then developed approximation algorithms that filled these gaps, quantified the expected performance of these algorithms in various scenarios, and implemented a simulator to test their algorithms numerically. This project gave them the opportunity to investigate new research ideas, learn about the process of conducting research projects and identifying research problems, and learn about the specific area of resource allocation in heterogeneous computer systems. The students further participated in department-wide seminars and networking events aimed at familiarizing students with research in the general area of electrical and computer engineering.

In addition to the CMU Ph.D. students, the PI worked closely with a Ph.D. student at the University of Colorado Boulder (Parisa Rahimzadeh) on activities related to the proposed research. The PI provided advice and guidance to this student in developing algorithms for handle failures in computing resources. This student presented her work with the PI at IEEE ICDCS 2019 and 2020, which gave the student experience in giving talks about her work and provided an opportunity for her to interact with other researchers in distributed computing systems. She has since graduated and now works at Amazon.

Results Dissemination: Several papers and posters have been presented at IEEE INFOCOM (the International Conference on Computer Communications) in 2020 and 2022, IEEE ICDCS (the International Conference on Distributed Computing Systems) in 2019 and 2020, and IEEE IWQoS (the International Symposium on Quality-of-Service) in 2021 as a result of this research. These paper presentations promoted our research results to the research communities in computer networking and distributed systems. They also allowed us to receive feedback on our work that informed subsequent versions of the papers. The specific papers presented at these venues are described below.

In addition to formal presentations at these conferences, the PI gave several seminar and invited talks based on the results of this research. These included seminars at academic institutions (the Pennsylvania State University, University of Illinois Urbana-Champaign, Stanford University and Johns Hopkins University), invited talks at companies (Google, KBR Wyle, and Ericsson), and invited talks at research conferences (the Conference on Information Sciences and Systems, the INFORMS Annual Meeting, and the Caching, Computing, and Delivery in Wireless Networks Workshop at WiOpt). While mostly centered around distributed computing and networking, the audience at INFORMS included operations researchers and that at KBR Wyle included real-time control experts. Thus, these seminars offered a chance to disseminate material outside of the research community that most directly benefits from our work.

The INFOCOM 2022 paper, "MoDEMS: Optimizing Edge Computing Migrations for User Mobility," presents algorithms to allocate edge computing resources to users. The main idea is that as users move, the latency of their connections to different edge servers will change. Thus, users should migrate to new edge servers as they move; however, in order to do so successfully the application they are using must also move to the new edge server. Such migrations consume edge server bandwidth, computing, and storage resources and thus introduce system overhead. Our work optimizes these migrations to take such overheads into account, using predictions of user mobility to optimally place applications at edge servers so as to minimize the total operational cost. Simulations on real-world traces and experimental testbeds verify the value of our migration algorithms. A preliminary version of this work appeared as a poster in IEEE IWQoS 2021, and an extended version has been submitted to the IEEE Journal on Selected Areas in Communications.

Our INFOCOM 2020 paper, entitled "Network-Aware Optimization of Distributed Learning for Fog Computing," presented our algorithms for distributing machine learning programs among heterogeneous computing devices. We provide the first formulation for this problem, and show that it can be solved using convex optimization and virtual pricing techniques. We then quantify, empirically and analytically, the benefits of decomposing ML algorithms in an intelligent way, compared to current practice that is blind to resource heterogeneity at different computing devices. An extension to this work appeared in the IEEE/ACM Transactions on Networking in 2021. These included generalizing the proposed algorithms to users with heterogeneous data and expanding the experiments to examine cases where devices might arrive and depart.

Our ICDCS 2020 paper, "SPARCLE: Stream Processing Applications over Dispersed Computing Networks," considers resource allocation algorithms for stream processing applications in dispersed computing networks with many different interconnected computing resources. We provide one of the first formulations of the problem that considers resource limitations on both the network and computing resources in the system, and propose scalable solution algorithms that assign the stream processing to different compute/network resources and optimize the input rate for each streaming application. Our experiments show a 3x increase in processing rate compared to state-of-the-art algorithms.

Finally, we presented a paper at the IEEE ICDCS (International Conference on Distributed Computing Systems) 2019. The paper, entitled "ECHO: Efficiently Overbooking Applications to Create a Highly Available Cloud," presented our algorithms for replicating jobs across multiple providers to ensure minimum availability when they fail. Although the paper focused on servers in cloud datacenters, the algorithms easily generalize to heterogeneous computing settings.

Two papers resulting from this work were recently submitted to NeurIPS (Conference on Neural Information Processing) 2022. One develops a novel reinforcement learning framework for placing components of a distributed application on a network of computing devices. While prior works have also utilized reinforcement learning to solve this problem, they generally cannot adapt to changes in the available devices, particularly if the number of devices change over time. We design an approach that can automatically adapt to different numbers and characteristics of devices, demonstrating its effectiveness relative to baselines in extensive simulations. Our second NeurIPS submission partially supported by this award considers the specific application of node classification in large graphs (e.g., identifying the ages of users in a social network). When the graph is distributed across multiple clients, this learning problem becomes challenging, since edges in the graph may cross different clients, who therefore must communicate with each other to jointly learn how to classify graph nodes. We therefore design a learning algorithm that minimizes communication between clients and optimizes the computations run at different clients to ensure that the model converges.

One paper is being prepared for a submission to the IEEE/ACM Transactions on Networking, “Learning with Side Information: Online Resource Allocation for 5G Networks”. The paper presents a method to learn how to send congestion signals to users so as to incentivize them to reduce their demands for limited network resources. The main novelty of this work is to exploit observations of user demands (“side information”) to better learn the congestion signals that would prompt them to reduce their demands. Similar ideas can be applied to help control user demands for heterogeneous computing resources. We are also in the midst of working on a submission to IEEE INFOCOM that utilizes multi-armed bandit algorithms to choose the devices on which to run computing jobs. The novelty of this work is that this choice may be subject to a budget constraint (e.g., on the amount of communicable data) that may change over time and must be learned.

Honors and Awards: One Ph.D. student partially supported by this award, Jinhang Zuo, was named a finalist in the 2021 Qualcomm Innovation Fellowship competition. The PI, Carlee Joe-Wong, received the Robert E. Doherty Career Development Professorship in Engineering from Carnegie Mellon University in 2021.

Protocol Activity Status:

Technology Transfer: Nothing to Report

PARTICIPANTS:

Participant Type: PD/PI
Participant: Carlee Joe-Wong
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Taejin Kim
Person Months Worked: 9.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Jinhang Zuo
Person Months Worked: 2.00 **Funding Support:**
Project Contribution:
National Academy Member: N

ARTICLES:

RPPR Final Report
as of 06-Jun-2022

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: IEEE/ACM Transactions on Networking
Publication Identifier Type: DOI Publication Identifier: 10.1109/TNET.2021.3075432
Volume: Issue: First Page #: 1
Date Submitted: 8/25/21 12:00AM Date Published:
Publication Location:
Article Title: Network-Aware Optimization of Distributed Learning for Fog Computing
Authors: Su Wang, Yichen Ruan, Yuwei Tu, Satyavrat Wagle, Christopher G. Brinton, Carlee Joe-Wong
Keywords: Federated learning, network optimization, fog computing
Abstract: Fog computing promises to enable machine learning tasks to scale to large amounts of data by distributing processing across connected devices. Two key challenges to achieving this goal are (i) heterogeneity in devices' compute resources and (ii) topology constraints on which devices communicate with each other. We address these challenges by developing a novel network-aware distributed learning methodology where devices optimally share local data processing and send their learnt parameters to a server for periodic aggregation. Unlike traditional federated learning, our method enables devices to offload their data processing tasks to each other, with these decisions optimized to trade off costs associated with data processing, offloading, and discarding. We analytically characterize the optimal data transfer solution under different assumptions on the fog network scenario. Our experiments on real-world data traces from our testbed confirm the benefits conferred by our algorithms.
Distribution Statement: 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: **Y**

CONFERENCE PAPERS:

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: IEEE International Conference on Distributed Computing Systems (ICDCS)
Date Received: 31-Aug-2020 Conference Date: 08-Jul-2019 Date Published:
Conference Location: Dallas, TX, USA
Paper Title: ECHO: Efficiently Overbooking Applications to Create a Highly Available Cloud
Authors: Parisa Rahimzadeh, Youngbin Im, Gueyoung Jung, Carlee Joe-Wong, Sangtae Ha
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 5-Submitted
Conference Name: IEEE INFOCOM
Date Received: 29-Aug-2019 Conference Date: 27-Apr-2020 Date Published:
Conference Location: Beijing, China
Paper Title: Optimizing Edge Computing For User Mobility
Authors: Taejin Kim, Siqi Chen, Youngbin Im, Xiaoxi Zhang, Sangtae Ha, Carlee Joe-Wong
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: IEEE ICDCS
Date Received: 25-Aug-2021 Conference Date: 29-Nov-2020 Date Published:
Conference Location: Virtual
Paper Title: SPARCLE: Stream Processing Applications over Dispersed Computing Networks
Authors: Parisa Rahimzadeh, Jinsung Lee, Youngbin Im, Siun-Chuon Mau, Eric C. Lee, Bradford O. Smith, Fater
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: IEEE INFOCOM 2020 - IEEE Conference on Computer Communications
Date Received: 31-Aug-2020 Conference Date: 06-Jul-2020 Date Published:
Conference Location: Toronto, ON, Canada
Paper Title: Network-Aware Optimization of Distributed Learning for Fog Computing
Authors: Yuwei Tu, Yichen Ruan, Satyavrat Wagle, Christopher G. Brinton, Carlee Joe-Wong
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: IEEE/ACM International Symposium on Quality-of-Service (Poster Track)
Date Received: 05-Jun-2022 Conference Date: 25-Jun-2021 Date Published: 28-Jun-2021
Conference Location: Virtual
Paper Title: MoDEMS: Optimizing Edge Computing MigrationsFor User Mobility
Authors: Taejin Kim, Siqi Chen, Youngbin Im, Xiaoxi Zhang, Sangtae Ha, Carlee Joe-Wong
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: 2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)
Date Received: 26-Aug-2021 Conference Date: 29-Nov-2020 Date Published:
Conference Location: Virtual
Paper Title: A Community Platform for Research on Pricing and Distributed Machine Learning
Authors: Xuanzhe Li, Samuel Gomena, Logan Ballard, Juntao Li, Ehsan Aryafar, Carlee Joe-Wong
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 2-Awaiting Publica
Conference Name: IEEE INFOCOM
Date Received: 05-Jun-2022 Conference Date: 02-May-2022 Date Published:
Conference Location: Virtual
Paper Title: MoDEMS: Optimizing Edge Computing Migrations for User Mobility
Authors: Taejin Kim, Sandesh Dhawaskar Sathyanarayana, Siqi Chen, Youngbin Im, Xiaoxi Zhang, Sangtae Ha,
Acknowledged Federal Support: **Y**

Partners

,

I certify that the information in the report is complete and accurate:
Signature: Carlee Joe-Wong
Signature Date: 6/6/22 3:30AM

Nothing to report in the uploaded pdf (see accomplishments).