My Desktop
Prepare & Submit Proposals
Prepare Proposals in FastLane
New! Prepare Proposals (Limited proposal types)
Proposal Status
Awards & Reporting
Notifications & Requests
Project Reports
Submit Images/Videos
Award Functions
Manage Financials
Program Income Reporting
Grantee Cash Management Section Contacts
Administration
Lookup NSF ID

Preview of Award 1826884 - Annual Project Report

Cover |
Accomplishments |
Products |
Participants/Organizations |
Impacts |
Changes/Problems

Cover

Federal Agency and Organization Element to Which Report is Submitted:	4900
Federal Grant or Other Identifying Number Assigned by Agency:	1826884
Project Title:	CSR: Small: Collaborative Research: Designing Hierarchical Edge Cloud for Mobile Computing
PD/PI Name:	Wei Gao, Principal Investigator
Recipient Organization:	University of Pittsburgh
Project/Grant Period:	09/01/2017 - 09/30/2019
Reporting Period:	10/01/2017 - 09/30/2018
Submitting Official (if other than PD\PI):	Wei Gao Principal Investigator
Submission Date:	10/15/2018
Signature of Submitting Official (signature shall be submitted in accordance with agency specific instructions)	Wei Gao

---

Accomplishments

*** What are the major goals of the project?**

Cloud computing can be leveraged to bridge the gap between the increasing complexity of mobile applications and the limited capabilities of mobile devices, by remotely executing mobile applications at the cloud. However, the efficiency of such remote execution is hindered by excessive network latency accessing data centers and significant overhead of provisioning and managing large amounts of Virtual Machines (VMs). Traditional solutions reduce the cloud access latency by deploying servers at the network edge, but ignore the impact of mobile users' workload patterns on the efficiency of cloud operation. Instead, this project aims to design the edge cloud as a tree hierarchy of geo-distributed servers, so as to efficiently exploit the cloud resources for handling the peak load from mobile users. This research will benefit end users with various mobile devices by facilitating practical integration of these devices into the cloud. The results from this research are likely to foster new research directions on edge cloud design and mobile cloud computing. The project will engage under-represented students in the research activities, and the scholarly discovery of this project will be disseminated broadly to the community.

This project aims to satisfy the performance requirements of remote program execution by designing the edge cloud in a

hierarchical manner and hence ensuring efficient utilization of cloud resources. More specifically, this project consists of three closely intertwined research thrusts: (i) developing algorithms and systems to optimize the placement of mobile workloads among edge cloud servers and efficiently serve the mobile peak load; ii) mitigating the impact of user mobility on the performance of remote program execution, by developing efficient mobility-aware VM migration techniques; iii) developing an experimental testbed, as a unique research facility, to emulate and investigate the impact of mobile workload peak on edge cloud operations.

*** What was accomplished under these goals (you must provide information for at least one of the 4 categories below)?**

Major Activities:

Based on the previous research accomplishments made by this project, in this reporting period we have focused on developing new mobile system technologies that facilitate efficient workload migration from mobile devices to the edge cloud. More specifically, our major research activities include the following:

First, we have developed new mobile software systems to enhance the efficiency of mobile code offload from resource-constrained mobile devices to the edge cloud, by improving the granularity of such code offload from the traditional process-level to a more fine-grained method level. Such improvement allows more precise identification on the memory contexts that are relevant to the corresponding remote execution of mobile programs, hence minimizing the amount of wireless data transmission for mobile code offload. To support such fine-grained mobile code offload, our basic approach is to parse the mobile application binaries offline in advance, and then apply this parsing result to selectively migrate heap data while allowing successful method execution remotely. We will also parse all the mobile OS kernel libraries as an one-time effort, and then reuse such parsing results over different mobile applications for more accurate identification of relevant memory contexts.

Second, we further extended the scope of edge computing to individual application-specific domains, and develop mobile system designs that efficiently support multi-user Virtual Reality (VR) applications from an edge cloud server that may has only limited computation and communication resources. Our key insight of addressing the excessive computation and communication burden of rendering and transmitting VR frames for multiple VR users is that these frames may be highly correlated to each other and hence redundant, and we hence aim to remove the performance constraint on highly dynamic VR applications by adaptively reusing the redundant VR frames being rendered for different VR users. Such redundancy in each frame is decided at run-time by the edge cloud, which is then able to memoize the previous results of VR frame rendering for future reuse by other users. After a VR frame is generated, the edge cloud further reuses its redundant pixels compared with other frames, and only transmits the distinct portion of this frame to mobile devices.

Education Activities:

One PhD student has worked on the project. Some of the research results have been integrated with the education curricula at University of Tennessee Knoxville and University of Pittsburgh.

Specific Objectives:

During the last reporting period, our research objectives are two-fold as described below:

Our first objective is to improve the efficiency of mobile code offload from resource-constrained mobile devices to the edge cloud. Current offloading schemes either require the programmer's annotations, which restricts its wide application; or transmits too much unnecessary data, resulting bandwidth, and energy waste. Instead, a viable solution to this challenge is to enhance the granularity of such code offload, so as to avoid any irrelevant memory context

from being transmitted to the edge cloud. Such fine-grained code offload, however, is challenging due to the complexity of mobile programs' memory space at run-time.

Our second objective is to fully unleash the computation power of the edge cloud for specific mobile applications, especially Virtual Reality (VR). The major challenge when applying the edge cloud for VR applications, however, is the large amount of computation and communication overhead for rendering and transmitting the high-quality VR frames to mobile devices. The fundamental reason of such failure is that existing mobile workload offloading techniques, when being applied to VR applications, serve each user independently: every VR frame for a user is separately rendered by the edge cloud and fully transmitted back to the mobile HMD. The computation and communication overheads of such remote VR frame rendering, hence, grow with the number of concurrent VR users. Hence, we aim to correlate such VR workload across multiple users, so as to fundamentally reduce the edge cloud's burden.

Significant Results:

We have designed a new mobile code offload system, which addresses the aforementioned challenges and performs automated method-level workload offloading with least context migration. Our basic idea of achieving the least context migration while ensuring the offloading appropriateness is to identify the memory contexts that may be accessed by a specific application method prior to its execution, through offline parsing of the application executables. The parsing results will be stored as metadata along with the application executables at local mobile devices, and will be utilized by the run-time application execution to screen the thread stack and heap contexts to migrate only the relevant memory contexts to the remote cloud. In order to further improve the efficiency of such offline parsing and avoid unnecessary redundancy during parsing, we also pre-parse all the OS libraries that may be invoked by mobile application methods and then reuse these parsing results for different user applications. We have implemented the proposed system design over practical Android OS, and the experimental results over realistic smartphone applications show that our system can migrate 70 percent less memory contexts compared to existing schemes, while maintaining the same offloading effectiveness. To the best of our knowledge, we are the first to exploit the inner characteristics of application binaries for workload offloading in mobile clouds.

We have also developed Multi-User Virtual Reality (MUVR), a systematic mobile VR framework that maximizes the efficiency of edge cloud's resource utilization to support multi-user VR. The key approach of MUVR is to adaptively reuse the previous results of VR frame rendering whenever necessary, by identifying and exploiting the aforementioned redundancy when the edge cloud renders VR frames and transmits these frames to the mobile HMD. In particular, MUVR eliminates redundant computations in VR frame rendering via frame memoization, which caches the invariant background view of rendered VR frames. These caches will be opportunistically reused when rendering frames for other users in the future, if they are at the similar camera locations in the virtual world. Furthermore, in order to reduce the amount of VR frame data being wirelessly transmitted to the mobile HMD, MUVR avoids transmitting full VR frames for every user. Instead, it only transmits a small portion of VR frames in full as reference frames. Then, for any other frame produced between reference frames, only its distinct portion will be transmitted to the mobile HMD as a delta image.

The major challenge of designing MUVR, however, lies in the complicated dynamics of user movements in the VR world, which make it difficult to maintain and utilize the cached VR frames. To address these challenges, our primary idea is to maintain a two-

level hierarchical cache at the edge cloud. In particular, the edge cloud maintains a central cache, which aggregates the VR frames rendered for different VR users and reuses these cached frames whenever necessary: for any new camera location being requested for VR frame rendering, the cached VR frame with the closest matching location will be transformed by image warping, so as to be reused with minimum image quality loss. On the other hand, when the VR user stays stationary in the virtual world, individual VR application locally maintains a distributed small-sized cache to reuse a precedent background image, and only requests to the central cache for rendering a new VR frame if the user movement results in perceivable change of the user view. In this way, by dynamically adapting the threshold of image warping, we are able to flexibly balance between using the central and distributed caches, so as to maximize the efficiency of cache utilization while providing satisfactory VR image quality to users. We have implemented MUVR over the Android OS and Unity VR application engine as a mobile middleware between VR applications and OS drivers, so as to ensure its generality over different VR applications with heterogeneous dynamics and computation demands.

Key outcomes or Other achievements: The results of our work have resulted in one journal paper and one conference paper being published or accepted.

*** What opportunities for training and professional development has the project provided?**

One PhD student has worked on the project, and the research results have been published at top-tier journals and conference proceedings.

*** How have the results been disseminated to communities of interest?**

Our research work in this project has resulted in one top journal paper and one top conference paper. The publications will help people better understand our novel designs on utilizing the edge cloud's resources for more efficient mobile code offload and mobile VR computation, and further implement these designs to improve the performance and efficiency utilizing the public cloud resources in practice. We have also given seminar and summer camp talks to high school students to stimulate their interest in engineering majors.

*** What do you plan to do during the next reporting period to accomplish the goals?**

In the next year of this project, we plan to build on our current work on supporting mobile code offload, and to further expand the scope of such code offload to a broader collection of computationally expensive mobile applications. For example, one promising yet challenging direction is to extend our current work on mobile VR to the cases of Augmented or Mixed Reality, where the virtual objects appear as additional overlay on top of physical backgrounds and are more difficult to be offloaded to the edge cloud for remote rendering. In addition, we will also explore the possibility of further enhancing the efficiency of workload offload at the edge cloud via hardware-software co-designs. For example, new GPU architecture designs could possibly allow more fine-grained sharing of the VR frame rendering results across different VR users and further reduce the edge cloud's computational burden.

Products

Books

Book Chapters

Inventions

Journals or Juried Conference Papers

Yong Li and Wei Gao (2017). Minimizing Context Migration in Mobile Code Offload. *IEEE Transactions on Mobile Computing*. 16 (4), . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

Yong Li and Wei Gao (2018). MUVR: Supporting Multi-User Mobile Virtual Reality with Resource Constrained Edge Cloud. *Proceedings of the 3rd ACM/IEEE Symposium on Edge Computing (SEC)*. . Status = ACCEPTED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

Licenses

Other Conference Presentations / Papers

Other Products

Other Publications

Patents

Technologies or Techniques

Thesis/Dissertations

Websites
Project website
http://www.pitt.edu/~weigao/reporting/edge_cloud.html

Project website

Participants/Organizations

What individuals have worked on the project?

Name	Most Senior Project Role	Nearest Person Month Worked
Gao, Wei	PD/PI	2
Li, Yong	Graduate Student (research assistant)	4

Full details of individuals who have worked on the project:

Wei Gao
Email: weigao@pitt.edu
Most Senior Project Role: PD/PI
Nearest Person Month Worked: 2

Contribution to the Project: Manage and lead the project.

Funding Support: This project.

International Collaboration: No
International Travel: No

Yong Li
Email: yli118@vols.utk.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 4

Contribution to the Project: Developed and implemented the mobile code offload and multi-user Virtual Reality techniques and system designs.

Funding Support: this project.

International Collaboration: No
International Travel: No

What other organizations have been involved as partners?
Nothing to report.

What other collaborators or contacts have been involved?
Nothing to report

Impacts

What is the impact on the development of the principal discipline(s) of the project?

Integration of mobile devices into the cloud dramatically extends the capacities of these devices and fundamentally transforms the way mobile computing applications and services are developed and operated. This integration, however, also imposes serious challenges on the cloud capacity and the efficiency of cloud resource utilization. The transformative nature of the proposed research is to rethink how the edge cloud should be designed to efficiently support remote execution of mobile programs, by turning various analytical modeling and optimization techniques into actionable system design strategies. The research can also spawn a new area of research on hierarchical designs of edge cloud. Finally, the analysis techniques, the evaluation methodology and systems developed in this research will be valuable for future undertakings.

What is the impact on other disciplines?

The edge cloud is a typical example of computer systems with heterogeneous types of workloads in the system's execution. Being able to efficiently improve the performance and reduce the expense of executing these workloads has a direct and immediate impact on a large variety of distributed computing and cyber-physical systems.

What is the impact on the development of human resources?

Many of the research results have been integrated into the undergraduate curricula at the University of Tennessee and University of Pittsburgh, by adopting many perspectives of the research results for undergraduate students' course projects and senior design topics. The project has supported one PhD student working on his dissertation. The involvement of the graduate and undergraduate students into this research will prepare them for leadership roles in computer science research, academia, and industry.

What is the impact on physical resources that form infrastructure?
Nothing to report.

What is the impact on institutional resources that form infrastructure?
Nothing to report.

What is the impact on information resources that form infrastructure?
Nothing to report.

What is the impact on technology transfer?
Nothing to report.

What is the impact on society beyond science and technology?
Nothing to report.

Changes/Problems

Changes in approach and reason for change
Nothing to report.

Actual or Anticipated problems or delays and actions or plans to resolve them
Nothing to report.

Changes that have a significant impact on expenditures
Nothing to report.

Significant changes in use or care of human subjects
Nothing to report.

Significant changes in use or care of vertebrate animals
Nothing to report.

Significant changes in use or care of biohazards
Nothing to report.