My Desktop
Prepare & Submit Proposals
Proposal Status
Proposal Functions
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 1527612 - 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:	1527612
Project Title:	CSR: Small: Collaborative Research: Designing Hierarchical Edge Cloud for Mobile Computing
PD/PI Name:	Wei Gao, Principal Investigator
Recipient Organization:	University of Tennessee Knoxville
Project/Grant Period:	10/01/2015 - 09/30/2017
Reporting Period:	10/01/2015 - 09/30/2016
Submitting Official (if other than PD\PI):	Wei Gao Principal Investigator
Submission Date:	09/30/2016
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: Modern cloud computing services, such as Amazon EC2 and Microsoft Azure, are solely hosted by data centers and incapable of efficiently executing mobile applications due to the following reasons. First, mobile applications require immediate response, and hence suffer from the excessive network latency accessing the remote data centers. Second, data centers provide virtualized cloud resources as Virtual Machines (VMs), each of which serves an enterprise user with high volumes of workloads or responds to a type of web requests. As a result, data centers also handle each mobile application using a separate VM no matter how small its amount of workload is, but incur significant overhead for global VM provisioning and management due to the huge number of mobile applications using the cloud. Such overhead may even exceed the expense of mobile program execution itself and overload the data centers during the peak hours. To address these challenges, an edge cloud, which is an intermediate cloud layer being deployed at the network edge, is indispensable to efficient remote execution of mobile programs. On one hand, mobile workloads are flexibly distributed among different tiers of edge cloud servers and data centers, so as to efficiently serve the peak load through hierarchical load aggregation. On the other hand, the status of the cloud's resource and workload conditions is shared among data centers, edge cloud servers, and mobile devices, allowing appropriate resource allocation and workload placement. To realize such vision of hierarchical edge cloud, we had the following major activities in the past year.

First, to better leverage cloud computing and migrate mobile workloads for remote execution at the cloud, we designed a hierarchical edge cloud architecture, which deploys cloud servers at the network edge as a tree hierarchy of geo-distributed servers. As a result, we are able to efficiently handle the peak load and satisfy the requirements of remote program execution, so as to efficiently utilize the cloud resources to serve the peak loads from mobile users. This hierarchical architecture of edge cloud enables aggregation of the peak loads across different tiers of cloud servers to maximize the amount of mobile workloads being served. To ensure efficient utilization of cloud resources, we further propose a workload placement algorithm that decides which edge cloud servers mobile programs are placed on and how much computational capacity is provisioned to execute each program. The performance of our proposed hierarchical edge cloud architecture on serving mobile workloads is evaluated by formal analysis, small-scale system experimentation, and large-scale trace-based simulations.

Second, we envision that current schemes offloading mobile workloads for remote cloud execution either require the programmer's annotations, which restricts its wide application; or transmits too much unnecessary data, resulting bandwidth and energy waste. To address this challenge, we developed a novel method-level methodology to offload local computational workload with as least data transmission as possible. Our basic idea is to identify the contexts which are necessary to the method execution by parsing application binaries in advance and applying this parsing result to selectively migrate heap data while allowing successful method execution remotely. To further improve the efficiency of such offline parsing of application binaries, our scheme also conducts one-time parsing to all the mobile OS libraries and reuses these parsing results for different user applications. We have implemented our design over the Dalvik Virtual Machine of Android OS. Our experiments and evaluation against applications downloaded from Google Play show that our approach can save data transmission significantly comparing to existing schemes.

Education Activities:

Two PhD students have worked on the project. Some of the research results have been integrated with the education curricula at University of Tennessee, Knoxville.

Specific Objectives:

Edge clouds have recently been suggested to address the aforementioned problems of data centers when serving mobile workloads, but the core challenge in efficiently utilizing the edge cloud’s resources to serve the peak load remains unaddressed. More specifically, the performance of remote program execution is determined by both the network latency accessing the cloud and the cloud delay processing mobile workloads. Existing work, however, is limited to reducing the network latency. Current research on mobile cloud computing (MCC) reduces the network latency by supporting mobile code offload via live VM synthesis and migration, but inappropriately considers the cloud as a black box and assumes unlimited computing resources at the cloud. Existing edge cloud solutions including cyber foraging and fog computing reduce the network latency by deploying cloud servers at the network edge, but organize these servers into a flat architecture which is incapable of handling the peak load exceeding the capacity of individual servers. To address the above challenges, our work incorporates the following two objectives.

Our first objective is to optimize the performance of mobile workload execution in the edge cloud. The performance of such remote program execution depends on the efficiency of utilizing cloud resources. Existing work serves the peak load from mobile users by provisioning sufficient capacity at each edge cloud server, but reduces the efficiency of resource utilization serving nonpeak workloads. Instead, we efficiently serve the peak load with less cloud capacity being provisioned, by optimizing the placement of mobile workloads over multiple tiers of edge cloud servers. Afterwards, we further take mobile devices also into account, and let these devices appropriately offload their mobile programs for remote execution according to the current status of cloud load. Systematic techniques are also developed to improve the efficiency of remote program execution in practical cloud systems across different hardware architectures and software platforms.

Our second objective is to efficiently support user mobility. The frequently used method for supporting user mobility is VM provisioning, but incurs a large amount of system overhead. Although VM migration can be adopted to reduce this overhead, it also increases the network traffic and transmission delay between edge cloud servers. We need to address this problem and reduce the expense of VM migration by incorporating user mobility into mobile workload placement. Such incorporation will be done through development of a probabilistic framework, which minimizes the amount of VMs being migrated and also improves the performance of remote program execution after such VM migration.

Significant Results:

First, to meet the pressing needs to redesign the cloud architecture to serve the mobile workloads with better efficiency and scalability, we proposed to organize the edge cloud servers into a hierarchical architecture. Instead of serving mobile users directly using a flat collection of edge cloud servers, our basic idea is to opportunistically aggregate and serve the peak loads that exceed the capacities of lower tiers of edge cloud servers to other servers at higher tiers in the edge cloud hierarchy. As a result, we are able to serve larger amounts of peak loads with the same amount of computational capacities being provisioned at the edge. We developed analytical models to compare the performance and resource utilization efficiency between flat and hierarchical designs of the edge cloud, and provided theoretical results showing the advantage of hierarchical edge cloud architecture. Furthermore, we ensure efficient utilization of the computing resources at different tiers of edge cloud servers to better serve the peak load, by developing optimization algorithms that adaptively place the mobile workloads at different edge cloud servers. The proposed workload placement algorithms focus on deciding which the edge cloud servers mobile programs are placed on and how much computational capacity is provisioned to execute each program. We conducted performance evaluation of our proposed architecture over both small-scale computing clusters and large-scale mobile workload traces. The results demonstrate that the proposed hierarchical edge cloud architecture outperforms the flat edge cloud architecture by more than 25% in terms of the average delay of program execution.

Second, we presented a novel design of workload offloading system which 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% 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.

Key outcomes or Other achievements:

The results of our work “A Hierarchical Edge Cloud Architecture for Mobile Computing,” has been accepted by the highly competitive *IEEE Conference on Computer Communications (INFOCOM 2016)*, which has an acceptance ratio of 18%.

The results of our work “Minimizing Context Migration in Mobile Code Offload,” has been accepted by IEEE Transactions on Mobile Computing.

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

Two PhD students have worked on the project, and the research results have been published at various academic journals and conference proceedings.

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

Our research work in this project has resulted in one journal paper and one conference paper. These publications will help people better understand our novel designs on hierarchical edge cloud, 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?**

We will further investigate techniques to support user mobility in the hierarchical edge cloud. In particular, when a user moves and connects to different edge cloud servers, synthesizing a new VM for serving this user at each server is expensive, and to migrate an existing VM among these servers is challenging in the hierarchical edge cloud, which may execute a mobile application at multiple servers. Instead, we will first reduce the expense of such VM migration by taking user mobility into account when placing mobile workloads onto edge cloud servers, and then improve the performance of remote program execution after VM migration. Furthermore, we also plan to implement our proposed hierarchical edge cloud designs, optimization algorithms, and systematic techniques as an experimental testbed over practical mobile platforms and cloud computing facilities, and to evaluate their effectiveness in real-world mobile computing scenarios. The main contribution of this testbed will be that it provides a unique research facility for emulating and investigating the impact of mobile workload peak on edge cloud operations.

Products

Books

Book Chapters

Inventions

Journals or Juried Conference Papers

Liang Tong, Yong Li and Wei Gao (2016). A Hierarchical Edge Cloud Architecture for Mobile Computing. *in Proceedings of the 35th IEEE Conference on Computer Communications (INFOCOM)*. . Status = PUBLISHED; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes

Yong Li and Wei Gao (2016). Minimizing Context Migration in Mobile Code Offload. *IEEE Transactions on Mobile Computing*. . Status = AWAITING_PUBLICATION; Acknowledgment of Federal Support = Yes ; Peer Reviewed = Yes ; DOI: 10.1109/TMC.2016.2586056

Licenses

Other Conference Presentations / Papers

Other Products

Other Publications

Patents

Technologies or Techniques

Thesis/Dissertations

Websites

Project website
http://web.eecs.utk.edu/~weigao/reporting/edge_cloud.html

On this project website, we provide details regarding this specific project (personnel, papers, software, etc.).

Participants/Organizations

What individuals have worked on the project?

Name	Most Senior Project Role	Nearest Person Month Worked
Gao, Wei	PD/PI	1
Aljumaily, Mustafa	Graduate Student (research assistant)	2
Li, Yong	Graduate Student (research assistant)	4

Full details of individuals who have worked on the project:

Wei Gao
Email: weigao@utk.edu
Most Senior Project Role: PD/PI
Nearest Person Month Worked: 1

Contribution to the Project: Manage the project and the research team. Design and evaluate the hierarchical edge cloud architecture. Design and implement the workload offloading scheme with minimal context migration.

Funding Support: This grant

International Collaboration: No
International Travel: No

Mustafa Aljumaily
Email: mlatief@vols.utk.edu
Most Senior Project Role: Graduate Student (research assistant)
Nearest Person Month Worked: 2

Contribution to the Project: Help implement and evaluate the hierarchical edge cloud architecture and workload offloading schemes.

Funding Support: This grant

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: Design, implement and evaluate the workload offloading scheme with minimal context migration. Help evaluate the hierarchical edge cloud architecture.

Funding Support: This grant

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, by adopting many perspectives of the research results for undergraduate students' course projects and senior design topics. The project has supported two PhD students working on their dissertations. 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.