

DISS. ETH. NO. 20692

Integrating Cloud Applications with Mobile Devices

Design Principles and Performance Optimizations

A dissertation submitted to
ETH ZURICH

for the degree of
Doctor of Sciences

presented by
IOANA GIURGIU
Eng. Dipl. in Computer Science, TU Cluj-Napoca
born 10 March, 1985
citizen of Romania

accepted on the recommendation of

Prof. Dr. Gustavo Alonso, examiner
Prof. Dr. Timothy Roscoe, co-examiner
Prof. Dr. Srdjan Capkun, co-examiner
Prof. Dr. Christian F. Tschudin, co-examiner

2012

Abstract

With the growth in popularity of mobile devices and the increase in processing power and connectivity given by current trends, users demand ever more ubiquitous and complex applications. They use their devices much like they use their personal desktops and expect similar levels of capabilities, while maintaining or even increasing battery life. Applications such as video and audio processing, 3D gaming or augmented reality are stretching the resources of current mobile devices, and user experience rapidly degrades when several such applications are run concurrently. In the opposite scenario where applications are hosted and run in the cloud, mobile devices act only as access points. Such an approach, however, often raises usability issues due to varying latency and cannot adapt dynamically to more powerful devices.

This thesis contributes towards overcoming such limitations by proposing a flexible approach to integrating cloud-based applications with mobile devices. Our premise is that by exploring the existing space between the two traditional alternatives, parts of the computing process of applications can be optimally outsourced from the cloud to mobile devices. By doing so, users benefit from the seamless integration between their devices and the cloud high performance. Running complex operations within the cloud reduces the cost of mobile computing and allows even low-entry models to take advantage of the cloud capabilities. Thus, users experience a plethora of unique features enhancing their devices and are able to interact with a wider range of rich applications. With the accelerated increase in application complexity, however, the cloud-located application parts frequently require multiple networked virtual instances. The distribution of these instances on the physical infrastructure impacts the way users are able to interact with their applications, and therefore an efficient solution to this problem is also desired.

This thesis is concerned with improving user experience, by combining application performance with extending the battery life of mobile devices and how resources are utilized on both sides. We address the problem from the mobile and the cloud perspectives in three main parts. AlfredO uses modularization to provide optimal distributions

of rich applications between the cloud and the mobile device, and its goal is to minimize overall user interactions and reduce power consumption. The system efficiently and dynamically adjusts the application distribution to the changing conditions one is likely to encounter, such as variations in network bandwidth, user inputs or CPU load. AlfredO is effective in achieving a better utilization of mobile resources and an optimal integration of cloud-based applications with such devices. The second part proposes a performance estimation model for application deployment, that accounts for the variations in workloads and environmental factors. Based on passive measurements only, the model uses queuing networks to accurately find the optimal or nearly optimal distribution of an application in a wide range of scenarios. The final part switches the focus from the mobile device to how to efficiently distribute the cloud-based application modules in large data centers. The problem is addressed in a comprehensive and scalable manner by factoring in network, compute and availability performance aspects into one single solution.

The three parts of the thesis show that both the mobile and cloud perspectives are paramount in improving user experience in the context of integrating cloud-based applications with their mobile devices.

Zusammenfassung

Mit der gestiegenen Beliebtheit von mobilen Geräten und dem Anstieg der Prozessorleistung sowie der Verbindung zu aktuellen Trends verlangen die Anwender immer mehr ubiquitäre und komplexe Anwendungen. Sie nutzen ihre Geräte so häufig, wie sie ihre persönlichen Arbeitsflächen nutzen, wobei sie eine ähnliche Leistungsfähigkeit erwarten, während die Akkulaufzeit erhalten bleibt oder sogar erhöht wird. Anwendungen wie Video- und Audioverarbeitung, 3D-Spiele oder Augmented Reality erweitern die Möglichkeiten von gegenwärtig verwendeten mobilen Geräten. Dabei verringert sich die Benutzerfreundlichkeit schnell, wenn mehrere dieser Anwendungen gleichzeitig ausgeführt werden. In einem entgegengesetzten Szenario, bei dem Anwendungen in der Cloud gehostet und betrieben werden, dienen mobile Geräte lediglich als Zugangspunkte. Dieser Ansatz führt jedoch häufig zu einem Anstieg von Anwendungsproblemen aufgrund von variierenden Latenzzeiten und kann nicht dynamisch an leistungsfähigere Geräte angepasst werden.

Diese Dissertation trägt dazu bei, dass solche Beschränkungen überwunden werden, indem ein flexibler Ansatz zur Integration von Cloud basierten Anwendungen bei mobilen Geräten vorgeschlagen wird. Wir setzen voraus, dass bei der Untersuchung des zwischen diesen beiden traditionellen Alternativen bestehenden Raums Teile des Rechenprozesses der Anwendungen optimal aus der Cloud auf mobile Geräte ausgelagert werden. Auf diese Weise profitieren die Anwender von der nahtlosen Integration zwischen ihren Geräten und der hohen Leistung einer Cloud. Laufende komplexe Operationen innerhalb einer Cloud reduzieren die Kosten des Mobile Computing und ermöglichen sogar Low-Entry-Modellen von der Leistungsfähigkeit einer Cloud zu profitieren. Daher machen die Anwender die Erfahrung einer Unmenge von einzigartigen Funktionen, durch welche ihre Geräte verbessert werden, wodurch sie in der Lage sind, mit einer grösseren Auswahl an leistungsstarken Anwendungen zu interagieren. Mit dem schnelleren Anstieg der Anwendungskomplexität erfordern die Cloud lokalisierten Anwendungsteile jedoch häufig multiple, vernetzte, virtuelle Vorgänge. Die Verteilung dieser Vorgänge auf die physische Infrastruktur wirkt sich auf die Art und Weise aus, wie die

Anwender mit ihren Anwendungen interagieren können. Aus diesem Grund wird eine effiziente Lösung für dieses Problem angestrebt.

Diese Dissertation befasst sich mit einer Verbesserung der Benutzerfreundlichkeit, indem die Leistung der Anwendung mit der Verlängerung der Akkulaufzeit des mobilen Geräts kombiniert wird. Sie bezieht sich zudem darauf, wie die Ressourcen auf beiden Seiten genutzt werden. Wir behandeln das Problem in drei Teilen aus der mobilen Perspektive und aus der Cloud-Perspektive. AlfredO nutzt die Modularisierung, um eine optimale Verteilung von leistungsstarken Anwendungen zwischen einer Cloud und dem mobilen Gerät bereitzustellen. Das Ziel besteht in der Minimierung aller Anwenderinteraktionen sowie in der Reduzierung des Energieverbrauchs. Das System passt die Verteilung der Anwendung effizient und dynamisch an die sich ändernden Bedingungen an, die einem wahrscheinlich begegnen, wie Änderungen der Netzwerkbandbreite, der Eingaben des Anwenders oder der CPU-Belastung. AlfredO arbeitet effektiv bei der Erreichung einer besseren Ausnutzung der mobilen Möglichkeiten sowie bei der optimalen Integration von Cloud basierten Anwendungen mit solchen Geräten. In dem zweiten Teil wird ein Leistungsschätzmodell für die Anwendungsbereitstellung vorgeschlagen, durch das die Änderungen im Bereich der Arbeitsbelastung und der Umweltfaktoren erklärt wird. Auf der Grundlage von ausschliesslich passiven Messungen werden in dem Modell Queuing-Networks genutzt, um eine optimale oder annähernd optimale Verteilung einer Anwendung in einer grossen Anzahl von Szenarien genau festzustellen. Im letzten Teil wird der Schwerpunkt von dem mobilen Gerät auf die Frage gerichtet, wie die Cloud basierten Anwendungsmodule in grossen Datenzentren zu verteilen sind. Das Problem wird in umfassender und skalierbarer Weise unter Berücksichtigung der Aspekte Netzwerk, Berechnung und Verfügbarkeitsleistung in einer einzigen Lösung bearbeitet.

Die drei Teile der Dissertation zeigen, dass sowohl die mobile Perspektive als auch die Cloud-Perspektive von grösster Wichtigkeit bei der Verbesserung der Benutzerfreundlichkeit im Zusammenhang mit der Integration von Cloud basierten Anwendungen in den mobilen Geräten sind.