Department of Informatics

A Cloud Storage Overlay to Aggregate Heterogeneous Cloud Services

Dissertation submitted to the
Faculty of Business, Economics and Informatics
of the University of Zurich

to obtain the degree of
Doktor der Wissenschaften, Dr. sc.
(corresponds to Doctor of Science, Ph.D.)

presented by
Guilherme Sperb Machado
from
Porto Alegre, RS, Brazil

approved in February 2016

at the request of
Prof. Dr. Burkhard Stiller
Prof. Dr. Filip De Turck

The Faculty of Business, Economics and Informatics of the University of Zurich hereby authorizes the printing of this dissertation, without indicating an opinion of the views expressed in the work.

Zurich, February 17, 2016

Chairwoman of the Doctoral Board: Prof. Dr. Elaine M. Huang

Berichte aus der Informatik

Guilherme Sperb Machado

A Cloud Storage Overlay to Aggregate Heterogeneous Cloud Services

Shaker Verlag
Aachen 2016

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

Zugl.: Zürich, Univ., Diss., 2016

Printed in Germany.

ISBN 978-3-8440-4883-4
ISSN 0945-0807

Shaker Verlag GmbH • P.O. BOX 101818 • D-52018 Aachen
Phone: 0049/2407/9596-0 • Telefax: 0049/2407/9596-9
Internet: www.shaker.de • e-mail: info@shaker.de

Abstract

Technology advances in the last decade, such as mobile phones, audio and video with high quality, and Online Social Networks (OSN), allowed users to produce, share, and consume data amounts as never seen before. This demand created the necessity of storing and retrieving users' content in a fast, secure, and reliable manner.

In order to bridge the gap for modern storage demands, which includes private and sharing purposes, Cloud computing and Peer-to-Peer (P2P) technologies emerged. The technology employed by Cloud providers is highly elastic, distributed to a high number of servers, geographically spread in different domains, all being interconnected with complex networks. However, from the user's perspective, Cloud providers and their Cloud services are centralized entities, since users have to entirely trust their non-transparent internal process, e.g., users are not aware of what kind of data redundancy and security measures are implemented, nor where their data is stored. Moreover, Cloud services offered by different providers present heterogeneous characteristics: they offer different Application Programming Interfaces (API), accounting and charging schemes, privacy and security levels, functionality, and data type restrictions of files. Despite such heterogeneity, and showing one common aspect among Cloud services, data is stored on Cloud services' servers independently of these various differences.

In contrast to Cloud Computing, Peer-to-Peer (P2P) data storage systems present a higher transparency from the user's perspective: it is known where user's data are stored (the routing and lookup algorithm are known), how it is secured, and under which conditions data will be available or unavailable. However, for data storage purposes, P2P systems also show drawbacks due to churn and lack of a strong identity, leading to a non-suitable storage solution due to its lacking reliability.

Therefore, this thesis proposes an overlay to aggregate Cloud services' storage following a hybrid approach, using centralized and decentralized entities, for upload, download, and sharing files, either publicly or privately. The overlay is referred as *PiCsMu*, the "Platform-independent Cloud Storage System for Multiple Usage". Such overlay is responsible to decide how and where to store user's data, as well to manage the generated metadata by keeping it in a central server (for private files) or in a P2P network (for shared files). The underlay is formed by different Cloud services which stores actual data. The hybrid approach used in PiCsMu takes advantage of the wide storage elasticity of Cloud services, the scalability and privacy of a P2P system, and the stability, reliability, and security of centralized entities.

PiCsMu stimulated three areas of research. First, in order to tackle the heterogeneity problem of storage in Cloud services, this thesis investigates the *data validation* process in Cloud services, to understand how the acceptance of specific file formats is handled for storage purposes. In turn, data can be adapted accordingly to be stored in almost any Cloud service – even those which present data format restrictions –, thus, enabling the aggregation of Cloud services' storage. Second, this thesis defines the system architecture and related processes for upload/download in case of private or sharing purposes. These processes employ data redundancy techniques, encryption schemes, and the placement of data fragments in different Cloud services. Third, this thesis explores the integration of existing OSNs to provide incentives for PiCsMu adoption and to enhance content sharing experience through the means of recommendations. These recommendations are based on data measured from existing OSNs, e.g., measuring which friends are geographically closest to a particular user *and* which of them interact the most with. Such information can give an indication that they might also be interested in adopting the PiCsMu system. In order to achieve that, interactivity- and location-based methods have been developed to measure data collected from existing OSNs.

The PiCsMu system was evaluated as a whole by evaluating the design implementation of its three researched areas. For each of these three research areas the evaluation observed the feasibility, scalability, overhead, functionality, and accuracy of the developed solutions. Additionally, a legal discussion was added, since storage solutions face legal and regulative dimensions upon operation.

The achieved results show that it is possible to build an overlay with Cloud services storage in the underlay, e.g., Google Picasa, Imgur, ImageShack, Amazon S3, and SoundCloud, showing that multiple services with different data format restrictions can be aggregated. Also, these results reveal that PiCsMu scales with respect to different file sizes, showing a moderate data overhead and a low metadata overhead added by the PiCsMu system in case of 1 GByte files. Finally, interactivity- and location-based methods in support of the PiCsMu social recommendations show that they can estimate, respectively, 2 out 5 OSN friends that an OSN user also perceives as he/she interacts most with, and 1.3 out 5 OSN friends that an OSN user also perceives as the geographically closest to.

Kurzfassung

Technologische Errungenschaften des letzten Jahrzehntes, wie beispielsweise Mobiltelefone, hochwertige Audio- und Videoaufnahmen und soziale Netzwerke (OSNs), erlauben Nutzern, niemals zuvor gesehene Datenmengen zu produzieren, zu teilen und zu konsumieren. Hierdurch ergibt sich die Notwendigkeit, Nutzerdaten schnell, sicher und verlässlich zu speichern und abzurufen.

Cloud Computing und Peer-to-Peer (P2P) Technologie etablierten sich, um diesen modernen Speicherbedarf für Privat- und Verteilungszwecke zu decken. Die von Cloud-Anbietern verwendete Technologie ist höchst und über eine Vielzahl von Servern und Domänen verteilt, welche durch komplexe Netzwerke verbunden sind. Dennoch sieht es für die Cloud-Nutzer so aus, als wären Cloud-Anbieter zentralisierte Entitäten, da die Nutzer vollständig den für sie nicht transparenten, internen Prozessen der Cloud-Anbieter vertrauen müssen. So wissen Nutzer beispielsweise oft weder, wie Datenredundanz oder Sicherheit implementiert werden, noch, an welchem Ort die Daten gespeichert sind. Sogar trotz Dienstgütevereinbarungen haben Cloud-Nutzer nach wie vor Risiken zu tragen, beispielsweise bezüglich Datenverlust, Datenlecks und Daten-Nichtverfügbarkeit, wenn sie sich auf einen einzelnen Cloud-Anbieter verlassen. Darüber hinaus haben Cloud-Dienste verschiedener Anbieter unterschiedliche Charakteristika: Sie sind durch unterschiedliche Programmierschnittstellen, Verrechnungs- und Bepreisungsschemata, Privatheits- und Sicherheitsstufen und Funktionalitäten und Einschränkungen bzgl. zulässiger Datentypen gekennzeichnet. Trotz dieser Heterogenität haben alle Cloud-Dienste gemein, dass die Daten auf den Servern des Cloud-Anbieters gespeichert werden.

Im Gegensatz zu Cloud Computing Systemen sind Peer-to-Peer (P2P) basierte Datenspeichersysteme dem Nutzer gegenüber transparenter: Es ist

bekannt, wo die Nutzerdaten gespeichert sind (der Routing- und Lookup-Algorithmus sind bekannt), wie sie gesichert sind und unter welchen Umständen die Daten verfügbar oder nicht verfügbar sind. Wegen des Churn (Peers, die das System verlassen) und des Fehlens starker Identitäten haben P2P-Systeme dennoch Nachteile für die Datenspeicherung, und letztendlich macht sie fehlende Verlässlichkeit als Speicherlösung ungeeignet. Aus diesem Grund entwickelt diese Dissertation ein Overlay, um den Speicherplatz von Cloud-Diensten mittels eines hybriden Ansatzes zu aggregieren, sodass zentralisierte und dezentralisierte Systemkomponenten für Upload, Download, und das öffentliche oder private Teilen von Daten verwendet werden. Dieses Overlay wird als *PiCsMu*, das "Platform-independent Cloud Storage System for Multiple Usage", bezeichnet. Dieses Overlay entscheidet wie und wo die Nutzer-Daten gespeichert werden und verwaltet die dabei anfallenden Metadaten auf einem zentralen Server (bei privaten Daten) oder in einem P2P-Netzwerk (bei geteilten Daten). Das Underlay setzt sich aus verschiedenen Cloud-Diensten zusammen und speichert die tatsächlichen Daten. Der von PiCsMu verwendete hybride Ansatz nutzt die Vorteile der grossen Speicherflexibilität von Cloud-Diensten, die Skalierbarkeit und Privatheit von P2P-Systemen, und die Stabilität, Verlässlichkeit und Sicherheit von zentralisierten Systemkomponenten.

PiCsMu forschte auf den folgenden drei Gebieten. Erstens untersucht diese Dissertation den Prozess der *Datenvalidierung* von Cloud-Diensten, um zu verstehen, wie die Abnahme von spezifischen Datenformaten gehandhabt wird und somit das Problem der Heterogenität von Speichermöglichkeiten verschiedener Cloud-Dienste zu lösen ist. Dieses ermöglicht die Aggregation von Cloud-Diensten, da Daten entsprechend angepasst werden können, um sie dann bei fast jedem Cloud-Dienst abzuspeichern – sogar bei solchen, die Einschränkungen bezüglich zulässiger Datentypen haben. Zweitens definiert diese Dissertation die Systemarchitektur und damit zusammenhängende Prozesse für den Upload und Download von privaten oder geteilten Dateien. Diese Prozesse verwenden Datenredundanztechniken, Verschlüsselungsschemata und das Verteilen von Datenfragmenten über verschiedene Cloud-Dienste. Drittens untersucht diese Dissertation die Integration von existierenden OSNs, um Anreize für die Annahme von PiCsMu zu schaffen und das Erlebnis beim Teilen von Daten durch Empfehlungen zu verbessern. So wird zum Beispiel in existierenden OSNs gemessen, welche Freunde einem PiCsMu-Nutzer geographisch am

nächsten sind *und* mit welchen von diesen er am meisten interagiert, was ein Indiz dafür ist, dass diese Freunde ebenfalls Interesse an des Benutzung des PiCsMu-Systems haben. Hierzu wurden interaktivitätsabhängige und ortsabhängige Methoden entwickelt, mit welchen in existierenden OSNs gesammelte Daten ausgewertet werden können.

Das Design und die Implementierung des PiCsMu-Systems wurden ganzheitlich im Kontext der drei untersuchten Gebiete evaluiert. Für jedes der Gebiete wurde die Durchführbarkeit, Skalierbarkeit, Overhead, Funktionalität und Genauigkeit der Lösungen evaluiert. Da Speicherlösungen gesetzliche und regulative Hürden zu nehmen haben, beinhaltet diese Dissertation eine entsprechende Diskussion der Gesetzeslage.

Die erzielten Ergebnisse zeigen, dass es möglich ist, ein Overlay mit Cloud-Diensten, wie z.B. Google Picasa, Imgur, ImageShack, Amazon S3 und SoundCloud, als Underlay zu entwerfen. Dieses wiederum belegt, dass Cloud-Dienste mit verschiedenen Einschränkungen bzgl. zulässiger Datentypen aggregiert werden können. Es wird bewiesen, dass PiCsMu mit der Dateigrösse skaliert, sodass das PiCsMu-System einen moderaten Daten-Overhead und niedrigen Metadaten-Overhead für 1 GByte grosse Dateien hat. Abschliessend wird gezeigt, dass die interaktivitätsabhängigen und ortsabhängigen Methoden, welche das PiCsMu-System unterstützen, 2 von 5 OSN-Freunden, mit denen der Nutzer glaubt am meisten zu interagieren und 1,3 von 5 OSN-Freunden, die der Nutzer geographisch am nächsten glaubt, prognostizieren können.

Contents