

Performance and mobility in the mobile cloud

Jakub Krzywda
Dept. of Computing Science
Umeå University
SE-901 87 Umeå, Sweden
Email: jakub@cs.umu.se

William Tärneberg
Dept. of Electrical and Information Technology
Lund University
Ole Römers väg 3, 223 63 Lund, Sweden
Email: william.tarneberg@eit.lth.se

Abstract—In an mobile cloud topology the cloud resources are geographically dispersed throughout the mobile network. Services are actively located with close proximity to the user equipment. Geographically migrating a service from data centre to data centre with its user equipment imposes a load on the affected data centres. Consequently, user equipment mobility provides a fundamental problem to the mobile cloud paradigm.

This paper determines the fundamental service performance issues in system of mobile users with dispersed data centres, in relation to the placement of the mobile cloud host nodes and explores the user equipment and provider utility of subscribing to an mobile cloud node at a certain network depth.

Keywords—Cloud, Mobility, Mobile infrastructure, User experience consistency, Omnipresent Cloud, Infinite cloud, Edge cloud, Latency, Throughput, Virtualization, Geo-distributed resources, VM migration

I. INTRODUCTION

Mobile services and user equipment functions are at an increasing rate being virtualized and augmented to the cloud. Applications are soon more often than not seamlessly executed, partially or fully in the cloud. Alongside applications, fundamental user equipment resources, such as storage and CPU, are being virtualized to the cloud. In this paradigm, the border between what is being executed locally and remotely is blurred as developers are given more powerful tools to tap into remote ubiquitous generic virtual resources. This resource paradigm, has overwhelmingly augmented the capabilities of mobile applications, simplifying hardware, and enabled collaborative computing. In the years to come, just short of all devices will contribute data to the cloud and/or utilize its resources.

As we begin to rely more on remote resources we also grow more dependant on the communication delay introduced by the intermediate WAN network and by the geographical separation of the user equipment and the data centre. Latency sensitive applications such as process controls, storage, and compute offloading will quickly falter if subject to a significant and varying communication delay.

The virtual resources are accessed through increasingly congested mobile access networks. More devices are crowding the mobile networks and applications generating and receiving more data, this congestion translates into delay. Additionally, the geographic distance to the data centre introduces a propagation delay, bounded by the speed of light.

The mobile cloud paradigm, put forward by [2], [8], attempts to remedy the aforementioned congestion and latency by locat-

ing cloud resources at the edge of and adjacent to the mobile access networks. In the ad-hoc scenario, resources are shared amongst user equipments as each user equipment surrenders its available resources generically to its peers. However, from a network perspective, at one extreme data centre resources can possibly be located in at the edge of the network, adjacent or integrated into an radio base station, catering for the user equipments residing within its cell. Alternatively, or complementarily, data centres can be integrated with resources in the proposed forthcoming virtualized radio access networks. The scale and the degree of dispersion can be optimized for each application, given the applications resource tiers and its users mobility behaviour.

The geographic proximity between the user equipment and the data centre is proportional to application service delay, to that effect, services hosted in the mobile cloud are migrated with the user equipment, through the network, to minimize this incurred latency. In practice, services, or rather the VMs that host the services will be migrated to the node that is available, provides the lowest delay, and incurs least global network congestion. However, by doing so might minimize the experience delay for the user equipment, but will incur a migration overhead in data centre and in the network a VM is migrated. Conceivably, various schemes and cost functions can be deployed to minimize both the delay experienced by the user and the added resource strain to the data centre and the network.

The topology paradigms of tomorrow's all-IP mobile networks (all-IP (Internet Protocol) [7], [10] are yet to be determined, but one can assume that they will be influenced by the notion of virtualized resources [5], [9]. Large portions of radio base stations can possibly be virtualized and centralized to a common local-geographic data centre, shared by several radio base stations, leaving the radio base stations, in principal, with just the radio interface [12]. The expanse of the centralization is geographically bounded by propagation delay and signal attenuation, and is resource hampered by the aggregate traffic that passes through the dedicated data centre. There is extensive research directed at exploring relevant economic and IT models [1], [7], [13].

The concept of geo-distributed cloud resources has been worked on for a few years, but with a clear focus on storage and sharded data. The authors of [4] present a method to geographically migrate shared data resources globally, not only to minimize the distance between the user equipment and the data centre, and thus service latency, but also to globally load-balance the hosting data centres. Their results reveal a

significant reduction in service latency, inter-data centre communication, and contributed WAN congestion. Their proposed control process runs over longer periods of time and operate on a global scale with geographically static users. Although sharing some fundamental dynamics, albeit at different scales, in contrast, in the mobile cloud paradigm, user equipment movement is much more rapid and proportional to the size of a session. Additionally, mobile cloud virtualized resources are assumed to be universal and do not just cover data, and vary in size and capabilities.

The field of mobile cloud bears much in common with geodistributed cloud resources but is dominated by the notions of augmenting user equipment through virtualizing their resources [3] and reducing service response times through geocascaded data caching [2], [15]. As a result, much of the research is concerned with coping with specific dynamics, and do thus not address the generic case of generic locally geo-distributed resources serving a local subscriber populous. There is large amount of work left to explore the fundamental dynamics of the mobile cloud in order to be able to begin to consider specific applications and use-cases.

User mobility is a key differentiator between traditional distant immobile clouds and the mobile cloud, and is a fundamental dynamic property of a mobile cloud. It is therefore essential to understand how user equipment mobility affects the perceived service performance and what load it imposes on the network in the generic case.

This paper contributes with models designed to examine the fundamental and generic resource problems in a mobile cloud of mobile user equipments. The models include a generic mobile network inhabited by user equipment subscribing to $N_{service}$ services, served by N_{dc} locally geo-distributed data centres.

This paper provides an investigation into the fundamental effects of user equipment in the mobile cloud in relation to the number of subscribers, the abstract placement of the servers, and the number of services. An optimal or reasonable technical bounds for the mobile cloud topology is not yet to be determined. This paper disregards the deeper technical and topological constraints of existing mobile systems in order to provide fundamental results that can be employed to shape the forthcoming mobile network generations.

II. DESIARD MODEL

As the topology of the mobile cloud and future mobile networks is yet to be determined, and due to the fact that we want to research the effects of mobility without a socio-economic model, the network

III. SIMULATION

Multiple runs for each DC/mobile cloud placement mode.

Service The traffic generated by and the usage pattern of a simple web application is characteristic of any smaller mobile application. The HTTP traffic model in [11] provides a small scale closed loop traffic model that is representative of light mobile traffic.

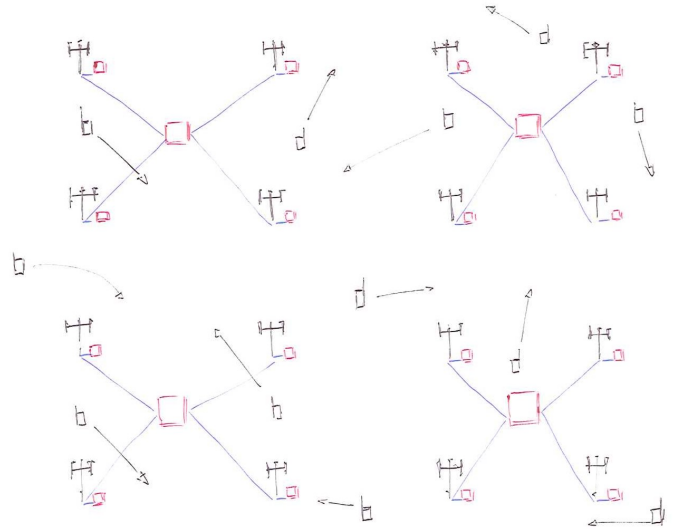


Fig. 1: Performance model

Mobility The 2 dimensional, multi model, mobility model [6] will provide the uniform mobile network with an relevant distribution of users.

Mobile Access Network Provided by Williams SIMJava framework. Handover are instantaneous and move

Core network No delay, no routing

Server The server provides VM and DC models that encompass, incurred VM migration performance degradation in DC and in VM, resulting in a different service time.

Possible service hosting schemes:

- One service model, one VM is employed to host that service for each user.
- One service model, each VM hosts multiple but each number of users, behaving as multiple services while still being compatible.

At all placement modes:

- Measure RTT for all packets at UE
- Measure DC load
- Measure ratio of requests generated vs. processed in mobile cloud node
- Identify the incurred VM migration load

IV. SIMULATION

Multiple runs for each DC/mobile cloud placement mode.

Service The traffic generated by and the usage pattern of a simple web application is characteristic of any smaller mobile application. The HTTP traffic model in [11] provides a small scale closed loop traffic model that is representative of light mobile traffic.

Mobility The 2 dimensional, multi model, mobility model [6] will provide the uniform mobile network with an relevant distribution of users.

Mobile Access Network by Williams SIMJava framework. Handover are instantaneous and move Core network delay, no routing

Server The server provides VM and DC models that encompass, incurred VM migration performance degradation in DC and in VM, resulting in a different service time.

Possible service hosting schemes:

- One service model, one VM is employed to host that service for each user.
- One service model, each VM hosts multiple but each number of users, behaving as multiple services while still being compatible.

At all placement modes:

- Measure RTT for all packets at UE
- Measure DC load
- Measure ratio of requests generated vs. processed in mobile cloudnode
- Identify the incurred VM migration load

V. RESULTS

VI. CONCLUSIONS

VII. FUTURE RESEARCH

- Optimal service/VM migration/placement in relevant topology
- Performance in LTE network topology using LTE-SIM [14]

REFERENCES

- [1] The telecom cloud opportunity. Whitepaper, Ericsson, 2012.
- [2] Ericsson AB. Ericsson and Akamai establish exclusive strategic alliance to create mobile cloud acceleration solutions. Press Release, February 2011. <http://www.ericsson.com/news/1488456>.
- [3] S. Abolfazli, Z. Sanaei, E. Ahmed, A. Gani, and R. Buyya. Cloud-based augmentation for mobile devices: Motivation, taxonomies, and open challenges. *Communications Surveys Tutorials, IEEE*, 16(1):337–368, First 2014.
- [4] Sharad Agarwal, John Dunagan, Navendu Jain, Stefan Saroiu, Alec Wolman, and Harbinder Bhogan. Volley: Automated data placement for geo-distributed cloud services. In *NSDI*, pages 17–32, 2010.
- [5] Fabio Baroncelli, Barbara Martini, and Piero Castoldi. Network virtualization for cloud computing. *annals of telecommunications-Annales des télécommunications*, 65(11-12):713–721, 2010.
- [6] Christian Bettstetter. Smooth is better than sharp: A random mobility model for simulation of wireless networks. In *Proceedings of the 4th ACM International Workshop on Modeling, Analysis and Simulation of Wireless and Mobile Systems, MSWIM '01*, pages 19–27, New York, NY, USA, 2001. ACM.
- [7] G. Caryer, T. Rings, J. Gallop, S. Schulz, J. Grabowski, I. Stokes-Rees, and T. Kovacicova. Grid/cloud computing interoperability, standardization and the next generation network (ngn). In *Intelligence in Next Generation Networks, 2009. ICIN 2009. 13th International Conference on*, pages 1–6, 2009.
- [8] Abhishek Chandra, Jon Weissman, and Benjamin Heintz. Decentralized edge clouds. *Internet Computing, IEEE*, 17(5):70–73, 2013.
- [9] NM Mosharaf Kabir Chowdhury and Raouf Boutaba. Network virtualization: state of the art and research challenges. *Communications Magazine, IEEE*, 47(7):20–26, 2009.
- [10] M.A.F. Gutierrez and N. Ventura. Mobile cloud computing based on service oriented architecture: Embracing network as a service for 3rd party application service providers. In *Kaleidoscope 2011: The Fully Networked Human? - Innovations for Future Networks and Services (K-2011)*, *Proceedings of ITU*, pages 1–7, 2011.
- [11] Zhen Liu, Nicolas Niclausse, and César Jalpa-Villanueva. Traffic model and performance evaluation of web servers. *Performance Evaluation*, 46(2):77–100, 2001.
- [12] Jordan Melzer. Cloud radio access networks.
- [13] S. Pal and T. Pal. Tsaas; customized telecom app hosting on cloud. In *Internet Multimedia Systems Architecture and Application (IMSAA), 2011 IEEE 5th International Conference on*, pages 1–6, 2011.
- [14] G. Piro, L.A. Grieco, G. Boggia, F. Capozzi, and P. Camarda. Simulating lte cellular systems: An open-source framework. *Vehicular Technology, IEEE Transactions on*, 60(2):498–513, Feb 2011.
- [15] L. Ramaswamy, Ling Liu, and A. Iyengar. Cache clouds: Cooperative caching of dynamic documents in edge networks. In *Distributed Computing Systems, 2005. ICDCS 2005. Proceedings. 25th IEEE International Conference on*, pages 229–238, June 2005.