Performance and mobility in the mobile cloud

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

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

***Abstract*—In a 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 a 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

User equipment mobility is a key differentiator between traditional cloud computing with distant data centres 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.

Mobile services and user equipment [1] functions are at an increasing rate being virtualized and augmented to the cloud. Applications will soon more often than not be 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, and enabled collaborative computing. In the years to come, at the dawn of the era of the Internet of things, 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 suceptible to the communication delay introduced by the the intermediate WAN network and by the geographical separation of the user equipment and the data centre [10]. Latency sensitive applications such as process controls, latency sensitive storage, real time video game rendering, and augmented reality video analysis will quickly falter if subject to a significant and varying communication delay.

[1] Any user client device accesing the service, such as a mobile phone

Virtual resources are accessed through increasingly congested mobile access networks. More devices are crowding the mobile networks and applications are generating and receiving more data, this congestion translates into delay or latency [15]. 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 [3], [9], [17], [25], [27], attemps to remedy the aforementioned congestion and latency performance inhibitors by locating cloud resources at the edged 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 proposedly be located at the edge of the network, adjacent or integrated into an radio base station, catering for the user equipments located within its cell coverage. Alternatively, or complimentary, 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.

Round trip time, is proportioanl to the geographic proximity between the user equipment and the data centre. 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 data centre that, is available, provides the lowest service latency, and incurs least global network congestion. Doing so might minimize the experienced service delay for the user equipment, but will incurr a migration overhead in the hosting data centre and in the network over which the 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 tomorrows all-IP (Internet Protocol) mobile networks [8], [14] are still to be determined, but one can assume that they will be influenced by the notion of virtualized resources [6], [11]. Large portions of radio base stations can proposedly be virtualized and centralized to a common data centrewith a localy-bounded service domain, shared by several radio base stations, leaving the radio base stations, in principal, with just the radio interface [20]. The expanse of the centralization is geographically bounded by propagation delay and signal attenuation, and is resource hampered by the aggregated traffic that passes through the dedicated data centre. There is to our knowledge, very little research exploring

future mobile telecom infrasturcture topologies with the mobile cloud in mind. There is on the other hand, extensive research directed at exploring relevant economic and IT models of how to integrate existing telecom servies to the cloud and how to apply telecom-grade SLAs to existing cloud services [1], [8], [21].

The concept of geo-distributed cloud resources has been worked on for a few years, but with a clear focus on storage and shared data. The authors of [5] 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 include data and vary in size and capabilities.

The field of mobile cloud has much in common with field of geo-distributed cloud resources, but is dominated by the notions of augmenting user equipments through virtualizing their resources [4] and reducing service response times through geo-cascaded data caching [3], [23]. As a result, much of the research is concerned with coping with specific dynamics, and do thus not address the generic case of small geo-distributed data centres, serving a local mobile subscriber populous. There is large amount of work left to explore the fundamental dynamics of the mobile cloud in order to be able to consider specific applications and use-cases.

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 equipments subscribing to a number of services, served by a number of 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 data centre, and the number of services.

II. Desired model

The desired model shall provide a setting for which we can explore fundamental resource and performance properties of the mobile cloud system paradigm with mobile user equipments. The mobile user equipments, radio access network, and service application will subject the data centres with a load characteristic for generic mobile phone traffic and the type of services that plausible might be deployed to the mobile cloud.

As the topology of any future mobile cloud or proposed forthcoming mobile networks is yet to be determined, in this paper we propose a generic telecom infrastructure model that disregards generational specific properties such as those found in the physical layer and radio resource load-balancing

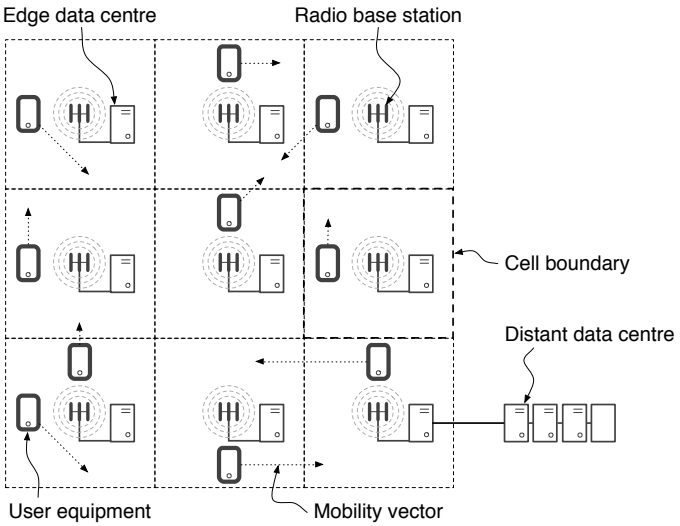

Fig. 1: System model

methods. These properties are not system variables at the abstraction level the mobile cloudneeds to be modelled in this paper. Nevertheless, conceivably, and in order to confine the geographic domain of the model, the model adheres to current general LTE cell planing practices [24], see Figure 1.

In order to be able to explore the fundamental effects of mobility on the performance of an mobile cloudservice in the generic case, the model does not adhere to any socidemographic patterns or urban topologies. With out any geographic bias, the mobile network base stations are uniformly distributed across its 2-dimensional domain.

Similarly, in order to represent the variety of possible services, the service model shall generate traffic that is characteristic for an active, generic, user equipment. Additionally, the generated traffic shall be provided by a stochastic process that is also independent of location.

Themobility model, the service model, and the uniformly distributed mobile network provides the modelled data centres withc a characteristic workload. It is worth reiterating that the traffic load is more relevant to our investigation than specific topological and network properties.

The data centre model will host multiple VMs that will process the arriving requests corresponding to its service commitment. Additionally, when a VM is migrated between data centres it shall incur a load on both data centres. Furthermore, the resources within a data centre are shared amongst the hosted VMs. The amount of compute resources dedicated to one service is thus proportional to the number of services hosted in that data centre. Minute memory management, interference, and cross-talk effects are not fundamental performance properties at this scale and are therefore not modelled.

III. Simulation model

A. Service

Most mobile applications use HTTP as a means to communicate with remote services [12], [19]. The HTTP traffic

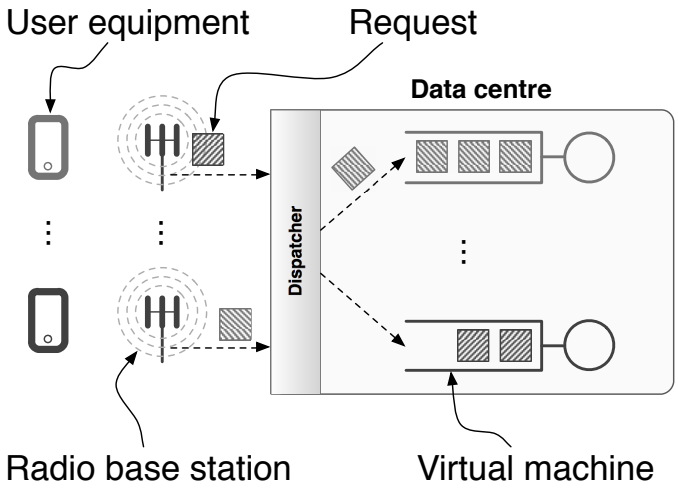

Fig. 2: Data centre model

model in [18] provides a small scale open loop traffic model that is representative of light mobile traffic.

B. Mobility

The 2-dimensional, multi modal, mobility model [7] will provide the uniform mobile network with a relevant distribution of users.

C. Mobile access network

Forthcoming cell plannig practices will aim increase area energy efficieny by favouring smaller cells in urban areas [13], [26]. The model will therefor employ a small homogenous mobile network composed of N_{rbs} radio base stations equidistantly distributed. The domain which the network serves is populated by a homogenous group user equipments, with a uniform service subscription distribution. A user equipment is handed over between base stations at the point where they cross the cell boundary distinguishing two independent radio base stations defined by the width of the rectangular cells d_{rbs}. The mobile access network model does take into account the physical layer, channel provisioning, and cell load balancing. Additionally, the radio access network functions as a mechanism to associate user equipments with data centres, propagation and system processing delays are thus not modelled.

D. Core network

The core network is modelled with a Weibull delay T_{net} in multiples of the number of network nodes between the source and the destination, in accordance with [22]. The distance between radio base stations is equal to the cell dimension d_{rbs}. Associated radio base stations are equidestant to their common data centre.

E. Data centre

As illustrated in Figure 2, data centre is modelled as series of parallel queues, one for each allotted VM, and thus, service $N_{i,vm}$. A dispatcher directs incoming requests to the corresponding VM based on which service S_j it carries. In normal operation, each request is served by each queue with a service time of $T_{i,vm}$, unique to the i_{th} data centre and is proportional to the number of VMs $N_{i,vm}$ running concurrently in the data centre.

To simplify the model of a data centre we will not consider CPU, memory, storage and intra data centre network separately. Instead, in this paper, we will use an abstraction of one dimensional computational resource.

1) Hosting VMs in a data centre: Hosting VMs in a data centre can be modeled in two ways: with or without competition for computational resource.

In the first approach, the resources of a data centre are aggregated in one pool that is continuously divisible. The pool of resources is divided evenly among all VMs. Hence, when the number of VMs hosted in the data centre increases, the amount of resources available for each VM shrinks. Consequently the service time of processing requests of each VM lengthens.

In the second approach, the resources of a data centre are discrete and each computational unit is used exclusively by one VM. Therefore, there is no influence of one VM on another. To incorporate the fact that the amount of resources is finite we put a limit on the maximal number of VMs that can be hosted in one data centre. Furthermore, on each data centreeach service is contained within one VM.

2) VM initialization and termination: When a decision of deploying a new service in a data centre is taken, a new VM will be started there to host that service. Due to the startup time, the newly admitted VM will not be able to start processing requests for a period of T_{vm_init}. Nevertheless, the new VM will start using resources of the data centre from the time of admission. Because of that, the service time, $T_{i,vm}$, for each of the VMs hosted in that data centre will be recalculated at that point. Similarly, after finishing serving the last request, VM will be still using the resources of the data centre for time $T_{vm\_release}$.

3) VM activation scheme: Definitions: service, session, client subscribed to service

Private VM — IaaS like, a VM is exclusively used by one user for offloading computations from his user equipment. User provides the executable program that is loaded to an edge data centre from an user equipment or a remote data centre.

> **per-session** — a VM is initialized upon receiving the first request from a client and terminated just after finishing processing the last request of the session.
> **client-within-cell** — a VM is initialized when a client that is subscribed to a service enters a cell and is kept alive as long as he stays within the cell.

Shared VM — SaaS like, a VM hosts a service that can be concurrently accessed by many users.

> **any-client-running** — a VM is initialized upon receiving the first request from a client and terminated when there are no more requests to serve

(waiting queue is empty and all sessions are finished).

any-client-within-cell — a VM is initialized when the first client that is subscribed to a service enters a cell and is kept alive as long as any subscribed client stays within the cell.

4) VM migration: Migration has an influence on the service performance as well as on the resource availability on both source and destination data centre. On the source side, apart from ordinary resource requirements due to serving requests, VM consumes resources for sending its image to the destination data centre. In a case of postcopy live migration, VM at the source side still uses some resources even after the workload is redirected to the new location. That is because the VM at the destination side pulls remaining memory pages from the source data centre.

To model the overhead of migration on the service performance additional delay in the response time should be introduced. Primarily, for the time of transferring the image of VM, $T_{vm\_transfer}$, the time of serving a request should be lengthen by $D_{vm\_transfer}$, because of using a part of resources for I/O operations. Moreover, during the time of switching the execution from the source to the destination, $T_{vm\_downtime}$, VM is not able to serve any requests. Additionally, in the case of the postcopy migration, delay D_{memory_pull} occurs for some number, N_{memory_pull}, of the first requests after redirecting the workload to the new data centre, due to the remote memory calls.

When VMs compete for resources, running additional VM on the destination side introduces an overhead by increasing the response time of other collocated VMs.

IV. Experiments

The aforementioned model was implemented in Java employing simjava [2] as the event driven framework.

A. User equipment, mobility and network

To reveal the effect of varying load on the data centres, the number of users, placement of the data centres, and the number for services were varied independantly thoughout as many simulation scenarios.

In proportion to the range of movemt and evenly divisibale by N_{dc}^2, the simulation domain spanns 16 radio base stations N_{rbs}. The cell dimension d_{rbs} adhere to a proposed maximum cel radius of 750m, as detailed in [26]. Although rectangular, its area equates to the area of a circular cell with a radius of 750m, with results in a cell width and hight d_{rbs} to 1300m. The population of user equipments N_{ue} ranged from from 10 to 1600 user equipments, doubleing with each increment. The number of data centres N_{dc} was varied between 16, 4 and 1, with each data centre serving 1, 4 or 16 radio base stations respectively. Additionally, for each run of N_{ue} the number of services N_{ser} was varied between 1 and 5.

The simulation time is set to 8 hours, $T_{sim} = 28800$ seconds, leaving sufficient time for each user equipmentto on average vist half of the radio base stations.

Parameter	Value
N_{ue}	10–500 (step: 10)
N_{rbs}	16
N_{dc}	$\{1, 4, 16\}$
N_{ser}	1–5
T_{ser}	
T_{sim}	28800 seconds
d_{rbs}	
T_{net}	
$N_{vm,limit}$	
T_{vm_init}	avg 82 s, min 69 s, max 126 s
$T_{vm\_release}$	avg 21 s, min 18 s, max 23 s
$T_{vm\_transfer}$	
$D_{vm\_transfer}$	
$T_{vm\_downtime}$	
D_{memory_pull}	
N_{memory_pull}	

TABLE I: Simulation parameter values

B. Data centreand VM parameters

Times for resource allocation and release and lognormal distributed with means T_{vm_init} and $T_{vm\_release}$ respectively, and are modelled after Amazon EC2 measurements for m1.small instance (1 vCPU, 1.7 GB of RAM and 160 GB disk) [16].

The service time for each data centre is set independently, proportional to the mean request generation rate over the number of radio base stations and the number of VMs running, see Equations **??**. Where λ_{sys} is the aggregate arrival rate to the system, and K is a scaleing factor set to 1. Each VM is provisioned to handle 50 users ($N_{ue} = 50$). The mean service time for each VM is thus $1/\mu$.

$$\mu = K \cdot \lambda_{ue} \cdot N_{ue} \cdot \frac{N_{services}}{N_{dc}} \tag{1}$$

$$\lambda_{ue} = \frac{\bar{N}_{requests}}{\bar{N}_{requests} \cdot \bar{T}_{request} + \bar{T}_{session}} \tag{2}$$

As resources are assumed to be obiqutous we wich to observe the effects of an overloaded data centre. For modeling simplicity, data centrecapacity if definef by a limit of how many VMs $N_{vm,limit}$ in can simultaniously host. Our experiment regards two fundamental provisioning scenarios. When the VM limit $N_{vm,limit}$ has been reached, either any new VMs are denied or the additionaly needed data centrecapacity is shared equaly amongst the $N_{i,vm}$ VMs by increasing the service time $T_{i,ser}$ with a factor K_{over},see Equation 3.

$$K_{over} = \frac{\max(N_{i,vm}, N_{vm,limit})}{N_{vm,limit}} \tag{3}$$

Simulation model parameters used in the experiments can be found in Table I, likewise the service parameters are declared in Table II.

C. Measurements

To ensure statistical accuracy, each simulation scenario was independently replicated 10 times.

Component	Distribution	Parameters
S_f	Pareto	K=133000 α =1.1
S_r	Pareto	K=1000
D_r	Weibull	α =1.46 β =0.382
D_s	Pareto	K=1 α=1.5

TABLE II: Service model components

For each abovementioned simulation scenario, every request is recorded with where it was prossed, if it was terminated, the time spent queing, processing, and propagating. Similarly, for each VM the proportion of time spent in each state is recorded.

V. Results

VI. Conclusions

VII. Future research

A. VM migration schemes

- **Precopy** — the whole memory of VM is copied preemptively before switching the execution to the new data centre.
- **Postcopy** — the memory of VM is copied after switching the execution to the new data centre as is needed to serve the incoming requests.

B. VM placement and data centre provisioning schemes

To understand the effects of various migration schemes on the data centreand service performance, the following migration schemes were deployed:

- A VM for service S_j, if active, residec in the data centrewith the largest number of subscribers. If this criteria were to change the the hosting VM will migrate to the resulting data centre.
- Each data centrethat hosts a user equipmentthat subscribes to S_j hosts an instance of a service S_j VM. If users disperse, the VM for service S_j will duplicat to the receiving data centre.

Differentiation between short term jobs (online processing) and long term (offline processing after upload, getting statistics, big data)

Use different service models or change parameters of distributions in current one. Model specific applications (youtube like, facebook like, etc.) and compare them, or show implications of different abstract and extreme configurations.

C. Multi-tierd service placemet schemes in the mobile cloud

References

[1] The telecom cloud oppertunity. Whitepaper, Ericsson, 2012.

[2] simjava, 02 2014. Available online at http://www.icsa.inf.ed.ac.uk/research/groups/hase/simjava/.

[3] 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.

[4] 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.

[5] 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.

[6] Fabio Baroncelli, Barbara Martini, and Piero Castoldi. Network virtualization for cloud computing. *annals of telecommunications-annales des télécommunications*, 65(11-12):713–721, 2010.

[7] 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.

[8] G. Caryer, T. Rings, J. Gallop, S. Schulz, J. Grabowski, I. Stokes-Rees, and T. Kovacikova. 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.

[9] Abhishek Chandra, Jon Weissman, and Benjamin Heintz. Decentralized edge clouds. *Internet Computing, IEEE*, 17(5):70–73, 2013.

[10] Baek-Young Choi, Sue Moon, Zhi-Li Zhang, Konstantina Papagiannaki, and Christophe Diot. Analysis of point-to-point packet delay in an operational network. *Computer networks*, 51(13):3812–3827, 2007.

[11] NM Mosharaf Kabir Chowdhury and Raouf Boutaba. Network virtualization: state of the art and research challenges. *Communications Magazine, IEEE*, 47(7):20–26, 2009.

[12] Hossein Falaki, Dimitrios Lymberopoulos, Ratul Mahajan, Srikanth Kandula, and Deborah Estrin. A first look at traffic on smartphones. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pages 281–287. ACM, 2010.

[13] AJ. Fehske, F. Richter, and G.P. Fettweis. Energy efficiency improvements through micro sites in cellular mobile radio networks. In *GLOBECOM Workshops, 2009 IEEE*, pages 1–5, Nov 2009.

[14] 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.

[15] Ningning Hu, Li Li, Zhuoqing Morley Mao, Peter Steenkiste, and Jia Wang. A measurement study of internet bottlenecks. In *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*, volume 3, pages 1689–1700. IEEE, 2005.

[16] A Iosup, S. Ostermann, M.N. Yigitbasi, R. Prodan, T. Fahringer, and D. H J Epema. Performance analysis of cloud computing services for many-tasks scientific computing. *Parallel and Distributed Systems, IEEE Transactions on*, 22(6):931–945, June 2011.

[17] Niko Kiukkonen, Jan Blom, Olivier Dousse, Daniel Gatica-Perez, and Juha Laurila. Towards rich mobile phone datasets: Lausanne data collection campaign. *Proc. ICPS, Berlin*, 2010.

[18] Zhen Liu, Nicolas Niclausse, and César Jalpa-Villanueva. Traffic model and performance evaluation of web servers. *Performance Evaluation*, 46(2):77–100, 2001.

[19] Gregor Maier, Fabian Schneider, and Anja Feldmann. A first look at mobile hand-held device traffic. In *Passive and Active Measurement*, pages 161–170. Springer, 2010.

[20] Jordan Melzer. Cloud radio access networks.

[21] 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.

[22] Konstantina Papagiannaki, Sue Moon, Chuck Fraleigh, Patrick Thiran, and Christophe Diot. Measurement and analysis of single-hop delay on an ip backbone network. *Selected Areas in Communications, IEEE Journal on*, 21(6):908–921, 2003.

[23] 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.

[24] J Salo, M Nur-Alam, and K Chang. Practical introduction to lte radio planning. *A white paper on basics of radio planning for 3GPP LTE in interference limited and coverage limited scenarios, European Communications Engineering (ECE) Ltd, Espoo, Finland*, 2010.

[25] Mahadev Satyanarayanan, Paramvir Bahl, Ramón Caceres, and Nigel Davies. The case for vm-based cloudlets in mobile computing. *Pervasive Computing, IEEE*, 8(4):14–23, 2009.

[26] Suhail Najm Shahab, Tiong Sieh Kiong, and Ayad Atiyah Abdulkafi. A framework for energy efficiency evaluation of lte network in urban, suburban and rural areas. *Australian Journal of Basic and Applied Sciences*, 7(7):404–413, 2013.

[27] William Voorsluys, James Broberg, Srikumar Venugopal, and Rajkumar Buyya. Cost of virtual machine live migration in clouds: A performance evaluation. In MartinGilje Jaatun, Gansen Zhao, and Chunming Rong, editors, *Cloud Computing*, volume 5931 of *Lecture Notes in Computer Science*, pages 254–265. Springer Berlin Heidelberg, 2009.