x-cloud challenges

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

William Tärneberg
and Montgomery Scott
Dept. of Electrical and Information Technology
Lund University
Ole Römers väg 3, 223 63 Lund, Sweden
Telephone: +46 (0)46 2229021
Email: william.tarneberg@eit.lth.se

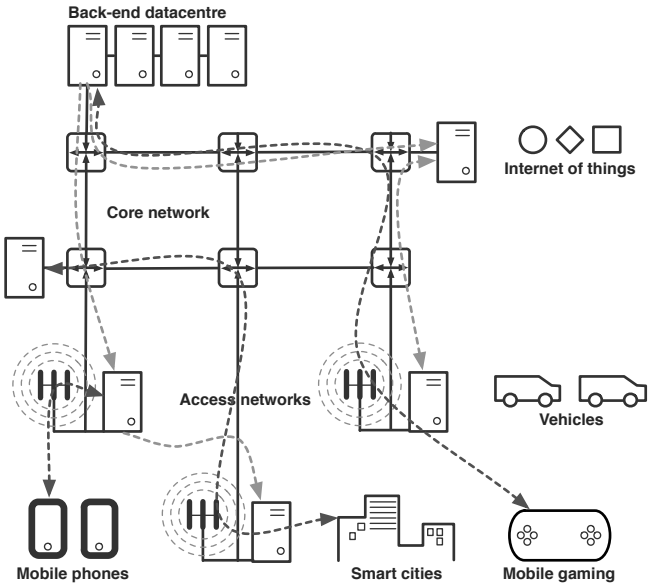

Fig. 1: x-cloud

***Abstract*—The abstract goes here.**

I. Introduction

II. The case for the x-cloud

- Internet structure, latency, and bandwidth: [6]

A. The bandwidth case for x-cloud

B. The latency case for x-cloud

The intermediate latency between a client and a data centre is a product of propagation, modulation, and network routing and traffic shaping. Propagation is a clear physical obstacle to reducing latency, and there is very little evidence to suggest that information will propagate faster than $\frac{2}{3}$ of the speed of light, at scale, in the near future. Furthermore, the delay in the backbone network is incurred to the most part by routing. A full point to point network where the propagation speed is the only limit, is not economically viable and would dissolve the fabric of the Internet. As such, we can always expect a certain amount of network contributed latency and jitter. At best, an LTE mobile access network adds about 5 ms of latency [2].

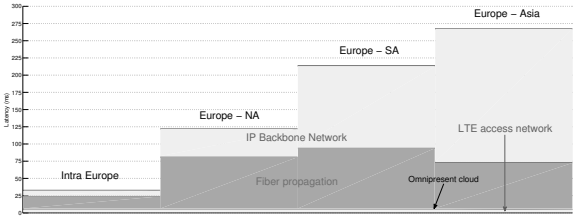

Fig. 2: IP Internet latency in western Europe [1] over LTE [2]

Radio access network latency can be expected to diminish over the next few generations of mobile networks.

Moving the cloud data centres closer to the IP backbone networks eliminates some of the additive latency on one side of the connection. Doing so, not only eliminate the propagation delay, but will over time, add more complexity to peripheries of the backbone as more severs nodes make their home there.

The x-cloud remedies this latency challenge in a more sustainable way. By moving compute resources to the mobile networks, IP backbone network propagation and routing delays are eliminate without disrupting the Internet topology. The resulting distributed infrastructure is capable of delivering content and services at latencies less than 10 ms.

The x-cloud will thus enable latency-sensitive services to be migrated to the cloud, such as, gaming, financial trading, process control, and most real-time human-machine interaction process.

C. The infrastructure case for x-cloud

Distributed virtualized mobile networks will rely on centralized compute nodes for higher level link management. One node will proposedly host multiple base stations, to which they connect over a network link, much like the Ericsson Radio Dot System [**?**], but at a larger scale. The size of these virtualization resource nodes is proportional to the maximum distance they can reside from the radio nodes, given the induced propagation delay. Supposedly these virtualization resource nodes will the placed in the vicinity of the core IP network. The virtualization resource nodes can be seen as to define geographic areas whos boundaries are defined by the reach of the mobile network which it serves. Depending on the level of desired provision and load balancing flexibility, these geographic domains will overlap to varying degrees.

The virtualization resource nodes are conceivably constructed of generic x86 or ARM servers, hosting VMs or containers within which the virtualized mobile network infrastructure is executed. Given the placement of the virtualization resource node, any free or designually excess capacity can be used hist other services.

The topology is designed to optimize the use of radio resources, the geographic domains which the virtualization resource nodes constitute do not necessarily overlap or map the demographic area which x-cloud services operate.

III. Simulation model

A. Models

1) Data Centre: [4]

2) Network:

- Latency
  - Point-to-Point core network delay model [3]
  - One-hop core network router queue delay model [5]

B. Python + [NS-3, Omnet, Matlab, Modelica]

C. Java

IV. Papers

A. Comparison of existing simulators from the perspective of x-cloud

A survey of existing simulators with comparison of their capabilities (and limitations) to simulate x-cloud .

Simulators of:

- Data centers,
- BTS,
- Network (BTS — DC),
- Mobile network,
- Mobile devices,
- Users (mobility).

What is different in operation/simulation of x-cloud ?

B. Limitations of current infrastructure & the setup/structure of x-cloud

1) Limitations of current infrastructure: Simulate the current infrastructure (mobile network + remote/big data centers) and show the limits of it.

- What will happen when the number of mobile devices increases by order(s) of magnitude? The influence on a network connection between a base station and a big/remote data center.
- How many mobile devices can be handeled by the current infrastructure (depending on a latency limits)?

2) The setup/structure of x-cloud : The x-cloud consists of antennas, small (edge) data centers and big (remote) data centers. Small data centers are located close to the antennas and can host both virtualized base station software and VMs with applications. Big data centrs are located far away from users. Small data centers have smaller amount of resources than big ones (maybe also performance is lower) and running applications there is more expensive. However, latency is is much lower than in a case of big (remote) data centers.

Questions about the setup/structure of x-cloud :

- how many antennas should be associated with one small (local) data center? (probably this will be limited by the latency between an antenna and a small data center)
- how big should small (local) data centers be (#CPUs etc)?

C. x-cloud model

D. Throughput and bandwidth limitations in the x-cloud

E. Virtual Machine placement and migration in x-cloud

Regarding placement of Virtual Machines (VMs) in the edge data centers:

- Should a VM that serves all users (even these outside of the range of the directly connected antennas) be placed in an edge data cener or should it be rather an additional instance that serves users that are in the close proximity (duplicating a VM in a big data center)?
- When a VM should be placed/duplicated in a small data center?
- While users are moving from one antenna to another when VM should be migrated from one edge data center to another one?

F. Other thoughts

- Different workload patterns?
- How to perform monitoring? (System is very distributed)
- Maybe new metrics to monitor (eg. distance from antenna, velocity, direction, etc.)
- Changes in the architecture of mobile applications

V. Conclusion

References

[1] Monthly network summary, 05 2014.

[2] T Blajić, D Nogulić, and M Družijanić. Latency improvements in 3g long term evolution. *Mipro CTI, svibanj*, 2006.

[3] 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.

[4] H. Khazaei, J. Misic, and V.B. Misic. Performance analysis of cloud computing centers using m/g/m/m+r queuing systems. *Parallel and Distributed Systems, IEEE Transactions on*, 23(5):936–943, May 2012.

[5] 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.

[6] Venugopalan Ramasubramanian, Dahlia Malkhi, Fabian Kuhn, Mahesh Balakrishnan, Archit Gupta, and Aditya Akella. On the treeness of internet latency and bandwidth. *SIGMETRICS Perform. Eval. Rev.*, 37(1):61–72, June 2009.