Proposal for Google Research Awards (PhD. level) - Latin America

1 Overview

Proposal Title:

Faculty Full Name: Luiz Fernando Bittencourt bit@ic.unicamp.br
PhD candidate: Eduardo de Souza Gama eduardogama@lrc.ic.unicamp.br
Google Contact name: Walfredo Cirne

2 Proposal Body

Abstract

The video streaming services already is responsible for the majority of the Internet traffic. A good cloud-level architecture partially solves some issues related to the live stream and Video on Demand (VoD) services. At same time, however, it introduces new ones such as higher latency and core network congestion. In order to improve on this matter, this work proposes a set of services to video streaming in a multi-tier fog computing environment. It also takes into consideration classified hierarchical tiers and the ETSI-NFV architecture. The main goal is to design and assess a reliable and high-quality multi-tier services architecture to be used in the fog/edge environments. We discuss the impact on the performance introduced by the multi-tier fog/edge computing, and also how the services may be used to improving the Quality of Experience (QoE) for end-users.

The infrastructure of the network edge has been growing rapidly. Led by globally known companies in the market. CDN providers like Akamai, LimeLight, Level 3 and MaxCDN have made major investments in leading edge servers in emerging markets such as Asia, South America and Africa. VoD service companies such as Netflix, Google and Amazon have interests in improving their CDN services, as a primary goal to save on bandwidth costs and operational services without violating QoE guarantees. In addition, this paradigm has attracted companies such as Intel, Qualcomm, IBM, Cisco, who usually mention a big growth on the edge of the network in the near future.

Research Goals

Currently the Internet is based on the characteristics of the best effort (Best Effort), that is, there is no guarantee of level of service. However, users often need to use the services made available in the cloud with a QoS guarantee, which is specified in a contract known as a Service Level Agreement (SLA). In other words, SLA is a document agreed between two or more entities in which the description, delivery and collection of contracted services are formally defined.

In addition, there are many pragmatic trade-offs in edge provisioning. In addition to meeting the purpose of proximity between users and edge servers, edge operators also consider cost budgets, the ability of edge sites, provisioning edge site interrupts, and constraints that users can work on. regional laws or policies of ISPs. These considerations usually fit well together, resulting in a huge research space. Edge operators should consider these offsets comprehensively and exploit the best provisioning plan under various constraints

In addition, a common problem with the video stream provisioning mechanism in CDNs, where CDN nodes in fog should be encouraged to share their resources and contribute to the content design. This can have a serious impact on the quality of streaming system service. Existing solutions address the problem by categorizing following way:

- *Monetary base*
- *Based on reciprocity*
- *Based on Reputation*

A uso da computação em névoa para a distribuição de conteudo CDN pode ser vital para a good QoS and QoE garantees. In these kinds of scenarios - that make up the fog network - may be any access points (AP) can be extended to also provide computing and storage services. The APs may be any device that offers wired/wireless network connection to the end-user, such as smart phones, tablets, laptops, connected vehicles.

To clarify the possibilities, we first introduce a multi-tier fog computing model detailedly in Figure **??**. As we can see, the higher tier is composed by the Provider that offers video streaming services (e.g. Netflix, Amazon, Youtube) located at the cloud, and makes either use of their own cache systems or leases one offered by CDN providers. These providers operate original content in datacenters at the WAN (long-haul netwoks). In fog network, the APs located at the fog are responsible to provisioning resources to CDN providers allocate their caches. Notice that the caches are to multi-hop away from the content provider/consumer. A very popular fashion to organize the cache placement content hierarchicaly, and bi-directional communication. In addition, each content pode ser splitted dynamically in a set of pieces in order to serve the end-users, taking into account different aspects, such as the mobility presented by the user and the possibility to predict the moving directing.

With new possibilities being created to offer better services and the working of the internet. As long as the network becomes more robustness, the problem become more complex and new challenges arises to be solved. To tailored a CDN systems with multi-tiers environment into the fog, different characteristics have to be studied such as cache allocation, placement, replacement and selection caches, usually, making real-time decisions. As different cache sizes being allocation over the tiers to storage a range of content to provisioning a region. Also, an cache size in an AP device have to be able of deal a set of different content pieces, in order to the end-users may have QoE garaties. Thus, different level of granularity into the APs devices arises, if the granularity of certain areas becomes too fine, scalability issues begin to appear. Dividing the content of the fog in the network over the APs at the right level of granularity is a complex problem in its own right.

For retrieve a data afetwards a VM (or content, service, etc) being allocated or allready exists. The operations feitas pela aplicação podem ser receiver-driven, such as used in Content-Centric Network (CCN) and named data network (NDN) paradgm. Organizadas de hierarquiquamente, one content object name could be retrive trough an url name "ucla/videos/demo.png". For a large video that is shareded em a set o chunks, thereby que uma requisiçao pode ser feita especificando um chunk do video da seguinte foma "ucla/videos/demo.png/1".

Um provedor de conteudo possui a set of repositorios CDN geograficamente distribuidos. As requisições dos usuarios para conteudos geralmente são feitas atraves de HTTP GET requests. To provisioning the content, the CDN redirects the request through either DNS redirection [38] or IP-layer anycast [41].

O conteudo provido por repositorios CDN geograficamente distribuidos, visando estar proximo ao usuario. Procura melhorar o troughput da rede. Mas para isso o provedor do conteudo alvo. CDN system

3 Business model perspective

A flexibilidade ao utilizar da edge networks infrastructure oferecem new oportunidades para melhorar o servico, abordando metricas e perspectivas que geralmente são analisadas de forma independente. It can be seen in three different perspectives:

- **End-user perspective:** uninterrupted connectivity and communication services, smoth consumer experience.
- **Content provider perspective:** connected intelligent transportation systems, road-side service units, sensors, and mission critical monitoring/traking services.
- **Network operator perspective:** scalable, energy- efficient, low-cost, uniformly-monitored, programmable, and secure communication infrastructure.

4 Distribution of Information Centric

5 CDN Slicing Architecture

This CDN service based on fog computing consider the scenarios presented in Fig. **??**. This section is interested in make up an architecture capable to provisioning this service in on-demand manner. In order to investigated aspects as scalability over the live migration, granularity of different contents and providers, aside from mobile predictability considering the social medias and moving direction. Initially, the model at the Fig **??** shows two layer components: the user device, fog nodes into the multi-tier APs and the cloud providers.

The usage of CDN multi-tier fog platform rather than the cloud CDN have to be consider, mainly, the cost reduction. The Edge server placement problem is responsavel de selecionar um nó na borda da rede. There will exists two kind of nodes in the network, which are both known and unknown nodes due to the flexibility presented by the edge. The edge node location could respresented by a tuple da $<< city, AS(Autonomous\ System) >>$. let us assume that in current scenario a set of files $S = \{s_1, s_2, ..., s_n\}$

The problem to cost minimization could be reducted to the problem "the optimal bandwidth allocation problem" (OBAP), described in [].

The cost of a solution could range of an agreement with a CSP. As the focus of CDN apps are high in/out of datas, this work leverage the concerns of price machines needed to the CDN utilization and the charge of fog networks, aside from the cost of download and upload files in order to sync different nodes and distribute the files to end-users.

For provide a high-quality experience sharing into the same regional APs, the model

For provide a high-quality experience general components have to take advantage of the moving direction presented by the user. This way, the monitoring and analytics modules may cooperate togheter with advanced machine learning, neural networks techniques for decision-making. Characteristics as users location, direction movement, users behaviour, content resquested, amount of stored data, data traffic needs, cloudlets capacity, newotrk capacity, and so on.

make location-based virtual machine migration feasible in the Fog computing paradigm.

to build pro a hierarchical, bi-directional computing infrastructure: edge devices communicate with cloudlets and cloudlets communicate with clouds. Cloudlets can also communicate with each other to perform data and process management in order to support application requirements, and to exchange Fog control/management data (such as user device and application state).

al-time management decisions, maintaining the service level agreement.

With the possibility to provision algorithms as a set of function trough VNFs to provision end-user some service, CDNaaS may be decouple in specific functions which will be orchestrate over the network to provision QoE garantees for end-users. Below presents an Architecture standard ETSI for CDN.

To deliver an efficient CDNaaS into an heterogeneos network environment with multi tier fog nodes, our envisioned architecture showed up in Fig 1 is composed by a cloud provider where is contented big datacenters, from here the when the user request a movie. The fog computing may be organized hierarquicaly by multiple levels from which serve como meio de prover o serviço de CDN. As can be seen in fig 1 an user 1 receiving a content from WiFi and the another user is receiving from both LTE and user 1.

The most recently CDN architectures highlight Network Function Virtualization (NFV) principle para permitir a execução de serviços especificos em servidores remotos. Accorndlying to [], NFV supports a wide range of services by orchestrating the VNF (virtual network function) deployment and operation across diverse computing, caching, and networking resources over the multi-tier fog nodes. Whereas SDN paradigm decouples the control plane from the underlying data plane in a centralized manner. It is independent from NFV, but it can add more flexibility to the SFs. Therefore, a hybrid concept may exploit both NFV and SDN at the same time. Virtualization concepts are integrated in different areas such as cloud and telecommunication for more services flexibility provisioning. This combination technology may be deployed as virtual machines capable to provision slices of CDN components on-the-fly in a dynamic manner [2][3]. This way, whenever CDN as a Service (CDNaaS) may be offered on demand for the end-users. In addition, the use of NFV and SDN jointly can take advantage of an platform prepared to deal with an multi-tier edge environment.

which delivery the distributed content,

it is needed instantiate VNFs that may replace virtualization services among de nodes in defirent tiers of the network. This way

With the Needed of download high quality multimedia content by the mobility users, this section introduces an architecture multi-tier composed by heterogeneous wireless network access (fig). Due to propogation of a In a network The devices Due to the diference of range among the devices such as Cloud, ISP, LTE, WiFi, D2D.

introduces an architecture composed of a cloud computing (Tier 1) together with multi-tier fog nodes (Tiers 2, 3, and 4), which work collaboratively to enable service migration for video distribution with QoE support. In such architecture, we consider fully connected and fully fog-enabled scenario, where fog nodes are hierarchically organized to provide video services for end-users. There may be widely distributed local fog nodes, e.g., mobile devices (i.e., Tier 4), where such fog node relays the video content via device-to-device (D2D) wireless communication for mobile devices with high and similar traffic demands could cooperate with each other to form a D2D network. The neighborhood fog node, e.g., Base Station (BS) or Access Point (AP) (i.e., Tier 3), supports a few dozen to perhaps a few hundred local fog nodes. Above these would be regional fog node, e.g., baseband unit (BBU) or Internet Service Provider (ISP) (i.e., Tier 2), managing city-wide coordination. On the top of such multi-tier architecture, there is the cloud (i.e., Tier 1).

6 Contributions

7 CDN Provisioning and Releated Work

This section presents the main previous work on the employment of CDN solutions on the Cloud computing and other realted technologies.

Hatem *et al.* [1][2][3][4] highlight Software Defined Network (SDN) and Network Function Virtualization (NFV) principles into the cloud. The SDN/NFV-based approach allow to virtualization specific functions in remote servers. This way, the migrations of CDN services can be virtualized over different datacenters. Hatem et al. addresses orchestration and cache problem, its work develope an exact algorithm for deciding the optimal locations to place CDN functions. The proposed algorithm including content caching and request redirection is introduced with operatong system, network, and quality of experience constraints. Therefore, for managing the CDN is made by a centralized way, and use end-user as target to make Device-to-device communication but not explores mobility. End-users requests will redirected to an optimal edge cloud location, whithout a different multi-tier level edge devices. Llorca *et al.* [ICC'15] propose a virtual cache network deployed fully in software over a programmable distributed cloud network infrastructure that can be elastically consumed and optimized using global information about network conditions and service requirements called SDvCDN. This approuch address placement (facility location), routing (flow network) and resource allocation (network design) problems.

Zhanikeev *et al.* [ICDCSW'16] proposes an caching technology distributed in 2-tier. Where the top layer is the original copy running in large-scale storage cloud, and the bottom tier are maintained by each participant of fog. The cloud has access to a number of fog nodes distributed regionally, and each regional network edge can balance the inter-cloud traffic load by keeping a portion of popular content at each local cloud. The network edge is splitted in two kinds of caches, the first one is an in-VM storage, implemented as files on virtual disk. The in-VM caching is volatile, either have to migrate with their caches or destroy them at each population upgrade. The second knd of cache is a storage facility outside of VM but inside a given fog. The pros of this cache are two-fold. First, it can be much larger than the in-VM cache. Secondly, the contents are persistent for the population in that fog cloud. The technology presented in previous work [Computer'15] is not limited to caching and storage in general, and can work for any generic service, including Hadoop environments, sensors, etc.

Frangoudis *et al.* [1] and Frank *et al.* [2] design an architecture for telco operator, which allows the interfaces and management tools to deploy a CDN infrastructure and lease it on demand. In [2] specifically design a prototype system called NetPaaS (Network Platform as a Service). In this case PaaS and IaaS services are provisioned by telco operator. The NetPaas support virtualization and physical CDN deployments, which allow the CDN operator can be in full control of the virtual resources. Whereas, in [1] offer a business model to the content provider lease a the CDN service in a Software-as-a-Service (SaaS) manner. Thus, the telco operator is capable to be on charge of the infrastructure and the CDN service. In Addition, the telco operator also may take better decision of the resource allocation. It should be noted that both works relies on the old and well-known technique called DNS forwarding [3] for load balancing – refer to [6] for a review of the various load balancing techniques.

Claeys *et al.* [1] propose an Integer Linear Programming (ILP) formulation of multi-tenant content placement and server selection problem. The scenarioes tailored of the work was the Internet Service Providers (ISPs). The main objective is to maximize the hit ratio of cache content into the ISPs servers, thus minimizing the bandwith consuoption. to become more realistic the proposed model take into account the migrations overhead introduced by the frequent contents requested. The proactive and reactive placement strategies are studied. The proative approach mortou ter um desempenho melhor na migração de conteudo durante horas de pico e mais cache hits on the first request of popular content. However, para alcançar este desempenho é preciso ter uma forte precisão na predição de popularidade do conteudo, o que torna um complicado devido as caracteristicas do trafego VoD, que ocorre grande variação de trafego durante o tempo. To deal with these limitations, this paper proposes a hybrid cache management system. The espaços em cache são distribuidos gegraficamente para a execução da estrategia proatica,

wherever occur unexpected pattern changes of request pattern the estrategia reativa é executada simultaneamente.

Qi Qi *et al.* [I'16] propose a orhestrator framework to play offload workflows in heterogeneous Mobile Cloud Computing Environments. Which the workflow tasks are ddistributed at achieving maximal perfomance experienced by end users and minimal cost os cloud resources. The app responsible for realizar o offladof the tasks work separately, sem a possibilidade de reaproveitar as tarefas de workflows já executados. Este trabalho aborda aspectos importantes de predição de mobilidade and distribuir as tasks da melhora maneira possivel em diferentes redes utilizando o conseito de virtualização. The work [I'16] focuses an orquestração de workflows levando em consideração o custo de alocação, consumo de energia e a mobilidade do dispositivo. In CDN systems exitem caracteristicas distintas entre offload de aplicações, such as cache hit ratio which is the compartilhar conteudo em comum entre usuarios finais, the timestamp de acordo com o popularity of content.

Rosario *et al.* [1] apresenta uma arquitetura para servicos de migração ao vivo de VM da nuvem para multiniveis da fog. O cenario experimental a nuvem distribui o conteudo de video para os diferentes niveis da fog. A arquitetura é baseada no paradigma sdn para, distribuição de video com suporte a QoE. The work split the multi-tier fog in three tier in order to their cover, storage, upload and download capacity. Important aspects could be tailored to support generic content and IoT environments, besides work with both private and public clouds. A divisão da nuvem em multiniveis se dá pelas caracteristicas do dispositivos conectado a nuvem, e não por qualquer interconexão entre esses aparelhos. The paper tem como focus prover tecnologias capazes de tornar este ambiente factivel, e melhorar o provisionamento de conteudo de servicos de stream de video.

Related Work						
Paradigma	Referencia	Tipos de Conteudo	Mobilidade	tipo	Mecanismos	Problemas
MEC	[8][15][16]	videos	Not	files	cache [8], cache+transcoder [15][16]	cache placement problem, content request load assignment
Muti-cloud	[6][7][14] [17][19]	generic, video [17]	Not	files	cache[6][7][14] [19], transcoder	

Li *et al.* [1] propose an Integer Linear Programming (ILP) formulation and heuristics for the problem of per user joint video quality and network selection in a multi- access heterogeneous network.

Woo *et al.* [MOBILECLOUD'14] propose an open platform for content delivery, namely vCDN, that can support a wide range of delivery patterns. It is envisioned that (delivery) control applications would be written by service providers using the platform for representing the delivery-related requirements on distributing their specific content w.r.t scale, responsiveness, security, and other properties. Specifically, for each control application, the platform can translate it to the form of an overlay network of cloud-based edge servers so as to satisfy the

Related Work								
Attack	Apache		Apache with seven		nginx		nginx with sev	
	Success Rate	TTS	Success Rate	TTS	Success Rate	TTS	Success Rate	T
Slowloris	0.0%	∞	98.7%	0.15s	15.3%	0.00s	96.5%	0
HTTP POST	0.0%	∞	97.3%	0.14s	–	–	–	
Slowread	13.8%	1.99s	97.2%	0.11s	5.2%	1.29s	96.7%	0
Resurrected Slowloris	31.9%	1.28s	95.6%	0.58s	4.3%	0.00s	99.6%	0

Table 1: Experimental results with sevenslow, sevenmem, and the combination of sevenslow, sevenmem and sevencpu. The duration of all experiments was of at least 30 mins. We measured the Success Rate, Time-to-Service (TTS), Stable Memory and the Time to Stabilize (TT Stab.).

application-specific requirements. In doing so, the centralized controller of software-defined networking (SDN) [3] is incorporated into the overlay network manage- ment, and thus building an overlay network and configuring its route, cache, and security functions can be transparently supported at runtime. We explore the combination of NDN and SDN in order to achieve programmability in the domain of content delivery

Em [6] proposes a function that allows CDN applications to discover local caching facilities dynamically, at runtime. For simplicity, analyzes the case when each location offers two caching option: VM-based for each app location-global shared by all local apps.

Retal *et al.* [ICC'17] propõe uma plataforma de *CDN as a Service (CDNaaS)* onde os usuário podem criar um *slice* de CDN incluindo cache, transcodificador e *streamers*, em ordem de gerenciar uma quantidade de videos para seus usuários. (Aborda CDN na nuvem) Benkacem *et al.* [JSAC'18] introduce a CDNaaS platform whereby a user can create a CDN slice defined as a set of isolated distributed network of edge servers over multi-cloud domains where a edge server hosts a single VNF such as virtual cache, virtual transcoder, virtual streamer and a CDN-slice-specific coordinator for the life cycle management of the slice resources and also for managing uploaded videos and subscribers. This platform is designed to have the maximum level of flexibility for scaling out of down a CDN slice on top of different public and private Infrastructure as a Service (IaaS) such as Amazon AWS service, Microsoft Azure, Rackspace, and OpenStack-managed cloud. Furthermore, the platform employs mechanisms and algorithms that create cost-efficient QoE-aware CDN slices, involving an optimal placement taking into account the desired QoE level. Therefore, the ob- jective of this paper is to find an efficient cost of CDN slice respecting, on one hand, the CDN owner requirements in terms of QoE, and on the other hand, the cloud infrastructure and its cost.

[SmartIoT'17] propose Elastic resource allocation for video surveillance systems. The elasticity comes from an algorithm they propose to handle some emergency surveillance event (like tracking a criminal) which requires a sudden increase of computation and communication resources to make sure that all the possible images are analyzed within a reasonable timeframe. When such an emergency event happens, network bandwidth allocation is reconfigured and computing resources are reallocated (by launching new VMs in the impacted zone and balancing the workload on nodes). When experimenting in their physical testbed, they verified that data propagation round-trip time is about 5 times lower with edge nodes close to the cameras compared to the cloud. They also found that the time for launching new VMs in the emergency mode is between one and two minutes, which they claim is acceptable in such a scenario.

The integrati Traditionally, system are mainly used to decrease the energy consumption of the infrastructure [?]. However, for from the IoT perspective, this is not a key point in the utilization of the technology.

8 Data Policy

References

[1] Luiz F. Bittencourt, Márcio Moraes Lopes, Ioan Petri, Omer F. Rana *Towards Virtual Machine Migration in Fog Computing*. 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing. 2015.

[2] Luiz F. Bittencourt, Omer F. Rana, and Ioan Petri,. *Cloud Computing at the Edges*. Springer International Publishing. 2016. pp. 3–12.

[3] Byers, Charles C. and Wetterwald, Patrick. 2015. *Fog Computing Distributing Data and Intelligence for Resiliency and Scale Necessary for IoT: The Internet of Things (Ubiquity Symposium)*. Ubiquity 2015, November, Article 4 (November 2015), 12 pages.

[4] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli *Fog computing and its role in the internet of things*. In workshop on Mobile cloud computing. ACM, 2012.

[5] I. Stojmenovic *Fog computing: A cloud to the ground support for smart things and machine-to-machine networks*. In ATNAC. IEEE, 2014.

[6] I. Stojmenovic and S. Wen *The fog computing paradigm: Scenarios and security issues*. In Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on, 2014.

[7] D. Evans 2005. *The Internet of Things: How the Next Evolution of the Internet Is Changing Everything*[online]. Disponível em: `http://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf`

[8] Shanhe Yi, Zhengrui Qin, and Qun Li. *Security and Privacy Issues of Fog Computing: A Survey*. The 10th International Conference on Wireless Algorithms, Systems, and Applications (WASA), 2015.

9 Attach CVs (maximum two pages for CVs)