Annals of Telecommunications (2018) 73:165–167
https://doi.org/10.1007/s12243-018-0630-y

EDITORIAL

Cloud communications and networking

Harry Perros[1] · Ioannis Papapanagiotou[2] · Guy Pujolle[3]

Special issue: cloud communications and networking

This special issue consists of eight papers and is dedicated to the topic of Cloud Communication and Networking. The authors of six of these papers were invited to submit an extended version of their paper that was presented either in (a) Smart Cloud Networks and Systems (SCNS 2016), or in (b) IEEE International Conference on Communications (ICC) - Cloud Communications and Networking Symposium, 2017. An additional paper was accepted through the open call, and we are also pleased to include an invited paper that provides a survey of the area.

Cloud Communications and Networking is a multi-faceted research area, and it was not possible to have representative papers from all relevant topics. Excluding the survey paper, the remaining papers of this special issue fall into the following general areas: scheduling of resources, wireless access and transient clouds, measurement tools, and TCP performance in data centers. Below, we briefly describe the papers in the order in which they appear.

1. "A Survey on Cloud Communications and Networking: State of the Art, Challenges and Opportunities", by G. P. Xavier and B. Kantarci.

The authors review the architectural challenges and solutions in today's leading cloud communication technologies, namely, network virtualization, Software-Defined Networking (SDN), Network Function Virtualization (NFV), and SDN-enabled NFV solutions. In addition, they review the Cloud-RAN architecture for radio access networks, along with its support for various existing and future wireless communication technologies including future 5G wireless networks. Furthermore, for all cloud communication concepts, the authors present a thorough discussion on the open issues and opportunities.

Scheduling of resources

2. "AutoSAC: Automatic Scaling and Admission Control of Forwarding Graphs", V. Millnert, E. Bini, J. Eker.

The authors develop a model of a service chain of network functions and use it to derive a service controller and an admission controller for the network functions, referred to as the Automatic Service and Admission Controller (AutoSAC). The service controller determines the number of virtual resources (e.g., virtual machines or containers) allocated to each network function using feedback information from instantiated network functions as well as feedforward information between the network functions. The admission controller is aware of the actions of the service controller and determines how many packets to reject. AutoSAC was evaluated using a real-world traffic trace from the Swedish University Network (SUNET).

3. "Dynamic VM allocation in a SaaS environment", B. Bouterse and H. Perros.

A number of forecasting models for predicting demand for virtual machines in a cloud-based software used as a SaaS are developed. These models are then used in a periodic-review provision model which determines how many virtual machines should be provisioned or de-provisioned at each inspection interval. A simple provisioning heuristic model is also proposed, whereby a fixed reserve capacity of virtual machines is continuously maintained. The performance of

✉ Harry Perros
hp@ncsu.edu

[1] NC State University, Raleigh, NC, USA

[2] Netflix, Los Gatos, CA, USA

[3] UPMC, Paris, France

these models is evaluated and compared with different model parameters using historical data from the Virtual Computing Laboratory (VCL) at North Carolina State University.

4. “Constrained Max-Min Fair Scheduling of Variable-Length Packet-Flows to Multiple Servers”, J. Khamse-Ashari, G. Kesidis, I. Lambadaris, B. Urgaonkar, and Y. Zhao.

A multi-server queueing system is studied wherein each user is constrained to get service only from a specified subset of servers. Fair packet scheduling in such a setting poses novel challenges. Specifically, the authors observed that max-min fair allocation of the available resource over different servers (notably bandwidth) in the presence of placement constraints results in different levels of fair service rates. To achieve max-min fair service rates, a novel packet scheduler is proposed which is inspired by the deficit-round robin (DRR) algorithm. The scheduler allocates tokens to flows in a round-by-round manner, where token allocation to flows at the beginning of each round is weighted max-min fair. The performance of the scheduler in terms of achieving fairness is shown through a worst-case performance analysis. In addition, numerical experiments were also carried out to illustrate that the scheduler provides service isolation and delay guarantees. The scheduler can be used in a cloud computing environment where virtual machines vie for different IT resources distributed over heterogeneous servers.

Wireless access and mobile clouds

5. “SDN-based Wi-Fi Direct Clustering for Cloud Access in Campus Networks”, T. M. Trang Nguen, L Hamidouche, F. Mathieu, S. Monnet, S. Iskoumen.

The mobile cloud paradigm is changing the way that teaching activities in a university campus are handled. Lectures and lab sessions can be carried out directly from tablets in a classroom by accessing a server in the cloud. In this paper, the authors address the problem of high-density cloud access by wireless devices in a campus network. They propose to use Wi-Fi Direct clustering to solve the problem of QoS degradation when a high number of wireless devices want to access a content in the cloud at the same time. A centralized software-defined network controller is used in the proposed architecture to capture the network state and organize the Wi-Fi Direct groups. The optimized number of clusters can be calculated as a function of the number of devices in the room. Through simulation, it is shown that the proposed solution provides a better QoS in terms of download time and applications’ throughput.

6. “Assignment and Collaborative Execution of Tasks on Transient Clouds”, M. Guiguis, Q. Gu, T. Penner, L. Tammineni, T. Langford, A. Rivera-Longoria, A. Johnson, B. van Slyke.

A Transient Cloud (TC) is a temporal cloud that enables nearby mobile devices to form an ad-hoc network and advertise their capabilities as cloud services. Through utilizing the collective power of the group, devices are no longer constrained by their local hardware and software capabilities. TC harnesses the ubiquitous nature of mobile devices along with their ever-increasing sets of capabilities in providing a rich computing platform. The authors present two instantiations of task assignment algorithms that achieve various goals, such as, balancing the load on devices and minimizing the cost of communication. In the first instantiation, the authors consider a centralized approach in which a cluster head is responsible for maintaining the list of capabilities and assigning tasks to devices based on their capabilities. The authors present a modified version of the Hungarian method that allows for balancing the load on devices. In the second instantiation, the authors consider a distributed approach in which devices advertise and find capabilities through an overlay network. The overlay network is designed to capitalize on locality and thus seeks to minimize the cost of finding devices with certain capabilities. The performance of the TC is evaluated through extensive simulation.

TCP performance

7. “Mitigating Incast-TCP Congestion in Data Centers with SDN”, A. M. Abdelmoniem, B. Bensaou, and A. J. Abu.

In a data center network (DCN), the presence of long-lived TCP flows tends to bloat the switch buffers. As a consequence, short-lived TCP-incast traffic suffers repeated losses that often lead to loss recovery via timeout. Because the minimum retransmission timeout (minRTO) in most TCP implementations is fixed to around 200 ms, interactive applications that often generate short-lived incast traffic tend to suffer unnecessarily long delays waiting for the timeout to elapse. The best and most direct solution to this problem would be to customize the minRTO to match DCN delays. However, this is not always possible, since in public data centers there are multiple tenants running different versions of TCP. In this paper, the authors propose to achieve the same result by using techniques and technologies that are already available in most commodity switches and data centers, and that do not interfere with the tenant’s virtual machines or TCP protocol. They rely on the programmable nature of SDN switches to design an SDN-based framework that uses an SDN network application in the controller and a shim-layer in the host hypervisor to mitigate incast congestion. The performance gains of the proposed scheme are demonstrated via a real deployment in a small-scale testbed as well as ns2 simulation experiments in networks of various sizes and settings.

Measurement tools

8. "COMIQUAL: Collaborative Measurement of Internet Quality", M. Ibrahim1, M. Chamoun, R. Kilany, M. El Helou, N. Rouhana.

With the continuous growth of both fixed and mobile Internet usage, measuring the Internet QoS (Quality of Service) becomes of vital interest for all involved Internet stakeholders, mainly consumers, operators, and regulators. In this paper, the authors describe in detail, COMIQUAL (COllaborative Measurement of Internet QUALity), a crowd-sourced large-scale Internet measurement platform that coordinates and collects measurements from measurements agents (MAs) installed on fixed and mobile end-user devices. Although the initial and main target of COMIQUAL is Lebanon, the platform is generically designed to measure the Internet access quality from the user's perspective anywhere on the globe. The MAs that execute mainly active measurements are jointly controlled by users and by a measurement center (MC); the latter sends measurement instructions to MAs and collects the measurement results. The communication protocol between MC and MAs uses JSON messages that are exchanged via HTTP through REST calls, and secured by HTTPS. Measurement results could be openly accessed in a raw format or viewed as an aggregation via a Google map. Moreover, an online statistical tool allows user-defined statistics computation and visualization. All these features, combined with the flexibility of the platform management, are the main drivers that will allow COMIQUAL to reach its ultimate goal, which is to create a collaborative, neutral, and transparent observatory of the Internet.

Guest Editors

- Harry Perros, NC State University, USA
- Ioannis Papapanagiotou, Netflix, USA
- Guy Pujolle, UPMC, France