Connectivity Requirements for Cloud-Based Services

A. Asensio*, M. Ruiz, and L. Velasco
Optical Communications Group (GCO), Universitat Politècnica de Catalunya (UPC), Barcelona, Spain
**e-mail: aasensio@ac.upc.edu*

ABSTRACT
The expected explosion of services offered in the cloud has revealed new paradigms in telecommunication networks and motivated research to satisfy the needs arising. Interconnection between facilities hosting computing resources or between computing resources and end-users may require service-specific Service Level Agreement (SLA) parameters. In this paper, we study the minimum connectivity requirements that transport networks need to satisfy to support cloud services and we map relevant Key Performance Indicators (KPIs) to them. Three clearly differentiated cloud-based scenarios are studied: datacenter (DC) interconnection in DC federations, live-TV distribution, and Cloud-Radio Access Networks (C-RAN) to support next generation mobile networks.
Keywords: cloud services, cloud radio access network, video distribution, telecom cloud.

1. INTRODUCTION

The expected explosion of services offered in the cloud has revealed new paradigms in telecommunication networks. In a recent study, in [1], Cisco forecast that traffic between datacenters (DCs) will reach 905 exabytes (EB) per year by 2019. Moreover, according to [2], 80% of the IP traffic will correspond to video traffic by 2019 and it is forecast that mobile data traffic will reach about 367 EBs per year by 2020. Aiming at interconnecting DCs and services in DCs and end-users, connectivity from the transport network is required.

From the network perspective, transport networks are currently configured with *big static fat pipes* based on capacity over-provisioning aiming at guaranteeing traffic demand and Quality of Service (QoS). *Cloud-ready transport network* [3] was introduced as an architecture to handle the dynamic cloud and network interaction. Moreover, to deal with applications requests for end-to-end (e2e) connectivity provisioning, the IETF has standardized the Application Based Network Operations (ABNO) [4]. ABNO's northbound interface can accept connection requests from the service-layer; thus, facilitating dynamic connection requests. Notwithstanding, the evolution towards cloud-ready transport networks is based not only on intelligent control architectures (e.g. ABNO-based architectures) but also on elastic data planes that can satisfy cloud requirements efficiently for both network and cloud operators. Hence, elastic optical networks (EONs) play an important role to support cloud services.

In this paper, we focus on the study of a number of use cases related to cloud services that require connectivity from the transport network. Specifically, three use cases are studied: *i*) DC interconnection in DC federations, *ii*) live-TV distribution on the telecom cloud, and *iii*) Cloud-Radio Access Networks (C-RAN) to support next generation mobile networks. From the results obtained in our previous works, we identify a set of minimum connectivity requirements that those representative services need and map them to Key Performance Indicators (KPIs).

2. CLOUD-BASED SERVICES

The first use case, related to DC interconnection, tackles the problem of data migration among geographically disperse DCs. By placing DCs in geographically diverse locations, cloud operators can move workloads among DCs aiming at optimizing some utility function, e.g. energy expenditures minimization, while ensuring good Quality of Experience (QoE) to end-users accessing to the services that they host. Medium-sized DC operators can cooperate by creating DC federations [5].

In DC federations, scheduling algorithms in cloud management run periodically taking decisions in advance on where to place workloads. Algorithms based on *follow-the-work* or *follow-the-sun* approaches have been attracting the interest of the research community. However, to carry out the computed workload placement, cloud management requires the collaboration from inter- and intra- DC networks. Coordination between the cloud and the network is required to finish data migrations within the required time. Fig. 1a represents a scenario where two federated DCs require connectivity from the transport network.

The second use case is related to video distribution. As introduced in the previous section, video traffic will represent a remarkable fraction of the total IP traffic. Moreover, video distribution is one of the stringent services that telecom networks need to support. Focusing on live-TV distribution services, uncompressed video streams coming from sources capturing live events are used before the video is produced. Once the video has been produced, individual flows with compressed video streams are sent to the users consuming that service. The quality of each compressed video stream corresponds to the one that fits better the user's device; i.e. standard Definition (SD), High Definition (HD) or Ultra-High Definition (UHD). Fig. 1b illustrates an example

representing the flow of an uncompressed UHD video stream and compressed video streams of different qualities.

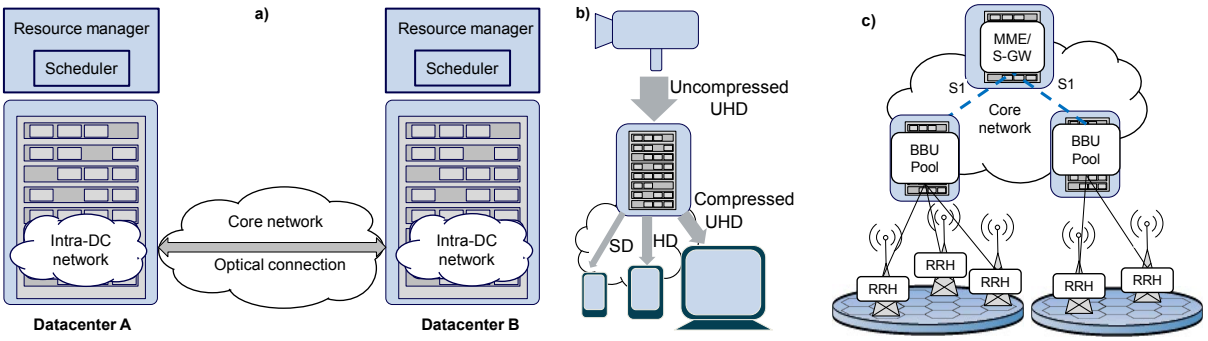

Figure 1: (a) DC interconnection, (b) live-TV distribution, and (c) C-RAN.

However, the evolution of video technologies towards UHD formats and the expected growth of end-users' devices consuming video services present some drawbacks when environments based on traditional Serial Digital Interfaces (SDI) are considered, since SDI capacity is limited to HD formats. Therefore, live-TV distribution requires from technologies different than SDI to support such required capacities, e.g. IP-based technologies. In line with this, the authors in [6] reported 4K UHD TV video streaming over an IP network thus enabling the migration from traditional Serial Digital Interfaces (SDI) -based transmission to all-IP environments.

Finally, the third use case is related to C-RAN, which is on the roadmap of certain mobile operators aiming at fulfilling part of the next generation mobile networks, e.g. 5G, requirements in a cost effective manner [7]. In C-RAN, Remote Radio Heads (RRHs), i.e. radio frequency processing hardware, are geographically distributed to cover certain areas; whereas Base Band Units (BBUs), i.e. baseband processing hardware, is centralized in BBU pools and can be shared among different sites along the time. In fact, C-RAN pooling gains have been demonstrated under certain cell's traffic assumptions [7]. Moreover, some of those geographically disperse RRHs, can be used during peak hours for offloading purposes, whereas they can remain unused during low demand periods, e.g. during night hours.

Focusing on the transport network, in the Long Term Evolution (LTE) and the LTE-Advanced (LTE-A) architectures, connections from BBUs to the corresponding mobile packet core elements, e.g. the mobility management entity (MME) or the serving gateway (S-GW), are needed to support the S1 interface. Therefore, considering LTE or LTE-A technologies, in C-RAN connections supporting S1 interfaces are required between the BBU pools, usually placed in the access or metro segment, and the core elements, i.e. MME and S-GW. Fig. 1c shows an example of C-RAN, where a set of RRHs cover different areas and are served from centralized BBU pools. For illustrative purposes, connections to support S1 interfaces are depicted.

3. CONNECTIVITY REQUIREMENTS AND KPI MAPPING

To study connectivity requirements for DC interconnection, we focus on the results obtained in our previous works [8] and [9]. In [8] we dimensioned a set of federated DCs according to realistic values. The amount of data to migrate was related to two main sources: *i*) workloads encapsulated in virtual machines (VMs) and *ii*) data for synchronizing databases (DBs) among federated DCs. In addition, based on a follow-the-work approach, we dimensioned static connections' capacity among DCs trying to reduce the over-provisioned bitrate while ensuring that data migrations finish in less than one hour in any case, since scheduling algorithms were set to run each hour. As a result, static connections of 150 Gb/s and 200 Gb/s were required for VM migration and DB synchronization, respectively. Figure 2a represents the *time-to-transfer* for VM migration and DB synchronization between two distant DCs. Clearly, data transferences finish in less than one hour at any time, being the peak time-to-transfer values above 50 minutes. Obviously, to finish data transferences in shorter times, higher capacities are required.

From the connections' capacities values shown, the first connectivity requirement is identified: *huge capacity*. However, even dimensioning connections to reduce the over-provisioned bitrate, static connections result in connections' bitrate underutilization. To illustrate such underutilization, especially during the hours where few data needs to be transferred, Fig. 2b shows the use of static connections between two DCs for VM migration and DB synchronization along the day. Therefore, aiming at minimizing the high costs that static connections entail due to their over-dimensioning and underutilization, dynamic connectivity can be considered. Results in [8] showed bitrate savings close to 60% when dynamic connections were considered against static ones. Moreover, we showed in [9] that communication costs need to be taken into account in DC federations. Impact of the

communications costs contributed to motivate dynamic connectivity. Dynamic connectivity can then be identified with the following connectivity requirement: *bandwidth-on-demand*.

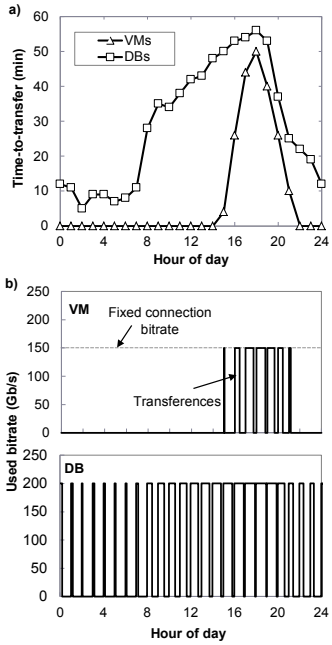

Figure 2. Time-to-transfer (a) and used bitrate (b) against hour of the day.

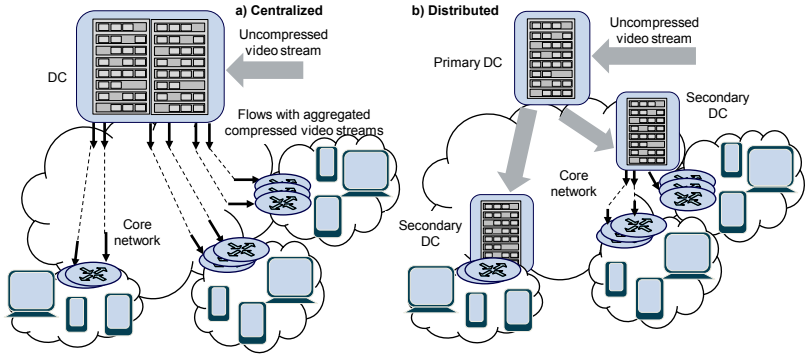

Figure 3. Live-TV distribution approaches: centralized (a) and distributed (b).

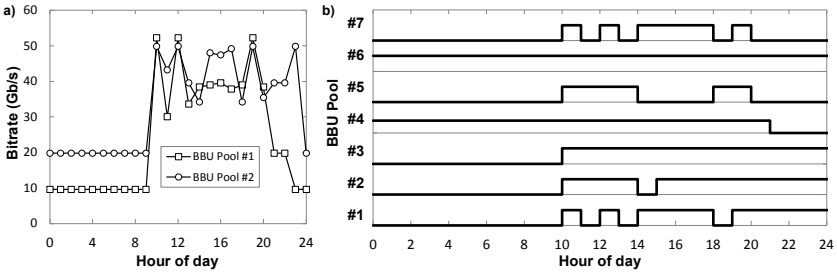

Figure 4. S1 capacity (a) and BBU pools requiring connectivity (b) against the hour of the day.

Regarding live-TV distribution, in [10] we compared a centralized approach against a distributed one, both requiring connectivity from the transport network. In the centralized approach (Fig. 3a), an uncompressed video stream is processed in a single DC and flows with aggregated compressed video streams are conveyed through the core network to metro areas. Differently, in the distributed approach (Fig. 3b), uncompressed video streams are conveyed through the core network from a primary DC to a set of secondary DCs, where they are processed. The resulting aggregated flows with compressed video streams are distributed to the users directly through the metro network when possible or using the core network when it is required.

Although network CAPEX savings as high as 32% were observed when the distributed approach was considered, common connectivity requirements can be found for both approaches. Flows transporting uncompressed and aggregated compressed video streams require huge capacities; e.g. in the realistic scenarios we tackled, flows of 100 Gb/s were considered to transport uncompressed video streams corresponding to 8 TV channels requiring 12 Gb/s each or aggregated compressed video streams according to different scenarios. However, larger values can be considered depending on several factors such as the aggregation level in flows conveying compressed video streams or the capacity required by uncompressed UHD video streams, which is close to 72 Gb/s for the 8K UHD format [11]. Therefore, *huge capacity* is required. In addition, the capacity of the connections supporting aggregated compressed video streams depends on the time, since live-TV consumption strongly depends on the hour of the day, presenting peaks at evening and certain night hours (*prime time*) and off-peak periods during office hours. Such connections may require *bandwidth-on-demand*. In addition to the above mentioned connectivity requirements, since live-TV services are delay sensitive (e.g. jitter highly impacts the service quality), strict QoS requirements, in terms of *delay*, need to be considered.

Next, to study C-RAN requirements we focus on the results obtained in our previous work in [12]. Considering the S1 interface (since it requires connectivity from the transport network), Fig. 4a represents the required capacity between two BBU pools and the corresponding location hosting the packet core functions (i.e. MME/S-GW) against the hour of the day when S1 capacity per each site is about 600 Mb/s. In addition, results in [12] showed scenarios, for S1 capacities about 1.5 Gb/s, where certain BBU pools are not used during low traffic hours; therefore, no connections are need between those BBU pools and the core elements during certain periods of time. From the results showed in [12], Fig. 4b represents, for a set of BBU pools, if they are in use or not against the hour of the day. Due to the bitrate required along the time and the variability on the connectivity required between BBU pools and the core elements, the following requirements are devised *huge capacity* and *bandwidth-on-demand*.

However, service-specific requirements to support 5G mobile networks may arise since 5G needs to support services targeting at their particular performance goals. Therefore, 5G performance goals and services supported in the mobile network may lead to strict *delay* constraints in connections supporting S1 interfaces. In addition, an

eventual failure of a link supporting a connection transporting S1 traffic from a BBU pool would impact several sites simultaneously. Considering *bitrate guarantees* by means of diversity in those connections can avoid mobile service interruption in a cost effective manner from the network provider side.

Finally, for the sake of completeness, a set of relevant KPIs that result helpful to quantify connectivity requirements' fulfilment are assigned to the different connectivity requirements described in this section. To quantify connectivity requirements related to *huge capacity* and *bandwidth-on-demand*, the unserved bitrate and the blocking probability can be considered. Regarding stringent QoS, in terms of *delay*, connection's e2e delay is a good indicator. The ratio between the bitrate lost and the required bitrate in the event of a failure of a link in the route of a connection can be considered to evaluate performance in the case that bitrate guarantees are required. Table 1 summarizes the use cases, their connectivity requirements and KPIs.

Table 1. Minimum connectivity requirements and their KPIs.

Use case	Requirements	KPIs
DC interconnection	Huge capacity Bandwidth-on-demand	Unserved bitrate and blocking probability
Live-TV	Huge capacity Bandwidth-on-demand Stringent QoS (delay)	Unserved bitrate and blocking probability e2e delay
C-RAN	Huge capacity Bandwidth-on-demand Stringent QoS (delay) Bitrate guarantees	Unserved bitrate and blocking probability e2e delay Bitrate lost/Required bitrate

4. CONCLUSIONS

In this paper, we identify the minimum connectivity requirements that transport networks need to satisfy to support cloud services, through the study of three clearly differentiated cloud-based scenarios: *i*) DC interconnection in DC federations, *ii*) live-TV distribution, and *iii*) C-RAN to support next generation mobile networks.

Specifically, huge capacity and bandwidth-on-demand have been identified as connectivity requirements in all three use cases. Stringent QoS, in terms of delay, has been described as a connectivity requirement to satisfy in live-TV distribution and C-RAN. In addition, C-RAN may require bitrate guarantees to avoid service interruption simultaneously in several sites, which can impact negatively to a wide range of users and services in mobile networks.

Finally, for the sake of completeness, a set of relevant KPIs that result helpful to quantify connectivity requirements' fulfilment have been identified with the different connectivity requirements.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the Spanish MINECO SYNERGY project (TEC2014-59995-R) and from the Catalan Institution for Research and Advanced Studies (ICREA).

REFERENCES

[1] CISCO Global Cloud Index (GCI), 2015.
[2] CISCO Visual Networking Index (VNI), 2016.
[3] L. M. Contreras, V. López, O. González, A. Tovar, F. Muñoz, A. Azañón, J.P. Fernández-Palacios, and J. Folgueira, "Towards cloud-ready transport networks," *IEEE Communications Magazine*, vol. 50, pp. 48-55, 2012.
[4] D. King and A. Farrel, "A PCE-based architecture for application-based network operations," IETF RFC7491, 2015.
[5] I. Goiri, J. Guitart, and J. Torres, "Characterizing cloud federation for enhancing providers' profit," in Proc. IEEE International Conference on Cloud Computing (CLOUD), 2010.
[6] F. Poulin, T. Kernen, and A. Kouadio, "Ultra high definition TV over IP networks," EBU Technical Review, 2014
[7] "C-RAN the road towards green RAN," China Mobile Research, 2011.
[8] L. Velasco, A. Asensio, J.Ll. Berral, V. López, D. Carrera, A. Castro, and J.P. Fernández-Palacios, "Cross-stratum orchestration and flexgrid optical networks for datacenter federations," *IEEE Network Magazine*, vol. 27, pp. 23-30, 2013.
[9] L. Velasco, A. Asensio, J. Ll. Berral, E. Bonetto, F. Musumeci, V. López, "Elastic operations in federated datacenters for performance and cost optimization," *Elsevier Computer Communications*, vol. 50, pp. 142-151, 2014.
[10] A. Asensio, L. M. Contreras, M. Ruiz, V. Lopez, and L. Velasco, "Scalability of telecom cloud architectures for live-TV distribution," in *Proc. IEEE/OSA Optical Fiber Communication Conference (OFC)*, 2015.
[11] S. Namiki, T. Kurosu, K. Tanizawa, J. Kurumida, T. Hasama, H. Ishikawa, T. Nakatogawa, M. Nakamura, and K. Oyamada, "Ultrahigh-definition video transmission and extremely green optical networks for future", *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 17, no. 2, pp. 446,457, 2011.
[12] A. Asensio, P. Saengudomlert, M. Ruiz, and L. Velasco, "Study of the centralization level of optical network-supported cloud RAN," accepted in *IEEE International Conference on Optical Network Design and Modeling* (*ONDM*), 2016.