An Overview of Application Traffic Management Approaches: Challenges and Potential Extensions

Ioanna Papafili[1], Krzysztof Wajda[2], Roman Łapacz[3],
Alessandro Predieri[4], Thomas Bocek[5], Michael Seufert[6]
[1] Department of Informatics, Athens University of Economics and Business, Greece
[2] AGH University of Science and Technology, Poland
[3] Instytut Chemii Bioorganiicznej PAN, Poland
[4] Interroute S.P.A., Italy
[5] Department of Informatics, University of Zurich, Switzerland
[6] Julius-Maximilians Universität Würzburg, Germany

***Abstract*— The Internet has seen a strong move to support overlay applications and services, which demand a coherent and integrated control in underlying heterogeneous networks in a scalable, resilient, and energy-efficient manner. To do so, a tighter integration of network management and overlay service functionality can lead to cross-layer optimization of operations and management, which is a promising approach and may offer a large business potential in operational perspectives for all players involved.**

Therefore, the objective of this paper is to present and discuss the impact of new paradigms such as cloud computing and software-defined networking which will play central role in the Future Internet, discuss major traffic trends and identify key challenges due to the adoption and operation of the new applications. Translating the key challenges to requirements for Future Internet traffic management mechanisms, the paper provides an overview of existing mechanisms in literature, assesses them w.r.t. the aforementioned requirements, and qualitatively estimates the expected optimization potential and gain, as well as provides hint for their potential extension and exploitation within the challenging environment of the Future Internet.

Keywords—cloud computing; global service mobility; traffic management; QoS/QoE; cost; economic awareness; social awareness; energy efficiency.

I. INTRODUCTION

There are important new services and applications influencing volume and patterns of Internet traffic and the satisfaction of the end-users [1]. These include cloud computing and applications served, such as video streaming platforms (e.g. YouTube, Vimeo), social networks (e.g. Facebook, Twitter), online storage systems (e.g. Dropbox, Google Drive), etc.

Moreover, the Future Internet's entertainment is foreseen to generate more traffic to come [2], while simultaneously, the mobile services and respective wireless access network demand is increasing as well [3], resulting in very different communication path quality levels and management tasks. In turn, traffic from such overlay applications on a very dramatically increasing number of devices and end-points (wired and mobile) is continuously and significantly increasing.

In this paper, we identify major key challenges of the Future Internet w.r.t. to the efficient management of traffic generated by new cloud services and applications mentioned above; where efficiency is considered in terms of cost, energy and operation. These key challenges set the requirements for new traffic management mechanisms that will sufficiently address them.

In order to perform efficient management of traffic generated by overlay applications such as video streaming, P2P file sharing and P2P video streaming, CDNs, VoIP, etc. a multitude of mechanisms has been proposed in literature. In this paper, we overview and analyze some categories of traffic management approaches in recent literature, which may already address partly or could be extended to fully address traffic generated by cloud services and applications w.r.t. cost, Quality of Service (QoS) / Quality of Experience (QoE) and energy requirements. In particular, we investigate how these approaches can effectively intervene in the default system behavior, and by which involved stakeholder they can be applied. Finally, we qualitatively assess their sufficiency and identify potential need for further elaboration.

The remainder of this paper is organized as follows: in Section II, we provide a brief description of cloud computing, its key features, and significant new paradigms stemming from it, i.e. Software-Defined Networking (SDN) and Network Function Virtualization (NFV). Then, in Section III, we identify cloud traffic trends and key challenges of the Future Internet in terms of traffic control and management, cost and energy consumption. In Section IV, which is central in our contribution, we provide an extensive overview of a multitude of traffic management mechanisms that may constitute partly sufficient solutions to the aforementioned challenges, we qualitatively assess these mechanisms w.r.t. to their expected efficiency, and provide hints on their potential extension and exploitation to address the traffic demand in the Future Internet. Finally, Section V concludes our contribution.

II. CLOUD COMPUTING

Cloud computing refers to the delivery of computing resources over the Internet. Most common examples of cloud services include online file storage, social networking sites, webmail, and online business applications. The key factor of the success of the cloud computing paradigm is the fact that cloud computing forecast a model that allows access to information and computer resources from anywhere that a network connection is available. The following definition of cloud computing has been developed by the U.S. National Institute of Standards and Technology (NIST) [5]:

"Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics, three service models, and four deployment models."

A key feature of cloud computing is the transparency property, e.g. the fact that when end-user purchases a cloud service, he is not aware of what synergies in terms of cooperation among all actors composing the cloud ecosystem such as Internet Service Providers (ISPs), Cloud Application Providers, Content Providers, or Cloud/Datacenter Operators, are necessary to compose the final service. This greatly enhances the customer QoE, but on the other hand, it imposes several limitations and requirements in the process that drives the formation of relationships among all cloud ecosystem actors both from a technical and a business point of view.

Apart from benefits, cloud computing is accompanied by certain serious concerns. The main one seems to be security. Potential cloud service customers are still reluctant and mistrustful to store their data externally. Other emerging concerns are as follow: business continuity when the customer wants to change the cloud provider, QoS or lack of standardized cloud computing solutions [6].

Existing concerns and identified challenges are being investigated by the researches all over the world. In the context of networking cloud computing is being analyzed in conjunction with other promising paradigms and technologies, like SDN [7] and NFV [8]. Well known cloud models – IaaS, PaaS and SaaS – are being followed by more advanced solutions like Inter-cloud. This federation approach, mobile nature of data and services nowadays and the big data [9] trend have a crucial influence over network traffic characteristics.

III. CLOUD TRAFFIC TRENDS

The scope of this section is to overview traffic trends recognized when analyzing implementation and current usage of clouds and to formulate motivation for possible optimization approaches related to enhancements in clouds proliferation.

Identification of traffic characteristics, mainly referring to generated traffic volumes, as well as traffic patterns, manifesting themselves by e.g., variability and burstiness, and QoE requirements, where available, is an important task done permanently by network operators and administrators. Besides theoretical aspects of traffic analysis, carried in order to study network-related phenomena and to follow evolution of services and user behavior, such analysis helps to solve urgent problems (such as failures or lack of resources) and support traffic management and long-term network planning. There are long-lasting efforts to characterize traffic in the Internet. These investigations profit from previous and ongoing efforts to characterize traffic in the network. The goal is to recognize traffic flows, characteristics of single flows and also characteristics of multiplexed traffic composed of many flows.

When analyzing evolution of clouds, their impact on market trends and business we are facing core concept and technologies being at the boundary between professionally offered services and end-users, i.e., consumers of such services.

Publicly available statistics about traffic in IP networks [1], [2], [3], unveil important trends in popularity of clouds and cloud-based services in recent years. Clouds have replaced recently the overlay networks by importance, popularity and market share. Clouds differ from P2P since they have legal character and also they profit from availability of rather sufficient network resources (bandwidth, router power) due to qualitative and quantitative enhancements in IT environment. However, even profiting from usage of sufficient resources do not preclude ISPs and Cloud Operators from optimization of traffic management which should lead to energy efficiency and increase of perceived QoE. Cloud traffic trends analyzed and forecasted until 2016 are presented in Table I, reaching 44% of Compound Annual Growth Rate[1] (CAGR).

TABLE I. CLOUD IP TRAFFIC TRENDS 2011-2016 [1].

	2011	2012	2013	2014	2015	2016	CAGR 2011-2016
By segment (EB per year)							
Consumer	559	992	1,426	1,960	2,692	3,659	**46 %**
Business	124	189	268	364	474	596	**37 %**
Total cloud traffic(EB per year)							
	683	1,181	1,694	2,324	3,166	4,255	**44 %**

Sandvine's reports on Internet traffic trends [4] are identifying main phenomena, short- and long term trends and they give wide vision of applications' popularity, traffic growths and market specificity for US, Europe and other markets.

[1] Compound annual growth rate represents growth over a period of years, with each year's growth added to the original value. It is calculated by taking the nth root of the total percentage growth rate, where n is the number of years in the period being considered.

The key elements of clouds are data centers and their efficiency and performance is crucial for the successful provisioning of cloud-based services. Traffic in data centers grows faster that traffic reported for the whole Internet since significant amount of traffic is exchanged internally in data centers. From Table II it can be seen that by 2016 traffic in data centers will quadruple to reach 6.6 ZBs per year. This translates into impressive value of a 31% CAGR.

TABLE II. FORECASTED GLOBAL DATA CENTER TRAFFIC GROWING RATE 2011-2016 IN ZETTABYTES/YEAR [1].

2011	2012	2013	2014	2015	2016	CAGR 2011-2016
1,8 ZB	2,6 ZB	3,3 ZB	4,1 ZB	5,2 ZB	6,6 ZB	**31 %**

In order to analyze the data center traffic with more insight into factors influencing performance of clouds, it is necessary to split traffic into three parts:

- traffic exchanged among different datacenters,
- traffic sent inside data centers,
- traffic sent from data centers to users.

Only the first part of traffic, accounting for about 7 % of total traffic is subject to traffic management mechanisms which can be employed by ISPs and Cloud Operators but in 2016 this is forecasted to reach traffic volume of 468 EB, with stable CAGR 32 %.

Traffic volumes sent inside clouds seen in traffic measurements and analysis justify the motivation for further inspection of cloud traffic and designing control mechanisms for managing traffic flows in order to minimize long distance traffic and inter-cloud transport, whatever the nature of cloud organization. The growth rate of traffic from cloud-based data centers (Table III), with CAGR equal to 44 %, gives the indication of potential for optimization of this traffic.

TABLE III. FORECASTED DATA CENTER TRAFFIC TRENDS FOR THE PERIOD 2011-2016 [1].

	2011	2012	2013	2014	2015	2016	CAGR 2011-2016
By type (EB per year)							
Data center to user	299	438	561	714	912	1,160	**31 %**
Data center to data center	118	173	222	284	365	468	**32 %**
Within data center	1,338	1,940	2,468	3,126	3,969	5,021	**30 %**
By type (EB per year)							
Traditional data center	1,072	1,370	1,557	1,800	2,080	2,394	**17 %**
Cloud data center	683	1,181	1,694	2,324	3,166	4,255	**44 %**

IV. FUTURE CHALLENGES

These ongoing trends in the cloud applications domain pose several major challenges for the current structure of the Internet, both from a technical and economic point of view:

The core characteristic of cloud applications is that they can be reachable from any place. Hence, traffic generated by cloud applications will have unpredictable patterns, crossing many and different domains (from time to time), due to the mobility of cloud users or the distribution of the cloud resources. Furthermore, the synchronization of data between multiple data centers providing the same application causes large amounts of traffic, both on intra and inter-domain links of several ISPs.

Additionally, emerging cloud applications such as video streaming platforms or Online Social Networks (OSNs) make traffic management difficult for ISPs, due to incomplete information available to ISPs about the operation of the cloud. As a result, ISPs may employ traffic management techniques in order to reduce their costs (e.g. transit) that may deteriorate the QoE perceived by the end-users of cloud applications.

On the other hand, the Cloud or Datacenter Operator, the Cloud Application Provider or the Content Delivery Network (CDN) Provider make decisions aiming to optimize their own operation, ignoring the impact of their decisions in the underlay, i.e., physical network of the ISP. Thus, two major challenges are: ***bridging the information gap between the cloud layer and the network layer***, as well as ***pursuing incentive compatibility*** so as to re-align the behavior of both the physical network and the cloud towards collaboration, and ultimately, a *win-win* situation.

Moreover, in order to achieve incentive compatibility, cost efficiency needs to be pursued for all involved players, e.g. Cloud Operators and ISPs. Cost efficiency though is closely related to energy efficiency as energy involves cost. While energy efficiency has strong roots in the embedded devices and wireless (sensor [10]) networks [11], energy efficiency is becoming important in clouds and data centers. Data centers typically measure their energy usage in power usage effectiveness (PUE). To gain some insight, while a PUE of 1.0 is ideal, Google claims that their data centers have a PUE of 1.12 [12]. However, this only shows how much energy was consumed by the equipment, but it does not show how efficient the equipment is.

Besides energy efficiency in data centers considering CPU and storage, the power consumption in the network of an ISP is also an important factor. The authors in [13] present that energy consumption in the network takes a significant proportion of the total power consumption for cloud storage services. They argue that in special situations, e.g., if files stored online are accessed frequently, the energy consumption is higher than accessing the file from the local disk due to network power consumption. Thus, as a conclusion, ***energy efficiency*** of both cloud and network is another significant challenge.

Furthermore, due to the rapidly increasing number of mobile devices, one can claim that in the future, a service should "follow" its receiver. For example, a user uses a service on a notebook or tablet and wants to receive high QoE regardless where he/she connects to the networks: at home in Europe, in the train, airport, or hotel. It raises a challenge, the so-called ***global service mobility***, which needs to be addressed not only by ISPs, but also by Application providers and, finally, Cloud Operators. To enable access to the service with the same level of QoE anywhere in the world, the service provider may need to use services of some number of clouds offering their resources in different parts of the network.

Finally**,** the usability of overlay applications does not only depend on the software and the end user device like for traditional applications. Instead, it strongly depends on network conditions and QoS/QoE guarantees for the path between the data center and the end-user, which might be controlled by several different ISPs. Furthermore, end-users of overlay applications are neither interested in networking issues nor in QoS parameters. They judge the usability of an application based on their own experience, which is referred to as the QoE. Thus, another important challenge is ***QoE awareness***; i.e., traffic management based on monitoring of a QoE metric (subjective metric) that can be done by and mapping QoS metrics into QoE using (objective ones, such as packet loss, jitter, number of stalling events for HTTP-based video streaming).

QoE-awareness is realized by thorough analysis of users' expectations. Originally it was done by experiments with users using content changed or corrupted in controlled way. For a very popular service category, i.e., video streaming, the user-perceived quality is based on the values of startup delays and occurring stalling events, and both parameters can be defined as QoE indicators. If the source of service degradation is known, or at least suspected, there is possibility of intervention by service provider and QoE improvements [14].

V. Optimization Approaches & expected gains

In this section, we overview and qualitatively assess traffic approaches that may address the traffic trends and challenges identified in Section III and IV, respectively. In particular, the approaches overviewed fall under four major categories: economic awareness, cloud caching, social awareness, and service adaptation.

A. Economic Awareness

Economic traffic management (ETM) [15] has been proposed by the project SmoothIT to deal with inefficiencies caused by excessive overlay traffic trespassing ISPs' physical networks by enabling collaboration between them, and ultimately, by bridging the information gap between them. ETM's philosophy is aligned with the *Design for Tussle* proposed in [16] and Future Internet Design Principles reported in [17], as ETM aims to address incentives of both the overlay and the underlay, and allows them to express their preferences. In particular, the inefficiency in both overlay's and underlay's operation due to traffic generated by overlay and specifically peer-to-peer (P2P) networks such as BitTorrent had been identified before in [18] and [19].

This inefficiency was interpreted by SmoothIT in high inter-connection costs for the ISPs due to the abundant traffic passing over expensive inter-domain links, and in poor QoE (e.g., low download completion times for file storage/sharing, high stalling times for video streaming) for end-users due to ISPs' practices such as deep packet inspection (DPI) and P2P traffic throttling or shaping. In the context of SmoothIT and following the ETM paradigm, several optimization approaches have been proposed. Two of the most representative ones are the BGP Locality Promotion (BGP-Loc) [20] and the Highly Active Peer (HAP) [21] mechanisms.

BGP-Loc enables peers of an ISP domain to receive "advice", practically ratings of their overlay neighbors, by an entity called SmoothIT Information Service (SIS). These ratings are calculated based on ISP-related factors, such as underlay proximity, and link congestion or BGP multi-exit descriptor values. Considering BGP-Loc in the context of cloud services and applications, it could be extended to provide underlay information to the end-users, i.e., customers of Cloud Operators such as end-users or Application Providers in order to make decisions on inter-datacenter communications; for instance, alternative resource and path selection could be employed so as to reduce operating costs or congestion for ISPs and improve performance for the user of the cloud service, e.g., in terms of latency.

A technical issue concerning the BGP-Loc mechanism is the intervention of the mechanism in the communication of the user (end-user or Application Provider) to the Cloud Operator. Specifically, in the P2P context, a centralized server called SIS (SmoothIT Information Service) was contacted by the alternated client of the end-user (peer) and provided its advice based on underlay information. In a cloud service scenario, the service might be accessed by the user through Web or an interface provided by the cloud provider itself. Thus, different incentives must be provided to the stakeholders to collaborate and make use of such service. Moreover, other types of underlay information could be employed to perform decision making except BGP values, such as round trip time, number of hops, number of ASes, as well as criteria related to the Cloud Operator (in the case of collaboration), e.g., energy efficiency, latency, etc.

Another mechanism employing economic awareness is HAP that involves the enhancement of the access rate of locality-aware, high-volume uploaders though a reward scheme, achieving both the promotion of locality and faster content distribution [21]. HAP, due to the offering of extra capacity resources, exploits the native self-organizing incentive-based mechanisms of the overlay to increase the level of traffic locality within ISPs. Considering a cloud environment, the deployment of a HAP-like mechanism could be developed again by ISPs to provide incentives to content and service providers, or Cloud Operators and CDNs, in order

to make ISP-friendly decisions, e.g. employ underlay criteria to perform content replication during off-peak hours. Related incentives would again include the reward of Cloud or Datacenter Operators and Application Providers with extra capacity in terms of bandwidth and potentially prioritization of their traffic. Moreover, concrete agreements could be established among different stakeholders in form of SLAs.

A major issue regarding the successful deployment of the HAP mechanism in a P2P context was the monitoring of the users' behavior. In the cloud context, the deployment seems to be easier as monitoring of an ISP-friendly behavior would target only a specific number of providers, with whom also specific SLAs could be established; thus, resorting to SLA monitoring and management.

B. Cloud Caching

Caching has appeared in literature numerous times, each time addressing a different type of traffic or application, e.g., web caching, or caching employed by P2P file sharing systems. Such an approach is presented in [22], where the insertion of a simple cache called ISP-owned Peer (IoP) is proposed that runs the native P2P protocol and exploits the inherent capability P2P to attract peers for data exchange. By definition, IoP is employed by the ISP, though alternatively, it could also be employed by the overlay provider in order to provide its customers better QoE.

IoP could be re-designed to assist content delivery over social networks such as Facebook. The IoP's operation could then be consisted by two modules: one performing caching of popular or user-generated content (UGC) which is expected to be requested soon, and one making decision on which content to cache and where. The target of the IoP mechanism would be to reduce inter-domain traffic by reducing redundant data transfers from remote domains, as well as the improvement of the QoE (e.g. latency) of the end-users. Content management between several IoPs and other related decisions could then be performed based on social information and meta-information, i.e. information extracted by online social networks.

Considering the extension of the IoP mechanism in a cloud context, e.g. to address video traffic generated by online social networks, some technical issues that should be resolved include the following: i) the identification of content items (videos) that should be cached in order to achieve a measurable and also significant impact, ii) the decision on where to cache each content item, iii) the extraction of necessary social information by the social network, and iv) the intervention of the IoP in the video dissemination process as it happens currently over the social network.

Another in-network caching approach, called CloudAngels is proposed in [23]. The CloudAngels are Virtual Machines (VMs) whose capacity is determined dynamically depending on the content provider's stated objectives regarding the swarms' needs. Evaluation of CloudAngels revealed that the minimum distribution time, a both individual and social optimization objective, is significantly reduced. CloudAngels is a highly intervening approach as it is imposed by the Cloud Operator, while neither the Application Provider or the end-users can make decision on whether to communicate with it or not. Although CloudAngels is a cloud-ready caching approach, it could be extended to incorporate also social information regarding the potential destination of specific content items, as well as to achieve lowest energy consumption by minimizing the number of VMs that are active at some time to achieve a certain level of QoS. Regarding the exploitation of social information by the CloudAngels approach, similar technical issues apply as those reported for the IoP mechanism. On the other hand, concerning the implementation of CloudAngels to address also energy efficiency, then the monitoring of new metrics, e.g., energy consumption by VMs, should be addressed, and the decision making should be updated to incorporate also related information.

Along with high benefits, caching has been related to some side-effects such as increased overhead due to content replication among cache servers and content updates (mostly for live data, e.g., news web pages). Especially, in the case of large data centers, the traffic volumes generated due to replication of content between different PoPs (Points-of-Presence) are enormous. In recent years, in order to address the increased overhead, approaches have been proposed in literature such as [26] which proposes a scalable wide-area protocol called Summary Cache (SC) and was shown to reduce overhead due to web caches to the half compared to the Internet Cache Protocol. SC is enforced by the CDN operator and thus, is considered to be highly intervening. The idea of keeping record of content items in other caches can be adapted to be used in inter-datacenter communication.

C. Scheduling

In [27], NetStitcher is presented, which employs a network of storage nodes to stitch together unutilized bandwidth whenever and wherever it exists; i.e., identifying bandwidth leftovers, taking advantage of time-zone differences among large data center PoPs, and utilizing a stop-and-forward algorithm. Evaluations showed that the NetStitcher outperforms other compared mechanisms (including the application of no mechanism) in terms of fault tolerance and users' QoE. NetStitcher is highly intervening as it can be enforced by the data center operator regardless of other involved stakeholders' preferences.

NetStitcher could be extended to incorporate also energy consumption in the decision making, e.g., decide on which data center to move data in order to reduce overall energy consumption. Additionally, NetStitcher could be extended to address not only bulk data transfer, but also dynamic content and workload. Specifically, a similar (yet more complicated) scheduling mechanism could be employed which would allow the movement of dynamic load, e.g., simulation data, from one data center (offering computational resources not only storage) to another data center in a different geographic region over the

night, so as to exploit time-zone differences and the fact that the price of power is lower during night (off-peak) hours.

Clearly, the extension of NetStitcher to address also energy efficiency requires the identification of suitable energy metrics to be monitored and as already mentioned the update of the decision making process to incorporate also such information. Moreover, the extension of NetStitcher to address dynamic load is much more complicated as not only raw data but also state variables should be replicated seamlessly.

D. Social awareness

Social awareness is a term used for analyzing the impact of interpersonal communication on organization and structure of networks and is derived from peer-to-peer networking. Recently, an evolutionary migration towards more complex social networks or applications, such as Facebook or Twitter, can be observed. A socially-aware application exploits information about social relationships in order to classify, store and transmit content requested by users.

In [24], TailGate is proposed; a centralized approach for distribution of so-called long-tail content among geographically separated participants exploiting information on social relationships, considering assignment of users to PoPs, and profiting from time-zone differences similarly to the NetStitcher approach [27] presented in Section V.B. Instead of caching and direct downloading on demand, the TailGate system collects data on social relationships and behaviors, while it distributes content among PoPs using off-peak hours. TailGate is a cloud-ready approach as it addresses video traffic generated by online social networks which is mainly served by cloud and CDNs. Efficiency and perceived usefulness of TailGate depend on achieved precision of estimation of content distribution. Such precision is related to how long traces can be stored, processed, and analyzed.

Moreover, TailGate could be extended to incorporate also energy consumption in the decision making, e.g., decide on which data center to move data in order to reduce overall energy consumption. Again, as in the case of NetStitcher, the identification of suitable energy metrics to be monitored and the update of the decision making process to incorporate such information is required in order for TailGate to become also energy efficient.

Next, SocialTube [25] is an example of a distributed video sharing system, offering a functionally similar service like Facebook's video streaming but with higher efficiency. SocialTube allows for uploading and downloading video content employing additional information about social distance among users. This social distance reflects similarity of interests for the same content. The design of the mechanism is based on extensive observations of Facebook users' behaviour. Technically SocialTube creates P2P overlay using information from social network and identifies followers and non-followers of a video content provider. The social network-based P2P overlay has hierarchical structure which connects a source node with its followers, and connects the followers with other non-followers. The source pushes the first chunk of each new video to its followers where it is cached because there is high probability that it will be requested to be watched.

SocialTube is a cloud-ready approach, while it could be extended though to incorporate not only “social distance” information, but also underlay distance metrics, e.g., round trip time, number of hops, or BGP values, in order to achieve reduction of inter-domain traffic and thus cost for the ISPs. Major issues to be considered by the extension of SocialTube to incorporate also underlay information include the extraction of such information by the network indirectly, e.g., by probing, in the case where no collaboration is established with the ISP, and the update of the decision making to take into account both social and underlay information.

E. Service Adaptation

Service adaptation is a general term and constitutes a wide scope of methods that can be split into three areas: adaptation of application clients running on terminals, influencing transfer of traffic streams across the network, and optimizing usage of data center or server resources. This general framework gives many opportunities to shape the service if feedback information is available about performance of elements which compose the whole service environment: users, network, and servers. The leading target from running network-based services is to meet user expectations regarding simultaneously service functionalities and perceived quality.

When analyzing user expectations, for example for a video on demand service, it is possible to split such expectations into objective (e.g., transmission bitrate) and subjective (e.g., low startup delay, low stalling) expectations. If service quality is not sufficient, there is possibility to monitor and possibly influence parts of service architecture, composed of network resources, cloud resources, and user device. In general, a service quality degradation monitored by service intelligence and communicated clearly to the end user gives better subjective opinion of the user (expressed by QoE) than uncontrolled degradation or sudden disruption [28].

Client-side adaptation is probably the simplest and most efficient way of shaping service provisioning. Clients can easily request the service level from the cloud which leads to desired QoE level. Often the required service level and the perceived quality depend on the used device (e.g., optimized web pages for mobile devices, impact of display size for video streaming [29]). Additionally, imperfect network conditions, manifesting themselves as decrease of available bandwidth at the client device can be monitored. Thus, the client can request a quality adaptation from the server when changing network conditions occur during transmission, e.g., in case of video streaming by adapting video bitrate. However, such approaches need the implementation of a loopback mechanism done, e.g., by a specially designed control plane [30].

The impact of network part performance within the process of service adaptation is a generally widely explored topic since

this leads to controlling of transport layer parameters and promoting WAN optimization aspects. As TCP is used typically for transport layer, there are plethora of methods of improving perceived bandwidth and latency of transferred streams.

If facing limited bandwidth, it is important to recognize link with the lowest bandwidth in the whole transport chain. Usually the bottleneck is in the access area, so the service adaptation should lead to a smart choice of access technology and parameters (e.g., switching from GSM to WiFi link). Thus, imperfect or limited performance of the network should be monitored (e.g., passive YouTube QoE monitoring for ISPs [28]) and the network should be optimized (e.g., WAN optimization solution Steelhead [31]). However, if the problem cannot be overcome in the network, feedback should be given to the server in order to perform service adaptation in the overlay application.

Clouds are inherently flexible solutions managing usage of computation power, storage, and network resources, so service adaptation can be easily implemented with respect to applications' requirements and available resources. For successful service provisioning it is necessary to implement a model how to adapt cloud resources and utilize them in order to provide relevant SLAs to the users of clouds. Application Providers can be supplied with different levels of control over architectural components, and thus can adapt the used cloud resources according to the current application conditions. A comprehensive vision of autonomic service provisioning with mechanisms necessary to design and implement services, using features and services derived from IaaS providers, is presented in [32].

From a technical point of view, cloud-side adaptation can be obtained by scaling the resources allocated to different services or virtual machines. Moreover, a cloud federation concept which promotes cooperation of clouds could be employed that enables the transfer of tasks or migration of virtual machines among clouds. A comprehensive view of virtual machines for large datacenters is provided in [33].

Efficient implementation of cloud-side adaptation needs resources for monitoring, analyzing, and controlling components of the cloud as well as application functionality and performance. The cooperation among involved stakeholders needs dedicated communication mechanisms and can be supported by a designated control plane. Such resources are usually available in today's adaptively managed clouds. Thus, we refer to cloud-side adaptation as adaptations performed by the service provider within the cloud. It can employ the feedback of client or network monitors, and adapts the delivered service level. For example, in the case of video streaming this can be achieved by functionalities of scalable coding: decrease of frame rate, reduction of video resolution or decrease of image quality. An example how QoE can be controlled by exploiting the parameters offered by H.264/SVC (Scalable Video Coding) is presented in [34].

Additionally, cloud-side adaptation can be necessary in case of high workloads on the servers in order to avoid service failures. However, the main technical issue arises from the fact that the service (e.g., video content) has to be available for different service levels. While this can be achieved rather easily for static content, it is unclear whether and how it is possible for applications like live video streaming, video conferencing, or gaming, for which the content is created and delivered in real time.

Finally, Energy-Aware Traffic engineering as described in [35] shows how to reduce considerably energy consumption with the following energy saving methods: sleep mode of network links, sleep mode of routers, and rate adaption.

VI. SUMMARY AND CONCLUSION

There are important new applications greatly influencing volume and patterns of Internet traffic and the satisfaction of users. These include cloud computing and applications served, e.g., video streaming platforms, social networks, etc. Nonetheless, respective network management and operation frameworks are missing today for heterogeneous technologies.

The contribution of this article includes the identification of major challenges related to traffic management of cloud applications. In particular, we identified that new management mechanisms should support effectively scale, agility, stable Quality-of-Experiences (QoE), and flexibility, in an integrated set of network and overlay service functionality.

Moreover, we provided an in-depth overview of applications traffic management approaches, we examined potential extensions of the overviewed approaches to efficiently address the challenges identified, and finally, we qualitatively assessed the optimization potential in terms of traffic reduction, energy savings, and improved user experience. We found that there is a huge diversity of cloud services which makes it impossible to find a one-fits-all optimization approach. Instead, flexible approaches which can be customized to the specific requirements of a service should be taken into consideration.

ACKNOWLEDGMENT

This work was funded in the framework of the EU ICT Project SmartenIT (FP7-2012-ICT-317846). The authors alone are responsible for the content. We thank all partners of the SmartenIT project for their collaboration and useful discussion on the subject of this paper.

REFERENCES

[1] Cisco: Cisco Global Cloud Index: Forecast and Methodology, 2010-2015, White Paper, 2011.

[2] Cisco: Hyper-connectivity and the Approaching Zetabyte Era, 2010.

[3] Cisco: Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2010–2015, White Paper, 2011.

[4] Sandvine, “Global Internet Phenomena Report 1H 2013”, “Global Internet Phenomena Report 2H 2013”, Tech. Rep., 2013 (www.sandvine.com).

[5] The NIST Definition of Cloud computing, Peter Mell, Timothy Grance, September 2011

[6] The Future of Cloud Computing. Opportunities For European Cloud Computing Beyond 2010, Public Version 1.0, Keith Jeffrey (ERCIM), Burkhard Neidecker-Lutz (SAP Research)

[7] IRTF, Software defined Networking research group, http://irtf.org/sdnrg, last visit jan 2014

[8] ETSI, World class standards, http://www.etsi.org/technologies-clusters/technologies/nfv, last visit jan 2014

[9] Wikipedia, http://en.wikipedia.org/wiki/Big_data, last visit jan 2014

[10] Akyildiz, Ian F., et al. "Wireless sensor networks: a survey." *Computer networks* 38.4 (2002): 393-422.

[11] Jones, Christine E., et al. "A survey of energy efficient network protocols for wireless networks." *wireless networks* 7.4 (2001): 343-358.

[12] https://www.google.com/about/datacenters/efficiency/internal/

[13] Baliga, Jayant, et al. "Green cloud computing: Balancing energy in processing, storage, and transport." *Proceedings of the IEEE* 99.1 (2011): 149-167.

[14] Hoßfeld, Tobias, et al. "Internet video delivery in youTube: From traffic measurements to quality of experience." *Data Traffic Monitoring and Analysis*. Springer Berlin Heidelberg, 2013. 264-301.

[15] Hoßfeld, Tobias, et al. "An Economic Traffic Management Approach to Enable the TripleWin for Users, ISPs, and Overlay Providers." *Future Internet Assembly*. 2009.

[16] Clark, David D., et al. "Tussle in cyberspace: defining tomorrow's internet." *ACM SIGCOMM Computer Communication Review*. Vol. 32. No. 4. ACM, 2002.

[17] Papadimitriou, Dimitri, et al. "Design principles for the future internet architecture." *The Future Internet*. Springer Berlin Heidelberg, 2012. 55-67.

[18] Liu, Yong, et al. "On the interaction between overlay routing and underlay routing." *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*. Vol. 4. IEEE, 2005.

[19] Karagiannis, Thomas, Pablo Rodriguez, and Konstantina Papagiannaki. "Should internet service providers fear peer-assisted content distribution?."*Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement*. USENIX Association, 2005.

[20] Racz, Peter, Simon Oechsner, and Frank Lehrieder. "BGP-based locality promotion for P2P applications." *Computer Communications and Networks (ICCCN), 2010 Proceedings of 19th International Conference on*. IEEE, 2010.

[21] Pussep, Konstantin, et al. "Cooperative traffic management for video streaming overlays." *Computer Networks* 56.3 (2012): 1118-1130.

[22] Papafili, Ioanna, Sergios Soursos, and George D. Stamoulis. "Improvement of bittorrent performance and inter-domain traffic by inserting isp-owned peers."*Network Economics for Next Generation Networks*. Springer Berlin Heidelberg, 2009. 97-108.

[23] Sweha, Raymond, Vatche Ishakian, and Azer Bestavros. "Angels in the cloud: A peer-assisted bulk-synchronous content distribution service." *Cloud Computing (CLOUD), 2011 IEEE International Conference on*. IEEE, 2011.

[24] Traverso, Stefano, et al. "Tailgate: handling long-tail content with a little help from friends." *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012.

[25] Li, Ze, et al. "Socialtube: P2p-assisted video sharing in online social networks."*INFOCOM, 2012 Proceedings IEEE*. IEEE, 2012.

[26] Fan, Li, et al. "Summary cache: a scalable wide-area web cache sharing protocol." *IEEE/ACM Transactions on Networking (TON)* 8.3 (2000): 281-293.

[27] Laoutaris, Nikolaos, et al. "Inter-datacenter bulk transfers with netstitcher."*ACM SIGCOMM Computer Communication Review*. Vol. 41. No. 4. ACM, 2011.

[28] Schatz, Raimund, Tobias Hoßfeld, and Pedro Casas. "Passive youtube QoE monitoring for ISPs." *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2012 Sixth International Conference on*. IEEE, 2012.

[29] McCarthy, John D., M. Angela Sasse, and Dimitrios Miras. "Sharp or smooth?: comparing the effects of quantization vs. frame rate for streamed video."*Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2004.

[30] Liu, Xi, et al. "A case for a coordinated internet video control plane."*Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication*. ACM, 2012.

[31] http://www.riverbed.com/products-solutions/products/wan-optimization-steelhead/

[32] Casalicchio E, Silvestri L: Architectures for autonomic service management in cloud-based systems. In Computers and Communications (ISCC), 2011 IEEE Symposium on, IEEE 2011:161{166.

[33] Bifulco R, Canonico R, Ventre G, Manetti V: Transparent migration of virtual infrastructures in large datacenters for cloud computing. In Computers and Communications (ISCC), 2011 IEEE Symposium on, IEEE 2011:179{184.

[34] Zinner T, Abboud O, Hohlfeld O, Hossfeld T, Tran-Gia P: Towards QoE Management for Scalable Video Streaming. In 21th ITC Specialist Seminar on Multimedia Applications - Traffic, Performance and QoE, Miyazaki, Jap 2010.

[35] Mahadevan, Priya, Puneet Sharma, Sujata Banerjee, and Parthasarathy Ranganathan. "Energy aware network operations." In *INFOCOM Workshops 2009, IEEE*, pp. 1-6. IEEE, 2009.