On the Feasibility of Using Current Data Centre Infrastructure for Latency-sensitive Applications

David Griffin, Truong Khoa Phan, Elisa Maini, Miguel Rio and Pieter Simoens

***Abstract*—It has been claimed that the deployment of fog and edge computing infrastructure is a necessity to make high-performance cloud-based applications a possibility. However, there are a large number of middle-ground latency-sensitive applications such as online gaming, interactive photo editing and multimedia conferencing that require servers deployed closer to users than in globally centralised clouds but do not necessarily need the extreme low-latency provided by a new infrastructure of micro data centres located at the network edge, e.g. in base stations and ISP Points of Presence. In this paper we analyse a snapshot of today's data centres and the distribution of users around the globe and conclude that existing infrastructure provides a sufficiently distributed platform for middle-ground applications requiring a response time of $20-200$ ms. However, while placement and selection of edge servers for extreme low-latency applications is a relatively straightforward matter of choosing the closest, providing a high quality of experience for middle-ground latency applications that use the more widespread distribution of today's data centres, as we advocate in this paper, raises new management challenges to develop algorithms for optimising the placement of and the per-request selection between replicated service instances.**

I. Introduction and Motivation

Cloud computing offers a low-cost and scalable computation platform for applications running off-premises. Several major cloud service providers have emerged who offer large regional data centres strategically located around the globe to be topologically close to large centres of user demand. The decision on where to locate service components running in the cloud depends on many factors, including costs and the legal jurisdiction of the provider, but the main aspect considered in this paper regards the performance objectives of the application considering the location of user demand. Selecting a single data centre to house an application may achieve the required performance for applications that are insensitive to latency or jitter, but more demanding applications will only deliver high quality of experience to a relatively small proportion of users who are sufficiently close in terms of network proximity to the selected cloud node.

Service providers can replicate service instances in multiple geographical locations offered by a cloud provider to deploy instances closer to regions of user demand, which also improves the resilience of the deployed service. However, for more demanding low-latency and high-bandwidth applications this may not be sufficient: 30% of the population of the USA, for example, has a too high latency to one of Amazons EC2 data centres for cloud-based gaming [30]. This study shows that even in a well connected continent, served by more than one large cloud location, a finer granularity of deployment is needed for interactive applications.

To overcome the problems associated with a coarse granularity of cloud locations, edge and fog computing have been proposed for edge analytics in the domain of the Internet of Things [14], where cloud nodes are located at the edge of the network and close to users to maximise network performance and reduce the network load to centralised data centres [31]. In general, edge computing is desirable for services that require extremely low latency or very high bandwidth flows such as those as envisioned for tactile internet services in 5G networks, requiring a response in the order of 1 ms [32].

A global deployment of edge cloud nodes requires a major investment by cloud service providers and ISPs to deploy data centres in ISP points of presence, at mobile base stations and in other locations close to users. Such an infrastructure has started to be deployed to support virtualised network functions running as software components in ISP-provided data centres [10]. However, even when edge cloud nodes are widely deployed there is a significant management overhead required by service providers to provision and then manage widely replicated services in numerous edge node locations [28]. Moreover, these edge nodes are envisioned to be positioned in locations such as street cabinets or mobile base stations that impose restrictions on the physical size, power consumption and, hence, computational complexity of the edge cloud clusters.

There is a spectrum of responsiveness required by different application types. For the purposes of this paper we classify current and future cloud-based services into three broad groups: *latency-tolerant*, *latency-sensitive* and *latency-demanding* applications. *Latency-tolerant* services such as shared document editing, image- and text-based social networks can tolerate response times of 200 ms or greater. *Latency-demanding* applications, including highly responsive games [35] and other augmented and virtual reality applications requiring tactile response times[1], control systems for autonomous vehicles and robots, and real-time health-care intervention services require response times of 20 ms or less. A large number of middle-ground *latency-sensitive* applications such as cloud gaming, interactive conferences, remote video and image editing require a response time between these two extremes, in the $20-200$ ms range: touch interaction requires a

D. Griffin, T. K. Phan and M. Rio are with University College London, UK.

E. Maini is with Vodafone, UK.

P. Simoens is with Ghent University - imec, Belgium.

[1]https://www.fastcodesign.com/1671685/the-magic-number-for-making-virtual-reality-feel-like-reality

latency of around 170 ms [34] and interactive gaming requires a response time in the range of $50 - 200$ ms [33] [35].

Centralised cloud deployments are efficient, low cost and highly scalable on-demand for *latency-tolerant* applications. Edge node deployments are needed for extreme *latency-demanding* applications, but at an operational cost of deploying and managing service instances in a huge quantity of locations around the globe. Our view is that there are many *latency-sensitive* services that require higher performance than can be delivered by centralised clouds but may not need to be deployed at the extreme edge.

In this paper we explore the performance that can be delivered by today's cloud computing infrastructure for services that require middle-ground responsiveness. We have taken a snapshot of today's non-edge data centres located around the globe and modelled the network performance that can be delivered from those data centres to the current population distribution of Internet users. We show that approximately 90% of users around the globe can reach at least one data centre within 10 ms and that all users in the globe can reach at least five data centres within a latency of 100 ms. We suggest that this is sufficient for middle-ground, latency-sensitive interactive applications.

One of the main goals of this paper is to provide an insight into the availability of data centres for hosting latency-sensitive applications for globally distributed users. The focus is on quantifying the number of data centres that are feasible, on latency grounds, for service deployment: a key criterion and starting point for service placement decisions. But, the identification of the set of data centres within a maximum latency radius from users is insufficient *by itself* as a complete service deployment strategy. Many other factors should influence service placement and selection, including: data throughput rates, cost - in terms of deployment and operational costs for computation and storage resources as well as traffic transport costs - and reliability of both the network and data centre infrastructure. However, latency is one of the primary concerns for service placement as even if a data centre is cheap or highly reliable but it is too distant to deliver required application responsiveness it should be excluded from consideration. Reducing the set of data centres to the feasible ones on latency grounds reduces complexity significantly for subsequent service placement algorithms. Specific service placement and selection algorithms that select between the subset of feasible data centres meeting latency constraints to trade-off performance and costs are out of the scope of this paper. Examples of algorithms making use of the knowledge of feasible data centres as an essential input to simplify optimisation techniques are in [38] [39] and [40].

The performance of any service placement and instance selection algorithm is bounded by the number and distribution of available data centres. Obviously, there can be a chicken-and-egg problem: if there are only a limited number of locations available, this is a significant barrier for the development of next-generation demanding end-user services. Conversely, if there are no services needing distributed deployment, there is no incentive for network operators or cloud infrastructure providers to invest in additional cloud capacity, especially bearing in mind that the operational expenditure of distributed facilities is higher than that of only a handful of centralised sites. For this reason, it is important to gain insight in the infrastructure that is already available today and how well these sites are reachable by end-users.

We show that the vast majority of users have multiple feasible data centres for middle-ground latency-sensitive applications. This is an important result of our study as it shows that service providers have the opportunity to optimise placement decisions on multiple criteria other than just latency, for example to maximise the number of users receiving satisfactory QoE within a maximum cost budget, minimizing the number of hosting data centres and hence the total deployment costs.

The remainder of the paper is organised as follows. In section II we discuss prior and related work, section III describes our model of data centre and user locations, section IV explains how we modelled the network performance between users and data centres, section V presents the main results of our performance modelling and our findings on the availability of data centres within maximum latencies of users around the globe, section VI focusses on application-layer delays and their impact on our network latency findings, section VII discusses the management and operations aspects of deploying and managing services across distributed data centres, and section VIII presents our conclusions.

II. Background and related work

The authors in [1] presented several measurement results of Amazon Web Services (AWS) like EC2 and CloudFront AWS services. They showed interesting observations on the evolution of AWS over a two-year-long period ($2012 - 2013$). For instance, with Amazon EC2, in 2012, 70% of users can access the data centres located in Virginia with less than 100 ms response time, while in 2013 they observed that up to 90% of users can get the services in less than 100 ms. For CloudFront 83% of requests in 2012 were satisfied in less than 20 ms in Frankfurt and the caches in Milan and Sweden can serve 80% of requests in less than 3 ms. CloudCmp [2] was developed to compare performance and cost of the four incumbent cloud providers (Amazon AWS [5], Microsoft Azure [6], Google AppEngine [8], and Rackspace CloudServers [7]). The latency is measured from 260 vantage points on PlanetLab [9] to instances deployed on cloud providers. The results show that by selecting an appropriate cloud provider, the average round trip time (RTT) is 74 ms and up to 95% of requests can be served in less than 50 ms response time. The authors in [4] provide an extensive set of experiments conducted based on a real-world QoS dataset collected on PlanetLab, comprised of $360k$ measurements from 200 users on $1,597$ Web services. The results show that the mean response time is about 70 ms and the minimum response time is only 0.008 ms. The authors of [3] study the impact of infrastructural bottlenecks and network protocols on latency and conclude that to achieve 30 ms response time for fetching HTML pages of popular websites around 2000 CDN locations are needed. In general, the results in the literature show that, in small-scale studies, users could access services at a latency of less than 100 ms.

In this paper, we perform an extensive evaluation of latency through a combination of a theoretical model and real-world measurements for a large data set consisting of 3116 data centres and 3×10^9 users around the world (see Section III, IV and V) and show that if the current data centre infrastructure can be used by service providers it is good enough to guarantee QoS for middle-ground *latency-sensitive* applications.

The authors of [11] confirm that proximity to data centres is critical for mobile applications that are highly interactive and resource intensive. In [13], by using data collection of more than 250 mobile end-hosts over a two-month period, the authors show that service replicas selection plays an important role in reducing latency. Mobile clients can see up to 400% differences in latency depending on which replicas are selected [13]. In this paper, we identify the need for intelligent service resolution and service placement algorithms for optimising the configuration of servers on the current cloud infrastructure to minimise the number of data centres locations used while keeping low latency for users.

Edge/fog computing extends the cloud computing paradigm to the edge of the network, which aims to bring cloud and IT functionalities at the edge of radio access networks (RAN) [10], [14]. With a wide-spread geographical distribution of micro-cloud servers, the edge/fog computing model can guarantee low latency as well as better support for high bandwidth and mobile users. The authors in [15] show that communication latency can be significantly improved by careful design and operation of the cloud radio access networks (C-RAN) especially in the high mobility case. In addition, small cells with optical wireless links can meet the latency requirements of 10 ms [16]. Moreover, end-to-end latency below 2 ms can be achieved by using a next-generation baseband chipset [16]. On the other hand, the C-RAN further helps reducing the access network latency by coordinating transmission schedules [17]. While these papers confirm that extreme edge computing infrastructure is necessary for *latency-demanding* applications we show that middle-ground applications can be deployed on today's data centres provided that smart management and control algorithms are used to optimise performance and cost.

Given the data centre infrastructure, Content Delivery Networks (CDN) and Service Centric Networking (SCN) are the two models proposed to improve QoS such as reducing latency and increasing bandwidth transmission. CDNs are globally distributed network servers deployed in multiple geo-locations around the world [18]. By moving content to the edge of the network, the goal of a CDN is to serve content to end-users with high availability and high performance. Results in [19] confirm that the CDN architectures can significantly improve response time in comparison with accessing data from the origin sites. SCN [20], [22] on the other hand has been proposed as a potential solution to managing services more efficiently using Information Centric Networking (ICN) principles [21]. SCN decouples the service from their origin location, meaning that the requests can be served directly by any node that currently has the service running. By locating a service replica closer to the users it is possible to significantly reduce latency compared to accessing the service located at the origin server. However, as we show in this paper, reducing latency is not the only objective. In Section VI we identify the benefits of the ISP being involved in the resolution of service requests to service instances to reduce network transit costs and improve QoS.

III. Data Centre and User Locations

We start our study by a characterisation of the location and number of DCs currently available worldwide. This provides an upper bound for the expected performance of any service placement algorithm to plan a service placement and replication strategy to optimise latency to geographically distributed users within maximum cost constraints.

The website datacentermap.com maintains a registry of data centres worldwide which is updated regularly by cloud service providers. At the time we crawled this website, we collected information on 3116 data centres in 116 countries as shown in Figure 1. Apart from geographical data centre coordinates, this database also provides details about the available services, the tenants and which carriers are providing IP transit (IPv4/v6).

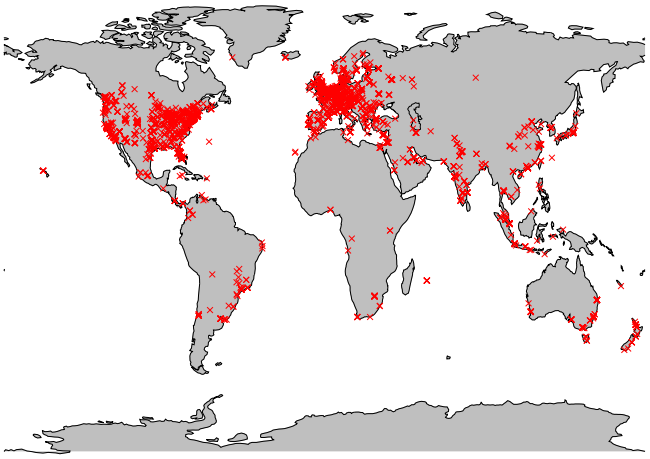


Fig. 1: Geographical location of the 3116 data centres.

We took the 3116 data centres as the baseline for our analysis work to determine the network performance between users to their n^{th} closest data centre and also to identify the number of data centres available within a maximum latency radius, as presented in section V. We also looked at the performance implications of limiting the number of available data centres below 3116. To do this we used a k-means clustering algorithm [23] for values of 1 to 500 on the full set of 3116 data centres, selecting the data centres closest to the centroid of each cluster. As an example, we show the results of the k-means clustering for 50 data centres in Figure 2.

We model demand after the demographic distribution of the worldwide population, as documented in the "cities1000" dataset provided by www.geonames.org. This file contains all cities worldwide with a population of more than 1000 inhabitants. These statistics were weighted with the internet penetration rate as reported by data.worldbank.org.

The dataset contains $112,106$ cities, with 3×10^9 users. According to the most recent statistics on the world population, this is an underestimation by a factor of 2.34. This can be

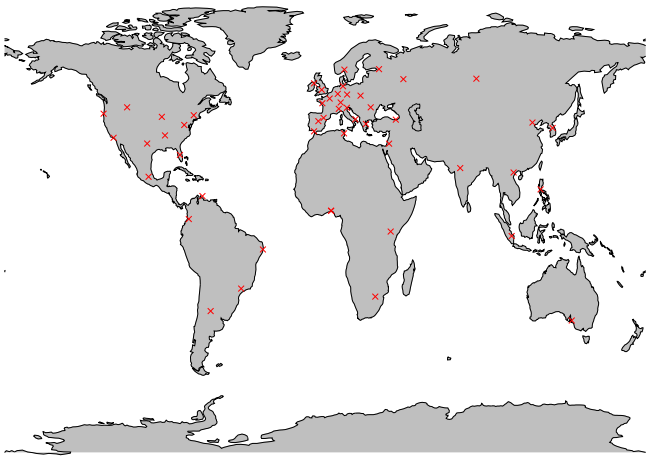


Fig. 2: Locations of 50 data centres using K-means algorithm.

explained as follows: first, the original dataset does not contain any cities with fewer than 1000 inhabitants; secondly, and possibly more importantly, is due to the age of the input data. For example, Ghent, Belgium has $231,000$ users in the dataset, whereas the most recent figures indicate a population of $250,000$ users. However, in our study the geographical distribution of users and their relative population density is more important than the precise value of current population and therefore does not have a major effect on our results or conclusions.

IV. Network model

Building upon the distribution of user and data centre locations we now model the network performance in terms of latency between users and data centres. In this section we first of all approximate the length of Internet paths and convert this to network latency in terms of round trip time (RTT) between users and data centres. The theoretical model is then refined with real-world measurements from probes around the globe to actual data centre locations. We show that the theoretical model is a good indicator of the lower bound of latency observed through our measurements, but that actual network and cloud resource load can influence the experienced performance. We show how the range of measured latency values affects our assumptions on the basic network model.

A. Mapping path length to latency

Analysing the expected performance between users and DCs requires an accurate prediction of network latency. An important key requisite for latency estimation is a topological model of the network path between pairs of endpoints on the Internet. While there are various ways to achieve this, through examining inter-Autonomous System (AS) topology maps and actual BGP routes or by performing exhaustive traceroutes from user locations to discover actual paths, a scalable approach is needed to model paths between all $112,106$ user locations and $3,116$ data centres. We propose a simple path model where the Internet routing geography between users and data centres is modelled by three segments: user location to the capital city of the country where the user is located; capital city of the users country to the capital city of the data centres location (this is omitted if the user and data centre are located in the same country); capital city of the country where the data centre is located to the data centre location. The rationale behind this model is that inter-domain routing is usually through public Internet Exchange Points, typically located in major cities of that country, or through direct peering between Autonomous Systems, which is also undertaken in large points of presence in major cities. We deviate from this model when the user and DC are located in the same AS, when we assume a direct path from user to DC which does not detour through the capital city.

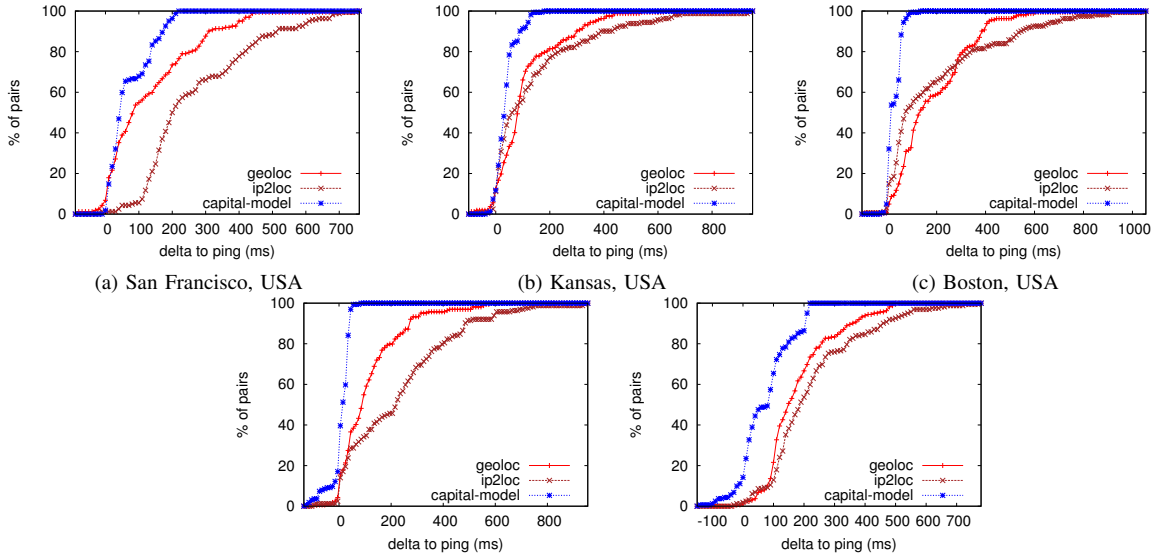
The great circle distance is calculated using the haversine formula for each of the three network segments and the three segments are summed for each user-DC pairing. Network latency, in terms of round-trip-time, can be estimated from the great circle distance using the conversion factor calculated in [24] where RTT can be approximated by 0.018 ms per km of great circle distance,[2] as determined by the analysis of measurements of global Internet traffic. This conversion factor reduces the latency compared to the speed of light due to two factors: the speed of light in fibre is reduced by 30% due to internal reflections within the optical fibre and, secondly, the factor accounts for typical deviations from a straight line path due to physical network topology within an AS.

To test for the potential inaccuracies in some cases we investigated the impact of not modelling the actual sequence of routers which form the path between user and DC; and secondly of any overestimations in path length in large countries, such as the USA, if all inter-AS traffic is assumed by our model to be routed via the capital city. In order to quantify the inaccuracies introduced by our simple model we compared the predicted latency to actual measurements. We also compared the latencies predicted by our model via capital cities to those predicted by following the actual path through the set of routers as identified by traceroute.

We chose representative locations in the USA, Europe and Australia as probe locations and took measurements to 209 CloudHarmony sites around the globe (see section VI for more details). The routers identified by traceroute between the probes and the DCs were geolocated using two different services: a commercial service, IP2Location[3] and a free service, GeoLite2.[4] The traceroute path length was calculated by summing the great circle distance between each pair of routers along the path and the total was converted to RTT latency using the conversion factor of 0.018 ms per km of great circle distance, as discussed above.

A set of latency measurements were taken between the probes and the DCs. For each (probe, DC) pair, ten measurements were taken every half an hour over a three day period. Analysis of the results showed that the mean and median RTT values were very close to the minimum RTT. The minimum measured RTT value was taken as the measured

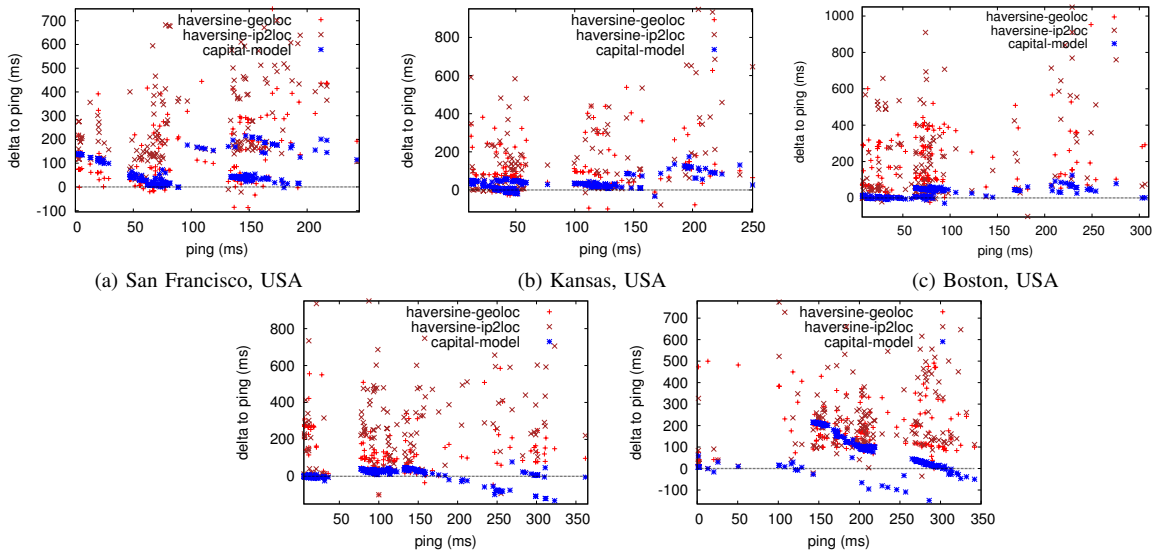
[2] See section V, Fig. 5 in [24]
[3] https://www.ip2location.com
[4] https://dev.maxmind.com/geoip/geoip2/geolite2/



(a) San Francisco, USA (b) Kansas, USA (c) Boston, USA

(d) Ghent, Belgium (e) Sydney, Australia

Fig. 3: CDF of difference in RTT between actual measurements and model predictions



(a) San Francisco, USA (b) Kansas, USA (c) Boston, USA

(d) Ghent, Belgium (e) Sydney, Australia

Fig. 4: Difference in RTT between actual measurements and model predictions, plotted against measured RTT

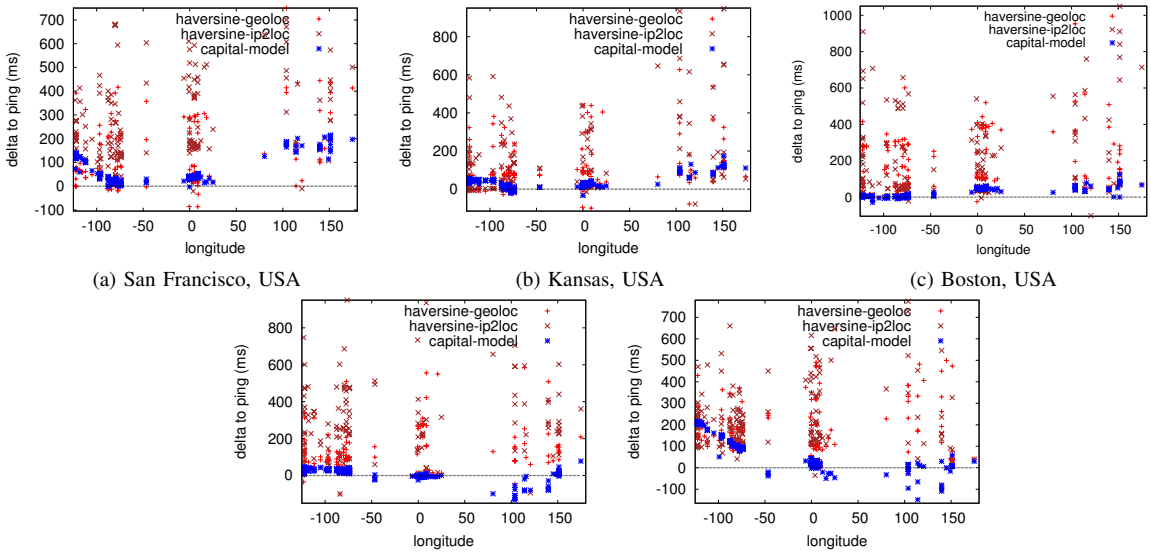
latency between each user-DC pair in the following analysis.

Figure 3 shows a comparison of the difference between the predicted latency and the actual measured RTT ("delta to ping") for the three models: our simple model routed via capital cities and that calculated from the actual sequence of routers revealed by traceroute, geolocated using the two different services: GeoLite2 (geoloc) and IP2Location (ip2loc).

Both of the geolocation services we used incorrectly located many routers, resulting, in some cases, of detours via distant continents, for example some paths within Belgium appeared to use intermediate routers in the USA. Overestimates of 100s of ms were common.

The CDFs show that the simple capital model is a more accurate approximation to measured latency than the geolocated traceroute models, however there are some over and under estimations of latency. Figure 4 plots the difference between the predicted and measured latency versus the measured latency to show how significant the errors are in proportion to the measured RTT. In the case of all five probes the latency predicted by the capital model is either very close to the measured value or is a small overestimate. The exceptions are from Ghent and Sydney where some underestimates appear to DCs greater than 200 ms away. However, this is not a problem in our case as we are concerned with middle-ground services with a maximum latency below 200 ms so any more distant DCs would not be suitable candidates for deploying service instances for users from these locations.

We earlier pointed out that our simplified model of a single



(a) San Francisco, USA (b) Kansas, USA (c) Boston, USA

(d) Ghent, Belgium (e) Sydney, Australia

Fig. 5: Difference in RTT between actual measurements and model predictions, plotted against longitude

IXP per country located in the capital city may introduce inaccuracies in larger countries such as the USA. To investigate this further Figure 5 shows the RTT delta plotted against the longitude of the DC location. The inaccuracy of the capital model within the USA is demonstrated quite clearly in these graphs. There are more overestimates in latency from the probe in the west coast of the USA (San Francisco) and the least in the east coast (Boston) which is geographically much closer to the capital city. The reason is that the model routes the path via Washington DC for any inter-AS traffic. A user in California connecting to DCs within the USA will detour via Washington DC. Nearby DCs will experience the greatest stretch, while DCs located in or near to the capital will experience very little stretch. This can be seen in Figures 4a and 5a where there is a close-to linear correlation with negative gradient of RTT delta with both actual RTT and longitude, respectively. This effect is reduced in the case of a probe located closer to the centre of the country, Kansas (Figures 4b and 5b) and is minimal in the case of Boston (Figures 4c and 5c), due to their relative location compared to Washington DC. The same negative correlation can be seen from Sydney (Figures 4e and 5e), due to the actual paths to DCs in the USA going via trans-Pacific cables and landing on the west coast of the USA. There is very little impact on traffic from Europe to DCs in the USA as shown in Figures 4d and 5d), for the probe in Belgium, due to Washington DC being relatively close to the direct path from Ghent to any DC location in the USA.

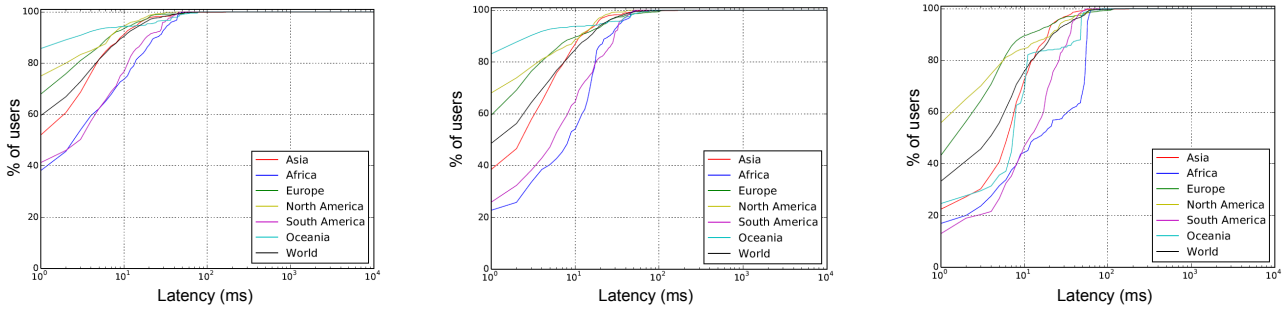
In summary, our capital model gives a good estimation of latency to DCs within 200 ms, while it is less accurate to DCs in distant continents. However, as we are focussing on middle-ground services with a maximum latency of 100 ms, services will not be located on the other side of the planet and so this inaccuracy of our capital model does not impact our findings in section V. The exception is in geographically large countries, such as the USA, where there are significant numbers of users and DCs located far from the capital city. In this case our capital model overestimates the latency within the country, by up to 120 ms in the most extreme cases. This makes our model conservative and therefore the results and conclusions we make on the suitability of today's DC location for low-latency applications are not invalidated. On the contrary, even more options for service placement within latency bounds are likely to be available than we report in this paper. Use of the capital model for predicting latency is very close to observed real-world RTT in the majority of user-DC pairings considered for nth-closest DCs and maximum latency radius in section V. Where the capital model deviates from measured values, e.g. in countries like the USA, it errors on over-estimating latency, which makes our findings conservative and does not affect our main conclusions.

V. Global Reachability of Data Centres

Having establishing the distribution of users and data centres around the globe in Section III and the method of estimating network latency in Section IV we now visualise the latency between users and data centres. We do this in two ways: by showing the latency to the closest data centres and by identifying the number of data centres available to users within a maximum latency radius.

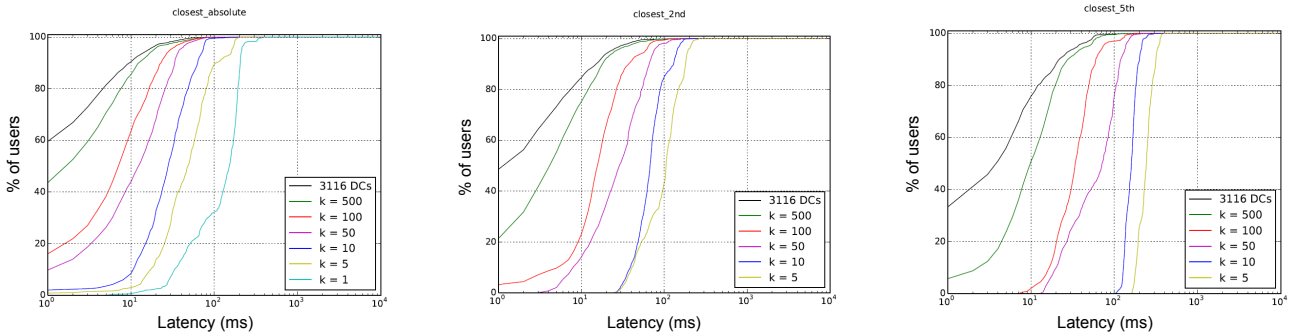
A. Closest data centres

We first model data centre availability in terms of the network latency to the closest, the 2^{nd} and the 5^{th} closest data centres shown as a CDF for users located in each continent and for the global population. Note that a log-scale is used for the x-axis in Figure 6 - Figure 9. As shown in Figure 6a, 95% of users worldwide can reach their closest data centre within approximately 20 ms. Users in Africa and South America perform slightly worse due to the lower density of data centres



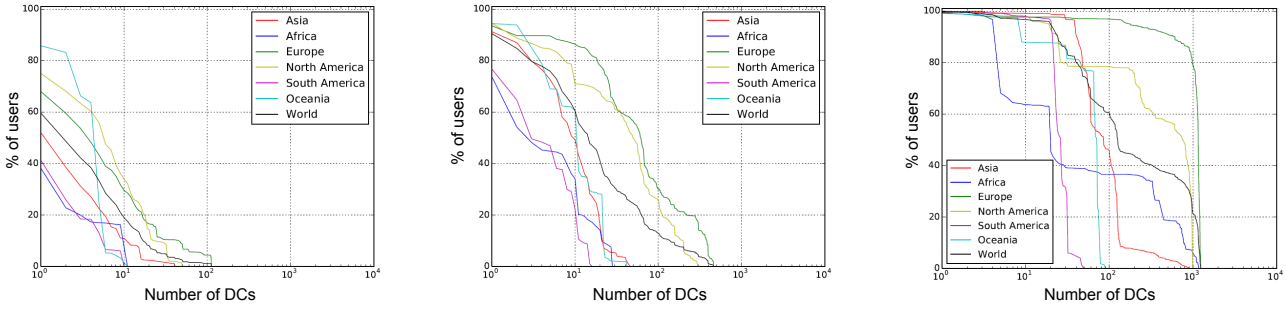
(a) Absolute closest data centre (b) The 2^{nd} closest data centre (c) The 5^{th} closest data centre

Fig. 6: CDF of RTT for all users, split by continent and for the global population



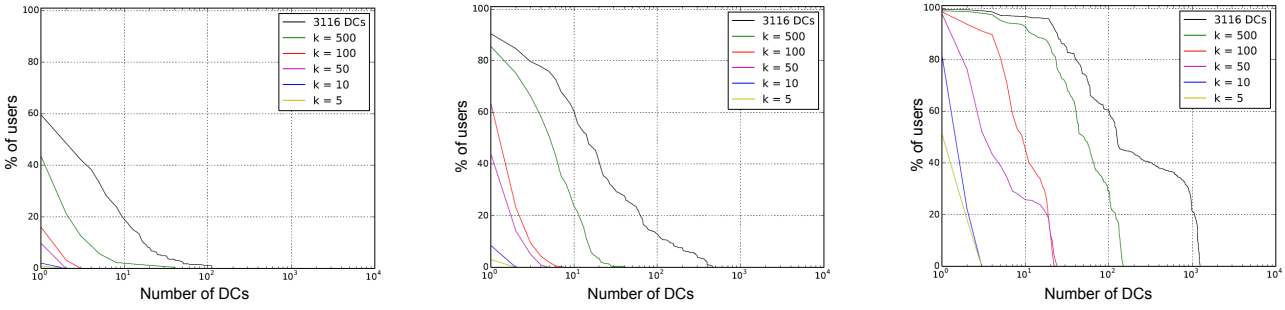
(a) Absolute closest data centre (b) The 2^{nd} closest data centre (c) The 5^{th} closest data centre

Fig. 7: Comparison of how RTT varies with the total population of data centres worldwide (from $K = 1$ to 3116)



(a) Within 1 ms RTT (b) Within 10 ms RTT (c) Within 50 ms RTT

Fig. 8: CCDF of number of data centres available split by continent and for the global population



(a) Within 1 ms RTT (b) Within 10 ms RTT (c) Within 50 ms RTT

Fig. 9: CCDF of number of data centres available for the worldwide population (from $K = 5$ to 3116)

in those continents, with the closest data centre being within $30-40$ ms for 95% of the population.

Figure 6b and Figure 6c show the CDF to the 2^{nd} and the 5^{th} closest data centres. This models the degree of flexibility in service placement and resolution to not simply place or select the absolute closest data centre for each user. Furthermore, having multiple data centres within the latency constraints of an application provides the opportunity to replicate service instances in multiple locations to balance load over multiple servers or to increase the resilience to failures of any single server. Those figures show that there are 90% and 85% of users worldwide can reach the 2^{nd} and the 5^{th} data centre within approximately 20 ms.

So far we have plotted the RTT to the n^{th} closest data centre for the full set of 3116 data centres around the world. In the following, we model a more limited deployment of services in a smaller set of data centres, as calculated by K-means clustering. We plot the results for $K = 1, 5, 10, 50, 100$ and 500 compared to the original results for 3116 data centres in Figure 7. Obviously, the more data centres we have, the more options there are for a user to reach a data centre within latency bounds. For instance: considering a RTT of 20 ms, 95% of users can reach their closest data centre in case of full 3116 data centres (Figure 7a). Given 500 data centres, the number of users slightly reduces to 94% which means that if we use less than $\frac{1}{6}$ of the current data centres, we still can achieve similar QoS for users. When $K \leq 100$, we see a significant drop in the percentage of users able to reach their closest data centre within the same latency bound. Similar observations can be found in the 2^{nd} and the 5^{th} closest data centres (Figure 7b and Figure 7c).

B. Quantity of data centres within a defined latency

After establishing a model to estimate the latency between users and DCs, we wanted to investigate how many DCs a user can reach within a given maximum latency. For all 3116 data centres again, we show the number of data centres split by continent. As can be seen in Figure 8 the density of data centres is not even over the globe: North America, Europe and Oceania have much more data centres than in Asia, Africa and South America. Especially, we found that around 85% of users in Oceania can reach at least 1 data centre within 1 ms while this is about 40% of users in Africa and South America (Figure 8a). When increasing the latency range to 10 ms, almost 80% of users in Africa and South America can reach their closest data centres while this number is 90% for the rest of continents (Figure 8b).

We also model a more limited deployment of data centres and show the fraction of users that can reach their closest data centres within a latency range. Given $RTT = 50$ ms, more than 98% of users can find a data centre even with only 50 data centres to be deployed around the world (Figure 9c). To find at least one data centre within $RTT = 10$ ms, we need to deploy more than 50 data centres (Figure 9b) while it is more than 500 data centres if we consider a latency range of 1 ms (Figure 9a). Note that, although we are assuming a latency range of $20-200$ ms for middle-ground applications we have investigated $RTT = 50$ ms to leave a sufficient time budget for server processing, middlebox traversal, etc.

Fig. 10: Location of the CloudHarmony cloud sites that were queried.

The main observation from the modelling reported in this section is that a large fraction of the worldwide population has already a large number of data centres within a relatively small RTT, sufficient for applications with a middle-ground response time requirement.

VI. Real-world measurements of application-level performance

So far we have considered the best-case network level latencies between users and data centres. However, in any real-world deployment, other factors may affect the actual response-time a user may experience when accessing a cloud-based service instance. Network congestion, queuing and processing delays at the remote data-centre may all increase latency and reduce the quality of experience of the users. In this section we report on a study of application-layer latency measurements from probes deployed around the world to actual data centres and conclude with an assessment of how delay variations may impact our overall findings on the suitability of today's data centre locations for deploying middle-ground services.

CloudHarmony[5] monitors data centres worldwide and provides APIs to access this data. In addition, it provides functionality to test network latency from any location to 209 data centres worldwide, as shown in Figure 10. We conducted tests by deploying a probe in 24 nodes that are part of three academic testbeds: InstaGENI, PlanetLab Europe and Kreonet. The geographical location of the probes is visualised in Figure 11. We collected measurements over a two month period from each of the 24 probes to all 209 data centre locations.

Rather than calculating solely the average latency the focus of this study was to examine the variations in delay that are due to variable network queuing and application-layer processing delays to get an insight into how the minimum network latency between users and data centres may increase

[5]www.cloudharmony.com

Fig. 11: Geographical location of the probes.

TABLE I: Linear regression results for EU probes to data centres in the different continents.

Continent	Percentile	Intercept	Slope	R^2
Europe	5th	0.0	0.035	0.71
	50th	13.84	0.04	0.73
	95th	43.25	0.041	0.58

in real-world scenarios. Figure 12 contains scatterplots of the standard deviation of the latency measurements for probes located in Ghent, Georgia Tech and Daejon. It can be seen that the standard deviation is independent from the great circle distance and it can be concluded that the major component of delay variation is not correlated with network path length.

In order to see how these measurement observations compare to the model we used in section IV-A, we performed a linear regression on the latency measurements collected. We performed a constrained linear regression on the X-th percentile of the (distance, latency) pairs, grouped in bin sizes of 10, 50 or 100 km. Table I shows the subset of latency measurements collected from the 12 European probes to all data centres in Europe. For reasons of brevity and clarity, the table only contains the results for the bin size of 50 km, and for the 5^{th}, 50^{th} and 95^{th} percentile. Similar results were obtained for other bin sizes and percentiles.

Figure 13 visualises these percentile-based regressions against the individual measurement points. Notably, as the percentile increases, mainly the intercept on the Y-axis increases, whereas the slope of the linear curve only slightly increases from 0.040 to 0.041 between the 50^{th} and 95^{th} percentile.

An example of how actual measurement observations compare to our basic theoretical model is shown in Figure 14 of the CDF of RTT to the 5^{th} closest data centre for users in Europe. This compares the network latency model used in section IV-A, labelled "our_model" in the graph to the measured service-level latency from 5^{th} to 95^{th} percentile as calculated above. This shows that the network level model is close to the minimum of the measured values and real-world measurements of service-level latency, including access and data centre networks and delays in the software stack in the data centres can be higher. For example median measured latency for 95% of users is ~ 60 ms compared to ~ 25 ms predicted core network-level latency. The difference between the predicted core network latency and the observed response time is due to additional delays introduced by application-layer processing, access and data centre networks. As can be seen in Fig. 13 the 5th, 50th and 95th percentile regression lines are close to parallel, indicating that the latency overhead is independent of distance between users and data centres. We have already shown that minimum measured latency is very close to our network model, and we can conclude that application-layer processing, access and data centre network latency can be considered constant for a particular application.

As an indication of the reliability of data centres and of the network path, Figure 15 shows the proportion of failed requests from our CloudHarmony measurements campaign, after filtering out the results from those data centres and probes which had very high failure rates due to software failures and other bugs. Figure 15(a) plots the failure rate between probe-data centre pairs (range from 0 to 20%) and Figure 15(b) shows the CDF of the proportion of failed requests per data centre tested. In this relatively small study it can be seen that, while the reliability is high across all data centres, some data centres are more reliable than others. A reliability metric for data centres is an important input into the placement and selection optimisation algorithms discussed in the following section. Failed requests can be due to backbone or data centre network problems. Core network reliability is studied further in [41] and data centre network reliability is investigated in [42].

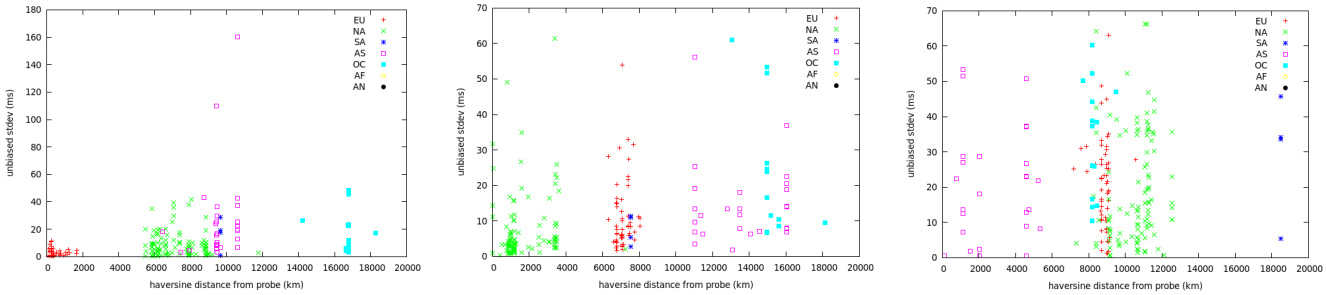
VII. Management and Operations Considerations

As we have shown in this paper, there is a sufficient distribution and availability of today's data centres to meet the demands of many middle-ground responsiveness applications i.e. those requiring a response time of around $20 - 200$ ms. However, optimising for multiple performance metrics and cost is not a simple matter of making decisions based on geographical proximity of data centres. Management and control algorithms are needed to optimise the placement and selection of service instances on multiple additional criteria (including network throughput, cost and reliability) within this rich landscape of potential locations. In turn, this requires a framework supporting the collaboration of all actors involved, including the ISPs, service providers, and cloud infrastructure.

Placement and deployment algorithms optimise the number and location of data centres where service instances are deployed, considering the anticipated demand levels from diverse geographical locations and monetary cost constraints. Deploying services in every data centre will provide the highest possible performance but this comes at a high cost in terms of the quantity of data centre resources to be reserved as well as the operational cost of configuring them everywhere. Selecting a single centralised location, on the other hand, minimises deployment cost but at the expense of reduced performance and increased network latency for the majority of users.

Service deployment and placement decisions are conducted off-line, prior to the invocation of services by users [38]. As such, these algorithms do not need to respond to user requests in real-time and can therefore be relatively heavy-weight, undertaking multi-objective optimisation.

In our view, deployment optimisation is best undertaken by the service provider, who has visibility of the predicted global



(a) Virtual Wall (b) Georgia Tech Univ (c) Daejon (Kreonet)

Fig. 12: Scatterplots on standard deviation of average latency

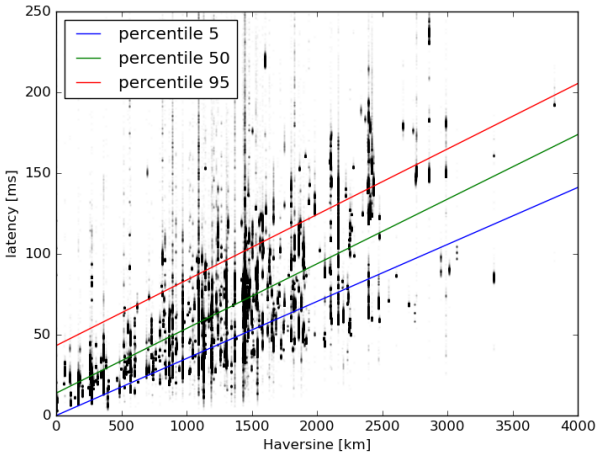


Fig. 13: Linear regression from clients in the EU to data centres in the EU.

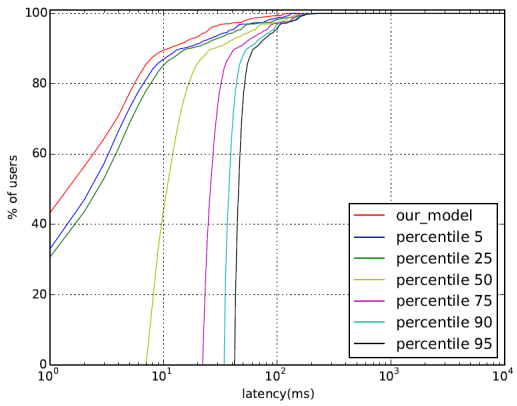


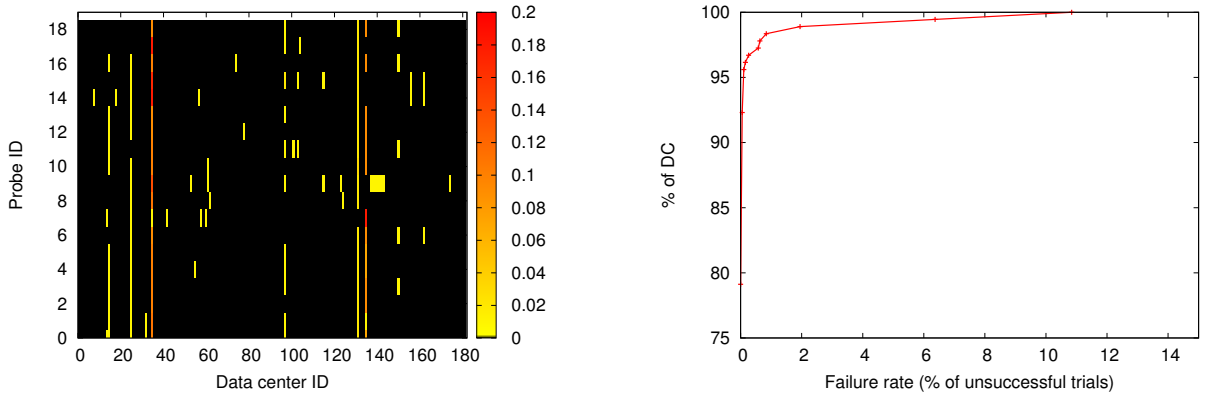
Fig. 14: CDF of RTT to the 5^{th} closest data centre for users in Europe, comparing the theoretical network level latency model against 5^{th} to 95^{th} percentile of measured RTT at the service-level.

usage patterns, the knowledge about how cloud infrastructure variations might affect application performance and sets out the tolerable deployment costs. The service provider will also have detailed knowledge about the implementation of the server logic, its computational complexity and, therefore, its performance and processing overhead. Knowledge of application-specific processing latency will determine the latency budget available for the core-networking component: an key input into the placement algorithms which chose between feasible data centres. Replicating services in many data centres around the globe requires service providers to build relationships with many cloud service providers, which can be a significant barrier in terms of operational overhead, especially for SMEs. This can be alleviated by means of a cloud-broker operating as an orchestrator on behalf of individual service providers/application developers, who also offers a software platform for management of the service life cycle across distributed cloud nodes [28]. The service provider needs to expose to the broker some application metrics and how these are weighted to evaluate the feasibility of a data centre to host an instance.

Selection algorithms optimise the resolution of service requests to map them to the most appropriate running service instance. Resolution algorithms should consider the user's network location, the current load on the servers and the actual network conditions between the requestor and the candidate instances and aim to reduce inter-domain bandwidth usage and hence minimise network costs.

Although service selection is needed per user request, the optimisation of the resolution process does not necessarily need to take place at each invocation time. Selection optimisation can be undertaken periodically, based on predicted usage patterns, with the result of the current optimisation epoch resulting in forwarding tables which can be modified by real-time load measurements [39] [40].

Today's CDNs do not take into account the current server load and the approximation of a client's location by the location of its DNS resolver negatively impacts the client's performance [37]. In our view, resolution of service requests to service instances is best undertaken by ISPs, as they have a detailed view on the client's location and on the BGP routing path towards data centres in remote Autonomous Systems [28]. Besides improvement customer satisfaction, ISPs may be incentified to act as a service resolver as this would allow them to have more fine-grained control on their inter-AS transit traffic costs. An important aspect of this conjecture is on the distribution of data centres across ASes, which determines



(a) Failure rate per probe-data centre pair

(b) CDF of the proportion of failed requests per data centre

Fig. 15: Reliability of data centres

how much choice an ISP may have. This is investigated in the next subsection VII-A where we investigate how much choice an ISP may have in selecting between service instances. In subsection VII-B we discuss the view ISPs have on the performance of end-to-end paths compared to over-the-top (OTT) providers.

A. Bandwidth and network cost considerations for service resolution

We have argued that only a limited set of highly latency-demanding applications might benefit from being deployed in edge nodes and thus the commercialisation of such a costly infrastructure may be rather limited. However, ISPs may have other motivations to deploy edge computing nodes, such as reducing traffic, as evidenced by the recent emergence of telco CDNs to reduce ingress traffic from remote Autonomous Systems [36]. Bandwidth costs for ISPs vary depending on whether destinations are within the local network, in peering domains or over paid-for transit links. We analyse the availability of data centres over these different locations for the ISP of the originating user and conclude that it is important for the ISP to be involved in the selection between alternative data centres running a service instance to optimise inter-domain traffic and hence reduce costs.

In this section we examine how much choice an ISP may have on selecting between instances deployed on the set of 3116 data centres. We used the IPv4 Routed /24 AS Links Dataset from CAIDA[6] to classify the routes between the ASes of user-data centre pairs. Directly connected ASes were classified as being interconnected by peer-to-peer connections and those routed over intermediate ASes were classified as using customer-provider transit links.

In Figure 16 we show the number of data centres available within 1 ms and 50 ms of users that are located in the users' own AS or in a neighbouring AS over a peer-to-peer link. The x-axis is sorted by rank. Figure 17 shows the ratio data centres within the local AS or over a peering link to all data centres within 1 ms and 50 ms of the users.

[6]http://www.caida.org/data/active/ipv4_routed_topology_aslinks_dataset.xml

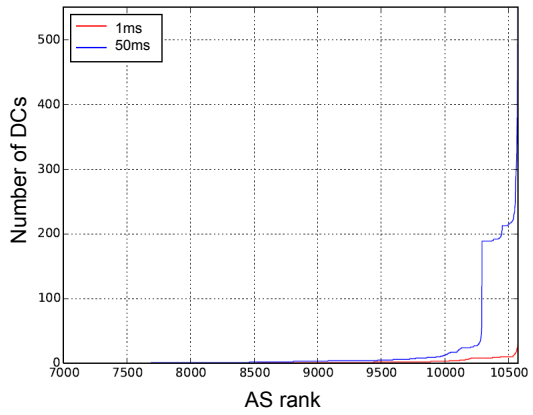


Fig. 16: Number of data centres available within the local AS or over a peering connection for data centres within 1 ms and 50 ms of the users, by AS rank.

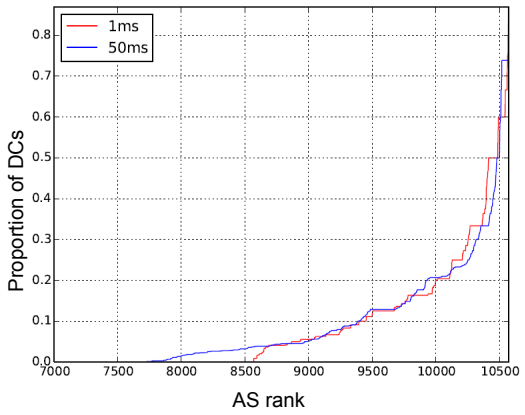


Fig. 17: Proportion of data centres available within the local AS or over a peering connection compared to total data centres including those over transit links for data centres within 1 ms and 50 ms of the users, by AS rank

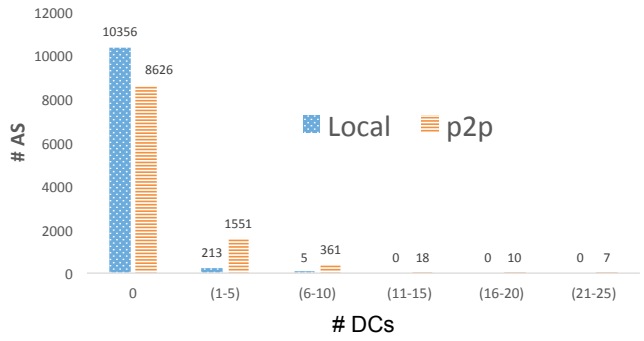


Fig. 18: Histogram of the number of data centres reachable within 1 ms RTT of users within the local AS or over a peering link

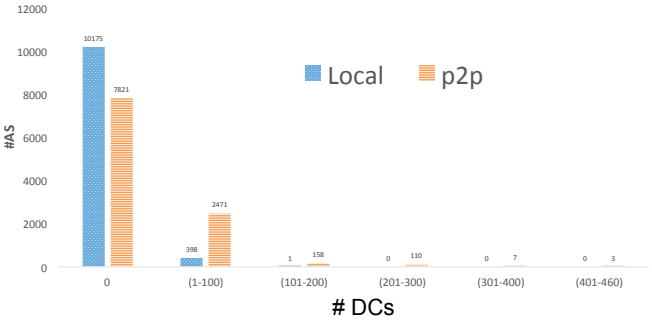


Fig. 19: Histogram of the number of data centres reachable within 50 ms RTT of users within the local AS or over a peering link

Figure 18 and Figure 19 show histograms of the number of data centres available in the local AS and over peer-to-peer links within 1 ms and 50 ms of the users.

As can be seen from the graphs, very few data centres are positioned within the users' local AS or in neighbouring ASes connected by peering inter-domain links if we only consider the snapshot of today's data centres in the dataset from datacentermap.com. If, in the future, ISPs deployed their own data centres within their core network or at edge nodes than this proportion could increase significantly. However, this would be at a higher deployment cost, from the service providers' point of view, for replicating service instances more widely in individual ISPs' data centres.

A conclusion from this analysis is that when considering services in locations that meet the performance constraints of a service, a large number of locations will not be within a user's AS or in neighbouring ASes. This means that minimising inter-domain bandwidth and transit costs is not simply a matter of selecting local or neighbouring data centres. Most candidate data centres are available in remote ASes over paid-for transit links and so minimising transit costs in the resolution process means taking into account the costs of individual transit links. Reducing transit costs is mostly in the interest of the ISP and, furthermore, detailed information on transit costs is unlikely to be shared with third parties, for business confidentiality reasons. Hence, resolution decisions that aim to reduce transit costs while maintaining performance targets are best made by ISPs themselves, or by trusted third parties that have access to full cost and network performance information.

The graphs show there is a limited choice between free (local or P2P) vs paid for (transit) data centres, where available this makes an obvious choice for ISPs. The majority of available data centres are via transit links and this is where the ISP can influence costs more, by selecting service instances in data centres available via cost-effective transit links that meet the performance constraints of the service.

B. Role of ISPs in monitoring and enforcing network performance

The knowledge of the length of inter-domain paths alone is not sufficient for service resolution if inter-domain delay is the criterion. Paths may change over time, because of link failures or updated traffic policies. This is hard to capture with OTT measurements, such as those provided by CloudHarmony, but can be monitored by ISPs via BGP route reflectors. For effective resolution with delay as a criterion, BGP information should be combined with direct measurements of delay. Observed variability of end-to-end- delays (confirmed by direct measurements) suggests that ISP-supported resolution has the most commercial potential for demanding services. Inter-domain delay changes often coincide with inter-domain routing events and in such cases BGP information can efficiently be used for service resolution.

Demanding services require resolution based on up-to-date knowledge of the network state. An important consideration is the impact of routing changes on end-to-end latency between users and data centres. In [28] we show that BGP updates can result in significant changes in network performance including latency to reach the same destination. For any major BGP storm there is an impact on RTT, which is transient as routing converges. This typically takes between 6 - 20 minutes. Therefore performance estimates can vary dynamically. Consequently, keeping track of respective changes by smaller service providers can be prohibitive for them. Monitoring BGP events on large scale can effectively be done by ISPs, especially in the vicinity of their AS. This in turn may open opportunities for ISPs to reclaim position in the value chain by providing resolution service and supporting orchestration for demanding services.

On the other hand an OTT approach to network performance monitoring means that a timely reaction to short-term deteriorations of connectivity is possible only when e2e measurements are taken very frequently. This however can be difficult to achieve in realistic scenarios. Therefore, monitoring BGP events in real time can be seen as an enabler that allows to increase the responsiveness of service resolution (and possibly service orchestration) platform.

As shown by the authors in [29], the positioning of ISPs in the service resolution process will be strengthened in the future when ISPs are involved in the edge cloud ecosystem under the growing popularity of demanding services distributed in the edge.

VIII. Conclusions

Future latency-demanding Internet services may require a new infrastructure of data centres at the edge of the network to deliver extremely low network latencies and high responsiveness. We have modelled the distribution of data centres available today and the locations of users around the globe and conclude that the current set of data centres are sufficient to deliver the necessary quality of experience in terms of response times for interactive, latency-sensitive applications. However, this requires a management and control framework and a set of placement and resolution algorithms that can fully take advantage of today's data centres and provide a federated platform between cloud-service providers. There isn't a one size fits all approach to service deployment and operations and management algorithms need to understand the requirements of different services and trade-off the costs of wide scale deployment versus the required QoE. Reducing latency is not the only goal of managing service deployment and bandwidth utilisation and inter-domain traffic costs need to be considered when selecting between service replicas. Hence there is a need for ISPs to be involved in service resolution and selection to react to real-time network performance degradations as well as to ensure that inter-domain bandwidth use is cost-effective. Cooperation between application service providers and ISPs is needed with service providers focussing on global deployment concerns with ISPs being involved in the dynamic selection between service replicas.

Acknowledgements

The research leading to the results in this paper has received funding from the European Commission in the FP7 FUSION and H2020 5G-MEDIA projects under grant agreements 318205 and 761699. The authors would particularly like to thank our colleagues Dariusz Burstzynowski, Folker Schamel and Frederik Vandeputte for their valuable contributions and insights into the modelling work presented here.

References

[1] I. Bermudez and S. Traverso and M. Munafo and M. Mellia, *A Distributed Architecture for the Monitoring of Clouds and CDNs: Applications to Amazon AWS*, in IEEE Transactions on Network and Service Management, 2014.

[2] A. Li and X. Yang and S. Kandula and M. Zhang, *CloudCmp: Comparing Public Cloud Providers*, in ACM Conference on Internet Measurement, 2010.

[3] A. Singla and B. Chandrasekaran and P. B. Godfrey and B. Maggs, *The Internet at the Speed of Light*, in ACM Workshop on Hot Topics in Networks, 2014.

[4] J. Zhu and Y. Kang and Z. Zheng and M. R. Lyu, *WSP: A Network Coordinate Based Web Service Positioning Framework for Response Time Prediction*, in IEEE International Conference on Web Services, 2012.

[5] Amazon Web Service. http://aws.amazon.com.

[6] Microsoft Windows Azure. http://www.microsoft.com/windowsazure.

[7] Rackspace Cloud. http://www.rackspacecloud.com.

[8] Google AppEngine. http://code.google.com/appengine.

[9] PlanetLab. http://www.planet-lab.org.

[10] Mobile Edge Computing. http://www.etsi.org/technologies-clusters/technologies/mobile-edge-computing

[11] K. Ha and P. Pillai and G. Lewis and S. Simanta and S. Clinch and N. Davies and M. Satyanarayanan, *The Impact of Mobile Multimedia Applications on Data Center Consolidation*, in IEEE International Conference on Cloud Engineering, 2013.

[12] S. Sundaresan and W. d. Donato and N. Feamster and R. Teixeira and S. Crawford and A. Pescape, *Broadband Internet Performance: a View from the Gateway*, in ACM SIGCOMM, 2011.

[13] J. P. Rula and F. E. Bustamante, *Behind the Curtain - The Importance of Replica Selection in Next Generation Cellular Networks*, in ACM SIGCOMM (poster), 2014.

[14] F. Bonomi and R. Milito and J. Zhu and S. Addepalli, *Fog Computing and Its Role in The Internet of Things*, in ACM workshop on Mobile cloud computing (MCC), 2012.

[15] Y. Cai and F. R. Yu and S. Bu, *Cloud Computing Meets Mobile Wireless Communications in Next Generation Cellular Networks*, in IEEE Network, 2014.

[16] D. Schulz and C. Alexakis and J. Hilt and M. Schlosser and K. Habel and V. Jungnickel and R. Freund, *Low Latency Mobile Backhauling using Optical Wireless Links*, in Broadband Coverage in Germany, 2015.

[17] S. Lien and S. Hung and K. Chen and Y. Liang, *Ultra-low-latency Ubiquitous Connections in Heterogeneous Cloud Radio Access Networks*, in IEEE Wireless Communications, 2015.

[18] E. Nygren and R. K. Sitaraman and J. Sun, *The Akamai Network: A Platform for High-Performance Internet Applications*, in ACM SIGOPS Operating Systems Review, 2010.

[19] B. Krishnamurthy and F. Park and Y. Zhang, *On the Use and Performance of Content Distribution Networks*, in ACM SIGCOMM Workshop on Internet Measurement, 2001.

[20] E. Nordstrom and D. Shue and P. Gopalan and R. Kiefer and M. Arye and S. Y. Ko and J. Rexford and M. J. Freedman, *Serval: An End-Host Stack for Service-Centric Networking*, in USENIX Symposium on Networked Systems Design and Implementation (NSDI), 2012.

[21] V. Jacobson and D. K. Smetters and J. D. Thornton and M. F. Plass and N. H. Briggs and R. L. Braynard, *Networking named content*, in ACM CoNEXT, 2009.

[22] A. Sathiaseelan and L. Wang and A. Aucinas and G. Tyson and J. Crowcroft, *SCANDEX: Service Centric Networking for Challenged Decentralised Networks*, in MobiCom Workshop on DIY Networking, 2015.

[23] J. B. MacQueen, *Some Methods for Classification and Analysis of Multivariate Observations*, in 5-th Berkeley Symposium on Mathematical Statistics and Probability, 1967.

[24] R. Landa and J. T. Araujo and R. G. Clegg and E. Mykoniati and D. Griffin and M. Rio, *The large-scale geography of Internet round trip times*, in IFIP Networking, 2013.

[25] D. Komosny and S. Pang and J. Pruzinsky and P. Ilko and J. Polasek, *PlanetLab Europe as Geographically-Distributed Testbed for Software Development and Evaluation*, Advances in Electrical and Electronic Engineering, 2015.

[26] G. P. Fettweis, *The tactile internet: applications and challenges*, in IEEE Vehicular Technology Magazine, 2014.

[27] L. Gao, *On Inferring Autonomous System Relationships in the Internet*, in IEEE/ACM Transactions on Networking, 2001.

[28] P. Simoens and D. Griffin and E. Maini and T. K. Phan and M. Rio and L. Vermoesen and F. Vandeputte and F. Schamel and D. Burstynowski, *Service-centric Networking for Distributed Heterogeneous Clouds*, in IEEE Communications Magazine, Volume: 55, Issue: 7, 2017.

[29] L. Vermoesen, et al. *On the Economical Benefit of Service Orchestration and Routing for Distributed Cloud Infrastructures: Quantifying the Value of Automation*, Alcatel-Lucent Bell Labs Business Modeling (White Paper), 2014.

[30] S. Choy, et al., *The brewing storm in cloud gaming: a measurement study on cloud to end-user latency*, Proc. of the 11th Annual Workshop on Network and Systems Support for Games, 2012

[31] M. Chen, Y. Hao, Y. Li, C. Lai, and D. Wu, *On the computation offloading at ad hoc cloudlet: Architecture and service models*, IEEE Commun., vol. 53, no. 6, pp. 1824, Jun. 2015.

[32] G. Fettweis et al., *5G: Personal mobile internet beyond what cellular did to telephony*, IEEE Comm. Mag., vol. 52(2), 2014.

[33] R. Shea, J. Liu, E. C. H. Ngai and Y. Cui, *Cloud gaming: architecture and performance*, in IEEE Network, vol. 27, no. 4, pp. 16-21, July-August 2013.

[34] W. Ritter, G. Kempter, and T. Werner, *User-Acceptance of Latency in Touch Interactions*, Proc. of International Conference on Universal Access in Human-Computer Interaction. Springer International Publishing, 2015.

[35] M. Dick, O. Wellnitz, and L. Wolf, *Analysis of factors affecting players' performance and perception in multiplayer games*, Proc. of 4th ACM SIGCOMM workshop on Network and system support for games (NetGames '05). ACM, New York, NY, USA.

[36] Frank, B., Poese, I., Lin, Y., Smaragdakis, G., Feldmann, A., Maggs, B., ... & Weber, R. (2013). *Pushing cdn-isp collaboration to the limit*. ACM SIGCOMM Computer Communication Review, 43(3), 34-44.

[37] Otto, J. S., Snchez, M. A., Rula, J. P., & Bustamante, F. E. *Content delivery and the natural evolution of DNS: remote dns trends, performance issues and alternative solutions*. In Proceedings of the 2012 ACM conference on Internet measurement conference (pp. 523-536). ACM.

[38] E. Maini, T. K. Phan, D. Griffin & M. Rio. *Hierarchical Service Placement for Demanding Applications*. In IEEE Globecom Workshops, 2016, doi 10.1109/GLOCOMW.2016.7848922

[39] T. K. Phan, D. Griffin, E. Maini, & M. Rio. *Utility-maximizing Server Selection*. In IFIP Networking, 2016.

[40] T. K. Phan, D. Griffin, E. Maini, & M. Rio. *Utility-centric Networking: Balancing Transit Costs with Quality of Experience*. In IEEE/ACM Transactions on Networking, 2018.

[41] A. Markopoulou, G. Iannaccone, S. Bhattacharyya, C. Chuah, Y. Ganjali, & C. Diot. *Characterization of Failures in an Operational IP Backbone Network*. In IEEE/ACM Transactions on Networking, Vol. 16, No. 4, 2008.

[42] P. Gill, N. Jain, & N. Nagappan. *Understanding Network Failures in Data Centers: Measurement, Analysis, and Implications*. In ACM SIGCOMM, 2011.

David Griffin is a Principal Research Associate in the Department of Electronic and Electrical Engineering, University College London. He has a BSc from Loughborough University and a PhD from UCL, both in Electronic and Electrical Engineering. His research interests include planning, management and dynamic control for providing QoS in multiservice networks and novel routing paradigms for the future Internet.

Pieter Simoens is assistant professor at the IDLab department of Ghent University, Belgium. He is also affiliated as senior research with imec. He holds a PhD from Ghent University. He teaches courses on system design and mobile development. His research interests include distributed real-time systems, with at a specific focus on the delivery of advanced services through distributed edge clouds. He has (co)authored more than 70 articles in journals and conference proceedings.

Truong Khoa Phan received his PhD degree from INRIA/I3S, Sophia, France. He is currently a Research Associate the Department of Electronic and Electrical Engineering, University College London. His research interests include network optimisation, cloud computing, multicast and P2P.

Elisa Maini is a SDN/NFV Product Architect at Vodafone, UK. She received her Ph.D. in Computer and Automation Engineering from the University of Naples Federico II. Her current research interests include network optimisation and modelling, software-defined networking, and network function virtualisation.

Miguel Rio is a Professor in the Department of Electronic and Electrical Engineering, University College London where he researches and lectures on Internet technologies. His research interests include on real-time overlay streaming, network support for interactive applications, Quality of Service routing and network monitoring and measurement.