Bandwidth Measurements within the Cloud: Characterizing Regular Behaviors and Correlating Downtimes

José Luis García-Dorado

Univeridad Técnica del Norte, Ibarra, Ecuador

NOTE: This is a version of an unedited manuscript that was accepted for publication. Please, cite as:
J.L. García-Dorado. Bandwidth Measurements within the Cloud: Characterizing Regular Behaviors and Correlating Downtimes. ACM TRANSACTIONS ON INTERNET TECHNOLOGY, 17 (4), Ar. 37, (2017).
The final publication is available at: `https://doi.org/10.1145/3093893`

Abstract

The search for availability, reliability, and quality of service has led cloud infrastructure customers to disseminate their services, contents, and data over multiple cloud data centers, often involving several Cloud service providers (CSPs). The consequence of this is that a large amount of data must be transmitted across the public Cloud. However, little is known about the bandwidth dynamics involved. To address this, we have conducted a measurement campaign for bandwidth between eighteen data centers of four major CSPs. This extensive campaign allowed us to characterize the resulting time series of bandwidth as the addition of a stationary component and some infrequent excursions (typically downtimes). While the former provides a description of the bandwidth users can expect in the Cloud, the latter is closely related to the robustness of the Cloud (i.e., the occurrence of downtimes is correlated). Both components have been studied further by applying factor analysis, specifically ANOVA, as a mechanism to formally compare data centers' behaviors and extract generalities. The results show that the stationary process is closely related to the data center locations and CSPs involved in transfers which, fortunately, makes the Cloud more predictable and allows the set of reported measurements to be extrapolated. On the other hand, although correlation in the Cloud is low, i.e., only 10% of the measured pair of paths showed some correlation, we found evidence that such correlation depends on the particular relationships between pairs of data centers with little connection to more general factors. Positively, this implies that data centers either in the same area or within the same CSP do not show

qualitatively more correlation than other data centers, which eases the deployment of robust infrastructures. On the downside, this metric is scarcely generalizable and, consequently, calls for exhaustive monitoring.
Index Terms: Public Cloud, Inter-cloud, TCP Bandwidth, ANOVA, Traffic Correlation.

1 Introduction

As an attempt to provide the best quality of service, Cloud service providers (CSPs), e.g., Amazon EC2, Microsoft Azure, Rackspace, or Google Cloud, have pointed to the dispersion of their data centers across the world. Such geo-distribution ensures both high availability and reliability and also reduces the final users' latency, given its significant impact in business revenue. For example, Amazon estimated that every 100 ms of latency costs 1% in sales [1].

There are numerous examples of research and industry efforts that exploit this paradigm. Netflix is a significant example of an application that disseminates its contents over a CSP infrastructure to successfully provide on-demand media services. Similarly, Cassandra [2] or Volley [3] applications deploy overlay architectures and distributed services in the Cloud to take shelter from both network failures and congestion. Moreover, geo-distribution has become a high-availability and safe way to store/backup data. Dropbox and Google Drive as well as bank networks are good examples of this. In addition to these examples, software distributions (such as popular operating system releases), distributed databases [4], and virtual machine clones [5] also share the task of moving a large amount of data to be disseminated over different data centers.

Importantly, while in the past the set of data centers where a Cloud customer deployed a service or application typically belonged to the same CSP, nowadays this scenario is becoming less commonplace. The deployment of both applications and services over different CSPs has proven to be a fundamental tool for providing the lowest latencies [6] and enhanced robustness [7] in the Cloud. Essentially, the limitations a particular CSP can present, e.g., spatially (a poorly covered geographical area) or temporally (a period of malfunction), can be compensated by others, making multi-CSP deployments a better approach than their single-CSP counterparts.

To give some representative figures of the importance of traffic between different points of presence, some Internet use surveys [8] have shown that more than 77% of data center operators run backup and replication applications among three or more sites, whereas more than 50% ones report over 1 PB of data in their primary data center. Furthermore, 70% of surveyed IT firms had between 1 and 10 Gb/s running between data centers, nearly half having 5 Gb/s or more (between 330 GB and 3.3 PB per month).

However, while there has been much effort in the research community directed at studying cost-effective bulk data transfers over data centers in the public Cloud [9, 10], and the process of Cloud benchmarking has gained in relevance [11, 12, 13], we emphasize that little is known about its bandwidth. In particular, we refer to general results aiming at the identification of invariants on this measure [14]. We believe that this lack of studies is due to the difficulties in measuring bandwidth from a large number of data centers on several CSPs over a significant period of time. This paper aims at filling this gap.

We deployed a multi-point-of-presence testbed that included eighteen data centers spread over four major CSPs, and measured the TCP bandwidth of all the paths between them. We noticed that bandwidth in the Cloud can be modeled by a principal Gaussian component and some infrequent excursions. We related two such behaviors to a stationary process that represents the typical state of a path, and a peak/downtime process that represents times when a path behaves unusually. The characterization of the stationary state allows us to answer questions related to what a user can expect of Cloud performance. Meanwhile, the study of unusual behaviors allows us to delve into the robustness of the Cloud. That is, bad performance at one data center can be alleviated by regular behavior from another; in other words, if the Cloud is correlated.

We apply factor analysis on these two components with factors such as the time of the day, the day of the week, the geographical area in which data centers are located, the CSP to which they belong, and the specific data centers involved in the measurement of a path. While the latter factor accounts for the peculiarities each data center or data center pair can have on behavior, other factors allow us to find generalities to explain the phenomenon; for example, how much of the bandwidth can be explained because a data center belongs to a given CSP, depending on the time of day. Similarly, when we study downtimes, such factor analysis shows whether data centers either in the same area or within the same CSP change their performance in unison, or whether performance changes in data centers individually. This exerts a clear impact on how to geo-distribute a deployment in the Cloud. For example, should all data centers decrease their performance at the same time, availability would be at risk; but if only data centers' performance with a specific CSP or in a specific area dip simultaneously, reliability can be found through other CSPs or areas.

In particular, we have applied the analysis of variance (ANOVA) as factor analysis. Interestingly, it shows that the stationary behavior of a path between data centers in the Cloud depends strongly on the CSP and the areas of the pair (source/destination) of data centers involved, and depends qualitatively much less on other factors such as the particular pair of data centers involved. This supports the generalization and extrapolation of the results shown herein, and even the use of smarter monitoring systems [15]. Similarly, other factors such as the time of day and the day of the week only showed moderate significance. That is, the Cloud is mostly insensitive to the time with some expanded capacity during weekends, providing a clear conclusion as to the impact on scheduling of bulk transfers in the Cloud [16].

On the other hand, we found no evidence of additional correlation of the bandwidth time series within areas or CSPs. In general, time series are weakly correlated and such correlation is only marginal, as explained by location and CSP factors. In this way, the correlation exhibited by some paths is mostly the result of particularities of the data centers involved. This points at simpler Cloud deployments, as data centers in the same area or within the same CSP may contribute in the search for availability and reliability equivalently to distant nodes of other CSPs. Unfortunately, this also implies that the generalization of this metric is challenging, as similar data centers behave differently, which calls for exhaustive and fine-grained monitoring.

We believe that a better understanding of the bandwidth dynamics in the Cloud is useful for most of the players in the Internet arena. For CSPs, as we provide them with a fair comparative description of bandwidth performance.

Also for practitioners, as they can find the description of the regular performance of the Cloud as a mechanism to choose between CSPs, areas and data centers when updating their deployments as well to plan future ones (for example, as parameter inputs for novel Cloud simulators [17]). And, importantly, the factorial study allows them to both understand why certain performance was achieved and estimate what can be expected if they modify their deployments. Similarly, conclusions about correlations can enable them to make better decisions in terms of both robustness and cost. Finally, for the research community, we believe that this work represents a further step along the path of characterizing the Internet, a task initiated by institutions such as RIPE and Caida years ago. For our part, we focus specifically on an important fraction of the Internet, the traffic within the Cloud (i.e, the wide-area traffic between CSPs' premises within the public Cloud, or, in other words, inter-cloud traffic), remarking on invariants that potentially lead to new models and ideas.

The rest of this work is organized as follows: Section 2 presents the foundations of the paper, including our testbed and how we measure and model bandwidth time series in the Cloud. Sections 3 and 4 present the core results of this work, where stationary and downtime processes are both described and discussed. Section 5 is devoted to a review of related work, and, finally, Section 6 concludes this paper by remarking on the main conclusions and pointing out some lines of future work.

2 Preliminaries

The goal of this section is to achieve a comprehensive characterization of Cloud bandwidth and its dynamics. In this light, the first two practical questions we pose are, first, for how much time a path between two data centers must be measured to obtain a significant sample of bandwidth; and, second, how to formally compare a number of bandwidth time series paths. Let us first explain the testbed used to answer these questions.

2.1 Testbed

We borrowed the approach presented in [18] and [19] where some virtual machines (VM) were started and measurements were gathered between them. Specifically, the authors in the former studied four CSPs, namely, Amazon EC2, Microsoft Azure, Rackspace Cloud, and GoGrid (the germ of Google Cloud). As of 2010, this set of CSPs comprised the two most popular ones and two promising newcomers. Currently, they have become the four dominant players in the Cloud arena [20]. In this work, the authors referred to CSPs as $C1 \cdots C4$ instead of using their names in an attempt to keep the focus on the conclusions rather than on very specific values. Following such approach, we study these same four major CSPs and use equivalent name terminology.

Specifically, in this set of CSPs, we started a VM in each of their eighteen data centers available at 2015, Fig. 1 places such data centers. Table 1 details different features (CSP, geographical area, and data center name) for each of them. In what follows, such features will be used as explanatory factors for bandwidth phenomena in the Cloud.

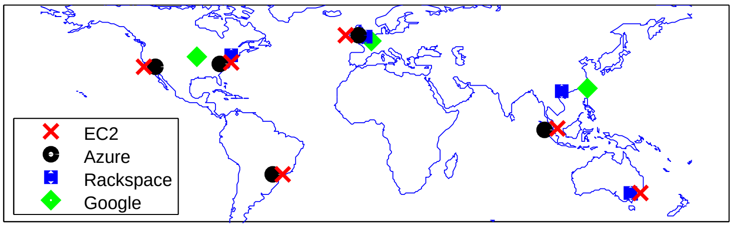


Figure 1: Multi-CSP and multi-data-center testbed deployment.

Table 1: Description of data centers under study.

CSP	Abrv.	Geo. area	Data center name
Amazon EC2	$C1$	Eastern US	Virginia$_{C1}$
		Western US	California$_{C1}$
		Northern Europe	Ireland$_{C1}$
		East Asia	Singapore$_{C1}$
		Australia	Sydney$_{C1}$
		South America	Sao Paulo$_{C1}$
Microsoft Azure	$C2$	Eastern US	Virginia$_{C2}$
		Western US	California$_{C2}$
		South America	Brazil$_{C2}$
		Northern Europe	Dublin$_{C2}$
		East Asia	Singapore$_{C2}$
Rackspace Cloud	$C3$	Eastern US	Virginia$_{C3}$
		Northern Europe	London$_{C3}$
		East Asia	Hong Kong$_{C3}$
		Australia	Sydney$_{C3}$
Google Cloud	$C4$	Central US	Iowa$_{C4}$
		Northern Europe	Belgium$_{C4}$
		East Asia	Taiwan$_{C4}$

To automate the process of monitoring the bandwidth between data centers, we developed *CloudB*: a scheduler and wrapper of other testing tools for Unix systems. In particular, it executes a set of provided tools with configurable duration and frequency between a list of IP addresses received as input. *CloudB* creates scripts for each testing tool observing the particular parameters for each of them. Then, it configures the task scheduler (UNIX's cron) with the requested frequency, and an auxiliary script to stop the tools execution according to the requested duration. *CloudB* has been successfully used with tools such as iPerf, hping3, wget, and Paris traceroute.

As we are interested in measuring bandwidth, our key measurement tool is iPerf [21]. It is a well-tested software program to measure the TCP[1] bandwidth between two nodes. Note that iPerf measures the TCP bandwidth and not the maximum capacity of the infrastructure [23]. This ties in with our approach to unveil the capacity a user can effectively achieve in the Cloud.

In measuring TCP bandwidth in a wide-area sense, both TCP flow-control and VM networking capacity (multiple tenants sharing a physical interface of a limited capacity) can act as undesirable bottlenecks for the goal of measuring

[1]Specifically, TCP Cubic as the default flavor for Unix-based operating systems beyond kernel 2.6.19, 2006 [22]

the Cloud capacity itself. Both [18] and [19] studied the former bottleneck. The send and receive windows sizes are set to 16 MB as larger window sizes did not result in higher measurements. Our initial tests confirmed this threshold as satisfactory.

The later bottleneck has not received so much attention as bandwidth outside data centers' infrastructure was supposed to be lower, at 1 Gb/s (often the virtualized VM-interface capacity for medium or large instances, but not necessarily for the best available VM flavors). Our initial tests suggested that this assumption is valid and, consequently, the VMs in our testbed were configured with at least 1 Gb/s of dedicated network capacity. The specific names for the flavors of VM used vary from one CSP to another (i.e., t2.xlarge in $C1$, A3 in $C2$, General1-8 in $C3$ and n1-standard-8 in $C4$), for further details we refer the reader to CSPs' datasheets. Otherwise, the total budget may reach prohibitive figures, as allocating better VMs (the best ones can even include hardware such as FPGAs) represented a cost increase. In fact, the measurements gathered throughout this paper have already incurred a cost of several thousands of dollars.

However, in such tests, we observed a few samples for a few paths were relatively closer to 1 Gb/s (but still several dozens of Mb/s below 1 Gb/s). For these cases, it is difficult to conclude if the measurements were limited by the TCP congestion control on the Internet (as desired), or by congestion on the virtualized VM interface. As a precaution against the latter, those paths with samples over 800 Mb/s were measured again with VMs resized by a higher level. Although it is worth remarking that these paths were less than 3% of the total number of paths. Specifically, this was done for an additional week after the measurement campaign was carried out.

The differences in the measured bandwidths were marginal (even, slightly better for cheaper units [19]). Later, the full measurement campaign confirmed this point (Section 3.1). That is, the conclusions of the forthcoming analysis did not change and the parameter estimates only varied in the least significant digits. Therefore, suggesting that our testbed is adequate for our purposes and that our initial concerns were excessive.

2.2 How to measure the bandwidth of a path

As TCP bandwidth oscillates with time, we focus on the duration of the process of gathering a bandwidth sample. In our extensive testbed, we started by measuring the TCP bandwidth between 100 randomly-chosen paths (this helped to keep the cost of the measurement campaign at reasonable figures). The duration was 15 minutes with 1-second granularity which could represent a download of 100-10 GB at a 1 Gb/s-100 Mb/s rate, which ties in with our scenario. Fig. 2 shows the results for four representative behaviors with 95% confidence intervals for the mean. They are representative in terms of spanning diverse coefficients of variation (CV) [24] for the bandwidth time series. The figure depicts CVs ranging from 0.09 to 0.46 (minimum and maximum values) and two intermediate examples.

Intuitively, the larger the CV, the more the duration of measurement must be aggregated. Fortunately, Fig. 3(a) proves that CVs tend to be small. Essentially, the empirical cumulative distribution function (ECDF) shows that all samples are below the rule-of-thumb threshold of 0.5 for significance of the mean, and

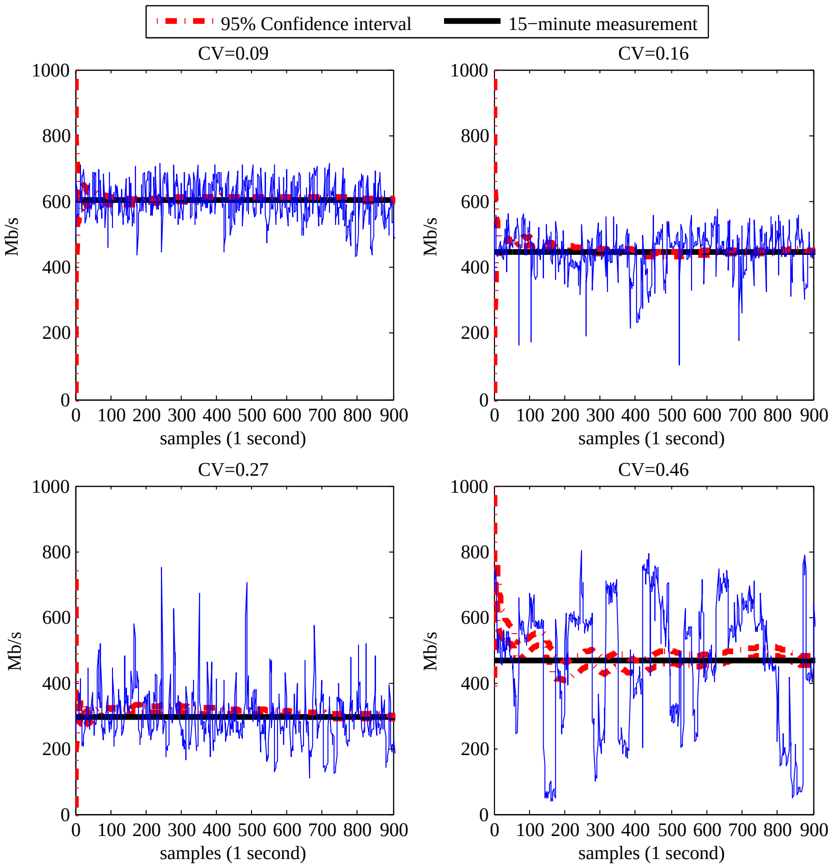
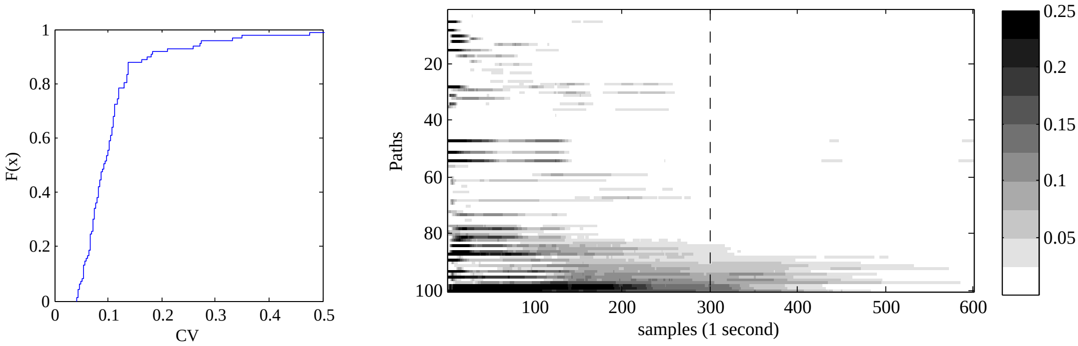


Figure 2: Examples of bandwidth time series behaviors according to their CV.



(a) Empirical cumulative distribution function for the CV of the sample.

(b) Error ratio according to aggregated times (paths in ascending order).

Figure 3: CV and duration for measuring bandwidth of paths.

more than 80% of the paths are below 0.2. This indicates that most of the paths' behaviors resemble the first examples.

Let us consider how this translates to time. Note that we seek the smallest interval of time (inter-cloud bandwidth is not cheap) that still provides a significant sample of the bandwidth a transfer achieves in the Cloud. By a *transfer in the Cloud*, we refer to tasks such as those carried out by backup and replication, distributed storage and search-indexing, or VM cloning among others tasks as the previous section remarked. In this light, to estimate the error in a shorter measurement, we consider as the ground-truth bandwidth of a sample its average throughput after the total duration of the measurement (as stated before, 15 minutes of aggregation).

In this way, Fig. 3(b) shows the error ratio per path as the difference between the average throughput after N seconds of aggregation (horizontal axis) and the average after 15 minutes. We use a progressive gray-scale where ratios larger than 0.25 are completely black. Assuming a ratio error threshold of 0.1-0.05, we note that more than 50% of paths needed more than one minute worth of data. Roughly 20% of paths required more than 200 seconds, and after 300 seconds only 5% of samples still showed ratios over 0.1 but were, in any case, below 0.15. As a trade-off between time aggregation and cost, we deem 5 minutes to be a good compromise. For the sake of completeness, we decided to compare the results using the median [25] instead of the mean. The results showed differences bounded by variations of 5% after 5 minutes of aggregation, making both metrics equivalent in practical terms.

2.3 How to model paths' behavior

To tackle the comparison of the bandwidth time series by means of a model, we measured another set of 100 randomly-chosen paths. This time, over three working days every other 15 minutes for 300 seconds, which translates into 288 samples in each time series. Fig. 4 illustrates with representative examples the four different patterns of behavior we found. Specifically, it shows the empirical probability density function for the samples that make up the bandwidth time

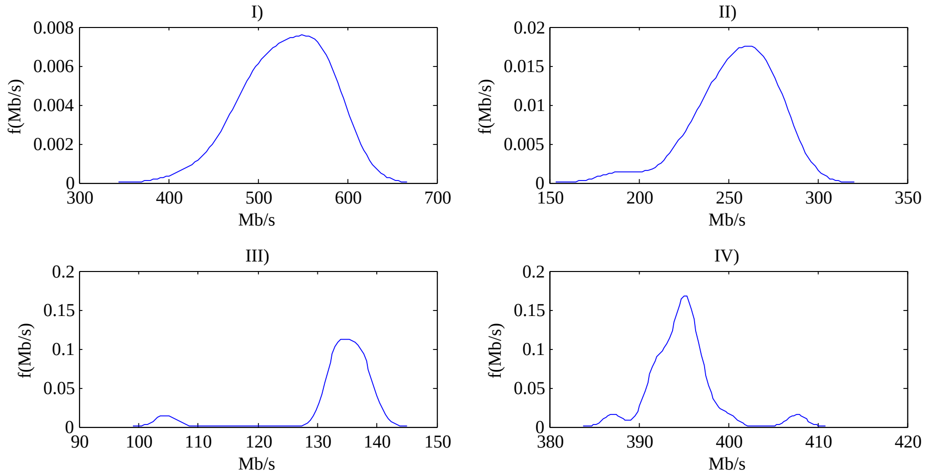


Figure 4: Examples of bandwidth time series behaviors.

series (after a Gaussian kernel softening process [26]).

The common denominator of all of these is a clear main component that is fairly Gaussian [27] with slightly negative skewness. Roughly 60% of the paths do not present any other component as in the first case of the figure. The other three cases illustrate a progressive increase in negative skewness (cases II and III) and the last one (IV), shows in addition a certain level of probability in the tail. We classify evenly the remaining 40% of paths into cases of types II and III, while less than 10% of cases exhibited a positive tail. Interestingly, we note that the percentage of samples falling into the Gaussian component for the full set of measured paths is over 85%. That is, most of the time the bandwidth time series follows a Gaussian distribution, oscillating softly over the mean; then, as an unusual event, bandwidth dips during a period (or, even more infrequently, it peaks). We will refer to the Gaussian component as the stationary behavior of the phenomenon, which represents what can be expected from a transfer in the Cloud from a path. On the other hand, excursions from the mean represent changes in regular behavior. Especially of interest is when a time series dips, to which we often refer as a downtime.

To formally split data into these two components, we simply apply a goodness-of-fit test on normality for the mean Gaussian component. In particular, we entrust Lilliefor's test at the 5% significance level with this task. Those samples that do not pass the test are considered as excursions.

2.4 The Measurement Campaign

So far, we have devised a testbed of eighteen data centers which, grouped in pairs, makes up 306 paths to measure. Also, some initial measurements allowed us to estimate how long to measure and how we can model bandwidth time series for subsequent comparisons. With these findings in mind, we executed *CloudB* for the full set of 306 paths parameterized with samples of 5-minute duration and sampling frequency of one sample per hour to balance cost and thoroughness [18]. Finally, the measurement campaign took place during the first three weeks of June, 2015.

In the rest of the paper, by elaborating on this resulting set of data, we describe the main Gaussian component of an extensive set of bandwidth time series in the Cloud and then relate such components to data centers' locations, CSPs, and peculiarities of specific data centers. Moreover, we assess how downtimes (and peaks) of bandwidth time series are correlated. The two following sections focus on these two issues.

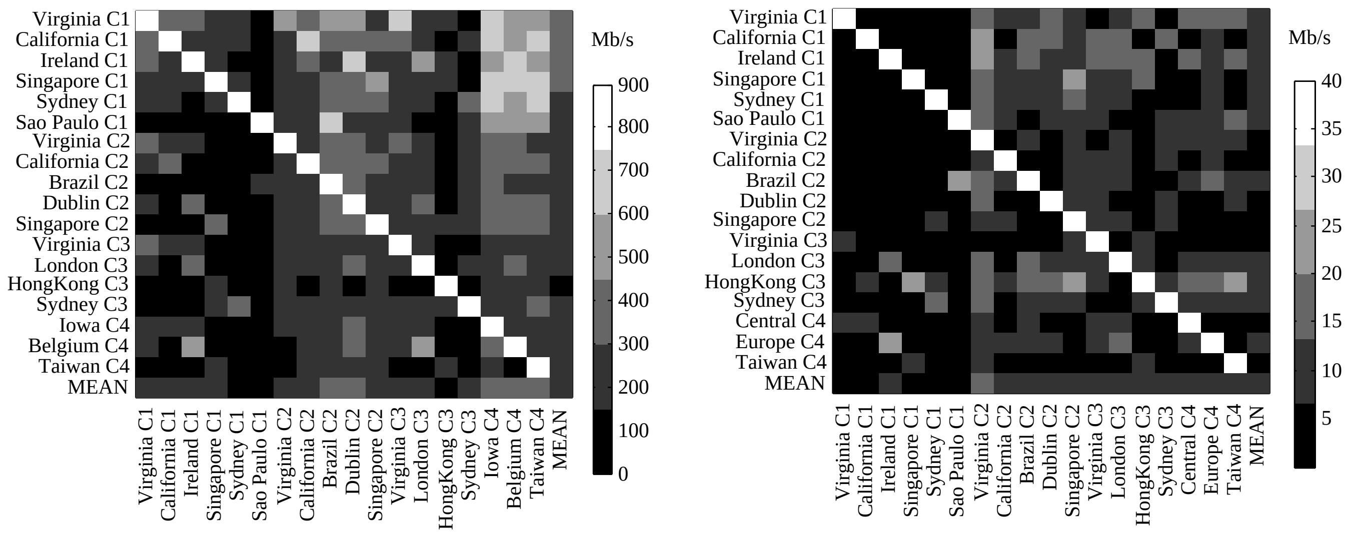
3 Stationary behavior

This section focuses on the stationary behavior of bandwidth in the Cloud. First, we pay attention to the mean of its principal Gaussian component. Then, we study the factors that explain this behavior. In other words, we first describe the performance and then generalize the results.

3.1 Overview

Figure 5(a) shows the TCP bandwidth mean for the 306 paths under study, as well as the mean by data center source (last column) and destination (last row), and overall mean in the Cloud (right- and bottom-most square). Next to it, Fig. 5(b) depicts the width of the 95% confidence interval for the mean to illustrate its significance. Several observations arise. On the large scale, the overall mean in the Cloud ranges between 150 and 300 Mb/s, specifically, roughly 250 Mb/s on average. However, the heterogeneity between different paths is clearly significant. On the one hand, some paths achieved rates close to 1 Gb/s, although none of the them achieved a mean over 750 Mb/s. In fact, no single measurement exceeded 930 Mb/s, which supports the idea that the capacity of the VMs in the testbed was adequate. On the other hand, several paths did not exceed a rate of 150 Mb/s. After a more detailed inspection, we assessed that all of the means surpassed 50 Mb/s.

We found some homogeneity by inspecting data centers as sources. Three data centers of $C1$ exhibit the best performance, with averages of roughly 400 Mb/s, whereas the data center of $C3$ located in East Asia is below 100 Mb/s. Regarding destinations, we identified more heterogeneity. $C2$ and $C4$ data centers stand out, and Brazilian and Australian data centers of $C1$ as well as the data center of $C3$ in East Asia obtained modest results. By observing the figure from a distance, some clusters (i.e., close squares with similar intensity) become apparent. To mention some of them: the good relationship between $C1$ and $C4$ (top right part of the figure), the good overall performance of $C2$ and $C3$ (columns 7-13), and the better performance of $C1$ outside its own infrastructure (rows 1-6, columns 7-18). To formally identify these interactions, next we apply factor analysis to the data. Regarding the confidence intervals, $C1$ shows narrow intervals which denotes flatter bandwidth time series. The other CSPs behave similarly with larger intervals, although even the broadest interval, Virginia_{C2} as data center destination, exhibits a width below 25 Mb/s.



(a) TCP bandwidth means. (b) Confidence intervals for the mean.

Figure 5: TCP bandwidth mean and its 95%-confidence-interval width per path between source (vertical) and destination (horizontal) data centers.

3.2 Factor analysis of bandwidth time series: Applying ANOVA to the data

We entrusted the analysis of variance (ANOVA) with factor analysis of the data [28]. ANOVA is a widely used statistical methodology whereby the observed variance of a given response variable is described in terms of explanatory factors, specifically as an addition of terms that account for the effect of such factors (or their interaction). In this way, ANOVA provides a mechanism to determine whether such factors have any statistical importance in explaining the response variable (if so, to what extent), and estimates the parameters for each of the different values a factor can take (levels) for each significant factor. The reader is referred to the appendix for further details and intuitions behind the application of ANOVA.

As a main objective of this paper, we pursue the identification of commonalities in the Cloud, allowing that the most particular factors participate only for variance that is not otherwise explained. Then, we propose to follow an ANOVA iterative approach whereby the most general factors are first used to explain variance, and the most specific ones are considered progressively (often named as ANOVA Type I).

Formally, in this study, general factors are the geographical location or area (factor $Area$) and the CSP (CSP) as Table 1 details, and also the interaction between a pair of well-connected areas (factor $Area^S * Area^D$) or CSPs (factor $CSP^S * CSP^D$). The interactions between factors allow us to evaluate whether the performance depends on the proximity of the source and destination ($Area$), whether connectivity is better inside CSPs' infrastructures (CSP) or, otherwise, whether each data center behaves somewhat independently. On the other hand, the most particular factors or interactions are those related to data center level:

factors that explain a time series according to the specific source data center (factor DC^S) or destination data center (factor DC^D), as well as factors that account for both ends (factor $DC^S * DC^D$).

In addition to these factors, we also add two intrinsic ones [29]. Specifically, the time (UTC) when the sample was gathered (factor $Time$) and the day of the week (factor $Weekday$). In fact, we consider these as the most general factors, as they apply to all the measurements. Note that they allow us to investigate whether the capacity in the Cloud is better at certain times and days of the week.

By ordering factors from more general to more specific, we construct the following ANOVA model:

$$\begin{aligned} BW_{twijki'j'k'p} = \mu + Time_t + Weekday_w + Area_i^S + CSP_j^S + DC_k^S + Area_{i'}^D + CSP_{j'}^D + DC_{k'}^D \\ + Area_i^S * Area_{i'}^D + CSP_j^S * CSP_{j'}^D + DC_k^S * DC_{k'}^D + \epsilon_{twijki'j'k'p} \end{aligned} \tag{1}$$

where $BW_{\bullet p}$ represents the p^{th} observation (a bandwidth sample) that results from the addition of terms according to the t^{th} level of factor $Time$ (e.g., 0 a.m., ... 12 p.m.), the w^{th} level of factor $Weekday$ (e.g., Mon, ... Sun), the i^{th} level of factor $Area$ as the source (e.g., a data center source is located in the Eastern US, Western US, ..., East Asia; see Table 1), the j^{th} level of factor CSP being the source (e.g., $C1$, $C2$, $C3$, and $C4$), the k^{th} level of factor DC being the source (e.g., Virginia_{C1}, California_{C1}, ... Taiwan_{C4}), and so forth for the levels for factors as destinations ($i'j'k'$). In addition, the two-way factors (represented by $*$) account for the impact that interactions exert: for example, the variance explained because of the source being in the Eastern US and the destination being in the Western US (including levels where both source and destination are the same). In more detail, an additional term for the interaction of i^{th} and i'^{th} levels of factor $Area$ (source and destination, respectively), for the interaction of j^{th} and j'^{th} levels of factor CSP (again, source and destination) and, finally, another for the interaction of k^{th} and k'^{th} levels of factor DC (also, source and destination).

In other words, BW represents each of the samples (p), samples that are indexed using t and w by the time and day of the week, as well as i, j, k and i', j', k' to index the geographical area, CSP, and data center, both source (S) and destination (D) of the path, respectively. The intercept term μ represents the overall mean response; that is, a constant figure over which the rest of the factors add terms. Finally, the difference between a sample p and the addition of the factor terms according to the levels of such a sample (often named the model value), is typified by $\epsilon_{ijki'j'k'p}$ (often named the random or experimental error). In our case, ϵ is linked to the variance of the Gaussian component, as Fig. 4 illustrated.

3.3 Results and discussion

Table 2 shows the results after applying the ANOVA test using SPSS [30] software over bandwidth time series measurements. First of all, the R^2 term is close to 1, so we can conclude that model explains the response variable (the bandwidth, in our case) with high accuracy. In this line, the last column exhibits the

Table 2: ANOVA table with $Time$, $Weekday$, $Area$, CSP, and DC (data center) and significant interactions as fixed factors, with bandwidth time series as the response variable ($\bar{R}^2$=0.92).

Factor	Sum of Squares	%Total	%Factors	df	Mean Square	F	p-value
μ	1009168371	49.2	·	1	1009168371	169459	0.00
$Time$	1105358	0.0	0.1	23	48059	8	0.01
$WeekDay$	7991054	0.4	0.8	6	1331842	224	0.00
$Area^S$	41652076	2.0	4.0	6	6942013	1166	0.00
$Area^D$	125808840	6.1	12.1	6	20968140	3521	0.00
CSP^S	80548490	3.9	7.7	3	26849497	4509	0.00
CSP^D	176277823	8.6	16.9	3	58759274	9867	0.00
DC^S	8274219	0.4	0.8	8	1034277	174	0.00
DC^D	40486106	2.0	3.9	8	5060763	850	0.00
$Area^S * Area^D$	192441930	9.4	18.5	35	5498341	923	0.00
$CSP^S * CSP^D$	125439547	6.1	12.0	9	13937727	2340	0.00
$DC^S * DC^D$	75897282	3.7	7.3	226	335829	56	0.00
Error	167088766	8.1	16.0	41491	4027		
Total	2052179862	100	100	41825			

p-value for the null hypothesis that supports the homogeneity of means; such a hypothesis can be rejected with significant confidence, i.e., all factors are able to explain some variance.

However, not all of them explain variance to a similar extent. The second column (sum of squares, in ANOVA terminology) represents the explained variance in absolute terms, while the third column shows these figures as percentages. This latter column can be interpreted as the percentages of variance that each factor or interaction helps to explain. To make the data easier to contrast, the fourth column shows the percentage for which each row accounts when the intercept term is not considered. Note that percentage in which the intercept term represents a significant fraction of the bandwidth of each path (the next section will estimate this term as roughly 40 Mb/s).

Next, we find that factors such as sources, $Area^S$ and CSP^S, depict modest significance in comparison to their destination counterparts $Area^D$ and CSP^D, which account for larger figures. This implies that the bandwidth is especially sensitive to where the path is destined. Moreover, the two-way factors involving pairs of $Area$ and CSP levels account for more than 30% of the explained variance apart from the intercept term.

So far, before even considering the data center ends, almost 88% of the total variance is explained. This highlights that the bandwidth behavior depends strongly on data center location and CSP rather than on specific behavior of data centers themselves. We speculate that this is due to the diverse routing agreements that govern the Internet, and to the well-known relationship of latency/TCP bandwidth. The closer two nodes are, the lower the latency and the higher the TCP bandwidth. The data center factors themselves explain less than 5% of variance, and only when both source and destination data centers are considered does this figure increase to roughly 6% of the total variance.

Finally, the factor $Time$ only exhibited marginal significance and no qualitative importance (less than 0.0%). This implies that the Cloud is almost insensitive to time; this is reasonable in the Cloud as it encompasses users and nodes widespread in the world. The $Weekday$ factor shows modest qualitative importance (0.4%), although, as will be shown, according to the parameter estimates, the difference between weekends and working days is clear. This contrasts with other commercial and academic networks with both strong daily and weekly patterns [31].

Given these results, the idea of extrapolating the bandwidth for non-measured paths based on factors makes sense. This is useful both for estimating performance of undeployed nodes and avoiding gathering measurements for each of the paths in a large deployment. This results in potential savings by reducing monitoring systems' costs.

3.3.1 Inspecting the levels

While the ANOVA table has unraveled how factors influence bandwidth time series, parameter estimates shed light on how their different levels interact. That is, ANOVA identified that the factors $Area$ and CSP can explain much of the bandwidth time series, but what specific values such factors take to increase or decrease the bandwidth is a matter of the parameter estimates. Intuitively, performance should peak when a path encompasses the same area and same CSP. Let us see to what extent this is true.

Table 3 summarizes the parameter estimates for the set of general factors μ, $Time$, $Weekday$, $Area$, CSP, and these two latter factors' interactions. The parameter estimates for such general factors allow us not only to discuss the interaction between levels, but their explicit inclusion provides the Internet community with an extrapolated set of data to consider in their research and commercial tasks with the appropriate adjustments to the particularities of each deployment in the Cloud. We do not include data center factors as there are 306 possible combinations, and because the addition of all the terms of the model is simply equal to the values depicted in Fig. 5(a).

The parameters for the factor $Time$'s levels unveil that the Cloud is practically insensitive to the time of day, with a minimum at 3 p.m. (UTC) and peaks at 2 a.m. and 11 p.m. It is difficult to relate this to properties of human activity as multiple time zones are involved; indeed, this is likely because the influence is low. By turning the focus on the day of the week, the homogeneity between two groups, working days and weekends, becomes apparent. The largest values (i.e., more capacity and intuitively less use of the Cloud infrastructure) are those from weekends with an additional term of about 30 Mb/s. It is worth remarking that Mondays behave similarly to weekends, likely because they include a fraction of the previous Sunday in some parts of the world at UTC time.

The $Area$ factor as a source and a destination represents an additional term of 85 Mb/s in some cases, among which the data centers located on both coasts of US stand out. On the other hand, the data centers in South America exhibit the lowest parameters. Regarding CSPs, $C4$ data centers share the largest capacities in both directions of traffic, and $C1$ and $C3$ show more moderate rates.

Some interesting conclusions arise out of the two-way factors, specifically when we pay attention to those rows where levels are equal (e.g., $C1*C1$). They represent the additional terms for paths whose ends are either in the same CSP or in the same area. The pairs within the same CSP do not show larger values; even for $C1$*$C1$ and $C4$*$C4$ pairs, the sum represented by such combinations is zero (i.e., no additional bandwidth because of such combinations). On the contrary, paths from $C1$ to $C4$ have an increment of more than 300 Mb/s,

Table 3: Parameter estimates for the bandwidth time series model.

Factor	Level	Mb/s
Intercept	μ	40
Time	0h	12
	1h	13
	2h	19
	3h	18
	4h	15
	5h	11
	6h	13
	7h	9
	8h	11
	9h	16
	10h	14
	11h	10
	12h	11
	13h	5
	14h	2
	15h	0
	16h	5
	17h	11
	18h	11
	19h	10
	20h	13
	21h	14
	22h	8
	23h	19
Weekday	Mon.	35
	Tu.	0
	Wed.	2
	Thu.	3
	Fri.	4
	Sat.	20
	Sun.	34

F	Level	Mb/s
$Area^S$	Western US	69
	Northern Europe	30
	Central US	43
	Eastern US	41
	Australia	55
	South America	0
	East Asia	20
$Area^D$	Western US	16
	Northern Europe	24
	Central US	35
	Eastern US	60
	Australia	23
	South America	0
	East Asia	40
CSP^S	$C1$	7
	$C2$	14
	$C3$	0
	$C4$	24
CSP^D	$C1$	0
	$C2$	75
	$C3$	49
	$C4$	78
$CSP^S * CSP^D$	$C1$ * $C1$	0
	$C1$ * $C2$	109
	$C1$ * $C3$	42
	$C1$ * $C4$	308
	$C2$ * $C1$	35
	$C2$ * $C2$	62
	$C2$ * $C3$	6
	$C2$ * $C4$	85
	$C3$ * $C1$	28
	$C3$ * $C2$	30
	$C3$ * $C3$	30
	$C3$ * $C4$	27
	$C4$ * $C1$	71
	$C4$ * $C2$	61
	$C4$ * $C3$	49
	$C4$ * $C4$	0

F	Level	Mb/s
$Area^S * Area^D$	Western US * Western US	368
	Western US * Northern Europe	110
	Western US * Australia	109
	Western US * Central US	137
	Western US * Eastern US	123
	Western US * South America	114
	Western US * East Asia	0
	Northern Europe * Western US	124
	Northern Europe * Northern Europe	337
	Northern Europe * Australia	94
	Northern Europe * Central US	125
	Northern Europe * Eastern US	157
	Northern Europe * South America	130
	Northern Europe * East Asia	0
	Central US * Western US	164
	Central US * Northern Europe	145
	Central US * Australia	67
	Central US * Eastern US	182
	Central US * South America	141
	Central US * Asia	0
	Eastern US * Western US	169
	Eastern US * Northern Europe	177
	Eastern US * Australia	93
	Eastern US * US	119
	Eastern US * Eastern US	365
	Eastern US * South America	186
	Eastern US * East Asia	0
	Australia * Western US	57
	Australia * Northern Europe	57
	Australia * Australia	275
	Australia * Central US	98
	Australia * Eastern US	57
	Australia * South America	73
	Australia * East Asia	0
	South America * Western US	132
	South America * Northern Europe	120
	South America * Australia	122
	South America * Central US	140
	South America * Eastern US	136
	South America * South America	433
	South America * East Asia	0
	East Asia * Western US	16
	East Asia * Northern Europe	0
	East Asia * Australia	12
	East Asia * Central US	8
	East Asia * Eastern US	20
	East Asia * South America	10
	East Asia * East Asia	105

although the increment in the inverse direction is much lower. In summary, we conclude that the pair of CSP involved in a transfer is important but, somewhat counter-intuitively, not because data centers belong to the same CSP.

The impact of the two-way $Area$ factor is just the opposite. The terms to add when two data centers are in the same area are large figures: specifically, they range from 105 to 433 Mb/s (East Asia and South America, respectively). In between, the rest of the areas connect internally with an addition term of about 300 Mb/s. By inspecting the cross relationships, it becomes apparent that East Asia is in some way isolated. Terms that represent paths to/from East Asia hardly have any additional bandwidth, while the other cross terms (connections between areas apart from East Asia) are above the range of 100 Mb/s. In conclusion, the bandwidth-delay product seems to play an import role in the Cloud, as proximity between data centers boosts performance.

Finally, regarding data center factors, we note that in all cases the additional term that represents such factors is lower than 100 Mb/s and is typically closer to zero than the 100 Mb/s rate. After ordering these factors by impact, pairs of data centers that stand out are those in Ireland and pairs of data centers where one is located in the Central US and the other is on one of the US coasts. In these cases, a term only slightly below 100 Mb/s is added.

3.3.2 The impact of routing in the findings

The rationale behind these findings may be explained by the routing policies and agreements each CSP follows. Intuitively, if CSPs tend to use the same or equivalent transit providers (ISP) regardless if source or destination belong to certain CSP, the fact that bandwidth between data centers of a same CSP does not show better results becomes coherent.

To assess this, we have analyzed one day of traceroute measurements gathered in parallel to the bandwidth measurement campaign, and extracted the transit hops of paths (i.e., those hops that interconnect the exit and enter of the CSP infrastructure where data centers are placed). Note that we did not consider $C2$ here as it blocks ICMP packets. Then per each path, we have manually identified if the transit hops belong to the own CSP infrastructure (working as a traffic carrier) or to public ISPs. To do so, we have exploited public database where IP addresses, Autonomous Systems and locations are related [32].

The results of this analysis are summarized in Table 4, where for each pair of data centers is shown the ISPs that carried the traffic between them (in bold, when the carrier is a CSP). To give an example, the route between Ireland$_{C1}$ and Sydney$_{C1}$ used the infrastructure of Level3 and NTT. Interestingly, we found significant heterogeneity. C_4 has an extensive transit infrastructure: C_4 carries all its internal traffic, most of its incoming traffic from other CSPs, and a significant fraction of its outgoing traffic to other CSPs). In contrast, C_1 only carries a few paths (partially, those involving Virgina, California, London and Dublin), and C_3 uses public ISPs.

The immediate conclusion is that the homogeneity between inter and intra CSP traffic, previously founded, cannot be explained because of the use of similar routes by CSPs. Most likely, it is because those CSPs that own infrastructure (C_4 and C_1) did not give priority to internal traffic with respect to traffic sourced/destination to other CSPs. This may also suggest that the differences in terms of infrastructures, equipment or load between public transit

Table 4: Transit ISPs used for paths per source/destination (vertical/horizontal) data-centers (asterisks represent marginal contribution as carrier, and ? several non-responding routers in the path).
Cn represents CSPs working as carriers (in bold). CTL is CenturyLink. EQ is Equinix. GTT is Global Telecom & Technology's Tinet. L3 is Level 3 Communications. NTT is Nippon Telegraph and Telephone. TAT is Tata Communications. TEF is Telefonica Wholesale. TIT is Telecom Italia's Seabone. TLS is Telstra global including Asia Netcom. TLSN is Telia Company. TPG is TPG Telecom. VOC is Vocus Communications. ZAY is Zayo Group including AboveNet. OT means other carriers.

		$C1$					
		Virginia	California	Ireland	Singapore	Sydney	Sao Paulo
$C1$	Virginia		**C1**	**C1**	NTT	NTT	L3
	California	**C1**		NTT	NTT	TLS	NTT/TIT
	Ireland	**C1**	GTT		**C1** */NTT	L3/NTT	GTT/TEF
	Singapore	NTT	NTT	TLS/GTT		NTT	NTT/TIT
	Sydney	TLS/CTL	NTT	TLS/GTT	NTT		TLS/TEF
	Sao Paulo	TEF/NTT	TIT/TEF/OT	TEF/NTT	TIT	TEF/TLS/OT	
$C3$	Virginia	**C1**?	**C1**?	L3	TAT	TEF/TLS/OT	TIT
	London	ZAY	L3	**C1**	TAT	TAT	TEF
	Hong Kong	TLS	TLS	TLS	TLS	TLS	TLS
	Sydney	TLS	VOC	TLS	VOC	TPG	TLS
$C4$	Iowa	**C4**/L3	**C4***/GTT	**C4***/L3	**C4**/OT*	CT/TLS	**C4**/TIT
	Belgium	**C4**/ZAY	**C4**/GTT	**C4**/**C1**	**C4**/OT*	**C4**/OT	**C4**/TIT
	Taiwan	**C4**/OT*	**C4**/GTT*	**C4**/L3	**C4**/TLS	**C4**/OT*	**C4**/TIT

		$C3$				$C4$		
		Virginia	London	Hong Kong	Sidney	Iowa	Belgium	Taiwan
$C1$	Virginia	**C1**	ZAY	TLS	OT/TPG	**C4**	**C4**	**C4**
	California	**C1**	TLS	TLS	VOC	**C4**	**C4**	**C4**
	Ireland	TLSN	**C1**	**C1***/TLS	**C1***/TLS	**C4**	**C4**	**C4**
	Singapore	NTT/L3	NTT/L3	TLS/EQ	TLS/EQ*	**C4** /OT	**C4**/EQ*	**C4** /OT
	Sydney	TLS/TAT	NTT/L3	TLS/EQ*	TPG	**C4**/EQ*	**C4**	**C4**
	Sao Paulo	TEF/TAT	TEF/TAT	L3	TIT/TLS	**C4**/OT*	**C4**/OT*	**C4**/OT*
$C3$	Virginia		TLSN/L3	TAT	VOC	L3	L3	L3
	London	L3/TLSN		TLS	TAT/TPG	**C4**	**C4**	**C4**
	Hong Kong	TAT	TLSN/L3		TLS	C4	**C4**	C4
	Sydney	VOC	TAT	TLS		VOC	**C4**	VOC*/**C4**
$C4$	Iowa	**C4***/L3	**C4**	TLS	TLS		**C4**?	**C4**?
	Belgium	TLS	**C4**	**C4**/OT*	**C4**/TLS	**C4**?		**C4**?
	Taiwan	**C4**/L3	**C4**	**C4**/OT	**C4***/TLS	**C4**?	**C4**?	

and CSP transit are little. Although such difference exists as the cells CSP^S and CSP^D in Table 3 indicate that $C4$, which exploits its own infrastructure, has extra terms simply because a path starts/ends on it.

4 Correlation between paths

We turn our attention to the correlation between the full set of measured paths. In other words, while the previous section focused on the principal Gaussian component in a stationary viewpoint, we now focus especially on the excursions. To this end, we have calculated the Pearson correlation coefficient (ρ) [24] between all the bandwidth time series (respecting the synchronism between signals).

Should excursions occur at the same time, ρ is significant; otherwise, the coefficient tends to be close to zero, indicating that bandwidth changes independently between paths. We note that a positive ρ implies that two time series change in the same way. That is, when one path performance peaks, the other does the same, and when one works poorly the other does, as well. On the other hand, non-significant coefficients and, especially, negatives ones, signify paths that can provide availability and reliability to others in the Cloud. Essentially, a downtime in a path can be resolved by good performance, or at least regular performance, of an alternative path.

In this light, we pose two questions: (i) whether or not the performance between data centers in the same area is correlated, which would support the approach of achieving redundancy by disseminating data in different areas; (ii) whether data centers in the same CSP are correlated, which would entail additional advantages (apart from lower latencies [6]) for multi-CSP deployments. To answer these questions, we follow an equivalent approach to that in the previous section.

4.1 Data analysis

Our testbed is comprised of 18 data centers with 306 paths between them, which in turn translates into 93,330 possible coefficients: one per each possible pair of paths in the set of 306 paths.

To provide an overview, Fig. 6 shows ρ as an ECDF for the full set of coefficients. Considering that values between -0.25 and 0.25 as uncorrelated, the figure shows that about 80% of the pairs of paths are not correlated. Roughly 10% of the tests yielded clearly positive correlations; for instance, figures lower than -0.5 and higher than 0.5. As an attempt to identify those paths, Fig. 7(a) shows the mean of ρ per each path indexed by source and destination data centers, and Fig. 7(b) depicts their corresponding 95% confidence interval widths.

For example, given the path from Virginia$_{C1}$ to California$_{C1}$, the cell (1,2) represents the average ρ of this path relative to the rest of the paths. In turn, the last column and row show average ρ by source and destination data center, respectively. In the figure, the coefficients are separated into five classes that we relate to strong correlation, significant correlation (both in the negative and positive directions), and marginal/no correlation. In general, these initial results suggest that the correlation is low and slightly positive.

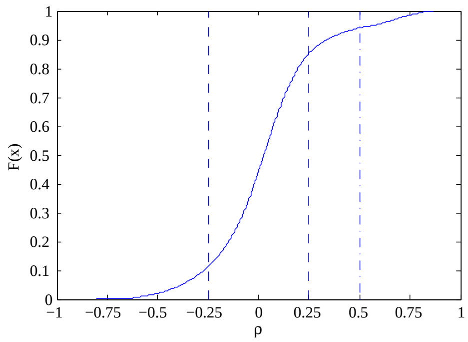
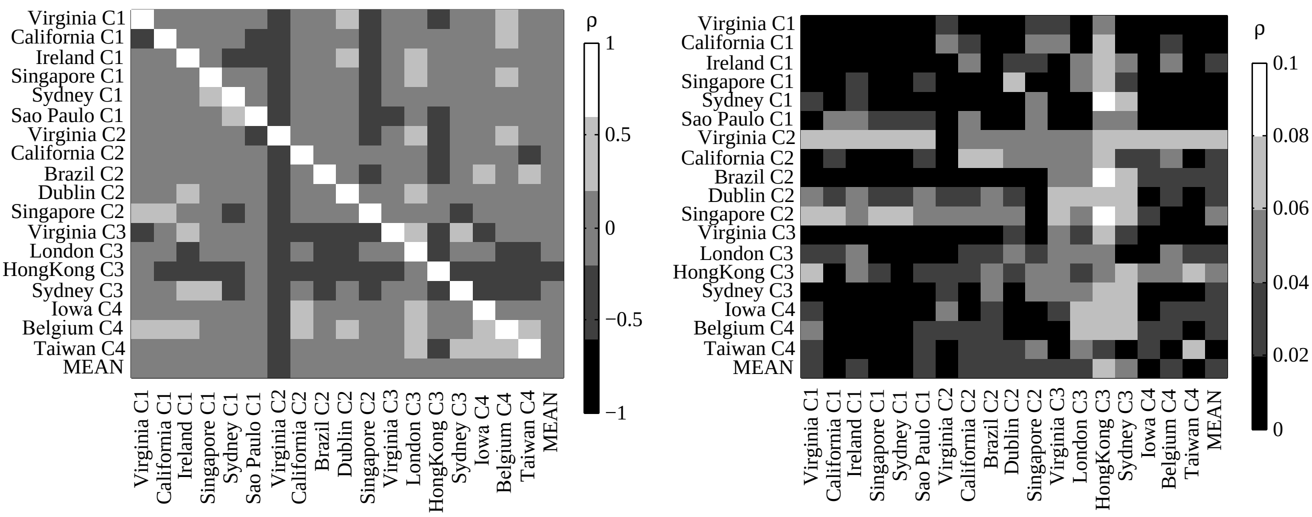


Figure 6: Empirical cumulative distribution function of ρ for all pairs of paths under study.



(a) Average ρ per path.

(b) Confidence-interval widths for averaged ρ.

Figure 7: Average ρ and its 95%-confidence-interval width per path (vertical: source data center, horizontal: destination data center).

By paying attention to the last column, on average it is apparent that none of the data centers stands out as particularly sensitive to other path changes. That is, in the larger picture there is no evidence that bandwidth changes in unison in the Cloud. However, the same figure shows that some paths had correlations, and importantly, the effect of averaging may have hidden details. Factor analysis highlights what the pair of correlated paths has in common. We note that a full-factorial approach to this problem would include 4-way factors, making the model hardly tractable (i.e., interaction of four factors, specifically both the source and destination of the path under study, and also both source and destination of the path to be compared). However, our focus is to shed light on whether a data center may expect correlation between the set of paths it can potentially use. That is, if a path in use is working poorly, another path can compensate for it, and this is only possible if the source of the paths share the same node. In other words, we study the correlation between paths in which the sources are at the same node in contrast to comparing all paths including disjoint ones.

We then simplify the problem, resulting in the following model:

$$\begin{aligned}
CORR_{ijki'j'k'i''j''k''p} = {} & \mu + Area_i^S + CSP_j^S + DC_k^S + Area_{i'}^{D1} + CSP_{j'}^{D1} + DC_{k'}^{D1} + Area_{i''}^{D2} + CSP_{j''}^{D2} + DC_{k''}^{D2} \\
& + Area_i^S * Area_{i'}^{D1} + Area_i^S * Area_{i''}^{D2} + Area_{i'}^{D1} * Area_{i''}^{D2} \\
& + CSP_j^S * CSP_{j'}^{D1} + CSP_j^S * CSP_{j''}^{D2} + CSP_{j'}^{D1} * CSP_{j''}^{D2} \\
& + DC_k^S * DC_{k'}^{D1} + DC_k^S * DC_{k''}^{D2} + DC_{k'}^{D1} * DC_{k''}^{D2} \\
& + Area_i^S * Area_{i'}^{D1} * Area_{i''}^{D2} + CSP_j^S * CSP_{j'}^{D1} * CSP_{j''}^{D2} + DC_k^S * DC_{k'}^{D1} * DC_{k''}^{D2} \\
& + \epsilon_{ijki'j'k'i''j''k''p} \qquad (2)
\end{aligned}$$

where the same terminology as in Equation 1 is followed, apart from the fact that we are modeling correlation coefficients ($CORR$) and that we are now considering three data center ends: a source data center (S) and the two destination data centers ($D1$ and $D2$). Then the correlation coefficient is calculated between the paths S-$D1$ and S-$D2$, for all possible combinations of data centers (or levels), where $'$ and $''$ are used to index levels within the two destination factors, respectively. Finally, the 2-way factors account for the correlation to add because of interactions of data center levels taken in pairs, and the 3-way factors for additional terms because of the combinations of levels as trios of data centers.

4.2 Results and discussion

Many observations arise by inspection of the percentage of variance that each factor explains, as in Table 5 (third column). First, the $Area$ and CSP as destinations are not significant whereas these factors understood as sources are clearly significant. This implies that certain areas and CSPs are less sensitive to changes in the Cloud (i.e., more robust).

However, the explained variance of 2-way and 3-way factors that involve $Area$ and CSP factors, some of which are statistically significant, is qualitatively marginal. In practical terms, we have found no evidence of that data centers within the same area or CSP present different correlations between them with

Table 5: ANOVA table with $Area$, CSP, DC, and significant interactions as fixed factors, and the correlation coefficient per pair of paths as the response variable ($\bar{R}^2$=0.99).

Factor	Sum of Squares	%Total	df	Mean Square	F	p-value
μ	51.9	9.7	1	51.9	752	0.00
$Area^S$	20.2	3.8	6	3.4	49	0.00
$Area^{D1}$	0.8	0.2	6	0.1	2.0	0.06
$Area^{D2}$	0.9	0.2	6	0.2	2.3	0.35
CSP^S	89.0	16.6	3	29.7	430	0.00
CSP^{D1}	1.6	0.3	3	0.5	7.8	0.00
CSP^{D2}	1.8	0.3	3	0.6	8.9	0.00
DC^S	40.9	7.6	8	5.1	74	0.00
DC^{D1}	2.6	0.5	8	0.3	4.6	0.00
DC^{D2}	2.9	0.5	8	0.4	5.3	0.00
$Area^S * Area^{D1}$	4.6	0.9	35	0.1	1.9	0.00
$Area^S * Area^{D2}$	5.3	1.0	35	0.2	2.2	0.00
$Area^{D1} * Area^{D2}$	8.3	1.5	35	0.2	3.4	0.00
$CSP^S * CSP^{D1}$	9.3	1.7	9	1.0	15	0.00
$CSP^S * CSP^{D2}$	10.5	2.0	9	1.2	17	0.00
$CSP^{D1} * CSP^{D2}$	4.1	0.8	9	0.5	6.6	0.00
$DC^S * DC^{D1}$	40.7	7.6	227	0.2	2.6	0.00
$DC^S * DC^{D2}$	56.6	10.6	236	0.2	3.5	0.00
$DC^{D1} * DC^{D2}$	40.7	7.6	236	0.2	2.5	0.00
$Area^S * Area^{D1} * Area^{D2}$	12.1	2.3	197	0.1	0.9	0.85
$CSP^S * CSP^S * CSP^S$	3.3	0.6	27	0.1	1.8	0.01
$DC^S * DC^{D1} * DC^{D2}$	141.8	26.5	3807	0.0	0.5	0.00
Error	0.0	0.0	0			
Total	535.5	100.0	4896			

respect to other data centers in the Cloud. In this way, the search for availability and reliability (note that latency or other QoS metrics are a different matter) in the Cloud can be achieved in the same manner by close and far-away data centers, and is only moderately improved by spreading data between different CSPs rather than relying on the same CSP.

On the other hand, the factors and interactions related to data centers are qualitatively very significant. In conclusion, each pair of paths is correlated on a data center basis with little contribution from the pair of locations and CSPs involved. The impact of this on current or future deployments arises in terms of simplifying reliability. Unfortunately, it also makes it more complex to infer the interactions between paths based on general factors, as we did with the bandwidth, as, in this case, pairs of similar paths behave heterogeneously. In general, however, the correlation between paths is relatively low in the Cloud, as the figures 6 and 7 proved.

Let us now pay attention to the parameter estimates of the subset of quantitatively significant factors or interaction of them. The parameter estimates for the factors $Area^S$ and CSP^S are shown in Table 6 as the only general factors with some qualitative importance. We found that both US costs present the smallest measured ρ, which translates into greater robustness to overall changes in the Cloud. Additionally, Australia and Northern Europe showed low estimates for ρ, while East Asia and Central US areas exhibited an additional term of more than 0.3. That is, these latter areas comprise data centers more sensitive to oscillation of bandwidth capacity in the Cloud, on the whole. Similarly, while

Table 6: Intercept, $Area^S$, and CSP^S parameter estimates in the correlation parameter model.

Factor	Level	ρ	Factor	Level	ρ
Intercept	μ	0.10	CSP^S	$C1$	0.27
$Area^S$	Asia	0.37		$C2$	0.26
	Brazil	0.02		$C3$	0.38
	Western US	0.00		$C4$	0.00
	Northern Europe	0.09			
	Central US	0.32			
	Eastern US	0.00			
	Australia	0.09			

CSP $C4$ presented an estimated additional correlation of zero when transfers are sourced in one of its data centers, $C3$ showed the largest positive ρ figure.

Finally, the parameter estimates for data center, pairs of data centers, and source and pairs of destination data centers, span several thousands of parameters. In the case of factor $DC^S * DC^{D1} * DC^{D2}$ involves, potentially, $18 \cdot 17 \cdot 17$ possibilities which renders its enumeration difficult. Moreover, note that the relevant finding here is not the set of specific values for each possible ρ given the significant heterogeneity founded in these levels. This calls for a fine and tailored monitoring for particular paths of interest in each deployment.

Even so, the most significant parameter estimates, after inspecting the datacenter factors and their interactions, are the low ρ for data centers within Ireland and the high estimates for Belgium_{C4} and Hong Kong_{C3}. Also the low correlation between data centers located in Virginia, and a generally high correlation of data centers located in Singapore stand out.

5 Related work

In this section, we first review studies that measured the bandwidth of infrastructures comparable to the public Cloud. Then, studies that reported some Cloud measurements or focused on measuring the Cloud from different points of view (among other metrics, VM flavors, storage capacities, latency, popularity, and CPU) are reviewed.

Regarding other infrastructures, the authors in [33] measured the bandwidth between most of the nodes of the Planetlab platform. Planetlab is a distributed federation where members can take control of probes across the world to test any novel algorithm or idea. In this case, the authors leveraged the platform to measure the bandwidth between different parts of the world: specifically, 250 probes were deployed and their connectivity tested. They entrusted this task to Pathrate, which is tool that estimates the maximum capacity of a path and not the available TCP bandwidth, as we devised. Unfortunately, they found a set of stricter bandwidth limits on probes that polluted the measurement campaign. That is, some of the measurements are simply bounded by limits without any link to the real capacity between end nodes. This justifies the careful selection of VM probes in our testbed. Apart from the capped samples, they found that paths typically range between 80 and 120 Mb/s, clearly below our measurements in the Cloud.

With the final aim of finding overlay routes in the Cloud to improve the quality of multimedia services, the authors in [34] showed some figures regarding TCP throughput measurements in the Cloud. Their solution was to disseminate contents but, importantly, with the constraint of not increasing costs. To do this, they limited the throughput of the processes to distribute data to the original 95^{th} percentile of the network (typically, the metric used to charge CSP). To test their ideas, they measured the bandwidth between data centers in Amazon's infrastructure (i.e., remote data centers of a CSP, here equivalent to intra-cloud). Specifically, between data centers in Northern California, Oregon, Virginia, Sao Paulo, Ireland, Singapore, and Tokyo. This set differs from ours in several data centers with Amazon apart from all the data centers of the other CSPs we study. They measured bandwidths for 3 minutes per path, but they did not specify which tool was used nor any other details, limiting the utility of comparison. In general, they found lower bandwidths than in our study.

Amazon's cloud infrastructure was also studied in [35], a paper that spans metrics such as latency and CPU performance using benchmarks. In contrast, in [36], such infrastructure was examined from the perspective of end users. Specifically, they found that many current deployments on Amazon are not yet exploiting the geo-distribution of contents for better quality of service. As an example, users from Italy are mostly served by the data center located in Virginia instead of its counterpart in Ireland, so the perceived quality for end users in Italy was considered poor.

Paying attention to the careful selection of probe capacities, the authors in [37] studied the impact of virtualization on TCP throughput at the data centers of Amazon located in the Eastern US. They found that virtualization exerts an impact when several VMs share the same CPU, as the sender is periodically taken out of the CPU and throughput decreases. To reach this conclusion, they compared at low-scale the TCP throughput of both small and medium instances. After this experiment, they found that a small number of instances achieved up to 500 Mb/s of TCP throughput, whereas medium probes achieve almost 900 Mb/s.

Similarly, the authors in [38] also paid attention to the relevance of VMs' capacities focusing on intra-region measurements (here equivalent to intra-data-center). Specifically, they measured inside Azure's data centers and, then, turned their attention to data centers of EC2 [25]. They reported bandwidth rates ranging from some hundreds to one thousand Mb/s, depending on the VMs' flavors. Interestingly, the authors remarked that short-duration captures of TCP bandwidth are useless, and defined thresholds of 5 and 8 minutes using the mean and median, respectively. We have checked that both mean and median present similar results for intervals of 5 minutes worth of data, and, in practice, their intervals are equivalent to ours.

For their part, the authors in [19] studied how to systematically pick a combination of CPU and VMs flavors to match the requirements of applications in the Cloud. They considered that bandwidth is one of the possible requirements an application must meet, and measured the bandwidth between different VM flavors at Amazon EC2 both inside and outside the same data center. Their results concluded that medium and large VMs achieve bandwidths close to 1 Gb/s when both ends are in the same data center. On the other hand, intra-cloud bandwidths exhibited far lower bandwidths typically bounded by 200 Mb/s. We found larger bounds for EC2, in general, 300 Mb/s, well below the capacity of the VM interface.

Then, the technical report this paper is based on was released [39] and some other Cloud measurement analyses have been carried out. Interestingly, this illustrates the still-increasing attention of the research community in the bandwidth in the Cloud. In particular, the authors in [40] extended their intra-region approach to include also intra-cloud measurements, specifically 12 paths between EC2 data centers and other 12 paths inside Azure's infrastructure. Several points arise. They empirically stated that assuming intra-cloud measurements, the impact of VM sizes, beyond certain level, can be considered negligible, differently from the intra-data center case also, previously, studied by them [38]. This level is below the VM sizes we used in our inter-cloud measurement campaign. In contrast with our approach, they measured both TCP and UDP capacities. Their results for TCP show lower numbers, none of the measured paths exceeded 300 Mb/s on average whereas samples over 800 Mb/s were gathered in our campaign. However, it is worth remarking that such samples often involved inter-cloud paths to/from Rackspace and Google Cloud that were not measured in their study. By comparing, specifically, results inside EC2 and Azure, in the case of EC2 they are fairly similar but in the case of Azure, the differences are still significant. Regarding UDP, they found far higher values with respect to TCP for both EC2 and Azure, i.e., with means over 600 Mb/s. Finally, as a later extension [41], the authors further studied the intra-cloud latency (not highly correlated with bandwidth but geographical distance), the impact of the availability zones inside regions (concluding that such impact is low), and the routing between EC2 data centers. Regarding this latter issue, we found similar conclusions reinforcing the discussion of Section 3.3.2 for the set of four CSPs we are studying.

An extensive effort to benchmark the Cloud was presented in [18]. The authors presented the results of applying the monitoring tool CloudCmp with several CSPs in 2010. They focused significant attention on metrics such as storage performance, CPU capacity, latency between different CSPs, and intra-data-center networking. However, they did not study the inter-data-center bandwidth between different CSPs across different areas. Specifically, they paid attention to pairs of data centers located in the US and belonging to the same CSP. They reported bandwidths ranging between 100 Mb/s and 500 Mb/s. Finally, as a further step in the same direction, the authors in [13] studied metrics such as CPU capacity, memory latency, java benchmarks [42] and disk I/O capacities following a more systematic approach after a deep measurement campaign.

In this paper, we have followed this later benchmarking-study so extending our study to a diverse set of data centers from several geographical areas and with various CSPs in attempt to provide general results for the public Cloud. But as a distinguishing characteristic from it, we note that we focused on the bandwidth in terms of available TCP throughput, as an approximation to the quality perceived by transfers, and, as a novelty, in both intra and inter-cloud scenarios. Moreover, we shed light on how to characterize the Cloud in order to provide not only a significant description but also a model for the phenomenon. Indeed, by leveraging a full-factorial approach (not so common in the Internet community), we were able to provide a novel comparison between data centers' behaviors based on specific factors (especially general ones). These factors revealed interesting interactions otherwise hidden in the data. Specifically, they explain why and how changes on Cloud deployments may impact on their performance instead of only describing such performance.

Finally, to the best of our knowledge, the study of the correlation of bandwidth time series has not been previously addressed despite its clear link to availability and reliability in the Cloud.

6 Conclusions and future work

Throughout this paper, we have shed light on the bandwidth within the public Cloud. We believe that we have provided Cloud customers with formal indications and generalizations of what they can expect when opening a TCP connection in their deployments and, importantly, both in terms of regular and malfunction behaviors.

That is, we have proposed a two-component approach to model time series of bandwidth in the Cloud. These components account for, on the one hand, the stationary behavior of data centers (i.e., how data centers tend to work) and, on the other hand, unexpected excursions (typically downtimes, but also peaks). While the importance of stationary behavior is immediate, the study of downtimes is no less transcendental. The question of whether changes in bandwidth capacity across the Cloud tend to be a synchronized process or if, conversely, data centers operate on their own has a dramatic impact on the robustness of any deployment in the Cloud.

After applying factor analysis in an extensive testbed, we concluded that the behavior of the stationary component can be considered as a homogeneous phenomenon, whereas the same does not apply to the other component. That is, while most of the bandwidth time series' variance can be described according to the locations and CSPs involved as source and destination in a transfer, the correlation between such time series (where the excursions are especially important) is a process that depends qualitatively on the specific data center ends involved.

By examining the specific parameters that increase the bandwidth of paths, we have found that transfers inside the same geographical area receive a significant additional term while the same does not apply inside CSPs. In more detail, the source CSP is significant in explaining the bandwidth of a path, but a path that involves the same CSP as both source and destination is of only marginal importance. As other interesting conclusions, we have found that the time of day during which the Cloud is measured is of little significance and that the day of the week is only marginally significant. That is, behavior differs from weekends to working days but not between days in these sets. This has an impact on how bulk transfers should be scheduled in the Cloud.

The study of correlations also exerts a direct impact on the planning of Cloud deployments. Fortunately, the low correlations found and the moderate significance of CSP and location data center interaction factors makes the achievement of robustness easier, as close data centers and data centers inside same CSP did not show greater correlation that those located far away or belonging to other CSPs. However, the peculiarities of data centers play an important role, so the generalization of results is more limited in respect of specific data centers. Consequently, continuous fine-grained monitoring at multiple probes is needed to estimate this metric.

We believe that all these lessons learned and measurements reported and made available upon request are of interest to both practitioners and researchers. The measurements and their descriptions are useful in making decisions for current and future deployments, and even to extrapolate figures to non-measured data center pairs with a certain confidence. Similarly, we see these conclusions and the identification of invariants as a step toward better knowledge of the dynamics of the Cloud and, consequently, of the Internet.

As future work, we plan to carry out a formal evaluation of error entailed by the extrapolation of measurements to non-monitored data centers. We will also assess the impact that can be exerted by potential heterogeneous performance, even for equally equipped VMs within the same CSPs, on the generalization of measurements in the Cloud. Finally, we plan to study the relationship between performance and routing changes on paths in more detail and especially over time.

Appendix. Analysis of Variance: ANOVA

ANOVA performs a contrast test using the ratio of the sum of squares within each factor (this is typically termed variance 'explained' by a factor or interaction) and between factors. Such a ratio follows a Snedecor-F distribution under the null hypothesis, which considers that the total sum of squares is due to the randomness of measurements (often experimental error), and not to differences in the population when grouped by factors and levels. However, if the null hypothesis is not accepted, it means that the factor used to build the groups is statistically significant according to the F-test. By comparing the percentage of variation that can be explained by the factors and interactions to the error, we obtain a notion of both the importance of each factor and the goodness of the model. The overall explained percentage is named the coefficient of determination, $\bar{R}^2$, which is typically considered relevant if above 0.85.

More intuitively, ANOVA compares the mean of a set of observations after and before they are grouped into factors and their interactions. If the difference is relatively large, the factor (or interaction) is considered significant, and otherwise it is irrelevant (simple experimental error). With the relevant factors ANOVA, poses a model where any observation is result of the addition of a constant plus the effect of any of the factors plus the effect of the interactions of factors and, finally, plus an experimental error. Hence, the simplest ANOVA univariate model for a response variable y with only one significant factor α is given by:

$$y_{ip} = \mu + \alpha_i + \epsilon_{ip}, \tag{3}$$

where y_{ip} represents the p^{th} observation on the i^{th} level ($i = 1, 2, \ldots, I$ levels) and μ represents a constant for all the samples (often the overall mean or the smallest value of the sample). On the other hand, α_i refers to the effect due to the i^{th} level of factor α and ϵ_{ip} is the deviation, random or experimental error, in the p^{th} sample on the i^{th} level. We also note that $\sum_{i=1}^{I} \alpha_i = 0$.

Similarly, The resulting model in the case of two significant factors, often referred to as 2-way, is:

$$y_{ijp} = \mu + \alpha_i + \beta_j + (\alpha_i * \beta_j) + \epsilon_{ijp}, \tag{4}$$

where, α_i and β_j represent the effect due to the i^{th} and j^{th} levels of factors α and β, respectively. Similarly, $(\alpha_i * \beta_j)$ represents the interactions between i^{th}

level of factor α and j^{th} level of factor β. Finally, ϵ_{ijp} represents the deviation in the p^{th} sample from the overall mean of the samples within the i^{th} level of factor α and the j^{th} level of factor β. Again, note that $\sum_{i=1}^{I} \alpha_i = 0$ and $\sum_{j=1}^{J} \beta_j = 0$ with J being the total number of levels of factor β.

This is extended similarly in the case of more than two factors. According to this, by replacing y by the samples of bandwidth time series we can determine how factors influence the bandwidth in the Cloud and to what extent. Assume a two-factor model where α represents the factor data center source and β the destination data center of a path. Then, each factor has a set of levels, as the last column of Table 1 details. This way, the value of α_i represents the overall effect for samples whose data center source is i, β_j represents the effect for paths destined to the j data center, and $\alpha_i * \beta_j$ the effect for paths sourced on i and destined to j, specifically.

As an example, let i be Virginia$_{C1}$ and j be Virginia$_{C2}$; then a bandwidth sample p from Virginia$_{C1}$ to Virginia$_{C2}$ results from the addition of the following values: First a constant term, then a second a term that accounts for the difference between the samples sourced at Virginia$_{C1}$ and the constant. Then, a term to account for the difference in mean between samples destined to Virginia$_{C2}$ and the two previous values. Fourth, a term that represents the difference between the addition so far and the samples for paths specifically from Virginia$_{C1}$ to Virginia$_{C2}$ (e.g., the bandwidth between these two data centers is higher/lower than between the rest of the pairs). Finally, the specific error of each sample, i.e., the difference of the specific sample and the previous addition is included. If such error can be considered statistically small ($\bar{R}^2$), ANOVA provides a simple regression based on the mean per factor and their interaction. Note that, in this simplified explanation, the order of application of factors may result in different conclusions and parameter estimates. This is called ANOVA Type I (also known as sum of squares of Type I); alternatively, Type II and III estimate all possible combinations, preserving those with lowest error.

Finally, ANOVA methodology requires the data to meet several requirements: first, the samples must be independent; second, data must be Gaussian-distributed; and third, data must fairly share the same intra-group variance (i.e., exhibit homoscedasticity). However, the results of ANOVA are generally accepted, provided that the number of elements in each group is large and similar between them and that there is not a large deviation from the homoscedasticity assumption in terms of mean-variance rate [43].

In this paper, the data under study meet the first two requirements, as we are characterizing the main Gaussian component of the bandwidth time series as described in Section 2.3, and a simple auto-correlation test proved that samples were independent. Regarding homoscedasticity, it is clear from the same section that the width of the Gaussian component differs between paths, which was confirmed by the Levene's test. However, note that the number of samples is large, the sample is balanced, and CVs are low, as the third assumption requires.

Acknowledgements

This research was carried out with the support of the Prometeo project of the Secretariat for Higher Education, Science, Technology and Innovation of the Republic of Ecuador. It was also partially supported by the UTN-CUICYT-177 project Characterization of Correlated Performance in the Cloud.

The author thanks professors Edgar A. Maya and Julio J. Armas for their assistance, as well as professor Sanjay G. Rao for his guidance in the Cloud Computing field. Finally, the author thanks the anonymous reviewers for their constructive feedback and comments.

References

[1] High Scalability, "Latency is everywhere and it costs you sales - how to crush it," http://highscalability.com/latency-everywhere-and-it-costs-you-sales-how-crush-it/, 2016.

[2] A. Lakshman and P. Malik, "Cassandra: a decentralized structured storage system," *SIGOPS - Operating Systems Review*, vol. 44, no. 2, pp. 35–40, 2010.

[3] S. Agarwal, J. Dunagan, N. Jain, S. Saroiu, A. Wolman, and H. Bhogan, "Volley: automated data placement for geo-distributed cloud services," in *USENIX Symposium on Networked Systems Design and Implementation*, 2010, pp. 17–32.

[4] J. C. Corbett, J. Dean, M. Epstein, A. Fikes, C. Frost *et al.*, "Spanner: Google's globally-distributed database," in *USENIX Conference on Operating Systems Design and Implementation*, 2012, pp. 251–264.

[5] T. Wood, K. K. Ramakrishnan, P. Shenoy, and J. van der Merwe, "Cloudnet: dynamic pooling of cloud resources by live WAN migration of virtual machines," *ACM SIGPLAN Notices*, vol. 46, no. 7, pp. 121–132, 2011.

[6] Z. Wu and H. V. Madhyastha, "Understanding the latency benefits of multi-cloud webservice deployments," *ACM SIGCOMM Computer Communication Review*, vol. 43, no. 2, pp. 13–20, 2013.

[7] P. Shankaranarayanan, A. Sivakumar, S. G. Rao, and M. Tawarmalani, "Performance sensitive replication in geo-distributed cloud datastores," in *IEEE/IFIP Conference on Dependable Systems and Networks*, 2014, pp. 240–251.

[8] Forrester Research, "The future of data center wide-area networking," http://www.forrester.com, 2016.

[9] Z. Wu, M. Butkiewicz, D. Perkins, E. Katz-Bassett, and H. V. Madhyastha, "Spanstore: Cost-effective geo-replicated storage spanning multiple cloud services," in *ACM Symposium on Operating Systems Principles*, 2013, pp. 292–308.

[10] J. L. García-Dorado and S. G. Rao, "Cost-aware multi data-center bulk transfers in the cloud from a customer-side perspective," *IEEE Transactions on Cloud Computing*, 2015. [Online]. Available: http://dx.doi.org/10.1109/TCC.2015.2469666

[11] G. Aceto, A. Botta, W. de Donato, and A. Pescapé, "Cloud monitoring: A survey," *Computer Networks*, vol. 57, no. 9, pp. 2093–2115, 2013.

[12] J. Scheuner, J. Cito, P. Leitner, and H. Gall, "Cloud WorkBench: Benchmarking IaaS providers based on infrastructure-as-code," in *IW3C2 Conference on World Wide Web*, 2015, pp. 239–242.

[13] P. Leitner and J. Cito, "Patterns in the chaos–a study of performance variation and predictability in public Iaas clouds," *ACM Transactions on Internet Technology*, vol. 16, no. 3, pp. 15:1–15:23, 2016.

[14] S. Floyd and V. Paxson, "Difficulties in simulating the Internet," *IEEE/ACM Transaction on Networking*, vol. 9, no. 4, pp. 392–403, 2001.

[15] E. Zohar, I. Cidon, and O. O. Mokryn, "The power of prediction: cloud bandwidth and cost reduction," in *ACM SIGCOMM*, 2011, pp. 86–97.

[16] N. Laoutaris, M. Sirivianos, X. Yang, and P. Rodriguez, "Inter-datacenter bulk transfers with Netstitcher," in *ACM SIGCOMM*, 2011, pp. 74–85.

[17] R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. F. De Rose, and R. Buyya, "Cloudsim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," *Software: Practice and Experience*, vol. 41, no. 1, pp. 23–50, 2011.

[18] A. Li, X. Yang, S. Kandula, and M. Zhang, "Cloudcmp: comparing public cloud providers," in *ACM SIGCOMM Conference on Internet measurement*, 2010, pp. 1–14.

[19] M. Hajjat, R. Liu, Y. Chang, T. S. Eugene Ng, and S. G. Rao, "Application-specific configuration selection in the cloud: impact of provider policy and potential of systematic testing," in *IEEE INFOCOM*, 2015, pp. 873–881.

[20] Alexa, "Top sites in the Cloud," http://www.alexa.com/topsites/category/Top/Computers/Internet/Cloud_Computing, 2016.

[21] A. Tirumala, M. Gates, F. Qin, J. Dugan, and J. Ferguson, "iPerf–the TCP/UDP bandwidth measurement tool," https://github.com/esnet/iperf, 2016.

[22] I. Abdeljaouad, H. Rachidi, S. Fernandes, and A. Karmouch, "Performance analysis of modern TCP variants: A comparison of Cubic, Compound and New Reno," in *IEEE Symposium on Communications*, 2010, pp. 80–83.

[23] M. Jain and C. Dovrolis, "End-to-end available bandwidth: Measurement methodology, dynamics, and relation with TCP throughput," in *ACM SIGCOMM*, 2002, pp. 295–308.

[24] M. Crovella and B. Krishnamurthy, *Internet measurement: infrastructure, traffic and applications*. John Wiley and Sons Inc., 2006.

[25] V. Persico, P. Marchetta, A. Botta, and A. Pescapé, "Measuring network throughput in the cloud: The case of amazon EC2," *Computer Networks*, vol. 93, no. 3, pp. 408–422, 2015.

[26] A. W. Bowman and A. Azzalini, *Applied smoothing techniques for data analysis*. Oxford University Press Inc., 1997.

[27] R. van de Meent, M. Mandjes, and A. Pras, "Gaussian traffic everywhere?" in *IEEE ICC*, 2006, pp. 573–578.

[28] O. J. Dunn and V. A. Clark, *Applied statistics: Analysis of variance and regression.* John Wiley and Sons Inc., 1974.

[29] J. L. García-Dorado, J. A. Hernández, J. Aracil, J. E. López de Vergara, and S. Lopez-Buedo, "Characterization of the busy-hour traffic of IP networks based on their intrinsic features," *Computer Networks*, vol. 55, no. 9, pp. 2111–2125, 2011.

[30] N. H. Nie, D. H. Bent, and C. H. Hull, *SPSS: Statistical package for the social sciences.* McGraw-Hill, 1975, vol. 227.

[31] F. Mata, P. Żuraniewski, M. Mandjes, and M. Mellia, "Anomaly detection in diurnal data," *Computer Networks*, vol. 60, pp. 187–200, 2014.

[32] I. Poese, S. Uhlig, M. A. Kaafar, B. Donnet, and B. Gueye, "Ip geolocation databases: Unreliable?" *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 2, pp. 53–56, 2011.

[33] S.-J. Lee, P. Sharma, S. Banerjee, S. Basu, and R. Fonseca, "Measuring bandwidth between Planetlab nodes," in *Passive and Active Network Measurement*, 2005, pp. 292–305.

[34] Y. Feng, B. Li, and B. Li, "Jetway: minimizing costs on inter-datacenter video traffic," in *ACM Conference on Multimedia*, 2012, pp. 259–268.

[35] E. Walker, "Benchmarking amazon EC2 for high-performance scientific computing," *Login: the magazine of USENIX & SAGE*, vol. 33, no. 5, pp. 18–23, 2008.

[36] I. Bermudez, S. Traverso, M. Mellia, and M. Munafò, "Exploring the cloud from passive measurements: The Amazon AWS case," in *Proceedings IEEE INFOCOM*, 2013, pp. 230–234.

[37] G. Wang and T. S. Eugene Ng, "The impact of virtualization on network performance of Amazon EC2 data center," in *IEEE INFOCOM*, 2010, pp. 1–9.

[38] V. Persico, P. Marchetta, A. Botta, and A. Pescapé, "On network throughput cariability in microsoft azure cloud," in *IEEE GLOBECOM*, 2015, pp. 1–6.

[39] J. L. García-Dorado, "Bandwidth in the cloud. *arxiv*," Aug. 2015. [Online]. Available: http://arxiv.org/abs/1512.01129

[40] V. Persico, P. Marchetta, A. Botta, and A. Pescapé, "A first look at public-cloud inter-datacenter network performance," in *IEEE GLOBECOM*, 2016, pp. 1–7.

[41] V. Persico, A. Botta, P. Marchetta, and A. Pescapé, "On the performance of the wide-area networks interconnecting public-cloud datacenters around the globe," *Computer Networks*, vol. 112, no. C, pp. 67–83, 2017.

[42] R. Zabolotnyi, P. Leitner, W. Hummer, and S. Dustdar, "Jcloudscale: Closing the gap between Iaas and Paas," *ACM Transactions on Internet Technology*, vol. 15, no. 3, pp. 10:1–10:20, 2015.

[43] G. V. Glass, P. D. Peckham, and J. R. Sanders, "Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance," *Review of Educational Research*, vol. 42, no. 3, pp. 237–288, 1972.