Computer Networks 000 (2016) 1–17

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Multidimensional cloud latency monitoring and evaluation

Ondrej Tomanek*, Pavol Mulinka, Lukas Kencl

Department of Telecommunications Engineering, Czech Technical University In Prague, Technicka 2, Prague, Czech Republic

ARTICLE INFO

Article history:
Received 27 November 2015
Revised 10 June 2016
Accepted 11 June 2016
Available online xxx

Keywords:
Multidimensional
Distributed
Cross-layer
Cloud computing
Network latency
Monitoring

ABSTRACT

Measuring or evaluating performance of a Cloud service is a non-trivial and highly ambiguous task. We focus on Cloud-service latency from the user's point of view, and, to this end, utilize the multidimensional latency measurements obtained using an in-house designed active-probing platform, CLAudit, deployed across PlanetLab and Microsoft Azure datacenters. The multiple geographic Vantage Points, multiple protocol layers and multiple datacenter locations of CLAudit measurements allow us to pinpoint with great precision if, where and what kind of a particular latency-generating event has happened. We analyze and interpret measurements over two one-month time-intervals, one in 2013 and one in 2016. As these traces are large, an automated interpretation has been appended to the data-capture process. In summary, we demonstrate the utility of the multidimensional approach and document the differences in the measured Cloud-services latency over time. Our measurements data is publicly available and we encourage the research community to use it for verification and further studies.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

The increasing reliance on Cloud Computing within the field of information and communication technologies, together with the lack of technical information disclosed, poses serious challenges to Cloud tenants and end-users, left on their own to monitor what they pay for. One-dimensional monitoring approach, i.e. deploying a single Vantage Point talking to an arbitrary Cloud target via a single protocol, only provides limited insights. For example, one might not be able to discover the cause of a network failure or identify path segments impacting the Cloud-service performance. These issues are amplified by the ever-increasing Internet path diversity, by the complexity of traffic patterns and by the number of devices involved.

In our work we focus on *multidimensional* monitoring of Cloud-service *latency*. We choose latency because of its direct impact on overall performance, and, consequently, on the end-user experience. Latency is difficult to model, predict or control, but measuring latency may provide a lot of useful information. Latency data can be relatively easily derived from communication flows. Methods of capture, analysis and interpretation of such measurements, such as the one presented in this paper, ought to be then especially useful to Cloud tenants, who may thus be able to overturn the aforementioned information deficit, even without any privileged access to the Cloud infrastructure.

The problem we thus address is rigorous monitoring and advantageous interpretation of Cloud-Service latency behavior. We build on our two previous works that focused on architecting a multidimensional Cloud Latency Auditing platform (CLAudit) [1], used to identify issues like Data Center (DC) failures or routing pathologies. Detection of such anomalous events is automated by inter-relating the measurements time-series across different dimensions [2].

In this paper, we present new insights derived from two comparable blocks of CLAudit measurements obtained in 2013 and 2016, when deployed across the PlanetLab [3] network and Microsoft Azure [4] DCs. We also extend the data post-processing pipeline with automated interpretation using an Interpretation tree, Impact Tree and an Interpretation Database. The resultant two-stage post-processing was ran offline on the comparable 2013 and 2016 data subsets and we present interpretation results together with what these have evolved into.

The main contributions of this work are:

- A comprehensive set of publicly available latency measurements of Cloud Services within multiple DCs, captured at multiple protocol layers, across a set of global Vantage Points;
- A detailed analysis of the collected measurements, including a comparison of the Cloud-service quality between the two periods;
- Interpretation of the detected suspicious events within the said time series, including breakdown of the events' likely root causes.

The presented insights into Vantage-Point to Data-Center latency are, to some extent, explainable by the already well

* Corresponding author.
E-mail address: ondrej.tomanek@fel.cvut.cz, tomanon1@fel.cvut.cz (O. Tomanek).

http://dx.doi.org/10.1016/j.comnet.2016.06.011
1389-1286/© 2016 Elsevier B.V. All rights reserved.

investigated Internet service latency. Our work adds on top a study of impact caused by technologies specific to public Cloud Computing, such as by DC middleware (which terminates different protocol layers at different locations) or by data storage in remote databases. The measurements and interpretation show improvements in most aspects of the Cloud-Service experience over time. The impact of infrastructure improvements on the side of the Cloud Service Provider (CSP) is clearly observable, most notably in the improved performance of its core backend networking (reduced rate of incidents from 2.7% to 0.1%), and in the general reduction of the DC-centered ("global-impact") events–incidents.

The rest of this paper is organized as follows: Section 2 presents the related work. We then describe the multidimensional monitoring architecture and setup in Section 3, followed by insights derived from the collected measurements in Section 4. Measurements post-processing pipeline is described in Section 5 and the interpretation of results in Section 6. We conclude the work by summarizing implications and suggesting ways to continue in Section 7.

2. Related work

2.1. Network latency foundation

Network latency may be understood as a sum of delays along the communication path. The majority of the past latency-studying and tackling efforts has come from the Internet traffic-engineering domain, striving to ensure sufficient traffic QoS (Quality-of-Service). Results exist in the form of theoretical concepts, industrial solutions, RFCs and standards. A theoretical foundation of latency engineering was given by Queuing theory [5,6] and Network calculus [7,8]. Specific solutions such as guaranteed service networks or end-to-end jitter bounds were discussed in [9,10] and [11]. Because of their specificity and limited applicability to today's computer networks, overprovisioning [12] remains the most popular way of achieving QoS.

Many negative latency-related observations have been published, e.g. that significant portion of Internet traffic suffers from routing pathologies or that variations in end-to-end latency indicate long congestion periods [13–15]. More bad news include latency unpredictability [16,17] and the deep-seated tail latency phenomenon [18–21].

2.2. Cloud network latency

Latency is of the utmost importance to the Cloud, but the complexity and diversity of this environment prevent the conventional Internet measurement techniques to be easily adjusted to fit Cloud Computing needs. The problem partly being the scale, because the larger the scale the greater the impact of latency variability and the need for reliably low latency [22,23]. The ever-increasing interdependence of traffic patterns; context dependency and various stochastic factors render the task of capturing latency analytically intractable [24]. There was some success in approximating latency of specific environments (Cloudlet-to-Cloud and cellular-to-Cloud latency can be approximated by Rayleigh distribution [25]), but Cloud Computing service latency both inside and outside the DC lacks a good fit so far.

Miscellaneous Cloud measurements and analyses were conducted [26], often on platforms and tools designed in academia (like *Fathom* [27] or *Flowping* [28], often using PlanetLab [29]). Deriving traffic characteristics of flows inside a DC was the focus of [30] or [31]. Specific measurements concerning Cloud performance include [24] and [32]. End-user-perceived Cloud-application performance measurements were discussed for example in [33–37]. General network delay tomography and specific blackbox latency predictions are the topic of [38] and [39,40], respectively. Observations from multiple Vantage Points have been used previously [41–43], but these do not measure back-end or latency at multiple protocol layers. Active probing of Cloud resources [44] confirms need for sophisticated measurement, but is availability-oriented and does not document trends.

Commercial and research testbed latency-tackling implementations focus on reducing latency in a certain part of the Cloud. WAN acceleration heuristics [45] or careful path selection algorithms represent WAN-centric optimization. Both industry and academia have independently converged towards moving the Cloud closer to the end-user and offloading strategic responsibilities from remote DCs. Concepts such as Femto cells, Cloudlets or Fog Computing are starting to be successfully deployed, in mobile clouds foremost [46,47]. Innovations inside the DC include data path management [48–50], transport optimization, adjusting TCP behavior within the DC network or proposing entire new protocols (*DCTCP* [51], *D3* [52], *D2TCP* [53], *PDQ* [54], *pFabric* [55] and *delay-based TCP* [56]).

2.3. Suspicious event detection

Anomaly detection is a well-known concept in telecommunications and networking. Previous Cloud and network studies focused on passive monitoring of: volumes of transferred data [57], CDN deployed on PlanetLab - *PlanetSeer* [58], or on active monitoring of ICMP and HTTP service availability [44]. Their common goal was to measure and aggregate suspicious events and anomalous behavior, whereas we focus on a framework for distinguishing suspicious events based on their geographic location and affected OSI layer.

Techniques were proposed for anomaly detection using a traffic-flow pattern analysis [59–61]. We focus on the user perspective specifically, studying Cloud latency measurements at multiple OSI layers and across multiple geo-dispersed Vantage Points. Cloud monitoring survey [62] summarizes the state of the art solutions in the field of Cloud monitoring, but only a little attention is given to detection of suspicious events in latency measurements. Latency as a metric is used for performance evaluation [63,64] and benchmarking [65], but not for suspicious event detection. Time-series studies [66–68] focused on similarity search using wavelets and statistical methods. In contrast, as we lack the definition of "normal" Cloud behavior, we are searching for deviations from its empirically derived parameter values.

Compared to the commercially-provided global monitoring software (like *Renesys* [69] or *ThousandEyes* [70]) we use advanced metrics, leverage co-incidences across all dimensions of measurement time series and provide automated interpretation of events, offloading much investigation from the end user. Furthermore, in contrast with commercial software, our measurements are publicly available to the research community for verification and further studies.

3. CLAudit measurement platform

CLAudit alias Cloud Latency Auditing Platform is a system for collecting and evaluating multidimensional measurements. By *measurements* we mean RTTs of individual protocol exchanges, processing times and overall latency, as shown on webpage retrieval processes in Fig. 1. By *multidimensional* we mean measurements capable of being looked at from point of view of Vantage Points, Data Centers and/or protocol layers (see examples in Figs. 3 and 4). This section provides a brief overview of CLAudit components and deployment together with examples of data it collects. For a detailed description, see [1].

Please cite this article as: O. Tomanek et al., Multidimensional cloud latency monitoring and evaluation, Computer Networks (2016), http://dx.doi.org/10.1016/j.comnet.2016.06.011

Download English Version:

https://daneshyari.com/en/article/4954931

Download Persian Version:

https://daneshyari.com/article/4954931

Daneshyari.com