User-level Data Center Tomography

Neil Alexander Twigg
University of Stirling
nat@cs.stir.ac.uk

Marwan Fayed
University of Stirling
mmf@cs.stir.ac.uk

Colin Perkins
University of Glasgow
csp@csperkins.org

Dimitrios Pezaros
University of Glasgow
dp@dcs.gla.ac.uk

Posco Tso
University of Glasgow
posco@dcs.gla.ac.uk

ABSTRACT

Measurement and inference in data centers present a set of opportunities and challenges distinct from the Internet domain. Existing toolsets may be perturbed or be mislead by issues related to virtualization. Yet, while equally confronted by scale, data centers are relatively homogenous and symmetric. We believe these may be attributes to be exploited. However, data is required to better evaluate our hypotheses. Therefore, we introduce our efforts to gather data using a single framework from which we can launch tests of our choosing. Our observations reinforce recent claims, but indicate changes in the network. They also reveal additional obfuscations stemming from virtualization.

Categories and Subject Descriptors

C.2.3 [**Computer-Communication Networks**]: Network Operations; C.4 [**Performance of Systems**]: Measurement Techniques

Keywords

Data Centers, Network Measurement, Tomography

1. INTRODUCTION

Data center network performance and behaviour is poorly understood in the public sphere. A better understanding facilitates research in network management, planning, design of protocols, and the ability to recognise abnormal events.

From inside the data center, operators have a direct line to network equipment and statistics [2]. In this context, there exists tools and projects that seek to measure, monitor, and diagnose data center network-related events. While informative, the output of these efforts must necessarily be kept confidential to maintain competitive advantage. When results can be shared, the experiments are irreproducible for the same reason. This poses challenges when attempting to characterize or improve upon network-level performance and security.

In this poster we present our efforts to provide a client-level perspective. We seek to identify the type and resolution of data available to the paying customer, as well as the inferences that might be made with this data. We take the position that, from a user-perspective, the only level of service that matters is the one that the user sees. For example, metrics such as bisection bandwidth describe network-wide characteristics that a single user may never see.

The task of measuring data center networks from a user perspective is challenging, and is only partially analogous to Internet measurement and tomography. Although issues of scale are present in both contexts, measurement in data center networks changes in two ways. The first is arguably to the benefit of any measurement infrastructure: Relative to the Internet, data centers are homogeneous and symmetric. This suggests that it may be possible to infer network characteristics without the need to evaluate complete sets of end-to-end paths. Alternatively, it suggests that characteristics inferred for a path of length x may hold for all internal paths of length x. The second difference is the possible perturbation caused by virtualization. Virtualization is known to lead to misleading or adverse data in existing toolsets [3]. For reasons different from previous work, this is confirmed by our own observations.

The client-level perspective is motivated by a series of questions in our own work:

- What is the effect of the data center on time-sensitive applications, and can shortcomings be predicted or compensated?
- Can a client detect and react to anomalies in their own 'slice' of the network?
- How do we obtain, and subsequently provide to the wider research community, useful data sets for the purpose of simulation and modeling?

Answers to such questions are impossible without traces and representative models of data center characteristics. Anecdotally, one of our own virtual machines was found to be participating in a DoS attack without our knowledge, after relaxing firewall rules for just a few hours. The opacity of the network only hindered our efforts to diagnose and repair.

2. FRAMEWORK

We are working to fill these gaps by constructing a flexible framework in which measurements can be selected and executed in a dynamic fashion. Our framework was built and tested on Amazon's AMI linux for EC2. It uses only standard C libraries and may be easily ported to other platforms.

When launched in the data center, clients first retrieve and build the latest source code from a git repository. Upon completion, clients then communicate with an off-site server to retrieve the list of expected measurements, and learn of the private IP addresses in use by other clients as they come

SIGCOMM'12, August 13–17, 2012, Helsinki, Finland.
ACM 978-1-4503-1419-0/12/08.

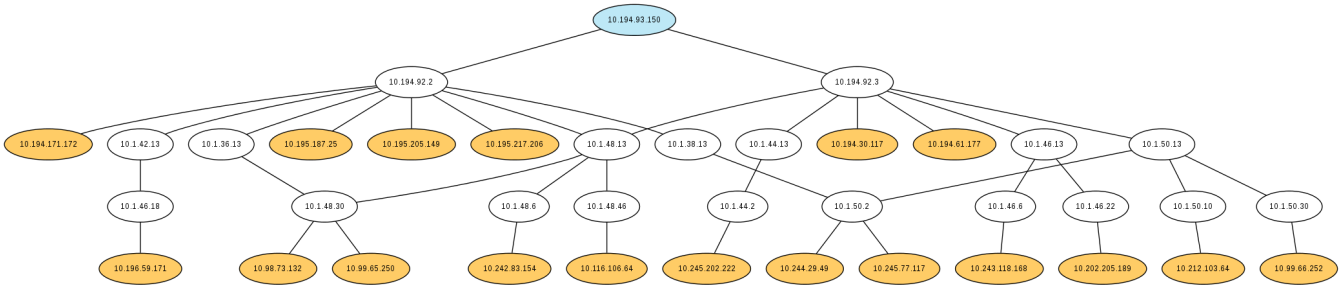
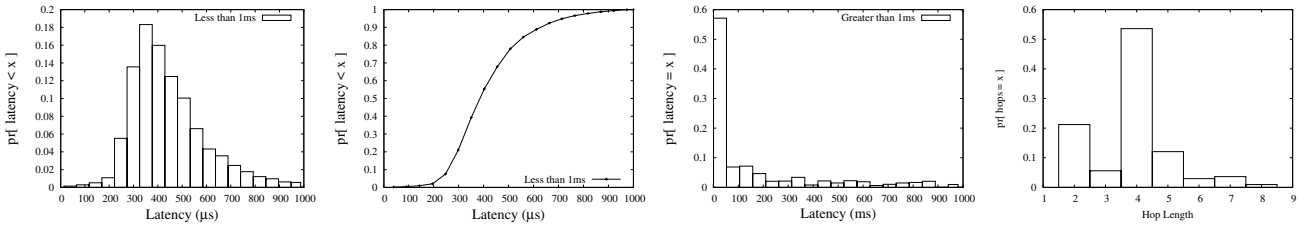


Figure 1: Sample topology from node at root (blue) to 19 other nodes (yellow).



(a) edf of RTTs $< 1ms$. (b) cdf of RTTs $< 1ms$. (c) edf of RTTs $> 1ms$. (d) edf of path lengths.

Figure 2: Summary of initial measurements

online. With a single point of distribution, experiments may be created, modified, and distributed quickly.

Current modules consist of ping- and traceroute-like measurements. In order to minimize timing issues, as well as embed relevant tags, the current modules build custom ICMP headers and use raw sockets. In the next section we validate the measurements taken using this framework by comparing against the work in [3]. We are also in talks with Amazon to determine accuracy. In future we will focus on bottleneck bandwidth estimation, as well as understanding the relationship between logical and physical paths.

3. INITIAL RESULTS

Initial tests were selected to address two concerns. First, the impact of layer-2 routing in the network with respect to its opacity at the IP layer. A multiple-hop layer-2 path between endpoints is presented to the user (at the IP layer) as a single hop. This makes inference of bottleneck links, for example, challenging. Second, fast round-trip times risk being too small to be reliably recorded given the granularity of the OS. Initial tests are positive, and reveal some truth in these assumptions, as discussed in following sections.

Network Maps

From the data gathered we can construct maps of the topology. A subset of a complete map appears in Figure 1 showing the topology from the view of the VM at the root to 19 other VMs in the network (leafs). The rich maps suggest there is sufficient IP-level detail irrespective of layer-2 routing. One interesting observation is that the host machine never responds to expired TTL. We have also been able to infer the practice of assigning to each physical machine a /24 network address by comparing VM with gateway addresses, and reconciling every VM always being reachable.

Round Trip Times

Our observations reinforce the high variability seen in [3] but differ in remaining respects. In that work it was observed that the first few RTT measurements in every new set of pings between a pair of nodes were abnormally high. Once those RTTs were removed, their observations revealed all RTTs to be less than $200\mu s$. We witnessed no such initial behaviour, as well as an increase in RTTs. Observations are summarized in Figures 2a to 2c. For clarity we have separated RTTs less than 1ms (the edf and cdf in Figures 2a and 2b respectively) from those more than $1ms$. The ratio of packets in the two categories is approximately 5:1. Interestingly, RTTs greater than $1ms$ appear to be fairly constant and persistent over spans of several seconds. We were also surprised by the appearance of negative RTT values in our measurements. We believe this to be a consequence of different timing as different cores revealed to the VMs.

Path Length

The distribution of path lengths also differs from the work in [3], which observes all path lengths to be either 3 or 4 hops, Figure 2d shows this no longer to be true. Most paths are 2, 4, or 5 hops in length, with a non-trivial number spanning as many as 8 hops.

Two years apart, our observations suggest that the means of the underlying distributions are similar, but that the variances are quite different. We see this as initial evidence of changes in the network and levels of service that need to be better understood. Observations also reveal negative pings and hidden nodes, both as a consequence of virtualization.

Network maps and measurements are viewable online at [1]. Members of the community will be encouraged to participate by injecting their own measurement modules. Recognizing that measurement is often a community effort, source code is expected to soon be available.

4. REFERENCES

[1] `http://d253108.cs.stir.ac.uk:89`.

[2] P. Gill, N. Jain, and N. Nagappan. Understanding network failures in data centers: Measurement, analysis, and implications. In *Proceedings of ACM SIGCOMM*, August 2011.

[3] G. Wang and T. E. Ng. The impact of virtualization on network performance of Amazon EC2 data center. In *Proceedings of the IEEE INFOCOM*, March 2010.