Location Matters:
Limitations of Global-Scale Datacenters

Yahel Ben-David, Shaddi Hasan, Paul Pearce

Abstract

Common knowledge states that cloud providers will move towards large centralized data centers to achieve economies of scale. However, this centralization presents challenges for certain classes of applications. We describe two of these key challenges – increased response times and regulatory restrictions – and describe how each could be overcome by a larger number of geographically disparate datacenters. We consider the implications for cloud application innovation, and propose a new model for cloud infrastructure called "carrier cloud" to augment the existing cloud infrastructure to overcome the above challenges.

1 Introduction

Cloud computing is taking over like a storm, replacing traditional information technology operations and business models the world over. Economies of scale enable cloud providers to reduce service costs by operating massive datacenters, comprising tens or even hundreds of thousands servers. Bulk hardware purchases, efficient large-scale cooling and environmental control systems, and volume discounted electrical power, coupled with centralized management tools and methodologies, provide for previously unattainable efficiency and cost reductions. The property of elasticity offered by the cloud is both an enabler for new business models as well as driver for reducing costs of existing businesses, and thus a key motivation for the adoption of cloud computing across many industries.

Current trends push large cloud providers to build large, centralized datacenters, but this approach is not without some inherent shortcomings. One of the most discussed and obvious is the issue of disaster tolerance. In order to meet availability and reliability goals, large cloud providers build multiple datacenters that are geographically distributed. For example, Amazon's AWS datacenters are divided into six "regions" corresponding to datacenters in Virginia, Northern California, Oregon, Ireland, Tokyo, and Singapore. Another key issue is response time. Geographic distance from end-users of cloud-hosted applications translates to increased latency, which is a key challenge for some applications and services wishing to leverage cloud computing. Latency sensitive applications like telephony, online gaming, and video-conferencing benefit from local datacenters that are closer to end-users, as do services like content distribution networks (CDNs) that aim to push static content towards the users at the edge of the network. While providers do indeed build geographically distributed datacenters, the relatively small number of datacenters needed to achieve adequate disaster tolerance does not mean that all users worldwide have low-latency access to datacenters (even ignoring issues such as poor last-mile connectivity and Internet-level routing failures). Finally, regulatory constraints often dictate security and especially privacy restrictions on the storage and transmission of information, traditionally expecting sensitive data to remain with the boundaries and control of a single sovereign country.

Given these and other shortcomings of the currently dominant cloud computing model, we initially set out to explore the case for *Nano Datacenters* [30], a new cloud infrastructure architecture to address some of these challenges. However, due to the forces outlined below, we have come to believe that a cloud operator will, in all cases, seek to build as large a datacenter as they can operate profitably. The question then becomes "at what cost do we centralize the world's cloud infrastructure?" In Section 3 we examine the response time to a major cloud service provider as seen from a large number of vantage points around the world. We also perform as survey of global regulations affecting cloud providers and applications in Section 4. We then consider the potential impact on application innovation in Section 5.

As part of our exploration of the space, we have identified an industry that is well positioned to offer localized, smaller-scale, cloud services competitively, in a manner that may alleviate the constraints of current cloud offerings. Telecommunication carriers are ideally positioned to provide end-to-end quality guarantees and therefore the SLAs (Service Level Agreements) desired by enterprise costumers. We discuss the business aspects of this ecosystem in more depth in Section 6, and consider how these *carrier clouds* may also enable services and offerings that are currently not feasible using today's centralized cloud infrastructure.

2 Why not Nano Datacenters?

The move towards large datacenters is enduring and industry-wide for reasons that are both compelling and universal. First, large datacenters allow operators to receive volume pricing from equipment vendors. For the largest datacenter operators, this even makes designing specialized hardware cost effective, allowing significant savings in power and cooling. Facebook recently launched the Open Connect Project [26] to open their custom, datacenter-oriented server, rack, and power designs; this standardization will help put the benefits of custom-designed hardware within reach of smaller cloud operators, perhaps even enterprises seeking to operate their own "private clouds". Secondly, the cooling and power infrastructure for a datacenter is more efficient, and thus less costly, at large scales. In addition, electric utilities charge lower per-unit rates to their largest industrial-scale customers and offer demand response incentives that are unavailable to smaller customers. Secondly, centralized datacenters simplify life for application developers, since partitions and failures are much less common within a single datacenter than across multiple datacenters. Even for developers using public cloud infrastructure, inter-datacenter communication is more expensive and slower than intra-datacenter communication. Finally, and perhaps most importantly, large datacenters allow operators to handle ever-increasing data storage and processing demands. In some ways, this is the key driver for datacenter expansion: modern applications require massive computing systems to provide high-quality results with the response times that users expect.

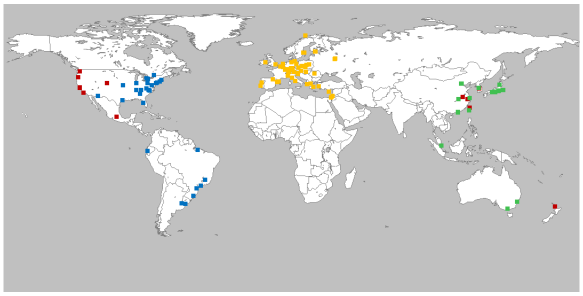

Figure 1: PlanetLab vantage points, colored by AWS datacenter with best median response time. Nodes colored blue, red, yellow, and green have the best response time to datacenters located in Virginia, Oregon, Ireland, and Singapore, respectively. Note the concentration of vantage points in the United States and Europe.

We initially set out to make a case for a new datacenter architecture called *Nano Datacenters* [30]. However, due to the forces outlined above, we have come to believe that a cloud operator will in all cases seek to build as large a datacenter as they can operate profitably. What this size is will be intimately coupled with the types of applications the operator runs on their datacenter, as well as what their revenue model is for those applications. We feel it is safe to assume that most operators' revenues are positively correlated with usage and computational capacity. Thus, because of the various savings associated with larger datacenters, operators should build as few large datacenters as possible in line with their tolerance for latency and disaster tolerance.

3 Cloud response times

In order to evaluate response time to cloud providers, we conducted a series of measurements from 124 PlanetLab [25] nodes around the world to four of Amazon's public cloud datacenters, located in Virginia, Oregon, Ireland, and Singapore [1] Each PlanetLab node issued an HTTP GET request to each of these datacenters every five seconds over a two-day period in December 2011. We used HTTP response time rather than `ping` because Amazon blocks incoming ICMP traffic to its datacenters; HTTP response time is additionally more meaning-

[1] Amazon has two additional data centers for their public cloud in Northern California and Tokyo which we did not include in our evaluation.

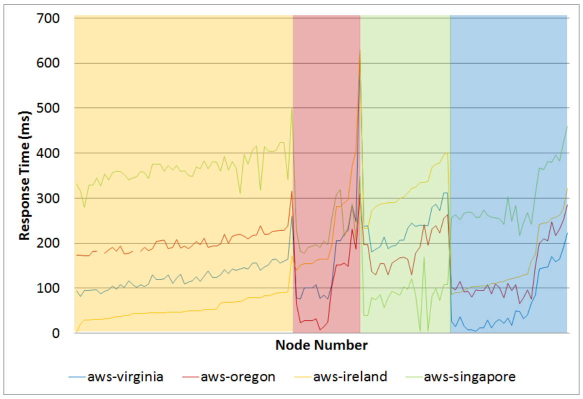

Figure 2: Median response times from all PlanetLab nodes to each Amazon datacenter. Shading the best data center for each block of nodes.

ful and interesting for most cloud users. We defined response time to be the time taken between completing the TCP handshake and successfully receiving the target file. The request always resulted in a 400 (“Bad Request”) status response due to a malformed URI (requesting a non-existent file). Thus, the body of the response was small enough to complete in a single round-trip, giving an approximation for network latency to the data center. For calibration purposes, we compared ping times from a separate machine in the same data center to a handful of PlanetLab nodes and found that our measured HTTP response times were roughly in line with the results from the `ping` utility.

Figure 1 shows the locations of the PlanetLab nodes we used, colored by the Amazon datacenter that provided the best median response time over the measurement period. Unfortunately, our vantage points are primarily located in the United States, Europe, and East Asia, all regions with excellent Internet connectivity and at least one nearby Amazon datacenter. Moreover, because most PlanetLab nodes are connected to well-provisioned research and education networks [2], the response times we see here may be significantly better than a user on a less-reliable commercial network might see. Nevertheless, our results are generally as we would expect: we see clear geographic segmentation, with PlanetLab nodes receiving the lowest response times from the datacenters geographically closest to them.

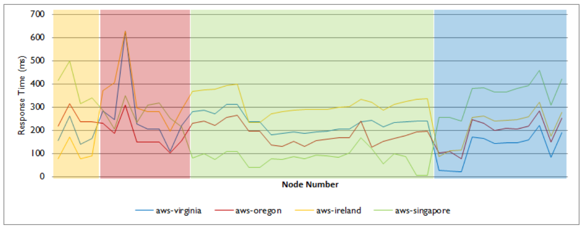

Figure 3: Median response times excluding all PlanetLab nodes in the US and Europe.

Looking more deeply, we compare the median response time from each PlanetLab node to each datacenter. Figure 2 shows the median response times for each PlanetLab node to each datacenter. Response times range between 5ms and over 600ms. The worst performance seems to have been for nodes located in China, with higher response times than nodes geographically further away from datacenters, suggesting other factors are at play beyond simply speed of light delay. Indeed, while we find a positive correlation between a node’s distance from a datacenter and its observed response time to that datacenter, the strength of the correlation varies across datacenters. Figure 4 shows that geographic distance accounts for almost 70% of the variation in response times for nodes communicating with the Virginia datacenter, while it accounts for only 18% of variation for the Singapore datacenter. In the Singapore case, we observe three clusters of points, suggesting a suboptimal or circuitous route between that datacenter and the nodes in the high response time cluster. All nodes in the high response time cluster are physically located in Europe or the Middle East; removing these nodes only increases R^2 to 0.51. Locating datacenters near users, then, is not sufficient; a cloud operator or user that wishes to control response times must be able to control the performance of the wide-area links between the client and the datacenter. Suboptimal wide-area performance can be caused by routing failures or slow BGP convergence times, both issues that other work has shown happens regularly on the Internet [24, 31, 18].

We next considered the distribution of response times for each datacenter, as seen in Figure 5. Every datacenter offered poor performance to some set of our measurement nodes, with the worst case re-

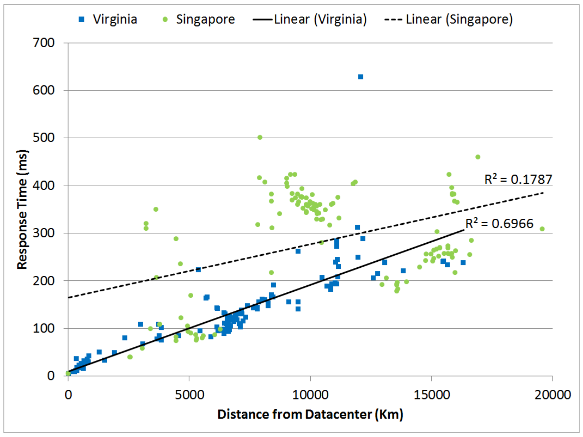

Figure 4: Regression analysis of geographic distance from datacenter and response time between PlanetLab and Amazon's Virginia and Singapore datacenters. In general we observe a positive linear correlation between distance and response time; however, distance alone is not the sole determinant of response time.

sponse time for any single datacenter ranging from 2-5 times greater than the median response time, with absolute response times reaching more than 600ms. This wide range of possible response times implies that no single datacenter can provide low response times to a set of end nodes. Particularly given the fact that geographic distance from a datacenter only accounts for 50-60% of the variance in response times across nodes, an application developer seeking to provide bounded response time to an unknown set of users must distribute their application onto as many datacenters as possible.

To see how adding datacenters could help, we additionally requested content from Amazon's Cloudfront CDN, which leverages 24 globally distributed points-of-presence (POP) to reduce response time to users. While our measurement workload is easily satisfied by existing CDNs, one could easily imagine hosting more complex applications in a similarly geographically distributed way given sufficient computing infrastructure at each POP. The last column in Figure 5 shows that using the geographically distributed Cloudfront infrastructure not only decreases median response time, but also decreases the maximum response time seen by any single measurement node.

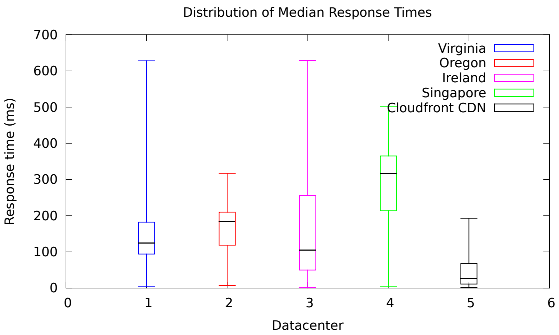

Figure 5: Distribution of median response times for all PlanetLab nodes to each Amazon datacenter, as well as the Cloudfront CDN. Adding additional POPs reduces both the median response time as well as the variance in response time across the fleet of measurement nodes.

4 Legal Concerns and Constraints

Legal constraints play a significant role in determining where regulated and private data can be deployed for companies across the globe. These constraints govern not only who has access to the data and how it is stored, but also in what jurisdictions the data can be located. Geographic location of data has been cited as the "numer one killer" of plans to use public cloud computing [33].

In examining the various legal restrictions that govern data, we divide regulations into two broad categories: privacy laws that govern *how* data is stored, and transborder data laws that govern *where* data can be stored. These issues are discussed in Section 4.1 and Section 4.2, respectively. We also examine the problem of *conflicting legislation* which may force a company governed in one jurisdiction to violate the laws governing data they house in another jurisdiction. Section 4.3 looks at the problem of conflicting legislation, particularly in the context of the USA PATRIOT Act [28] and how it impacts the choices of European cloud customers as well as American cloud providers. We then discuss the how these legislative constraints may influence datacenter placement in Section 4.4.

4.1 *How* Data Can Be Stored: Privacy Laws

The first class of regulations we look at are those that govern *how* data can be stored. In the United States regulations such as HIPAA [14], FERPA [9], and GLBA [12] control how data can be stored, as well as who may access that data. In the European Union (EU) the Data Protection Directive (EUDPD) [7] governs sensitive private data. This Data Protection Directive has no limitations on the movement of data within the EU, so we address this law further in the next section.

HIPAA governs confidential patent informations for healthcare providers, FERPA governs student records for academic institutions, and GLBA governs sensitive financial data. These legal regulations existed before the advent of cloud computing, and legal interpretations of their applicability to cloud providers are still evolving, as are the policies themselves [10, 27].

Since these laws primarily govern how the data is stored and who has access to it, of relevance to the problem of datacenter placement is if these privacy regulations have any bearing on the legal jurisdictions where the data can be stored. Unfortunately this question has not been fully addressed legally [10, 21, 27]. As recently as December 1, 2011, the US Department of Education was unsure if FERPA governed data could be housed in a cloud provider outside the US, stating:

> "Several commenters [on the proposed FERPA updates] sought clarification on whether the proposed regulations would permit cloud computing, where data can be hosted in a different State or country. [...] The Department has not yet issued any official guidance on cloud computing, as this is an emerging field." [10]

Since these privacy regulations do not have jurisdictional restriction inherent in their language, we must instead look at separate laws that govern transborder data flow.

4.2 *Where* Data Can Be Stored: Transborder Data Laws

One of the primary challenges companies face when moving to a cloud-computing environment is determining where and in what jurisdiction the data can be placed [3]. The movement of data originating in one jurisdiction to another jurisdiction is regarded as a *transborder data flow*, and is governed by special legislation. Depending on the country of origin and the type of content, moving data to a datacenter outside of the original jurisdiction may be difficult or even illegal. The problem of data location is daunting enough that at least one cloud-based software-as-a-service provider has created redundant datacenters within each jurisdiction they support, ensuring that customer data never leaves their chosen jurisdiction [20]. To better understand the problem, we examine the existing regulations that govern transborder data movement globally (Section 4.2.1), with a particular focus on EUDPD (Section 4.2.2), and then look at legislation designed to help with transborder concerns (Section 4.2.3).

4.2.1 Existing Global Transborder Data Regulation

As of October 2010, 68 countries and economic zones have enacted, or are in the process of enacting, legislation that controls how data may be moved across borders [17]. These regulations are outlined in at least 40 different pieces of legislation. The list of countries enforcing this legislation spans the globe from the Americas, to Europe, to Africa, to Asia. Figure 6 shows geographically which countries already have legislation in force, and which countries have drafted legislation but not yet enacted it.

Some legislation, such as the EUDPD, flatly forbids the transfer of data to other jurisdictions not explicitly approved [5]. Besides explicitly enumerating jurisdictions, many pieces of legislation have language permitting data to reside in other jurisdictions when "adequate" or "sufficient" (depending on the wording of the legislation) levels of protection are used. Further, other jurisdictions have restrictions permitting transborder data crossing only when the other jurisdiction has equivalent or better levels of protection. An example of such language can be seen in Israeli law:

> "A person shall not transfer, nor shall he enable, the transfer abroad of data from databases in Israel, unless the law of the country to which the data is transferred ensures a level of protection no lesser, mutatis mutandis, than the level of protection of data provided for by Israeli Law [...]" [17]

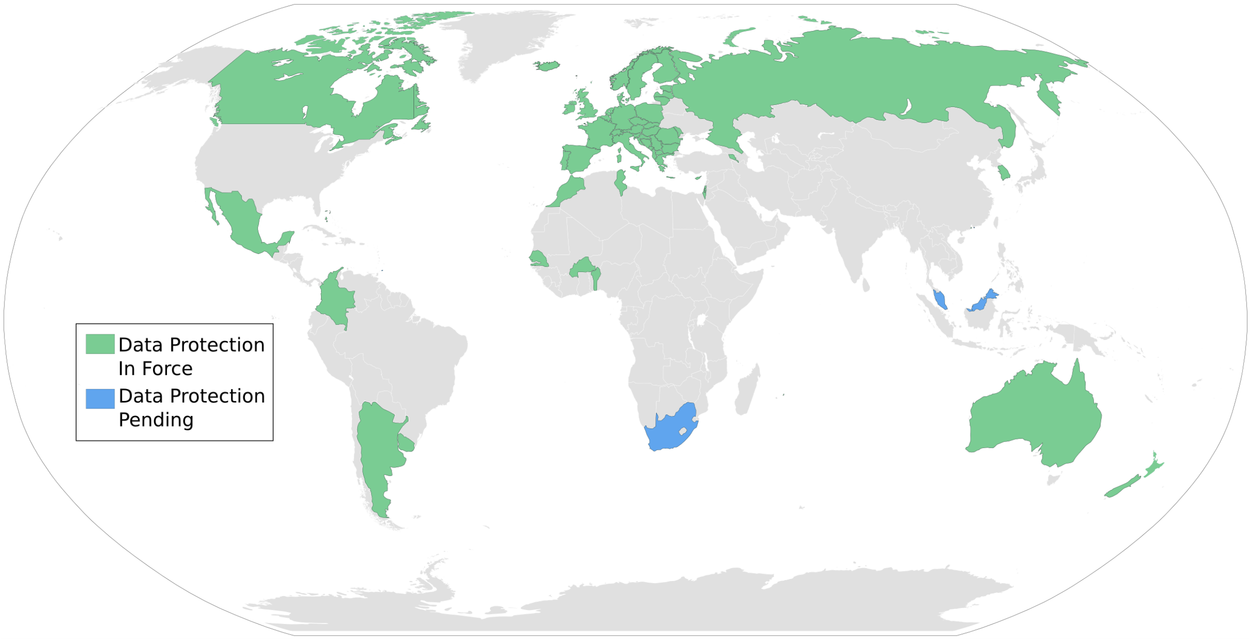

Figure 6: Map showing countries with regulations governing transborder data flow. Countries in green have regulations already enforced, while countries in blue have legislation pending. Some countries are too small to be visible in this graphic.

The sheer quantity and diversity of these laws makes it difficult to establish centralized datacenters that conform to all regulations. One study claimed that 85% of European companies surveyed cited international regulations as being a "major issue" [33]. To better understand the scope of these laws, we now look at one specific example, the EUDPD.

4.2.2 The European Union Data Protection Directive (EUDPD)

The European Union's Data Protection Directive (EUDPD) [7] explicitly forbids storing an EU citizen's personal information in a country outside the EU that does not provide "adequate protection" for the privacy of that information. The European Commission has deemed that only four non-EU or non-European Economic Area (EAA) countries provide this level of protection: Switzerland, Canada, Argentina, and the Isle of Man [7]. EU data can be stored in the US under certain circumstances with the United States Safe Harbor rules, which we discuss in Section 4.2.3. Figure 7 shows geographically where EU data can be stored. Such restrictions on location mean EU companies housing regulated data must ensure that their cloud-services never replicate data or move it to a datacenter outside of this very limited area.

4.2.3 Responses To Transborder Regulation: US Safe Harbor Frameworks

As a response to the rise in transborder data regulation, the US has enacted the Safe Harbor Frameworks [8]. These regulations are designed to allow US companies to store and operate on EUDPD-protected data. A US company under the jurisdiction of the Federal Trade Commission (FTC) that wishes to be part of the Safe Harbor program must apply and agree to conform to the "adequate" privacy standards laid out by the agreements. Companies can be certified as conforming to either the regulations of the EUDPD for EU data, or Switzerland for Swiss data.

There are two fundamental limitations with the Safe Harbor Frameworks. First, the scope of companies that can be certified as part of the program is very limited. Only those companies under the jurisdiction of the FTC can be Safe Harbor members [8]. Companies that have an interest in cloud computing that are ineligible include banks, telecommunication carriers, and non-profit organizations. Second,

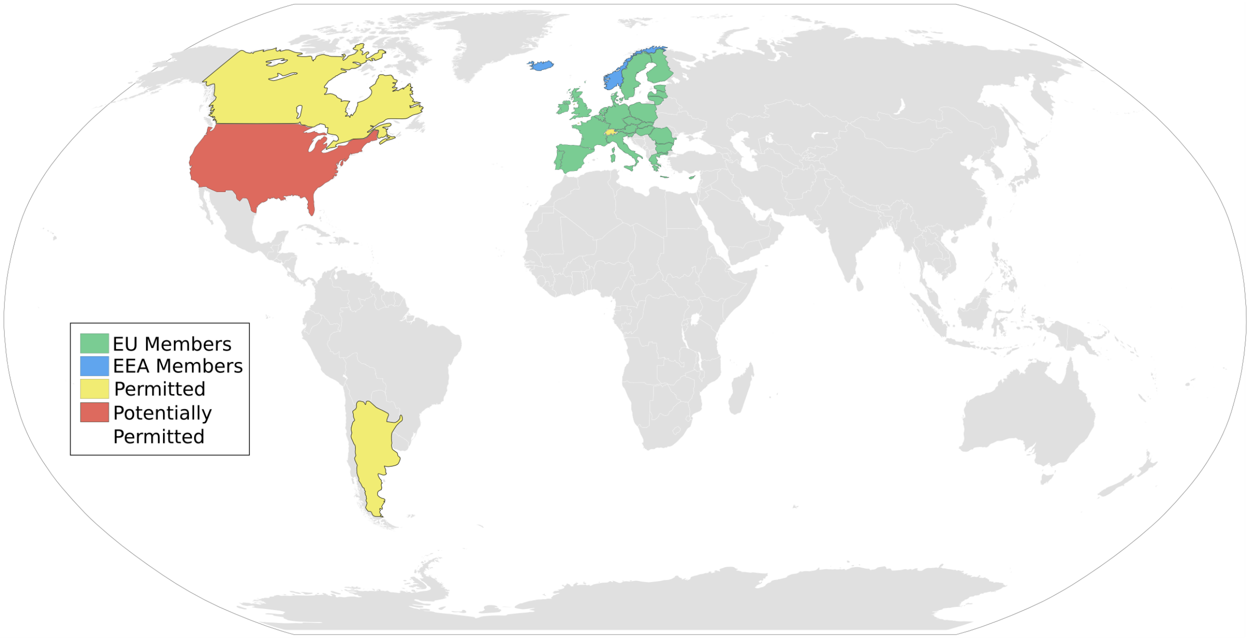

Figure 7: Map showing where EU private data can be stored globally. Countries in green are EU members and have unrestricted data flow. Countries in blue European Economic Area members, and also have unrestricted flow. Yellow countries are foreign countries approved by the EU, and the United States, red, is permitted in some cases (see Section 4.2.3)

there are extenuated conflicting legislative concerns, which we discuss in the next section.

The scope of the Safe Harbor Frameworks is quite limited; this is an agreement between three jurisdictions only. Although a step in the right direction, the complicated legal environment of numerous regulating bodies and legislation still points to per-jurisdiction datacenters as an attractive alternative.

4.3 Conflicting Legislation: The USA PATRIOT ACT

Unfortunately, the wide variety of regulation governing protected data sometimes conflicts with other legislation for areas such as counter-terrorism. This problem applies to not only what jurisdictions the data is located in, but also *which* company owns the datacenter that houses their data.

Data stored in the US. The USA PATRIOT Act gives the US government broad powers to collect the private data of foreign companies and individuals that is stored on computers inside the United States [32]. This law also supersedes the US Safe Harbor Framework, meaning EU companies doing business with Safe Harbor certified cloud service providers can not be sure their data is safe when housed in the US [32]. This power has given countries pause about storing data in US datacenters [13, 29]. One Canadian government report recommended the government “prohibit personal information being stored or sent outside Canada” as a result of the PATRIOT Act [29].

Data stored by US companies in other countries. Unfortunately, the powers of the USA PATRIOT Act extend beyond the jurisdiction of the US. The act gives the US the power to obtain private data stored in other jurisdictions (such as the EU) if the data is stored by a US governed company [32, 34]. Microsoft has publicly stated that they, nor any other US company could guarantee that data they housed in the EU could not be obtained by the US government [34]. The expanded reach of the US government has resulted in companies abandoning European cloud solutions offered by US companies, including a major British defense contractor reportedly abandoning Microsoft’s Office 365 service, citing such fears [33]. This means that cloud services offered by companies such as

Microsoft and Amazon can not be used by companies with regulated data, even if it is kept in their own jurisdiction. The problem of conflicting legislation also points to more geographically dispersed, yet locally-owned, datacenters.

Concerns over the PATRIOT Act are so great that the EU is expected to release an updated data directive to combat and US law by January 2012 [35].

4.4 Resulting Limitations

These jurisdictional regulations combine to form a quagmire of legal concerns for companies wishing to use cloud services. Depending on your location and the content of your data, regulation of data location can be unclear (FERPA, HIPAA), destination dependent, flatly prohibitive (EUDPD), or even in conflict with your own laws (US PATRIOT ACT).

A result of this legal red-tape is that companies want to keep their data in specific jurisdictions [20, 33]. This is very prohibitive for the notion of *only* global-scale datacenters, and supports the notion of smaller more geographically and organizationally dispersed datacenters. Some companies have already started toward this tend of jurisdictionally dispersed and aware datacenters [20] and, we expect others will follow.

5 Impact on Application Innovation

Despite the challenges outlined in the previous two sections, cloud providers seem to provide useful services to their many customers. Does this mean these challenges are not practically significant? While they may not be problems for most applications currently deployed on the cloud infrastructure, we believe that they have a potential to impact innovation in cloud applications. One of the key benefits that the current cloud infrastructure provides its users is flexibility over capacity; we believe that a future cloud infrastructure that also offered flexibility over location would offer similar advantages.

One application we have considered that could benefit from location flexibility is leveraging software-defined networking (SDN) to outsource network management. SDN allows network operators to build a network of dumb switches and manage those switches from a central controller with a complete view of the network. Once the view and management of the network has been centralized, outsourcing the management task itself becomes a possibility. For rural Internet service providers, this would be particularly appealing: urban migration saps talent from rural regions, meaning that skilled network administrators are likely to move to a city where they can receive higher wages. By outsourcing the network management task, rural carriers could benefit from the talents of skilled network administrators without the overhead of training and personnel churn.

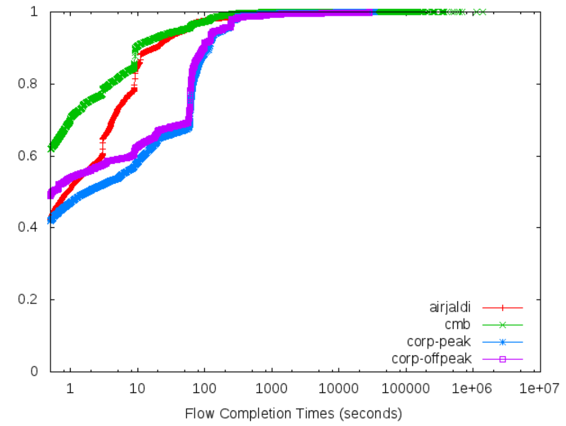

Figure 8: Flow completion time cumulative distributions as seen in four network traces.

Suppose such an ISP attempted to outsource the SDN controller to a cloud-hosted network management provider. Because policy is implemented at the controller, traffic in the provider's network will suffer the overhead of round-trip communication to the controller whenever a policy decision needs to be made for delivering traffic. While most policies do not require fine-granularity decision making (e.g., routing decisions generally need to update a switch's flow table when routing state changes, which occurs infrequently), some, particularly security and traffic engineering policies, require making decisions for every flow. Thus, the performance overhead of high cloud response times would potentially nullify any operational benefits of controller outsourcing.

To understand how well the current cloud infrastructure could support this outsourcing model on today's network traffic, we considered four network traces, ranging in length from four hours to a full week, and evaluated the flow completion times for every flow in each trace. We obtained traces from a medium-sized rural ISP in India, a university network in Sri Lanka, and an enterprise in Israel; all traces were captured at the border gateway for each network. Our metric for performance overhead was increase in flow completion times, since this is the metric that is most relevant for end-users. If flow completion times are generally short, user-perceived performance would be sensitive to response times for the cloud controller. As seen in Figure 8, we observe that flow completion times are in fact quite short, with between 40-60% of flows completing in under 500ms. For policies that require per-flow controller decisions, a cloud-hosted SDN controller would add significant performance overhead if hosted on the current cloud infrastructure. In the case of the rural ISP trace, for example, if the cloud controller were located at a datacenter with a response time of 200ms, well within the typical response time range we observed in section 3, 20% of flows would take at least 50% longer to complete.

6 Discussion

Given the limitations discussed thus far, we believe that the current cloud infrastructure model of ”one datacenter fits all” in fact does not. While small and medium businesses (SMBs) across the globe convert their recently-vacated server rooms into conference rooms, it appears that large enterprises are hesitant to join the global cloud party. Enterprises expect tight service level agreements (SLAs) for outsourced services, and cloud providers are reluctant to provide these bullet-proof guarantees, which are common in the telecom and co-location industries [15, 1, 23]. Even as cloud services evolve into public and private clouds, with the later positioned to offer a more refined set of SLAs, reliance on the best-effort service of the public Internet remains a critical element beyond the control of the cloud provider.

In our exploration of the space, we have identified telecom carriers as ideally positioned to provide these end-to-end guarantees. They control the communication medium between the enterprise offices and the datacenter, and at times even connect many end-users external to enterprise branches, especially within a single country where a particular carrier may have substantial market share. At the very least, this model greatly minimizes the number of intermediate ISPs between the localized cloud and subscribers, potentially mitigating the poor performance observed in Section 3 that was not explained by geographic distance. We refer to this model as the *carrier cloud*, signifying cloud infrastructure distributed among and hosted by telecom carriers.

The geographical proximity of the carriers' datacenters to end-users is essential to lowering response times, which further complements the high-quality SLAs that the carrier can provide. In many cases, a country-wide datacenter operated by a carrier would ensure the jurisdictional properties required for data privacy outlined in Section 4, while at the same time being large enough to leverage traditional cloud computing economies of scale. We acknowledge that country-wide datacenters, particularly in small countries, may not be as efficient as current datacenters, but by overcoming the key limitations mentioned above, they could drive enterprise customers to adopt cloud offerings, leading to increased scale.

As mentioned earlier, localized datacenters may open the space for novel service offerings and business models. Apart from being an ideal place on the network for deploying CDNs and caches as done traditionally, we envision carrier clouds would drive novel business models and innovative offerings that are latency sensitive or constrained by regulatory concerns. Services as Internet telephony and video-conferencing solutions, enabled by these localized clouds, may drive small entrants to competitively play in this emerging market, as well as provide for reduce costs to incumbents through adoption of modern solutions and migration away from ossified and costly technologies. Online multiplayer games also fall under this group of latency sensitive applications that would benefit from localized clouds.

Another field in which we envision substantial growth is that of network management outsourcing. Modern networks are difficult to manage and their operation takes up an ever growing chunk of IT budgets. Moreover, maintaining a high perspective over network management is of value and leads to improved efficiencies. For example, Network Intrusion Detection Systems (NIDS) benefit from observing network traffic from a network-wide vantage point as the ability to correlate and extract malice from benign traffic improves. Blocking malicious traffic is also best done higher towards the Internet backbone, thereby saving limited resources towards the edges. As discussed in Section 5, SDN is a very promising technology that we envision would trickle into network management services on every scale. The expressiveness of SDN would make these attractive for outsourcing of network management solutions, while the business relations between carriers and their subscribers make the former ideally suited to offer such outsourcing services. Furthermore, as discussed in Section 5, many SDN-based policies are latency sensitive and thereby depend on low-latency access to the datacenters where the SDN controllers reside. Finally, the carriers themselves are in need of cloud-based service to manage their own operations. From a business perspective that may be a key driver motivating carriers to enter this space, with their own IT departments being the first beneficiaries of the service as well as a valuable anchor customer.

Nevertheless, carriers traditionally avoided the setup and operation of complex systems and often outsource their core IT operations. While carriers are expanding their business offerings and reducing their outsourcing partnerships in favor of greater inhouse control, we do not expect most large players could make the shift from being almost totally dependent upon their technology providers in time to compete in this emerging market. It appears, however, that the large IT outsourcing companies that currently provide technology to the carriers have noticed the opportunities we present here and are transitioning to offer carriers the ability to sell these very same services. A good example is the new "Cloud-Band" [4] offering from Alcatel-Lucent.

7 Related Work

Prior work on cloud provider measurement has evaluated several facets of cloud performance. Garfinkel evaluated the usability and throughput of Amazon's web services, focusing primarily on throughput [11]. Li et al.'s comparison of cloud providers takes a similar tack, evaluating a number of performance dimensions beyond just response time or latency [19]. Our work differs from these in that we focus on the architectural question of the impact on datacenter location on performance, whereas these works aim to assist users of cloud services in choosing a cloud provider. The most similar study to our own is Kagan's a large-scale response time and reliability comparison of several cloud providers from thousands of network vantage points [16]. While the methodology of this work is somewhat opaque (being an industry presentation), and this work takes advantage of a significantly richer set of network vantage points, our results are in agreement.

Prior work analyzing the legal constraints associated with cloud computing come from a variety of academic, news, and legal sources. Numerous laws govern the space of data protection, chief among them are HIPAA [14], FERPA [9], GLBA [12], the USA PATRIOT Act [28], and the EU Data Protection Directive [7]. In 2010 Kuner surveyed global data protection and transborder dataflow laws, summarizing the relevant portions of each [17]. Whittaker has published a series of articles detailing the USA PATRIOT act and its impact on cloud computing data locality [33, 34, 35]. Mather et al. [27] looked at the various legal regulations effecting enterprises wishes to outsourcing data, focusing on risk and compliance. Multiple legal scholars [21, 6, 22] have weighed in on outsourcing data across borders, as well as the general issues with cloud computing and compilance [10]. Our contribution in this area is to aggregate and discuss these sources, rather than analyzing the laws ourselves.

8 Conclusions

We show that the current model of large-scale, centralized datacenters presents challenges for a variety of important and useful applications. In particular, applications that are latency-sensitive or which handle data subject to jurisdictional regulation are better served by a large number of smaller datacenters. Finally we identify carriers as ideally positioned to offer cloud-services out of their localized datacenters, predominantly on a state- or country-wide scale. These would overcome both the jurisdictional constrains as well as improve response times. These *carrier clouds* may be able to provide the strong SLAs as required by enterprise customers, thereby leading to attractive economies of scale, the key to competitive cloud computing services.

References

[1] A. Andrzejak, D. Kondo, and S. Yi. Decision model for cloud computing under sla constraints. In *Modeling, Analysis Simulation of Computer and Telecommunication Systems (MASCOTS), 2010 IEEE International Symposium on*, pages 257 –266, aug. 2010.

[2] S. Banerjee, T. Griffin, and M. Pias. The interdomain connectivity of planetlab nodes. *Passive and Active Network Measurement*, pages 73–82, 2004.

[3] D. Binning. Top five cloud computing security issues. *Computer Weekly*, April 24, 2009.

[4] B. Casemore. Alcatel-lucent banks on carrier clouds, 2011. Nov 25, 2011. Date retrieved: December 9, 2011.

[5] Commission decisions on the adequacy of the protection of personal data in third countries. `http://ec.europa.eu/justice/policies/privacy/thridcountries/index_en.htm`, Feburary 14, 2011.

[6] M. P. Eisenhauer. Privacy and Security Law Issues in Off-shore Outsourcing Transactions. `http://www.outsourcing.com/legal_corner/pdf/Outsourcing_Privacy.pdf`, Feburary 15, 2005.

[7] EU Data Protection. `http://ec.europa.eu/justice/policies/privacy/index_en.htm`, October 29, 2010.

[8] Export.gov - Safe Harbor. `http://export.gov/safeharbor/`, 2011.

[9] Family Educational Rights and Privacy Act. `http://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html`.

[10] ”FERPA Final Regulations Note”. `http://www.ofr.gov/OFRUpload/OFRData/2011-30683_PI.pdf`.

[11] S. Garfinkel. Technical report tr-08-07: An evaluation of amazons grid computing services: Ec2, s3 and sqs. Technical report.

[12] GrammLeachBliley Act. `http://www.gpo.gov/fdsys/pkg/PLAW-106publ102/content-detail.html`, November 12, 1999.

[13] T. Greene. The U.S. Patriot Act has an impact on cloud security. *Network World*, September 29, 2009.

[14] The Health Insurance Portability and Accountability Act of 1996. `http://www.gpo.gov/fdsys/pkg/PLAW-104publ191/content-detail.html`, August 21, 1996.

[15] P. Hofmann and D. Woods. Cloud computing: The limits of public clouds for business applications. *Internet Computing, IEEE*, 14(6):90 –93, nov.-dec. 2010.

[16] M. Kagan. Global Cloud Performance Data. `http://www.cloudconnectevent.com/2011/presentations/free/76-marty-kagan.pdf`, 2011.

[17] C. Kuner. Regulation of Transborder Data Flows under Data Protection and Privacy Law: Past, Present, and Future. *TILT Law & Technology Working Paper*, Oct. 1, 2010.

[18] N. Kushman, S. Kandula, and D. Katabi. Can you hear me now?!: it must be bgp. *ACM SIGCOMM Computer Communication Review*, 37(2):75–84, 2007.

[19] A. Li, X. Yang, S. Kandula, and M. Zhang. Cloudcmp: comparing public cloud providers. In *Proceedings of the 10th annual conference on Internet measurement*, pages 1–14. ACM, 2010.

[20] Mimecast Email Archiving. `http://www.mimecast.com/What-we-offer/Email-Archiving/`, 2011.

[21] R. A. Z. Monson. HIPAA and Foreign Outsourcing. *HIPAAlert*, February 23, 2004.

[22] ”Offshore Outsourcing of Data Services by Insured Institutions and Associated Consumer Privacy Risks”. `http://www.fdic.gov/regulations/examinations/offshore/`.

[23] P. Patel, A. Ranabahu, and A. Sheth. Service level agreement in cloud computing. *Cloud Workshops at OOPSLA09*, pages 1–10, 2009.

[24] V. Paxson. End-to-end routing behavior in the internet. In *Conference proceedings on Applications, technologies, architectures, and protocols for computer communications*, SIGCOMM ’96, pages 25–38, New York, NY, USA, 1996. ACM.

[25] L. L. Peterson, T. Anderson, D. Culler, and T. Roscoe. A blueprint for introducing distruptive technology into the Internet. In *Proceedings of the 1st ACM Workshop on Hot Topics in Networks (HotNets-I)*, Princeton, NJ, Oct. 2002.

[26] The Open Compute Project. `http://opencompute.org/`, 2011.

[27] S. L. Tim Mather, Subra Kumaraswamy. *Cloud Security and Privacy: An Enterprise Perspective on Risks and Compliance*. O’Reilly, 2009.

[28] ”Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism Act”. `http://www.gpo.gov/fdsys/pkg/PLAW-107publ56/content-detail.html`, October 26, 2001.

[29] USA Patriot Act comes under fire in B.C. report. *CBSNews*, October 30, 2004.

[30] V. Valancius, N. Laoutaris, L. Massoulié, C. Diot, and P. Rodriguez. Greening the internet with nano data centers. In *Proceedings of the 5th international conference on Emerging networking experiments and technologies*, 2009.

[31] F. Wang, Z. Mao, J. Wang, L. Gao, and R. Bush. A measurement study on the impact of routing events on end-to-end internet path performance. In *ACM SIGCOMM Computer Communication Review*, volume 36, pages 375–386. ACM, 2006.

[32] Z. Whittaker. Summary: ZDNet’s USA PATRIOT Act series. *ZDNet*, April 27, 2011.

[33] Z. Whittaker. Defense giant ditches Microsofts cloud citing Patriot Act fears. *ZDNet*, December 7, 2011.

[34] Z. Whittaker. Microsoft admits Patriot Act can access EU-based cloud data. *ZDNet*, June 28, 2011.

[35] Z. Whittaker. Updated European law will close Patriot Act data access loophole. *ZDNet*, November 8, 2011.