Research challenges for cross-cloud applications

Cuadrado, F; Navas, A; Duenas, JC; Vaquero, LM

2014. IEEE

Research Challenges for Cross-Cloud Applications

(Invited Paper)

Félix Cuadrado
School of Electronic Engineering and Computer Science
Queen Mary University of London, UK
Email: felix.cuadrado@qmul.ac.uk

Álvaro Navas, Juan C. Dueñas
Center for Open Middleware
Universidad Politécnica de Madrid, Spain
Email: alvaro.navas@centeropenmiddleware.com
juancarlos.duenas@centeropenmiddleware.com

Luis M. Vaquero
HP Labs Bristol, UK
Email: luis.vaquero@hp.com

***Abstract*—Federated clouds can expose the Internet as a homogeneous compute fabric. There is an opportunity for developing cross-cloud applications that can be deployed pervasively over the Internet, dynamically adapting their internal topology to their needs. In this paper we explore the main challenges for fully realizing the potential of cross-cloud applications. First, we focus on the networking dimension of these applications. We evaluate what support is needed from the infrastructure, and what are the further implications of opening the networking side. On a second part, we examine the impact of a distributed deployment for applications, assessing the implications from a management perspective, and how it affects the delivery of quality of service and non-functional requirements.**

I. Introduction

Cloud computing is transforming Internet applications, as new requirements are identified for Internet services. Initial Cloud computing offerings were concerned mostly about exploiting elasticity so that operational costs were adjusted to the real demands of the application. Services have matured, requiring stronger fault tolerance capabilities, as well as the ability to cope with increasingly strict regulations of governmental bodies regarding cloud applications and services.

Additionally, the ease of distribution of applications and services through the web and application stores has increased the potential audience for any service to a global scale. In order for Internet-scale applications to be competitive, they need to provide a satisfactory experience to users all around the world. These drivers have pushed cloud applications to adopt a geographically distributed approach, where multiple instances of the application are deployed across data centers in strategically placed locations.

The solution to this situation is migrating from the single cloud model to a cross-cloud environment. A Cross-cloud infrastructure is defined as the federation of multiple datacenters, offered by potentially multiple providers, with homogeneous APIs for acquiring virtual resources on demand. This model benefits application providers, as increasing competition would likely lower prices and let them avoid vendor lock-in. Even more, this freedom to choose where to deploy application with minimum management cost increase would allow them to easily comply with the increasingly strict regulations on user data management [1].

Federated cloud infrastructures are emerging to provide support to these reliability and distribution needs, with several models being currently adopted by industry:

- Major public cloud providers offer multiple data centers (in Amazon's terminology, multiple availability regions), where services can be deployed. The management APIs for accessing the infrastructure are homogeneous in these cases, although the level of integration of cloud services across regions (even billing) is often limited.
- The cloud-bursting model can be seen as another form of cloud federation. In this approach small private clouds overflood to an external cloud provider (usually a public cloud) when required.
- Multiple admin domains (*a la* Grid Virtual Organisations [2]): pre arranged agreements between partners to aggregate network, create provisioning, monitoring and security services across all otherwise independent sites.

Cloud infrastructure federation is progressing substantially A federated cloud extends the IaaS model across the boundaries of multiple data centers. The same format for packaging Virtual Machines (VMs) can be deployed on any of the federated data centers, and there is a common API that can be used for managing deployed VM instances across any data center. Moreover, barring the first instances mentioned, these associations combine the infrastructure from multiple providers, which brings the added advantage of avoiding vendor lock-in, and potentially enabling smaller players to compete in the cloud area by bringing their distinct advantage. However, deployment-level compatibility is not sufficient for fully addressing the Quality of Service (QoS), security and reliability needs of cross-cloud applications.

On the one hand, cloud federation efforts are not addressing the networking component of cross-cloud applications enough. Without first-class network awareness in the infrastructure offerings, there are numerous performance and reliability challenges that cannot be overcome for cross-cloud applications. Network communication through different clouds presents multiple challenges such as secure communications, the management of dynamic channels and the collection of network analytics. The final concerns directly involve cloud providers, as a federation would need an offering of different cross-cloud applications and services that ease the deployment and management of applications.

On the other hand, current applications are not designed to take advantage of a Cross-cloud environment. Applications

must be aware of the specific cloud platforms where they can be deployed. Members of a cross cloud infrastructure must declare some level of individual information (such as name, and location) that can be used by applications. At application-logic level, there are several further aspects that need to be adapted to work adequately on a cross-cloud environment, And such, applications must be able to adapt to the different characteristics of each cloud, with its own virtual instances and QoS.

The aim of this paper is to identify the main challenges and concerns that need to be addressed in order to make Cross-cloud infrastructures a reality. The next section examines the networking dimension of the management of cross-cloud environments. Section 3 examines how such an architecture affects applications and describes the main design principles a developer needs to know in order to get the most from this environment. The paper concludes with a reflection on the fundamental challenges for cross-cloud applications.

II. Cross-cloud Networking

Cloud applications can scale both horizontally and vertically [3], often within the boundaries of a data center. Cloud-bursting does scale out to an external cloud, in order to overcome the limited capacity of a private data center. The cross-cloud dimension enables applications to become pervasive across the internet. Therefore, scalability crosses data center boundaries, in order to exploit the different location of each data center, and further increase application resiliency. In this scenario, the networking dimension of applications can no longer be ignored for management purposes, as it becomes a fundamental element of the runtime behavior of cloud applications. We analyze in this section the main network aspects related to management that open new possibilities in a cross-cloud infrastructure.

A. Network Support for Applications

A distributed cloud application can be viewed as an application-level overlay, with multiple computing nodes. Applications have some visibility on the overlay nodes, but there is no information available about the network links connecting the overlay. Applications have acquired some control over their topology by employing proxies and middleboxes that reroute requests, as well as using the DNS service for tasks such as load balancing. However, the lack of access to network-level information (neither network topology information, nor potential of current network performance figures) can severely limit the effectiveness of these techniques. Additionally, ignoring the status of the underlying network can also create inefficiencies on the networking infrastructure [4] (the impact of P2P applications in the network infrastructure of an ISP being a clear example [5]).

As the application overlay adapts dynamically across the cross-cloud infrastructure, the impact on the network will vary. However, the networking fabric is statically provisioned and configured, so even if applications became network aware, the lack of network control impedes their performance.

Software-Defined Networking (SDN) is an emerging paradigm that aims to decouple the control and management planes of the networking fabric, enabling software-based control over the network elements. SDN technologies (such as the Openflow specification for flow forwarding management), can orchestrate dynamic management of networking resources, whereas traditionally network management has been handled with minimum dynamic action.

SDN allows to combine virtual networking, and tunneling technologies with the ability to modify the forwarding behavior of the networking elements. SDN-enabled networks allow cross-cloud application deployment to also control the networking infrastructure that links cloud instances together, as well as the links with end users. The extent of SDN impact for cross-cloud applications will be strongly dependent on its deployment across the Internet. Currently SDN is experiencing significant success inside data centers, where the added flexibility can be easily exploited. However, at this point it is not clear how much it will be pushed outside data centers. The multi-tenant nature of the Internet complicates achieving wide area guarantees, with some research initiatives pursuing collaboration points that allow to agree on resource reservation with certain guarantees [6].

We believe SDN is a necessary component for enabling Internet-wide elasticity, although there is room for discussing the specific technologies that will enable the SDN concepts. Today there are significant doubts about the scalability of Openflow for enabling network dynamics at the WAN level [7], although recently a major Internet player has disclosed its use of Openflow for implementing inter data center traffic engineering [8], with a relatively simple topology.

B. Network Analytics

Management decisions of cloud applications are based on the collected runtime information at the computing nodes. Statistics about the use of computing resources, such as memory, or processor, as well as application-level metrics such as number of requests per second are taken to decide whether application deployment should be adapted. Another implication of a deployment of SDN across the network would be the possibility of feeding network-level information to the management plane of applications, building computing and network combined models of applications [9].

The networking element of a distributed application plays a fundamental role, in particular when supporting distributed application aspects such as replication, synchronization, and migration. Moreover, the network usage of the application incurs on additional costs that should be considered along the cost of computing resources. Factoring in the networking dimension of the application can substantially help on this analysis.

In order to help applications management, network monitoring tools must be able to operate at a fine-grained scale to obtain per-application information. However, there is a tradeoff between the scalability of the collection technique, and the granularity and accuracy of the solution. There is a range of networking monitoring specifications, including Netflow, S-Flow and OpenFlow 1.3 (with the possibility to compute per-flow counters), but their scalability for application-level monitoring needs to be assessed first.

C. Cross-Cloud Wide Services

The separation of concerns principle has inspired middleware platforms to address non-functional requirements of applications with the provisioning of services that can be easily used by the deployed components. Cloud platforms provide a rich catalog of services including security options, billing or load balancing that can be easily integrated into deployed applications. However, in a cross-cloud scenario, supporting these services would require a significantly more complex common management API, and would greatly complicate data center management, with low capacity data centers potentially being forced out of the federation.

We believe these support services can be provisioned on demand to support cloud applications. Whenever an application with a supporting service (e.g. policy-based user access control) is deployed to a cloud, the supporting cross-cloud services would also be deployed either at the same location, or at a location close enough to provide the service seamlessly to the new instance of the application. Dynamic provisioning of these services would greatly alleviate its overhead on the whole infrastructure. However, the co-deployment of these services shares most of the challenges inherent to the management of cross-cloud applications themselves, with the added constraints of modifying the network flow for the applications, as well as the challenge of selecting the most adequate location for them.

D. Internet-wide VM migration

The networked nature of a cross-cloud infrastructure also restricts the speed at which network-wide migrations can be executed. Migrations inside a data center are not instantaneous; there is a minimum time for loading the virtual machine image, copying the application state, configuring and starting the new instance. The situation is considerably more unpredictable across distributed clouds, as virtual appliances and potentially the application state need to be transported across the network, which will greatly contribute to the total migration time, with the variable nature of the network complicating predictions on the total expected time. The availability of a SDN fabric across the cross-cloud infrastructure can handle the negotiation of virtual networks to transport the information, as well as provide the means for ensuring the security requirements of the application over the process.

The transition time while changes are being applied to the deployment can potentially disrupt the live operation of the application. Requests might need to be temporarily held, and rerouted, disrupting the level of service aimed at by the service. The networking side of the infrastructure can orchestrate the handover of requests during the transition, with either dynamic DNS [10] or level 2-3 in-network processing [11] [12] for transparently handling the migration. There is significant work ahead to improve how applications handle these transient states in a transparent way, as these transitions will redirect incoming requests, which may substantially increase the overall latency of the application during the migration.

E. Towards a Networked Cloud Marketplace

We have seen in this section that the networking component of applications is going to become substantially more relevant as cross-cloud applications become popular. However, currently inter-network connectivity providers (i.e. Internet Service Providers) and computing infrastructure providers (e.g. cloud providers) operate infrastructure silos that don't share any information with each other. Cloud federation will pave the way for interoperability at the computation and storage side of the infrastructure, but as long as the networking elements remain independent, there are severe limitations to what can be achieved by cross cloud applications. The concept of a marketplace for cloud resources has been analyzed by the research community [13], but we believe an integrated approach with the networking infrastructure is required.

We envision the creation of a networked resources marketplace, joined by both cloud providers and Internet Service providers. The marketplace would provide cross cloud applications access to the combined substrate of resources from multiple network domains, comprising networking, and computing infrastructure. The addition of network topology information for each data center would allow smaller players to provide computing resources that have unique location features that might be desirable for some applications. The marketplace could incorporate algorithms that take as input application requirements, and automatically negotiate, compose and create compute and network overlays, while hiding the infrastructure details from the applications view.

An integrated marketplace would not only simplify substantially cross cloud applications management, but also would provide substantial benefits for the network infrastructure providers. Instead of performing traffic engineering based on packets and traffic matrices, the marketplace would inform them of the real use each application makes of their infrastructure. That would enable them to consolidate traffic along their infrastructure, ensuring the real requirements from applications while at the same time achieving higher utilization rates (for a sample study on the benefits of content providers and infrastructure providers see [14]).

III. Cross-cloud Application Management

The distributed nature of cross-cloud applications allows to improve Quality of Service aspects of the runtime system, but at the same time it brings many challenges to the runtime management of the application. We start the analysis by evaluating the direct implications of the choice of location, and further on discuss the implications of a distributed deployment for the non-functional characteristics of applications.

A. Minimizing Global User Latency

A cross-cloud infrastructure offers a potentially diverse number of options for deploying cloud applications. Each data center will have a unique geographical location, as well as distinct network connectivity characteristics. Traditionally application design has ignored the networking side, as well as their connectivity to the rest of the Internet. However, network latency (from client to server) can be an important factor of the application latency that can substantially impact the perceived Quality of Experience for application users. As many cloud applications will have a global user base, it will no longer be possible to select a single location where network latency meets the set requirements for all users.

The problem of moving closer to clients has been traditionally addressed by Content Distribution Networks (CDNs). CDNs deploy a large number of caches embedded in access networks, that are used for serving static content to clients from a closer location, and additionally reduce the overall traffic on Internet Service Providers. Applications offload static content downloads to a CDN without any code modification. However, there is no way to automatically transfer computing elements close to the users in a similar way to what is done with static files. There was some initial interest on implementing such a edge computing model [15], but widely applicable solutions have not been found.

Automated edge computing deployment similar to what CDNs achieve would directly address the requirements on user latency. A similar approach has been advocated by the mobile computing community with the concept of cloudlets [16], VMs that can be dynamically deployed close to where the mobile user is. On an extreme case, it would be possible to minimize latency by deploying pico data center instances of the application inside the user browser [17], and have a fundamentally better architecture for providing strong user data management and privacy guarantees.

The closer the computation is to end users, the higher the number of edge servers that need to be deployed. This does not only increase the cost of the application deployment, but also complicates the internal architecture of the application, as instance synchronization might become the bottleneck of the application. These aspects are seldom considered on the existing literature. Moreover, in extreme cases latency problems are shifted from user-to-application, to intra-application restrictions from the edge servers to the core elements. In the latter case, the tighter control from the network that SDN promises would in principle allow to have working application overlays with network guarantees.

B. Dynamic Application Deployment

The fundamental deployment decision of cross-cloud applications is to determine what is the right geographical distribution of the application logic. The decision has to determine the right number of instances of the application virtual instance profiles, the infrastructure cost, the intended performance (including estimated network latency with users in case the service has strict requirements in that front), and resiliency. This challenge comprises several research problems that have been thoroughly investigated by the distributed systems and networking communities, such as replica placement [18], or server selection [19]. A differentiating factor for cross-cloud applications when compared to single cloud autonomic adaptation techniques is that the cost of geographical migrations can be significantly higher.

Cross-cloud deployment decisions are significantly more challenging because the optimal solution changes over time. User latency directly depends on the geographical distribution of the user workload, and normal daily patterns, shifts in popularity, and region-specific preferences will shape different scenarios that applications might have to adapt to. These shifts do not only impact the established companies; a breakout app from a small startup can grow to have millions of daily active users, and it might not have the internal expertise to manage the required infrastructure to support a global user base.

The time scale at which workload changes render the current deployment obsolete will influence the type of algorithms that can be implemented. Slow changes allow for offline, more costly reasoning techniques, which can provide accurate results, whereas highly volatile conditions require online methods that can decide on changes as quick as possible, trading off quality of the results for timeliness.

We need to take all these factors in consideration in order to find the right timescale for cross-cloud applications dynamic deployment. The answer will depend on the characteristics of the support infrastructure, as well as the variations in the user workload.

Indeed, a pressing matter for the research community is the lack of available datasets about realistic Internet service workloads. Researchers have to rely on assumptions, synthetic data and simulations, which can significantly lessen the impact of the academic contributions to these industrial problems.

C. Data Management in the Face of Constraints

Cloud computing has brought also significant challenges related to user privacy. Cloud abstractions do not provide strong user data privacy guarantees, with the lack of knowledge about data center location easily bypassing country-specific regulations for citizen data protection (with the revelations about the PRISM program exposing the limits of current solutions).

In order to prevent these risks, governments are creating stricter data protection regulations that impose constraints on how citizen data must be managed, such as keeping the data in a data center located at the country, or region [1]. These restrictions bring some reassurance on the validity of the law, but they only cover the final storage destination. It is not clear that is a strong enough requirement, as user data travels over application components, networks, and middle-boxes before being stored or read from its persistent location.

For both of these challenges, one possible approach is to move towards privacy and data management protection by design [20], explicitly modeling privacy constraints, and embedding them into the design states of applications to be able to provide guarantees that are not violated at any time. However, in a cross-cloud environment, these requirements won't be possible to hold without having complete end-to-end guarantee, involving both the computing and networking elements that are used at runtime by the cloud application.

D. Keeping a Consistent State

Web applications have become increasingly stateless, following the SOA and RESTful architecture design principles [21] for better scalability and portability. The move to a cloud environment has further consolidated that trend, as horizontal scalability requires adding and removing application instances, which is greatly simplified for stateless services.

While statelessness is a desirable quality of application components, Internet services do have a state that needs to be maintained, usually at a persistent data store. In the last years, many cloud services are opting for highly scalable storage solutions that forgo the strong consistency guarantees

of traditional SQL databases in favor of eventual consistency models [22].

However, the rise of weaker consistency models has substantially complicated the design of cloud applications, as applications can experience temporal inconsistencies that need to be handled directly by the application logic. Cross-cloud applications will rely even more on these partitioned data stores due to the distributed nature of the infrastructure, making this problem more pressing.

The rise of eventual consistency was motivated by Brewer's Theorem, which states that a distributed system can only achieve at the same time two out of these three properties: Data Consistency, Availability, and Partitioned information (CAP). However, the three dimensions of CAP are not binary; there is a wide spectrum of intermediate values for each of these characteristics that can provide more convenient solutions for the real needs of cloud applications [23]. The distributed systems and database communities are actively researching on novel consistency models that can provide strong consistency qualities under certain restrictions, which would substantially benefit cross-cloud applications. As an example, the Red/Blue consistency model [24] allows the coexistence of two consistency models (strong and weak) within an application. While this type of models requires analyzing the invariants of the specific application, it substantially improved performance and consistency guarantees.

E. Reliability Aspects

Virtual cloud resources are provided by physical servers from data centers, which are bound to experience failures, even in the case of the most mature vendors. Hardware and software errors can either crash the virtual machine or significantly degrade the performance of the hosted services [25].

Cross-cloud applications can engineer solutions that improve reliability by replication. A cross-cloud infrastructure allows selecting data centers with different location, characteristics and provider. While individual providers might provide lesser reliability guarantees than large public cloud platforms, its combination can potentially achieve a level of resiliency better than any single cloud provider.

Cross-cloud applications are composed at runtime by multiple services: application-specific functional components, cross-cloud services that take care of non-functional aspects (e.g. security), and network overlay management functions. In order to achieve high availability all the system components need to be protected against failures, but keeping track of the reliability status of the application can be costly. There is a need for new techniques and tools that test the overall reliability of a cross-cloud application, shutting down parts of the deployed application and infrastructure, and verifying on demand the reliability of the application [26].

Cloud application resiliency does not only apply to total failures of a component or the whole service, but also performance degradations that may break the QoS targets of the application. In the case of large-scale services, with strict time requirements for completing a service request, latency is dominated by the long tail of results [27]. Moreover, when running applications on virtual infrastructure, there is substantial variability in the effective performance obtained by the acquired instances [28]. The real time software engineering discipline has developed many architectural solutions for critical systems that can also be incorporated to these services to safeguard response time objectives.

F. Applications Management Architecture

From the early days [29] of the warehouse-scale computing era [30], services have been developed with a logically centralized management plane. On the other hand, fully decentralized, Peer to Peer architecture models have significantly faded in popularity, with centralized approaches allowing finer control over the computing resources, and scaling up to the needs from the major Internet-scale applications.

However, the multi-tenant, highly distributed nature of a cross-cloud environment presents a different environment where the centralized versus distributed debate needs to be reassessed. The hosting of the central management point of a cross-cloud application might be complicated, as it will be subject to trust and reliability requirements that might not be offered by any of the available data centers.

For geo-distributed applications, a fully centralized management plane presents some difficulties. Decisions must consider all the relevant runtime information collected at the distributed instances. Therefore, decision algorithms can either be run on the distributed environment [31], potentially incurring on significant penalty from the network component, or run on a centralized location [32], after having moved all the information to a single points. In any of these cases, the decision will require the use of parallel computing techniques for large-scale services. The cited papers report total computation time in the order of hours and days, making it not well suited for a highly dynamic environment. A decentralized management scheme can process most of the monitoring information locally, and only exchange a subset of that information with its peers. Some research studies show how decentralized decision making can be taken to geo-distributed application, combining the ability to enforce user-defined policies with an added level of resiliency over a central solution [33].

G. Security Challenges

Security-related aspects are one of the main factors hampering further cloud adoption. The inherent multi-tenancy of the environment, and the existence of three different actors in a normal scenario (the end user, the application provider, and the cloud provider), raise a significant number of security issues [34].

A cross-cloud environment presents additional security challenges. Instead of operating in a single data center, controlled by one company, cross-cloud applications can operate over a multi-tenant infrastructure from different providers. There is a need for extending security best practices to a federated environment. Federated identity and trust have been some of the key elements of research and industry since the early days of service-oriented architecture applications [35]. These advances can be applied to both the applications accessing the infrastructure, and also to inter-cloud communications.

Cross cloud solutions also need to potentially consider the networking aspect of the federated infrastructure. Data must

not only be stored securely, but communications need to be protected with the adequate means.

IV. Conclusions

The main differentiating aspect of cross-cloud applications is the nature of their physical distribution as an overlay across networks and data centers. This characteristic makes the network element critical for an effective deployment and management. In the paper we have highlighted how SDN-enabled networks can become a key element for fully realizing the vision behind these applications, with the ability to provide virtual links with certain guarantees, support to seamlessly integrate network-wide services and support for the low-level activities related to WAN-scale virtual machine migration. We believe a networked cloud marketplace might provide incentives for the different infrastructure stakeholders to collaborate, including the potential for better infrastructure management based on using information about the real needs of the applications that use the infrastructure.

On the application side, the location dimension greatly impacts application architecture and management. Decisions become substantially more complex; a cross-cloud infrastructure provides the required means for applications to achieve satisfactory performance for a changing workload of users around the world. The architecture of these applications can resemble a dynamic, distributed overlay, that raises several challenges regarding how to manage the internal application state, provide service reliability, and ensure security requirements.

References

[1] Z. N. Peterson, M. Gondree, and R. Beverly, "A position paper on data sovereignty: The importance of geolocating data in the cloud," in *USENIX NSDI*, 2011.

[2] I. Foster, C. Kesselman, and S. Tuecke, "The anatomy of the grid: Enabling scalable virtual organizations," *International journal of high performance computing applications*, vol. 15, no. 3, pp. 200–222, 2001.

[3] L. M. Vaquero, L. Rodero-Merino, and R. Buyya, "Dynamically scaling applications in the cloud," *ACM SIGCOMM CCR*, vol. 41, no. 1, pp. 45–52, 2011.

[4] L. M. Contreras, V. López, O. G. De Dios, A. Tovar, F. Munoz, A. Azanon, J. P. Fernandez-Palacios, and J. Folgueira, "Toward cloud-ready transport networks," *IEEE Communications Magazine*, vol. 50, no. 9, pp. 48–55, 2012.

[5] V. Aggarwal, A. Feldmann, and C. Scheideler, "Can isps and p2p users cooperate for improved performance?" *ACM SIGCOMM CCR*, vol. 37, no. 3, pp. 29–40, 2007.

[6] S. S. Lor, L. M. Vaquero, D. Audsin, P. Murray, H. Puthalath, B. Melander, A. Sefidcon, T. Edwall, J. Soares, M. Melo *et al.*, "Scalable network-aware data centre federation," in *IEEE ICON*, 2012, pp. 167–172.

[7] A. R. Curtis, J. C. Mogul, J. Tourrilhes, P. Yalagandula, P. Sharma, and S. Banerjee, "Devoflow: Scaling flow management for high-performance networks," in *ACM SIGCOMM CCR*, vol. 41, no. 4, 2011, pp. 254–265.

[8] S. Jain, A. Kumar, S. Mandal, J. Ong, L. Poutievski, A. Singh, S. Venkata, J. Wanderer, J. Zhou, M. Zhu *et al.*, "B4: Experience with a globally-deployed software defined wan," in *ACM SIGCOMM*, 2013, pp. 3–14.

[9] M. Alicherry and T. Lakshman, "Network aware resource allocation in distributed clouds," in *IEEE INFOCOM*, 2012, pp. 963–971.

[10] R. Bradford, E. Kotsovinos, A. Feldmann, and H. Schiöberg, "Live wide-area migration of virtual machines including local persistent state," in *ACM VEE*, 2007, pp. 169–179.

[11] F. Hao, T. Lakshman, S. Mukherjee, and H. Song, "Enhancing dynamic cloud-based services using network virtualization," in *ACM VISA workshop*, 2009, pp. 37–44.

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

[13] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Generation Computer Systems*, vol. 25, no. 6, pp. 599–616, 2009.

[14] B. Frank, I. Poese, Y. Lin, G. Smaragdakis, A. Feldmann, B. Maggs, J. Rake, S. Uhlig, and R. Weber, "Pushing cdn-isp collaboration to the limit," *ACM SIGCOMM CCR*, vol. 43, no. 3, pp. 34–44, 2013.

[15] S. Sivasubramanian, G. Pierre, M. van Steen, and G. Alonso, "Analysis of caching and replication strategies for web applications," *IEEE Internet Computing*, vol. 11, no. 1, pp. 60–66, 2007.

[16] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, "The case for vm-based cloudlets in mobile computing," *Pervasive Computing, IEEE*, vol. 8, no. 4, pp. 14–23, 2009.

[17] J. Howell, B. Parno, and J. R. Douceur, "Embassies: radically refactoring the web," in *USENIX NSDI*, 2013, pp. 529–546.

[18] L. Qiu, V. N. Padmanabhan, and G. M. Voelker, "On the placement of web server replicas," in *IEEE INFOCOM*, vol. 3, 2001, pp. 1587–1596.

[19] R. Krishnan, H. V. Madhyastha, S. Srinivasan, S. Jain, A. Krishnamurthy, T. Anderson, and J. Gao, "Moving beyond end-to-end path information to optimize cdn performance," in *ACM IMC*, 2009, pp. 190–201.

[20] A. Cavoukian, "Privacy by design," *Take the Challenge. Information and Privacy Commissioner of Ontario, Canada*, 2009.

[21] R. T. Fielding, "Architectural styles and the design of network-based software architectures," Ph.D. dissertation, University of California, 2000.

[22] W. Vogels, "Eventually consistent," *Communications of the ACM*, vol. 52, no. 1, pp. 40–44, 2009.

[23] E. Brewer, "Cap twelve years later," *IEEE Computer*, vol. 45, no. 2, pp. 23–29, 2012.

[24] C. Li, D. Porto, A. Clement, J. Gehrke, N. Preguiça, and R. Rodrigues, "Making geo-replicated systems fast as possible, consistent when necessary," in *USENIX OSDI*, 2012.

[25] T. Do, M. Hao, T. Leesatapornwongsa, T. Patana-anake, and H. S. Gunawi, "Limplock: Understanding the impact of limpware on scale-out cloud systems," in *ACM SoCC*, 2013, p. 14.

[26] A. Cockcroft, "Dystopia as a service," in *USENIX HotCloud*.

[27] J. Dean and L. A. Barroso, "The tail at scale," *Communications of the ACM*, vol. 56, no. 2, pp. 74–80, 2013.

[28] A. Iosup, R. Prodan, and D. Epema, "Iaas cloud benchmarking: approaches, challenges, and experience." in *HotTopiCS*, 2013, pp. 1–2.

[29] S. Ghemawat, H. Gobioff, and S.-T. Leung, "The google file system," in *ACM SIGOPS OSR*, vol. 37, no. 5, 2003, pp. 29–43.

[30] L. A. Barroso and U. Hölzle, "The datacenter as a computer: An introduction to the design of warehouse-scale machines," *Synthesis Lectures on Computer Architecture*, vol. 4, no. 1, pp. 1–108, 2009.

[31] S. Agarwal, J. Dunagan, N. Jain, S. Saroiu, A. Wolman, and H. Bhogan, "Volley: Automated data placement for geo-distributed cloud services." in *USENIX NSDI*, 2010, pp. 17–32.

[32] J. Ugander and L. Backstrom, "Balanced label propagation for partitioning massive graphs," in *ACM WSDM*, 2013, pp. 507–516.

[33] P. Wendell, J. W. Jiang, M. J. Freedman, and J. Rexford, "Donar: decentralized server selection for cloud services," in *ACM SIGCOMM CCR*, vol. 40, no. 4, 2010, pp. 231–242.

[34] S. Subashini and V. Kavitha, "A survey on security issues in service delivery models of cloud computing," *Journal of Network and Computer Applications*, vol. 34, no. 1, pp. 1–11, 2011.

[35] E. Maler and D. Reed, "The venn of identity: Options and issues in federated identity management," *Security & Privacy, IEEE*, vol. 6, no. 2, pp. 16–23, 2008.