Live Migration in Emerging Cloud Paradigms

Massimo Ficco, Second University of Naples, Italy
Christian Esposito, University of Salerno, Italy
Henry Chang, University of Hong Kong
Kim-Kwang Raymond Choo, University of South Australia

Cloud Computing, IEEE, vol.3, no.1, pp.12-19, March-April 2016

Virtualization, Elasticity, and Resource Sharing enable new levels of flexibility, convenience, and economic benefits, but they result in new challenges for performance and privacy, as well as new potential security vulnerabilities. Security, privacy, and performance are major concerns for both public and private organizations that wish to shift their business-critical and sensitive data to the cloud.

According to the existing cloud model, cloud customers use the services offered by cloud service providers (CSPs) without knowing exactly where their virtual machines (VMs) and data are stored or located and where they'll be migrated. Such problems are exacerbated in emerging cloud paradigms, such as cloud federation and fog computing. Specifically, cloud federation potentially represents a new business model, in which different (geographically distributed) CSPs aggregate resources to create large-scale distributed virtual computing clusters, operating as though they're within a single cloud organization.[1] This offers greater guarantees in terms of resilience and scalability required by data- and media-intensive critical applications. Fog computing is the evolution of cloud computing toward the Internet of Things, in that it extends the cloud's elastic resource provisioning to the edge of the network, such as portable devices, smart objects, and wireless sensors.[2]

To ensure that customer expectations can be fulfilled, services offered in such emerging cloud paradigms should be formally defined by service- level agreements (SLAs) and total cost of ownership (TCO). An SLA is part of the service contract between a CSP and customer that specifies each party's obligations and describes the desired quality of service (QoS) terms. As IEEE Cloud Computing Editor-in-Chief Mazin Yousif remarked, an SLA usually doesn't include cost elements, which are determined through a separate pricing document. If the services listed in the SLA don't meet customers' expectations, the penalty is imposed on the CSP. Upon conclusion of the SLA negotiation and the brokering of the cloud resources, computing and storage resources must be allocated within the cloud infrastructure (Figure 1). This process can be modelled as an optimization problem of resource allocation with the objective of determining the allocation plan that minimizes the number of used physical host machines under the constraints of the requirements expressed in the SLA.[3] Several metaheuristic approaches address this problem.[4]

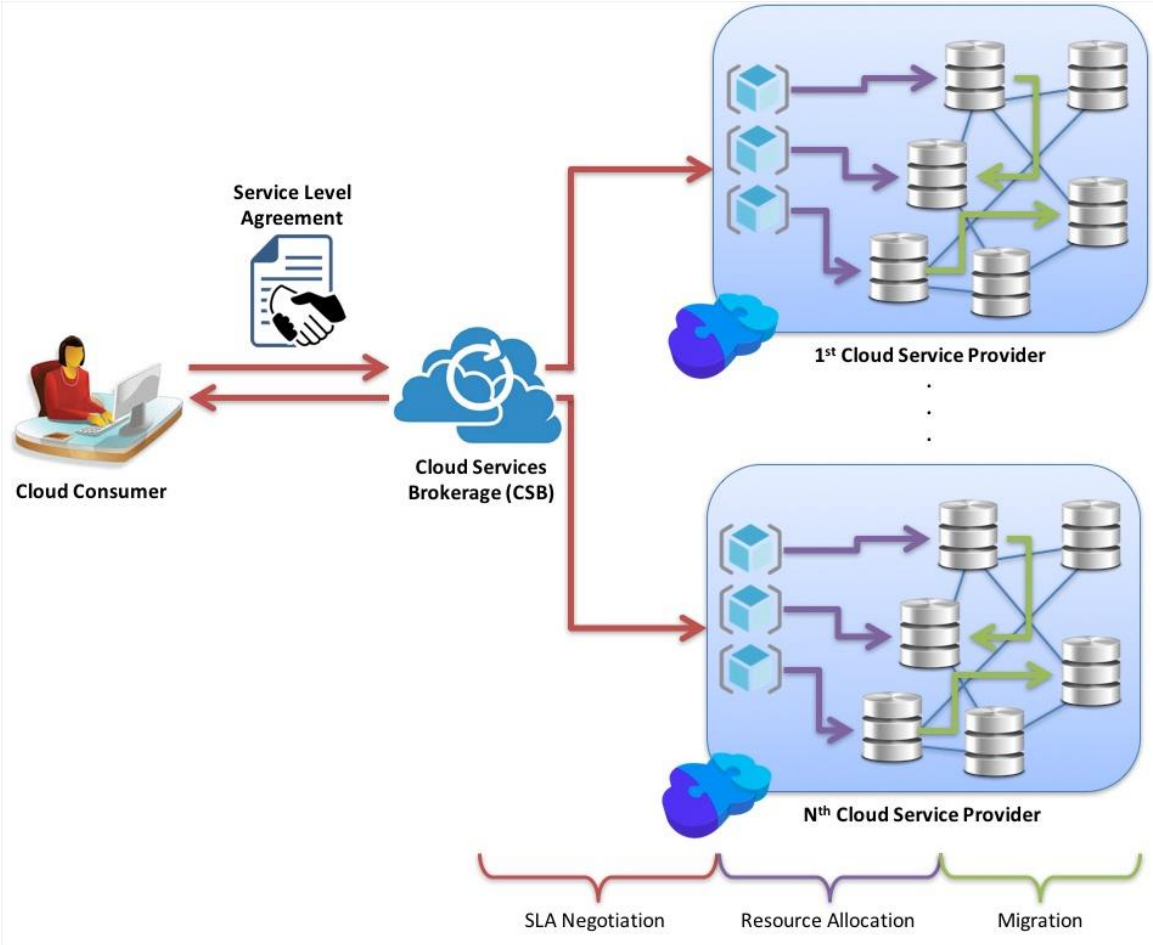

FIGURE 1. Elastic resource provisioning is managed in three stages. In the first stage, the cloud service provider (CSP) and cloud customer negotiate the service-level agreement (SLA). In the second stage, the CSP enforces the SLA by allocating virtual resources on the cloud infrastructure. In the final stage, the CSP can trigger live migration of virtual resources to adjust the planning with respect to changing conditions and resource usage.

However, while consumers are using the virtual resources, applications, and services, CPU load, memory capacity, and network traffic volume can increase. Replication, migration, and resizing represent different strategies and approaches for handling resource reallocation in the cloud infrastructure to provide the required scaling and elasticity capabilities.[5] For example, in fog computing, computational tasks are executed close to the edge nodes of the network as needed. Specifically, user mobility and possible network disconnections require frequent migration of applications and data to meet end-to-end latency restrictions, preserve resources in user devices, and conserve bandwidth in the infrastructure.

Elastic Cloud Resource Allocation

Migration in cloud computing represents the key mechanism for providing elasticity and is a pivotal success factor in cloud computing adoption. Migration can be triggered in the scenarios shown in Figure 2.

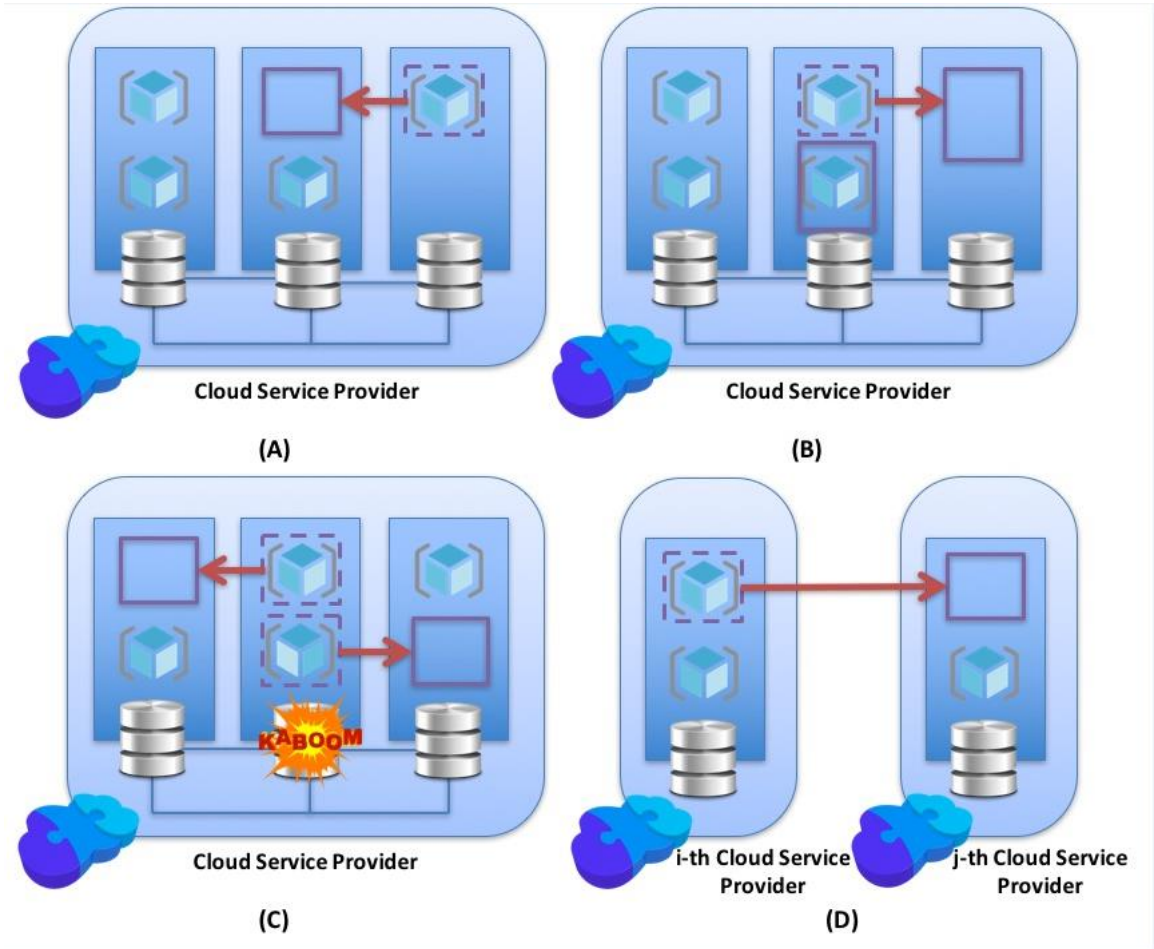

FIGURE 2. Cloud service providers can migrate resources to save energy, resolve hotspots, support fault tolerance, or consolidate resources between cloud services.

In case A, a virtual resource is moved from a hosting entity to another belonging to the same cloud federation to improve the cloud infrastructure's energy efficiency. In fact, power consumption due to running the host machines and the cooling system is one of the main cost factors in a datacenter. Therefore, the optimal usage of the physical hosts is of pivotal importance. A CSP can use migration to resolve the situation of underutilized machines, relieving them of their workloads and turning them off to reduce costs.

In case B, during the virtual resource life cycle, the number of resources needed can drastically change, and more resources might be needed to satisfy the contracted SLA. However, the hosting machine might not be able to accomplish such a resizing request because of resource competition with other co-resident tenants. Thus, one of the running entities will need to migrate to another location.

In case C, during a machine life cycle, a fault at the hardware, software, or networking level could occur, and a cyberattack on a coresident entity can affect SLA satisfaction, leading to its failure or consequent QoS degradation. Such an event doesn't have to compromise the availability and responsiveness of the virtual entities running on top of the machine affected by the failure or the cyberattack. Under these circumstances, migrating the virtual entities to the available physical machines can achieve high availability and fault tolerance.

In case D, to cope with the scalability limits of a single cloud provider, virtual resources can be migrated within the infrastructure of another federated CSP. For multiple geo-distributed datacenters, live migration over the wide area network (WAN) enables power saving across datacenters. Moreover, migrating several VMs or workloads from heavy-loaded datacenters to light-loaded datacenters allows CSPs to consolidate resources from a small datacenter into a large datacenter.

Optimization Challenges

Although metaheuristics-based approaches to cloud resource allocation have been highly successful, they need to be significantly extended with new objective functions and constraints that can model the new dynamic nature of cloud migration. In particular, when one of the previously introduced cases arises, the CSP must

- determine which virtual entities must be involved in the migration;
- identify the best action to take to resolve the situation that triggered the migration, such as migrating, resizing, or replicating the involved entities; and
- compute the best destination of the migration or replication action.

Such a decision-making process must consider the objective of maximizing the cloud infrastructure's efficiency and effectiveness by, respectively, reducing power and resource consumption and augmenting consumer satisfaction, within the constraints imposed by the agreed SLA. Given the cloud infrastructure's current state, such a process must take the necessary decision to reach a new state which leads to an optimal replacement of the virtual resources. On the other hand, we have to assume that the optimization problem underlying the migration is characterized by a large size and vast solution space due to the presence of hundreds of physical computing nodes (each running multiple VMs), and eventually also distributed datacenters. The size can be further increased in the emerging cloud paradigms, imposing additional constraints for optimal placement of virtual entities during migration.

In a cloud federation, the migrating or replicating virtual entity can be allocated anywhere within the federation infrastructure, even within a different cloud from where the original entity was located, as Figure 3 illustrates. However, unlike live VM migration in a local area network, the live WAN migration of VMs includes the workload transfer of not only the VM memory state, but also its disk image and the network connections. This could result in a VM migration time, application downtime, and a large amount of network

traffic, further degrading performance of the migrating VMs as well as VMs in the migration source and destination datacenters.

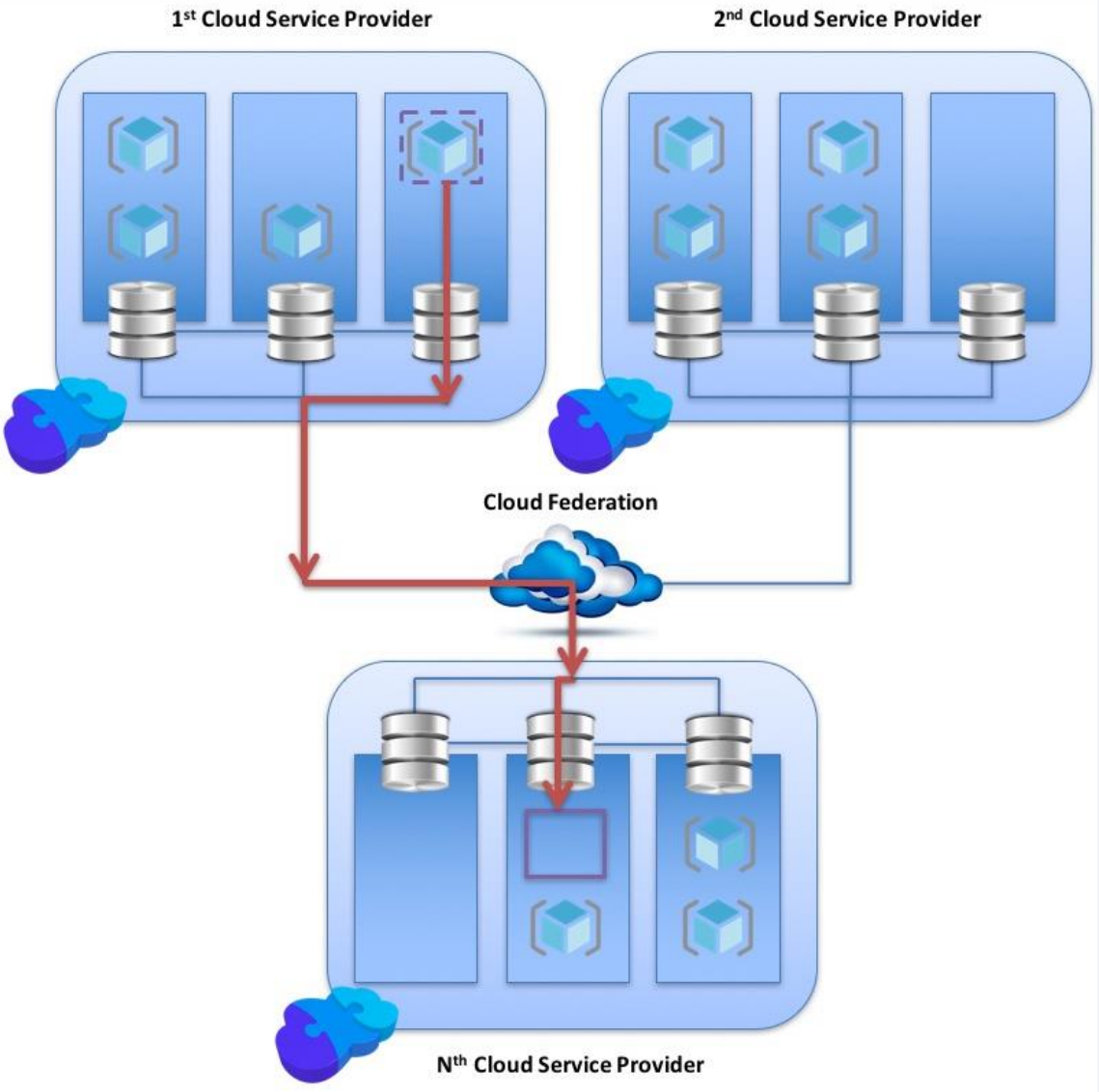

FIGURE 3. During migration within a federation of clouds, virtual resources are transferred to the destination across a wide area network, facilitated by middleware in the different cloud administration domains.

The underlying idea of fog computing is to host data and applications in an interchangeable and osmotic manner, at both cloud commodities and neighboring users. To this end, virtual entities can be migrated on-demand or at the CSP's convenience within the layered architecture, as Figure 4 illustrates.

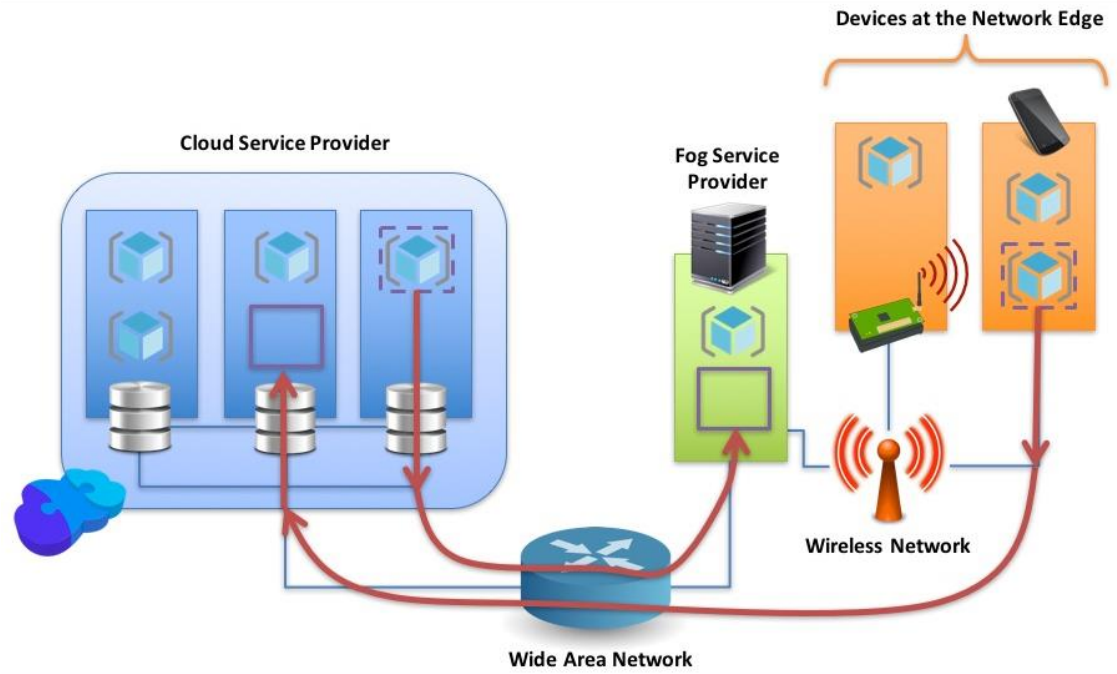

FIGURE 4. Migration in a fog computing ecosystem includes moving virtual resources from the cloud to the network edge and/or vice versa.

Personal Data Privacy Challenges

Both cloud federation and fog computing impose additional limitations on resource placement, which need to be considered when planning migration actions. Therefore, we must not lose sight of the relevant regulatory challenges if cloud or fog computing is used to store personal data that's subject to privacy or similar legal protection.[5] When it comes to privacy protection, clouds face a number of challenges, but the most relevant challenge for cloud migration is storage location restriction.[6] Many data protection laws include a provision stating that personal data collected in one jurisdiction must not be transferred to another jurisdiction for processing or storage, unless the data is afforded an adequate or similar level of protection to that it would receive if residing in the original jurisdiction.[7]

For example, Articles 25 and 26 of the European Union (EU) Data Protection Directive EC/95/46 require that, if the personal data of EU citizens is being transferred to a third country for processing, the third county must afford an "adequate level of protection" such as provided by EU member states.[8] Article 40 of the new EU General Data Protection Regulation, set to replace Data Protection Directive EC/95/46 in two years, also contains a transborder restriction to the same effect (see www.statewatch.org/news/2015/dec/eu-council-dp-reg-draft-final-compromise-15039-15.pdf). Although this requirement can be exempted if consent is obtained from all the individuals whose personal data is involved, this scenario can't be achieved realistically. Therefore, in practice, only three other options are available:

- the data is transferred to a jurisdiction that has been deemed by the European Commission as adequate;
- if the transfer is within different branches or locations of the same data controller (a term given to organizations that collect/process personal data for their own purposes), there should be formal internal "binding corporate rules" to ensure that personal data transferred outside of the EU is afforded the same level of protection by the data controller (the now-infamous safe harbor scheme can be considered a specific form of binding corporate rules for certain US companies transferring EU personal data to their US locations); and
- if the transfer is to a third party outside of the EU, the contract must contain a set of model contract terms as specified by the European Commission (see http://ec.europa.eu/justice/data-protection/international-transfers/index_en.htm).

Similar cross-border restrictions and exemptions applicable to the cloud can be found in other jurisdictions, such as Hong Kong, Malaysia, Macau, Singapore, and Taiwan.[9,10] In some of these jurisdictions, the laws allow data controllers to exercise due diligence to ensure that data transferred to a third jurisdiction will be afforded the same protection. This usually means either the data controller carries out a legal comparison between the data protection laws of the two jurisdictions to ensure compatibility, or model contract terms are used to ensure protection.[11,12]

Because CSPs are essentially contractors, the second option, which is relevant only to cross-border personal data transfer within the same company or entity, isn't applicable to them. On the other hand, most CSPs wouldn't contemplate customizing contract terms, so the third option isn't viable either. As such, data controllers would have to ensure that their collected personal data is only stored or processed in overseas jurisdictions that are either approved by the relevant regulatory authorities or have undergone a due diligence process.

The situation is further complicated by another trend in data protection law led by Russia—data localization laws. Data localization laws can read similarly to transborder restrictions, but are often written in such a way that they require the personal data collected from the citizens of one jurisdiction to be stored and processed in that jurisdiction. The twist here is the silence as to whether transborder data flow is restricted or not. For example, legal experts believe that with the Russian data localization law, personal data collected from Russian citizens will be allowed to be transferred outside of the country so long as the primary database is in Russia.[13] Other jurisdictions with data localization laws enacted include Brazil (for "connection" providers and ISPs), Vietnam (for ISP data), and Indonesia (for all businesses).[14]

Because of the increased regulation on personal dataflow across multiple jurisdictions and the trend in using clouds, the International Organization for Standardization has developed ISO 27018 for protection of personally identifiable information (PII) in public clouds acting as PII processors (www.iso27001security.com/html /27018.html).[15] The standard outlines the controls required by CSPs that will be handling personal data. According to its control on geographical location, CSPs are required to make known,

prior to contract signing and any time there's a change, the identities of countries where customers' personal data might be stored or where its subcontractors might reside, so customers can decide, object to, or otherwise terminate the contract because of transborder restrictions or data localization laws.

CSPs targeting clouds for customers to store their collected personal data must therefore consider such arrangements by providing their customers with the storage location information and allowing them the following options to determine how their personal data may be stored: white-listing (jurisdictions where personal data may be stored); black- listing (jurisdictions where personal data may not be stored); and/or sticky-listing (jurisdictions where personal data must be stored at all times).

Conclusion

In considering these challenges, our driving idea is to resolve the optimization problem (underlying migration in traditional and more advanced cloud computing architectures) in a distributed manner. Such an approach is significantly different from the state of the play, which generally consists of a centralized resolver collecting monitoring information on the system state and running metaheuristics to determine optimal planning.[16] In fact, a distributed resolution of the optimization problem should be able to handle its large dimension and quickly determine the solution. We base our solution on game theory, [17] modelling each virtual entity requiring a migration as a player. Each player is characterized by a set of possible strategies (such as migrating to a given physical machine or remaining at the current machine), each having properly determined costs and payoffs as well as location constraints and preferences.

By modelling a live migration problem in terms of a set of interacting rational and strategic players, we can deal with the problem's dynamic nature. Therefore, as triggering events occur, the player can be involved in a game interaction. Each player aims to select and execute the strategy exhibiting the lowest cost and offering the highest payoff. In addition, when applicable, players might also consider regulatory requirements in terms of which locations data may occupy, must avoid, or should "stick" with to provide the optimal gain from each game interaction. In reality, it might mean that a cloud CSP will offer its customers online tools that let them choose, in terms of performance and cost, whether to have dedicated, shared, or redundant VMs, and, in terms of regulatory requirements, locations where their data may reside, may not reside, or must reside. The underlying optimization game interaction will then consider all these customer choices before migrating data and applications. Our conceptual approach doesn't involve cooperation among players, leading to a non-cooperative formulation of the migration game. This significantly simplifies the game's implementation and formalization, but can result in a suboptimal solution.

References

1. S. Nepal, R. Ranjan, and K.-K.R. Choo, “Trust- worthy Processing of Healthcare Big Data in Hy- brid Clouds,” IEEE Cloud Computing, vol. 2, no.2, 2015, pp. 78–84.

2. H. Chang, “Privacy Regulatory Model for the Cloud: A Case Study,” IEEE Cloud Computing, vol. 2, no. 3, 2015, pp. 67–72.

3. C. Esposito et al., “Smart Cloud Storage Service Selection Based on Fuzzy Logic, Theory of Evi- dence and Game Theory,” IEEE Trans. Computers, 2015, doi: 10.1109/ TC.2015.2389952.

4. C.-W. Tsai and J.J.P.C. Rodrigues, “Metaheuristic Scheduling for Cloud: A Survey,” IEEE Sys- tems J., vol. 8, no. 1, 2014, pp. 279–291.

5. C. Hooper, B. Martini, and K.-K.R. Choo, “Cloud Computing and Its Implications for Cybercrime Investigations in Australia,” Computer Law and Security Rev., vol. 29, no. 2, 2013, pp. 152–163.

6. H. Chang, “Data Protection Regulation and Cloud Computing,” Privacy and Legal Issues in Cloud Computing, A.S.Y. Cheung and R.H. We- ber, eds., Edward Elgar, 2015, pp. 26-42.

7. H. Chang, “Privacy Regulatory Model for the Cloud: A Case Study,” IEEE Cloud Computing, vol. 2, no. 3, 2015, pp. 67–72.

8. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data, Official J. European Communities, no. 281, 23 Nov. 1995; http://ec.europa.eu/ justice/policies/privacy/docs/95-46-ce/dir1995-46_part1_en.pdf.

9. H. Chang, “Privacy Accountability Management Framework for Data Controllers Operating across Asia,” LLM dissertation, University of Strathclyde Law School, 2014.

10. Chronicle of Data Protection: Privacy and Information Security News and Trends, Hogan Lovells; www.hldataprotection.com/tags/russia.

11. Guidance on Personal Data Protection in Cross-border Data Transfer, guidance note, Office of the Pri- vacy Commissioner for Personal Data, Hong Kong, Dec. 2014; www.pcpd.org.hk //english/resources _centre/publications/files/GN_crossborder_e.pdf.

12. Personal Data Protection (Enforcement) Regulations 2014, part III, Singapore, 2014; https:// w w w.pdpc.gov.sg / leg islat ion-a nd-g u idel i nes / legislation.

13. “Hogan Lovells Partner Highlights Compliance Goals with Russia’s Data Localization Law,” Chronical of Data Protection: Privacy and Information Security News and Trends, 8 Sept. 2015; w w w.h ld at aprotect ion.com / 2015/09/a r t icles / international-eu-pr ivacy/ hogan-lovells-partner
-highlights-compliance-goals-with-russias-data-localization-law.

14. J. Dhont and K.Woodcock, “Data Localization Requirements: Growing Trends and Impact for Company Compliance,” Compliance & Ethics Professional, Feb. 2015, pp. 57–59; www.lorenzlaw.com/w p-content/uploads/Data-localization-requirements-Grow ing-t rends-and-impact-of-compnay-compliance1.pdf.

15. Information Technology—Security Techniques— Code of Practice for Protection of Personally Identifiable Information (PII) in Public Clouds Acting as PII Processors, ISO/IEC 27018:2014, International Organization for Standardization, 2014; www.iso.org /iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=61498.

16. G. Galante and L.C.E. de Bona, “A Survey on Cloud Computing Elasticity,” Proc. IEEE/ACM
5th Int’l Conf. Utility and Cloud Computing, 2012, pp. 263–270.

17. M. Osborne and A. Rubinstein, A Course in Game Theory, MIT Press, 1994.

MASSIMO FICCO is an assistant professor at the Second University of Naples. His research interests include security, cloud computing, and pervasive systems. Ficco has a PhD in computer engineering from the University of Napoli “Parthenope,” Italy. Contact him at massimo.ficco@unina2.it.

CHRISTIAN ESPOSITO is an adjunct professor at the University of Napoli “Federico II” and a research fellow at the University of Salerno. His research interests include information security and reliability, middleware, and distributed systems. Esposito has a PhD in computer engineering from the University of Napoli “Federico II,” Italy. Contact him at christian.esposito@dia.unisa.it.

HENRY CHANG is an adjunct associate professor in the Law and Technology Centre at the University of Hong Kong. His research interests include impacts of technology on privacy, privacy by design, and accountability. Chang has a DBA in IT management from Southern Cross University, Australia, and an LLM from the University of Strathclyde, United Kingdom. Contact him at hcychang@hku.hk.

KIM-KWANG RAYMOND CHOO is an associate professor of cybersecurity and forensics at the University of South Australia, and a visiting expert at INTERPOL Global Complex for Innovation, Singapore. Choo has a PhD in information security from Queensland University of Technology, Australia. He’s a senior member of IEEE. Contact him at raymond.choo@fulbrightmail.org.