A Manifesto for Future Generation Cloud Computing: Research Directions for the Next Decade

RAJKUMAR BUYYA, University of Melbourne, Australia
SATISH NARAYANA SRIRAMA, University of Tartu, Estonia
GIULIANO CASALE, Imperial College London, UK
RODRIGO CALHEIROS, Western Sydney University, Australia
YOGESH SIMMHAN, Indian Institute of Science, India
BLESSON VARGHESE, Queen's University Belfast, UK
EROL GELENBE, Imperial College London, UK
BAHMAN JAVADI, Western Sydney University, Australia
LUIS MIGUEL VAQUERO, University of Bristol, UK
MARCO A. S. NETTO, IBM Research, Brazil
ADEL NADJARAN TOOSI, Monash University, Australia
MARIA ALEJANDRA RODRIGUEZ, University of Melbourne, Australia
IGNACIO M. LLORENTE, Universidad Complutense de Madrid, Spain
SABRINA DE CAPITANI DI VIMERCATI and PIERANGELA SAMARATI, Università degli Studi di Milano, Italy
DEJAN MILOJICIC, Hewlett Packard Labs, USA
CARLOS VARELA, Rensselaer Polytechnic Institute, USA
RAMI BAHSOON, University of Birmingham, UK
MARCOS DIAS DE ASSUNCAO, INRIA, France
OMER RANA, Cardiff University, UK
WANLEI ZHOU, University of Technology Sydney, Australia
HAI JIN, Huazhong University of Science and Technology, China
WOLFGANG GENTZSCH, UberCloud, USA
ALBERT Y. ZOMAYA, University of Sydney, Australia
HAIYING SHEN, University of Virginia, USA

S. N. Srirama co-led this work with first author; Co-First author; Also with University of Melbourne.

Authors' addresses: R. Buyya (corresponding author) and M. A. Rodriguez, University of Melbourne, Parkville, Victoria, 3010, Australia; email: {rbuyya, marodriguez}@unimelb.edu.au; S. N. Srirama (corresponding author), University of Tartu, Ulikooli 17 – 324, Tartu 50090, Estonia; email: srirama@ut.ee; G. Casale, Imperial College London, Huxley 432, UK; email: g.casale@imperial.ac.uk; R. Calheiros and B. Javadi, Western Sydney University, Locked Bag 1797, Penrith, NSW 2751, Australia; email: {r.calheiros, b.javadi}@westernsydney.edu.au; Y. Simmhan, Indian Institute of Science, CDS 206, Nilgiri Marg, Mathikere, Bangalore 560012, India; email: simmhan@iisc.ac.in; B. Varghese, Queen's University Belfast, Northern Ireland, BT9 5BN, UK; email: b.varghese@qub.ac.uk; E. Gelenbe, Imperial College London, South Kensington Campus, London SW7 2AZ, UK; email: e.gelenbe@imperial.ac.uk; L. M. Vaquero, University of Bristol, Merchant Venturers Building, Woodland Road, Clifton, Bristol UK; email: luis.vaquero@bristol.ac.uk; M. A. S. Netto, IBM Research, Sao Paulo, Brazil; email: mstelmar@br.ibm.com; A. N. Toosi, Monash University, Clayton, VIC 3800, Australia; email: Adel.N.Toosi@monash.edu; I. M. Llorente, Universidad Complutense de Madrid, 28040 Madrid, Spain; email: imllorente@ucm.es; S. De Capitani di Vimercati and P. Samarati, Universitá degli Studi di Milano, Via Celoria 18, 20133 Milano, Italy; email: {sabrina.decapitani, pierangela.samarati}@unimi.it; D. Milojicic, Hewlett Packard Labs, Palo Alto, California, USA; email: dejan.milojicic@hpe.com; C. Varela, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180, USA; email: cvarela@cs.rpi.edu; R. Bahsoon, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK; email: r.bahsoon@cs.bham.ac.uk; M. Dias de Assuncao, INRIA, 46 allée d'Italie, 69364 Lyon, France; email: marcosdiasassuncao@gmail.com; O. Rana, Cardiff University, 5 The Parade, Newport Road, Cardiff, CF24 3AA, UK; email: ranaof@cardiff.ac.uk; W. Zhou, University of Technology Sydney, 15 Broadway, Ultimo NSW 2007, Australia; email: Wanlei.Zhou@uts.edu.au; H. Jin, Huazhong University of Science and Technology, Wuhan, 430074, China; email: hjin@hust.edu.cn; W. Gentzsch, UberCloud, 2310 Homestead Rd. Suite:C1-301, Los Altos, California USA; email: wgentzsch@gmail.com; A. Y. Zomaya, University of Sydney, Building J12, Sydney, NSW 2006, Australia; email: albert.zomaya@sydney.edu.au; H. Shen, University of Virginia, 85 Engineer's Way, P.O. Box 400740, Charlottesville, VA 22904-4740, USA; email: hs6ms@virginia.edu.

0360-0300/2018/11-ART105 $15.00
https://doi.org/10.1145/3241737

The Cloud computing paradigm has revolutionised the computer science horizon during the past decade and has enabled the emergence of computing as the fifth utility. It has captured significant attention of academia, industries, and government bodies. Now, it has emerged as the backbone of modern economy by offering subscription-based services anytime, anywhere following a pay-as-you-go model. This has instigated (1) shorter establishment times for start-ups, (2) creation of scalable global enterprise applications, (3) better cost-to-value associativity for scientific and high-performance computing applications, and (4) different invocation/execution models for pervasive and ubiquitous applications. The recent technological developments and paradigms such as serverless computing, software-defined networking, Internet of Things, and processing at network edge are creating new opportunities for Cloud computing. However, they are also posing several new challenges and creating the need for new approaches and research strategies, as well as the re-evaluation of the models that were developed to address issues such as scalability, elasticity, reliability, security, sustainability, and application models. The proposed manifesto addresses them by identifying the major open challenges in Cloud computing, emerging trends, and impact areas. It then offers research directions for the next decade, thus helping in the realisation of Future Generation Cloud Computing.

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Computer systems organization** → **Cloud computing**; • **Information systems** → *Cloud based storage*; Data centers; • **Security and privacy** → *Security services*; • **Networks** → *Cloud computing*; • **Software and its engineering** → *Cloud computing*;

Additional Key Words and Phrases: Cloud computing, scalability, sustainability, InterCloud, data management, Cloud economics, application development, Fog computing, serverless computing

ACM Reference format:
Rajkumar Buyya, Satish Narayana Srirama, Giuliano Casale, Rodrigo Calheiros, Yogesh Simmhan, Blesson Varghese, Erol Gelenbe, Bahman Javadi, Luis Miguel Vaquero, Marco A. S. Netto, Adel Nadjaran Toosi, Maria Alejandra Rodriguez, Ignacio M. Llorente, Sabrina De Capitani di Vimercati, Pierangela Samarati, Dejan Milojicic, Carlos Varela, Rami Bahsoon, Marcos Dias de Assuncao, Omer Rana, Wanlei Zhou, Hai Jin, Wolfgang Gentzsch, Albert Y. Zomaya, and Haiying Shen. 2018. A Manifesto for Future Generation Cloud Computing: Research Directions for the Next Decade. *ACM Comput. Surv.* 51, 5, Article 105 (November 2018), 38 pages.
https://doi.org/10.1145/3241737

1 INTRODUCTION

Cloud computing has shaped the way in which software and IT infrastructure are used by consumers and triggered the emergence of computing as the fifth utility [24]. Since its emergence, industry organisations, governmental institutions, and academia have embraced it and its adoption has seen a rapid growth. This paradigm has developed into the backbone of modern economy by providing on-demand access to subscription-based IT resources, resembling not only the way in which basic utility services are accessed but also the reliance of modern society on them. Cloud computing has enabled new businesses to be established in a shorter amount of time, has facilitated the expansion of enterprises across the globe, has accelerated the pace of scientific progress, and has led to the creation of various models of computation for pervasive and ubiquitous applications, among other benefits.

Until now, there have been three main service models that have fostered the adoption of Clouds, namely Software, Platform, and Infrastructure as a Service (SaaS, PaaS, and IaaS). SaaS offers the highest level of abstraction and allows users to access applications hosted in Cloud data centres (CDC), usually over the Internet. This, for instance, has allowed businesses to access software in a flexible manner by enabling unlimited and on-demand access to a range of ready-to-use

applications. SaaS has also allowed organisations to avoid incurring in internal or direct expenses, such as license fees and IT infrastructure maintenance. PaaS is tailored for users that require more control over their IT resources and offers a framework for the creation and deployment of Cloud applications that includes features such as programming models and auto-scaling. This, for example, has allowed developers to easily create applications that benefit from the elastic Cloud resource model. Finally, IaaS offers access to computing resources, usually by leasing Virtual Machines (VMs) and storage space. This layer is not only the foundation for SaaS and PaaS, but has also been the pillar of Cloud computing. It has done so by enabling users to access the IT infrastructure they require only when they need it, to adjust the amount of resources used in a flexible way, and to pay only for what has been used, all while having a high degree of control over the resources.

1.1 Motivation and Goals of the Manifesto

Throughout the evolution of Cloud computing and its increasing adoption, not only have the aforementioned models advanced and new ones emerged, but also the technologies in which this paradigm is based (e.g., virtualization) have continued to progress. For instance, the use of novel virtualization techniques such as containers that enable improved utilisation of the physical resources and further hide the complexities of hardware is becoming increasingly widespread, even leading to a new service model being offered by providers known as Container as a Service (CaaS). There has also been a rise in the type and number of specialised Cloud services that aid industries in creating value by being easily configured to meet specific business requirements. Examples of these are emerging, easy-to-use, Cloud-based data analytics services and serverless architectures.

Another clear trend is that Clouds are becoming increasingly geographically distributed to support emerging application paradigms. For example, Cloud providers have recently started extending their infrastructure and services to include edge devices for supporting emerging paradigms such as the Internet of Things (IoT) and Fog computing. Fog computing aims at moving decision making operations as close to the data sources as possible by leveraging resources on the edge such as mobile base stations, gateways, network switches and routers, thus reducing response time and network latencies. Additionally, as a way of fulfilling increasingly complex requirements that demand the composition of multiple services and as a way of achieving reliability and improving sustainability, services spanning across multiple geographically distributed CDCs have also become more widespread.

The adoption of Cloud computing will continue to increase and support for these emerging models and services is of paramount importance. In 2016, the IDG's Cloud adoption report found that 70% of organisations have at least one of their applications deployed in the Cloud and that the numbers are growing [89]. In the same year, the IDC's (International Data Corporation) Worldwide Semiannual Public Cloud Services Spending Guide [88] reported that Cloud services were expected to grow from $70 billion in 2015 to more than $203 billion in 2020, an annual growth rate almost seven times the rate of overall IT spending growth. This extensive usage of Cloud computing in various emerging domains is posing several new challenges and is forcing us to rethink the research strategies and re-evaluate the models that were developed to address issues such as scalability, resource management, reliability, and security for the realisation of next-generation Cloud computing environments [149].

This comprehensive manifesto brings these advancements together and identifies open challenges that need to be addressed for realising the *Future Generation Cloud Computing*. Given that rapid changes in computing/IT technologies in a span of 4–5 years are common, and the focus of the manifesto is for the next decade, we envision that identified research directions get addressed and will have impact on the next two or three generations of utility-oriented Cloud computing

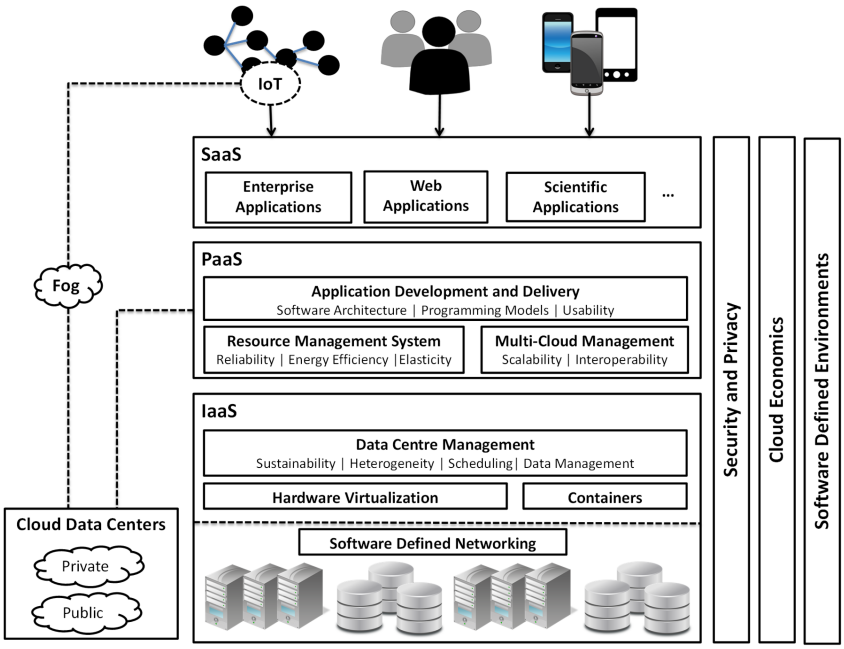

Fig. 1. Components of the Cloud computing paradigm.

technologies, infrastructures, and their applications' services. The manifesto first discusses major challenges in Cloud computing, investigates their state-of-the-art solutions, and identifies their limitations. The manifesto then discusses the emerging trends and impact areas, that further drive these Cloud computing challenges. Having identified these open issues, the manifesto then offers comprehensive future research directions in the Cloud computing horizon for the next decade. Figure 1 illustrates the main components of the Cloud computing paradigm and positions the identified trends and challenges, which are discussed further in the next sections.

The rest of the article is organised as follows: Section 2 discusses the state of the art of the challenges in Cloud computing and identifies open issues. Section 3 along with online Appendix A discusses the emerging trends and impact areas related to the Cloud computing horizon. Section 4 along with online Appendix B provides a detailed discussion about the future research directions to address the open challenges of Cloud computing. In the process, the section also mentions how the respective future research directions will be guided and influenced by the emerging trends. Section 5 provides a conclusion for the manifesto.

2 CHALLENGES: STATE OF THE ART AND OPEN ISSUES

As Cloud computing became popular, it has been extensively utilised in hosting a wide variety of applications. It posed several challenges (shown within the inner ring in Figure 2) such as issues with sustainability, scalability, security, and data management among the others. Over the past decade, these challenges were systematically addressed and the state of the art in Cloud computing has advanced significantly. However, there remains several issues open, as summarised in the outer ring of Figure 2. The rest of the section identifies and details the challenges in Cloud computing and their state of the art, along with the limitations driving their future research.

Fig. 2. Cloud computing challenges, state of the art, and open issues.

2.1 Scalability and Elasticity

Cloud computing differs from earlier models of distributed computing such as grids and clusters, in that it promises virtually unlimited computational resources on demand. At least two clear benefits can be obtained from this promise: First, unexpected peaks in computational demand do not entail breaking service level agreements (SLAs) due to the inability of a fixed computing infrastructure to deliver users' expected quality of service (QoS), and, second, Cloud computing users do not need to make significant up-front investments in computing infrastructure but can rather grow organically as their computing needs increase and only pay for resources as needed. The first (QoS) benefit of the Cloud computing paradigm can only be realised if the infrastructure supports *scalable* services, whereby additional computational resources can be allocated, and new resources have a direct, positive impact on the performance and QoS of the hosted applications. The second (economic) benefit can only be realised if the infrastructure supports *elastic* services,

whereby allocated computational resources can *follow demand* and by dynamically growing and shrinking prevent over- and under-allocation of resources.

The research challenges associated with *scalable services* can be broken into hardware, middleware, and application levels. Cloud computing providers must embrace parallel computing *hardware* including multi-core, clusters, accelerators such as Graphics Processing Units (GPUs) [160], and non-traditional (e.g., neuromorphic and future quantum) architectures, and they need to present such heterogeneous hardware to IaaS Cloud computing users in abstractions (e.g., VMs, containers) that while providing isolation, also enable performance guarantees. At the *middleware* level, programming models and abstractions are necessary, so that PaaS Cloud computing application developers can focus on functional concerns (e.g., defining *map* and *reduce* functions) while leaving non-functional concerns (e.g., scalability, fault tolerance) to the middleware layer [92]. At the *application* level, new generic algorithms need to be developed so that inherent scalability limitations of sequential deterministic algorithms can be overcome; these include asynchronous evolutionary algorithms, approximation algorithms, and online/incremental algorithms (see e.g., Reference [43]). These algorithms may trade off precision or consistency for scalability and performance.

Ultimately, the scalability of the Cloud is limited by the extent to which individual components, namely compute, storage and interconnects scale. Computation has been limited by the end of scaling of both Moore's law (doubling the number of transistors every 1.5 year) and Dennard scaling ("the power use stays in proportion with area: both voltage and current scale (downward) with length"). As a consequence, the new computational units do not scale any more, nor does the power use scale. This directly influences the scaling of computation performance and cost of the Cloud. Research in new technologies, beyond Complementary Metal-Oxide-Semiconductor, is necessary for further scaling. Similar is true for memory. Dynamic Random-Access Memory is limiting the cost and scaling of existing computers, and new non-volatile technologies are being explored that will introduce additional scaling of load-store operating memory while reducing the power consumption. Finally, the photonic interconnects are the third pillar that enables the so called silicon photonics to propagate photonic connections into the chips improving performance, increasing scale, and reducing power consumption.

However, the research challenges associated with *elastic services* include the ability to accurately predict computational demand and performance of applications under different resource allocations [91, 141], the use of these workload and performance models in informing resource management decisions in middleware [93], and the ability of applications to scale up and down, including dynamic creation, mobility, and garbage collection of VMs, containers, and other resource abstractions [147]. While virtualization (e.g., VMs) has achieved steady maturity in terms of performance guarantees rivalling native performance for CPU-intensive applications, ease of use of containers (especially quick restarts) has led to the adoption of containers by the developers community [51]. Programming models that enable dynamic reconfiguration of applications significantly help in elasticity [146], by allowing middleware to move computations and data across Clouds, between public and private Clouds, and closer to edge resources as needed by future Cloud applications running over sensor networks such as the IoT.

In summary, scalability and elasticity provide operational capabilities to improve performance of Cloud computing applications in a cost-effective way, which are yet to be fully exploited. However, resource management and scheduling mechanisms need to be able to strategically use these capabilities.

2.2 Resource Management and Scheduling

The scale of modern CDCs has been rapidly growing and as of today they contain computing and storage devices in the range of tens to hundreds of thousands, hosting complex Cloud

applications and relevant data. This makes the adoption of effective resource management and scheduling policies important to achieve high scalability and operational efficiency.

Nowadays, IaaS providers mostly rely on either *static* VM provisioning policies, which allocate a fixed set of physical resources to VMs using bin-packing algorithms, or *dynamic* policies, capable of handling load variations through live VM migrations and other load balancing techniques [116]. These policies can either be reactive or proactive, and typically rely on knowledge of VM resource requirements, either user-supplied or estimated using monitoring data and forecasting.

Resource management methods are also important for PaaS and SaaS providers to help managing the type and amount of resources allocated to distributed applications, containers, web-services and micro-services. Policies available at this level include for example: (1) auto-scaling techniques, which dynamically scale up and down resources based on current and forecasted workloads; (2) resource throttling methods, to handle workload bursts, trends, smooth auto-scaling transients, or control usage of preemptible VMs (e.g., micro VMs); (3) admission control methods, to handle peak load and prioritize workloads of high-value customers; (4) service orchestration and workflow schedulers, to compose and orchestrate workloads, possibly specialised for the target domain (e.g., scientific data workflows [115]), which make decisions based on their cost-awareness and the constraint requirements of tasks; and (5) multi-Cloud load balancers, to spread the load of an application across multiple CDCs.

The area of resource management and scheduling has spawned a large body of research, and some recent surveys include References [7, 112, 117, 136]. However, several challenges and limitations still remain. For example, existing management policies tend to be intolerant to inaccurate estimates of resource requirements, calling for studying novel tradeoffs between policy optimality and its robustness to inaccurate workload information [94]. Further, demand estimation and workload prediction methods can be brittle and it remains an open question whether Machine Learning (ML) and Artificial Intelligence (AI) methods can fully address this shortcoming [26]. Another frequent issue is that resource management policies tend to focus on optimising specific metrics and resources, often lacking a systematic approach to co-existence in the same environment of multiple control loops, to ensure fair resource access across users, and to holistically optimise across layers of the Cloud stack. Novel resource management and scheduling methods for hybrid Clouds and federated Clouds also need to be devised [91]. Risks related to the interplay between security and resource management are also insufficiently addressed in current research work.

2.3 Reliability

Reliability is another critical challenge in Cloud computing environments. Data centres hosting Cloud computing consist of highly interconnected and interdependent systems. Because of their scale, complexity, and interdependencies, Cloud computing systems face a variety of reliability-related threats such as hardware failures, resource missing failures, overflow failures, network failures, timeout failures, and flaws in software being triggered by environmental change. Some of these failures can escalate and devastatingly impact system operation, thus causing critical failures [76]. Moreover, a cascade of failures may be triggered leading to large-scale service disruptions with far-reaching consequences [96]. As organisations are increasingly interested in adapting Cloud computing technology for applications with stringent reliability assurance and resilience requirements [134], there is an urgent demand for new ways to provision Cloud services with assured performance and resilience to deal with all types of independent and correlated failures [41]. Moreover, the mutual impact of reliability and energy efficiency of Cloud systems is one of the current research challenges [152].

Although reliability in distributed computing has been studied before [126], standard fault-tolerance and reliability approaches cannot be directly applied in Cloud computing systems. The

scale and expected reliability of Cloud computing are increasingly important but hard to analyse due to the range of inter-related characteristics, e.g., their massive-scale, service sharing models, wide-area network, and heterogeneous software/hardware components. Previously, independent failures have mostly been addressed separately; however, the investigation into their interplay has been completely ignored [73]. Furthermore, since Cloud computing is typically more service oriented rather than resource oriented, reliability models for traditional distributed systems cannot be directly applied to Cloud computing. So, existing state-of-the-art Cloud environments lack thorough service reliability models, automatic reliability-aware service management mechanisms, and failure-aware provisioning policies.

2.4 Sustainability

Sustainability is the greatest challenge of our century, and ICT in general [61] utilises today close to 10% of all electricity consumed worldwide, resulting in a CO_2 impact that is comparable to that of air travel. In addition to the energy consumed to operate ICT systems, we know that substantial electricity is used to manufacture electronic components and then decommission them after the end of their useful lifetime; the amount of energy consumed in this process can be four- to fivefold greater than the electricity that this equipment will consume to operate during its lifetime.

CDC deployments until recently have mainly focused on high performance and have not paid enough attention to energy consumption. Thus, today a typical CDC's energy consumption is similar to that of 25,000 households [103], while the total number of operational CDCs worldwide is 8.5 million in 2017 according to IDC. Indeed, according to Greenpeace, Cloud computing worldwide consumes more energy than most countries and only the four largest economies (USA, China, Russia, and Japan) surpass Clouds in their annual electricity usage. As the energy consumption, and the relative cost of energy in the total expenditures for the Cloud, rapidly increases, not enough research has gone into minimising the amount of energy consumed by Clouds, information systems that exploit Cloud systems, and networks [21, 125].

However, networks and the Cloud also have a huge potential to save energy in many areas such as smart cities or to be used to optimise the mix of renewable and non-renewable energy worldwide [135]. However, the energy consumption of Clouds cannot be viewed independently of the QoS that they provide, so that both energy and QoS must be managed in conjunction. Indeed, for a given computer and network technology, reduced energy consumption is often coupled with a reduction of the QoS that users will experience. In some cases, such as critical or even life-threatening real-time needs, such as Cloud support of search and rescue operations, hospital operations or emergency management, a Cloud cannot choose to save energy in exchange for reduced QoS.

Current Cloud systems and efforts have in the past primarily focused on consolidation of VMs for minimising energy consumption of servers [13]. But other elements of CDC infrastructures, such as cooling systems (close to 35% of energy) and networks, which must be very fast and efficient, also consume significant energy that needs to be optimised by proper scheduling of the traffic flows between servers (and over high-speed networks) inside the data centre [68].

Because of multi-core architectures, novel hardware based sleep-start controls and clock speed management techniques, the power consumption of servers increasingly depends, and in a non-linear manner, on their instantaneous workload. Thus new ML-based methods have been developed to dynamically allocate tasks to multiple servers in a CDC or in the Fog [155] so that a combination of violation of SLA, which are costly to the Cloud operator and inconvenient for the end user, and other operating costs including energy consumption, are minimised. Holistic techniques must also address the QoS effect of networks such as packet delays on overall SLA, and the energy effects of networks for remote access to CDC [154]. The purpose of these methods is to

provide online automatic, or autonomic and self-aware methods to holistically manage both QoS and energy consumption of Cloud systems.

Recent work [159] has also shown that deep learning with neural networks can be effectively applied in experimental but realistic settings so that tasks are allocated to servers in a manner that optimises a prescribed performance profile that can include execution delays, response times, system throughput, and energy consumption of the CDC. Another approach that maximises the sustainability of Cloud systems and networks involves rationing the energy supply [60] so that the CDC can modulate its own energy consumption and delivered QoS in response, dynamically modifying the processors' variable clock rates as a function of the supply of energy. It has also been suggested that different sources of renewable and non-renewable energy can be mixed [62].

2.5 Heterogeneity

Public Cloud infrastructure has constantly evolved in the past decade. This is because service providers have increased their offerings while continually incorporating state-of-the-art hardware to meet customer demands and maximise performance and efficiency. This has resulted in an inherently heterogeneous Cloud with heterogeneity at three levels.

The first is at the VM level, which is due to the organisation of homogeneous (or near homogeneous; for example, same processor family) resources in multiple ways and configurations. For example, homogeneous hardware processors with N cores can be organised as VMs with any subset or multiples of N cores. The second is at the vendor level, which is due to employing resources from multiple Cloud providers with different hypervisors or software suites. This is usually seen in multi-Cloud environments [109]. The third is at the hardware architecture level, which is due to employing both CPUs and hardware accelerators, such as GPUs and Field Programmable Gate Arrays (FPGAs) [137].

The key challenges that arise due to heterogeneity in the Cloud are twofold. The first challenge is related to resource and workload management in heterogeneous environments. The state of the art in resource management focuses on static and dynamic VM placement and provisioning using global or local scheduling techniques that consider network parameters and energy consumption [35]. Workload management is underpinned by benchmarking techniques that are used for workload placement and scheduling techniques. Current benchmarking practices are reasonably mature for the first level of heterogeneity and are developing for the second level [98, 148]. However, significant research is still required to predict workload performance given the heterogeneity at the hardware architecture level. Despite advances, research in both heterogeneous resource management and workload management on heterogeneous resources remain fragmented, since they are specific to their level of heterogeneity and do not work across the VM, vendor, and hardware architecture levels. It is still challenging to obtain a general purpose Cloud platform that integrates and manages heterogeneity at all three levels.

The second challenge is related to the development of application software that is compatible with heterogeneous resources. Currently, most accelerators require different (and sometimes vendor specific) programming languages. Software development practices for exploiting accelerators for example additionally require low-level programming skills and has a significant learning curve. For example, CUDA or OpenCL are required for programming GPUs. This gap between hardware accelerators and high-level programming makes it difficult to easily adopt accelerators in Cloud software. It is recognised that abstracting hardware accelerators under middleware will reduce opportunities for optimising the source code for maximising performance. When the Cloud service offering is only the "infrastructure," the onus is on individual developers to provide source code that is targeted to the hardware environment. However, when services, such as "software" and "platforms" are offered on the Cloud, the onus is not on the developer, since the aim of these

services is to abstract the low-level technicalities away from the user. Therefore, it becomes necessary that the hardware is abstracted via a middleware for applications to exploit. Certainly, this comes at the expense of performance and fewer opportunities to optimise the code. Hence, there is a tradeoff between performance and ease of use, when moving from VMs at the infrastructure level and on to using software and services available higher up in the computing stack. One open challenge in this area is developing software that is agnostic of the underlying hardware and can adapt based on the available hardware [100].

2.6 Interconnected Clouds

Although interconnection of Clouds was one of the earliest research problems that was identified in Cloud computing [14, 23, 131], Cloud interoperation continues to be an open issue, since the field has rapidly evolved over the last half decade. Cloud providers and platforms still operate in silos, and their efforts for integration usually target their own portfolio of services. Cloud interoperation should be viewed as the capability of public Clouds, private Clouds, and other diverse systems to understand each other's system interfaces, configurations, forms of authentication and authorisation, data formats, and application initialisation and customisation [139].

Within the broader concept of interconnected Clouds, there are a number of methods that can be used to aggregate the functionalities and services of disparate Cloud providers and/or data centres. These techniques vary on who are the players that engage in the interconnections, its objectives, and the level of transparency in the aggregation of services offered to users [144].

Existing public Cloud providers offer proprietary mechanisms for interoperation that exhibit important limitations as they are not based on standards and open-source, and they do not interoperate with other providers. Although there are multiple efforts for standardisation, such as Open Grid Forum's Open Cloud Computing Interface, Storage Networking Industry Association's Cloud Data Management Interface, Distributed Management Task Force's (DMTF) Cloud Infrastructure Management Interface, DMTF's Open Virtualization Format, IEEE's InterCloud and National Institute of Standards and Technology's (NIST) Federated Cloud, the interfaces of existing Cloud services are not standardised and different providers use different APIs, formats and contextualization mechanisms for comparable Cloud services.

Broadly, the approaches can be classified as federated Cloud computing if the interconnection is initiated and managed by providers (and usually transparent to users) as InterCloud or hybrid Clouds if initiated and managed by users or third parties on behalf of the users.

Federated Cloud computing is considered as the next step in the evolution of Cloud computing and an integral part of the new emerging Edge and Fog computing architectures. The federated Cloud model is gaining increasing interest in the IT market, since it can bring important benefits for companies and institutions, such as resource asset optimisation, cost savings, agile resource delivery, scalability, high availability and business continuity, and geographic dispersion [23].

In the area of InterClouds and hybrid Clouds, Moreno et al. notice that a number of approaches were proposed to provide "*the necessary mechanisms for sharing computing, storage, and networking resources*" [119]. This happens for two reasons. First, companies would like to use as much as possible of their existing in house infrastructures, for both economic and compliance reasons, and thus they should seamlessly integrate with public Cloud resources used by the company. Second, for all the workloads that are allowed to go to Clouds or for resource needs exceeding on premise capabilities, companies are seeking to offload as much of their applications as possible to the public Clouds, driven not only by the economic benefits and shared resources, but also due to the potential freedom to choose among multiple vendors on their terms.

State-of-the-art projects such as Aneka [20] have developed middleware and library solutions for integration of different resources (VMs, databases, etc.). However, the problem with such

approaches is that they need to operate in the lowest common denominator among the services offered by each provider, and this leads to suboptimal Cloud applications or support at specific service models.

Regardless of the particular Cloud interconnection pattern in place, interoperability and portability have multiple aspects and relate to a number of different components in the architecture of Cloud computing and data centres, each of which needs to be considered in its own right. These include standard interfaces, portable data formats and applications, and internationally recognised standards for service quality and security. The efficient and transparent provision, management and configuration of cross-site virtual networks to interconnect the on-premise Cloud and the external provider resources is still an important challenge that is slowing down the full adoption of this technology [87].

As Cloud adoption grows and more applications are moved to the Cloud, the need for satisfactory solutions is likely to grow. Challenges in this area concern how to go beyond the minimum common denominator of services when interoperating across providers (and thus enabling richer Cloud applications); how to coordinate authorisation, access, and billing across providers; and how to apply InterCloud solutions in the context of Fog computing and other emerging trends.

2.7 Empowering Resource-Constrained Devices

Cloud services are relevant not only for enterprise applications, but also for the resource constrained devices and their applications. With the recent innovation and development, mobile devices such as smartphones and tablets, have achieved better CPU and memory capabilities. They also have been integrated with a wide range of hardware and sensors such as camera, GPS (Global Positioning System), accelerometer, and so on. In addition, with the advances in 4G, 5G, and ubiquitous WiFi, the devices have achieved significantly higher data transmission rates. This progress has led to the usage of these devices in a variety of applications such as mobile commerce, mobile social networking and location based services. While the advances in the mobiles are significant and they are also being used as service providers, they still have limited battery life and when compared to desktops have limited CPU, memory and storage capacities, for hosting/executing resource-intensive tasks/applications. These limitations can be addressed by harnessing external Cloud resources, which led to the emergence of Mobile Cloud paradigm.

Mobile Cloud has been studied extensively during the past years [45] and the research mainly focused at two of its binding models, the *task delegation* and the *mobile code offloading* [53]. With the task delegation approach, the mobile invokes web services from multiple Cloud providers, and thus faces issues such as Cloud interoperability and requirement of platform specific API. Task delegation is accomplished with the help of middlewares [53]. Mobile code offloading, on the other hand, profiles and partitions the applications, and the resource-intensive methods/operations are identified and offloaded to surrogate Cloud instances (Cloudlets/swarmlets). Typical research challenges here include developing the ideal offloading approach, identifying the resource-intensive methods, and studying ideal decision mechanisms considering both the device context (e.g., battery level and network connectivity) and Cloud context (e.g., current load on the Cloud surrogates) [52, 161]. While applications based on task delegation are common, mobile code offloading is still facing adaptability challenges [52].

Correspondingly, IoT has evolved as *"web 4.0 and beyond"* and *"Industry 4.0,"* where physical objects with sensing and actuator capabilities, along with the participating individuals, are connected and communicate over the Internet [140]. There are predictions that billions of such devices/*things* will be connected using advances in building innovative physical objects and communication protocols [48]. Cloud primarily helps IoT by providing resources for the storage and distributed processing of the acquired sensor data, in different scenarios. While this *Cloud-centric*

IoT model [75, 140] is interesting, it ends up with inherent challenges such as network latencies for scenarios with sub-second response requirements. An additional aspect that arises with IoT devices is their substantial energy consumption, which can be mitigated by the use of renewable energy [62], but this in turn raises the issue of QoS as the renewable energy sources are generally sporadic. To address these issues and to realise the IoT scenarios, Fog computing is emerging as a new trend to bring computing and system supervisory activities closer to the IoT devices themselves, which is discussed in detail in online Appendix A.2. Fog computing mainly brings several advantages to IoT devices, such as security for edge devices, cognition of situations, agility of deployment, ultra-low latency, and efficiency on cost and performance, which are all critical challenges in the IoT environments.

2.8 Security and Privacy

Security is a major concern in ICT systems and Cloud computing is no exception. Here, we provide an overview of the existing solutions addressing problems related to the secure and private management of data and computations in the Cloud (confidentiality, integrity, and availability) along with some observations on their limitations and challenges that still need to be addressed.

With respect to the confidentiality, existing solutions typically encrypt the data before storing them at external Cloud providers [80]. Encryption, however, limits the support of query evaluation at the provider side. Solutions addressing this problem include the definition of *indexes*, which enable (partial) query evaluation at the provider side without the need to decrypt data, and the use of *encryption techniques* that support the execution of operations or the evaluation of conditions directly over encrypted data. Indexes are metadata that preserve some of the properties of the attributes on which they have been defined and can then be used for query evaluation (e.g., References [2, 37, 80]). The definition of indexes must balance precision and privacy: Precise indexes offer efficient query execution but may lead to improper exposure of confidential information. Encryption techniques supporting the execution of operations on encrypted data without decryption are, for example, Order Preserving Encryption that allows the evaluation of range conditions (e.g., References [2, 153]), and fully (or partial) homomorphic encryption that allows the evaluation of arbitrarily complex functions on encrypted data (e.g., References [18, 70, 71]). Taking these encryption techniques as basic building blocks, some encrypted database systems have been developed (e.g., References [6, 128]), which support SQL queries over encrypted data.

Another interesting problem related to the confidentiality and privacy of data arises when considering modern Cloud-based applications (e.g., applications for accurate social services, better healthcare, detecting fraud, and national security) that explore data over multiple data sources with cross-domain knowledge. A major challenge of such applications is to preserve privacy, as data mining tools with cross-domain knowledge can reveal more personal information than anticipated, therefore prohibiting organisations to share their data. A research challenge is the design of theoretical models and practical mechanisms to preserve privacy for cross-domain knowledge [163]. Furthermore, the data collected and stored in the Cloud (e.g., data about the techniques, incentives, internal communication structures, and behaviours of attackers) can be used to verify and evaluate new theory and technical methods (e.g., References [83, 143]). A current booming trend is to use ML methods in information security and privacy to analyse Big Data for threat analysis, attack intelligence, virus propagation, and data correlations [82].

Many approaches protecting the confidentiality of data rely on the implicit assumption that any authorised user, who knows the decryption key, can access the whole data content. However, in many situations there is the need of supporting *selective visibility* for *different users*. Works addressing this problem are based on *selective encryption* and on *attribute-based encryption* (ABE) [151]. Policy updates are supported, for example, by *over-encryption*, which, however, requires the help

of the Cloud provider, and by the *Mix&Slice* approach [10], which departs from the support of the Cloud provider and uses different rounds of encryption to provide complete mixing of the resource. The problem of selective sharing has been considered also in scenarios where different parties cooperate for sharing data and to perform distributed computations.

Alternative solutions to encryption have been adopted when associations among the data are more sensitive than the data themselves [33]. Such solutions split data in different fragments stored at different servers or guaranteed to be non linkable. They support only certain types of sensitive constraints and queries and the computational complexity for retrieving data increases.

While all solutions described above successfully provide efficient and selective access to outsourced data, they are exposed to attacks exploiting frequency of accesses to violate data and users privacy. This problem has been addressed by *Private Information Retrieval* (PIR) techniques, which operate on publicly available data, and, more recently by *privacy-preserving indexing techniques* based on, for example, Oblivious RAM, B-tree data structures, and binary search tree [44]. This field is still in its infancy and the development of practical solutions is an open problem.

With respect to the integrity, different techniques such as digital signatures, Provable Data Possession, Proof Of Retrievability, let detecting unauthorised modifications of data stored at an external Cloud provider. Verifying the integrity of stored data by its owner and authorised users is, however, only one of the aspects of integrity. When data can change dynamically, possibly by multiple writers, and queries need to be supported, several additional problems have to be addressed. Researchers have investigated the use of authenticated data structures (*deterministic* approaches) or insertion of integrity checks (*probabilistic* approaches) [39] to verify the correctness, completeness, and freshness of a computation. Both deterministic and probabilistic approaches can represent promising directions but are limited in their applicability and integrity guarantees provided.

With respect to the availability, some proposals have focused on the problem of how a user can select the services offered by a Cloud provider that match user's security and privacy requirements [38]. Typically, the expected behaviours of Cloud providers are defined by SLAs stipulated between a user and the Cloud provider itself. Recent proposals have addressed the problem of exploring possible dependencies among different characteristics of the services offered by Cloud providers [40]. These proposals represent only a first step in the definition of a comprehensive framework that allows users to select the Cloud provider that best fits their needs and verifies that providers offer services fully compliant with the signed contract.

Hardware-based techniques have also been adopted to guarantee the proper protection of sensitive data in the Cloud. Some of the most notable solutions include the *ARM TrustZone* and the *Intel Software Guard Extensions* (SGX) technology. ARM TrustZone introduces several hardware-assisted security extensions to ARM processor cores and on-chip peripherals. The platform is then split into a "secure world" and a "normal world," each of which has different privileges and a controlled communication interface. The Intel SGX technology supports the creation of trusted execution environments, called *enclaves*, where sensitive data can be stored and processed.

Advanced *cyberattacks* in the Cloud domain represent a serious threat that may affect the confidentiality, integrity, and availability of data and computations. In particular, Advanced Persistent Threats (APTs) deserves a particular mention. This is an emerging class of cyberattacks that are goal oriented, highly targeted, well organised, well funded, technically advanced, stealthy, and persistent. The notorious Stuxnet, Flame, and Red October are some examples of APTs. APTs poses a severe threat to the Cloud computing domain, as APTs have special characteristics that can disable the existing defence mechanisms of Cloud computing such as antivirus, firewall, intrusion detection, and antivirus [158]. Indeed, APT-based cyber breach instances and cybercrime activities have recently been on the rise, and it has been predicted that a 50% increase in security budgets will be

observed to rapidly detect and respond to them [19]. In this context, enhancing the technical levels of cyber defence only is far from being enough [57]. To mitigate the loss caused by APTs, a mixture of technical-driven security solutions and policy-driven security solutions must be designed. For example, data encryption can be viewed as the final layer of protection for ATP attacks. A policy to force all sensitive data to be encrypted and to stay in a "trusted environment" can prevent data leakage—even if the attack can successfully penetrate into the system, all they can see is encrypted data. Another example is to utilise one-time password for strong authentication, providing better protection to clouds.

2.9 Economics of Cloud Computing

Research themes in Cloud economics have centred on a number of key aspects over recent years: (1) pricing of Cloud services—i.e., how a Cloud provider should determine and differentiate between different capabilities they offer, at different price bands and durations (e.g., micro, mini, large VM instances); (2) brokerage mechanisms that enable a user to dynamically search for Cloud services that match a given profile within a predefined budget; and (3) monitoring to determine if user requirements are being met, and identifying penalty (often financial) that must be paid by a Cloud provider if values associated with pre-agreed metrics have been violated. The last of these has seen considerable work in the specification and implementation of SLAs, including implementation of specifications such as WS-Agreement [3].

SLA is traditionally a business concept, as it specifies contractual financial agreements between parties who engage in business activities. Faniyi and Bahsoon [50] observed that up to three SLA parameters (performance, memory, and CPU cycle) are often used. SLA management also relates to the supply and demand of computational resources, instances and services [17, 22]. A related area of *policy-based approaches* is also studied extensively [25]. Policy-based approaches are effective when resource adaptation scenarios are limited in number. As the number of encoded policies grow, these approaches can be difficult to scale. Various optimisation strategies have been used to enable SLA and policy-based resource enforcement.

Another related aspect in Cloud economics has been an understanding of how an organisation migrates current in-house or externally hosted infrastructure to Cloud providers, involving the migration of an in-house IT department to a Cloud provider. Migration of existing services needs to take account of both social and economic aspects of how Cloud services are provisioned and subsequently used, and risk associated with uptime and availability of often business critical capability. Migrating systems management capabilities outside an organisation also has an influence on what skills need to be retained within an organisation. According to a survey by RightScale [156], IT departments may now be acting as potential brokers for services that are hosted, externally within a data centre. Systems management personnel may now be acting as intermediaries between internal user requests and technical staff at the CDC, whilst some companies may fully rely instead on technical staff at the data centre, completely removing the need for local personnel. This would indicate that small companies, in particular, may not need to retain IT skills for systems management and administration, instead relying on pre-agreed SLAs with CDCs. This has already changed the landscape of the potential skills base in IT companies. Many Universities also make use of Microsoft Office 365 for managing email, an activity that was closely guarded and managed by their Information Services/IT departments in the past.

The above context has also been motivated with interest in new implementation technologies such as sub-second billing made possible through container-based deployments, often also referred to as "serverless computing," such as in Google "functions," AWS Lambda, amongst others. Serverless computing is discussed further in online Appendix A.4.

Licensing is another economics-related issue, which can include annual or perpetual licensing. These can be restrictive for Cloud resources (e.g., not on-demand, limited number of cores, etc.) when dealing with the demands of large business and engineering simulations for physics, manufacturing, and so on. Independent Software Vendors (ISVs) such as ANSYS, Dassault, Siemens, and COMSOL are currently investigating or already have more suitable licensing models for the Cloud, such as BYOL (bring your own license), or credits/tokens/elastic units, or fully on-demand.

Another challenge in Cloud economics is related to choosing the right Cloud provider. Comparing offerings between different Cloud providers is time consuming and often challenging, as providers do not use the same terminology when offering computational and storage resources, making a like-for-like comparison difficult. A number of commercial and research grade platforms have been proposed to investigate benefit/limits of Cloud selection, such as RightScale PlanForCloud , CloudMarketMaker [97], pricing tools from particular providers (e.g., Amazon Cost Calculator , and SMI (Service Measurement Index) for ranking Cloud services [59]. Such platforms focus on what the user requires and hide the internal details of the Cloud provider's resource specifications and pricing models. In addition, marketplace models are also studied where users purchase services from SaaS providers that in turn procure computing resources from either PaaS or IaaS providers [4].

2.10 Application Development and Delivery

Cloud computing empowers application developers with the ability to programmatically control infrastructure resources and platforms. Several benefits have emerged from this feature, such as the ability to couple the application with auto-scaling controllers and to embed in the code advanced self-* mechanisms for organising, healing, optimising, and securing the Cloud application at runtime.

A key benefit of *resource* programmability is a looser boundary between development and operations, which results in the ability to accelerate the delivery of changes to the production environment. To support this feature, a variety of agile delivery tools and model-based orchestration languages (e.g., Terraform and OASIS TOSCA) are increasingly adopted in Cloud application delivery pipelines and DevOps methodologies [12]. These tools help automating lifecycle management, including continuous delivery and continuous integration, application and platform configuration, and testing.

In terms of *platform* programmability, separation of concerns has helped in tackling the complexity of software development for the Cloud and runtime management. For example, MapReduce enables application developers to specify functional components of their application, namely *map* and *reduce* functions on their data; while enabling the middleware layers to deal with non-functional concerns, such as parallelisation, data locality optimisation, and fault tolerance. Several other programming models have emerged and are currently being investigated, to cope with the increasing heterogeneity of Cloud platforms. For example, in Edge computing, the effort to split applications relies on the developers [31]. Recent efforts in this area are also not yet fully automated [101]. Problems of this kind can be seen in many situations. Even though it is expected that there will be a wide variety and large number of edge devices and applications, there is a shortage of application delivery frameworks and programming models to deliver software spanning both the Edge and the CDC, to enable the use of heterogeneous hardware within Cloud applications, and to facilitate InterClouds operation.

Besides supporting and amplifying the above trends, an important research challenge is application evolution. Accelerated and continuous delivery may foster a short-term view of the application evolution, with a shift towards reacting to quality problems arising in production rather than avoiding them through careful design. This is in contrast with traditional approaches, where the

application is carefully designed and tested to be as bug-free as possible prior to release. However, the traditional model requires more time between releases and thus it is less agile than continuous delivery methods. There is still a shortage of research in Cloud software engineering methods to combine the strengths of these two delivery approaches. For example, continuous acquisition of performance and reliability data across Cloud application releases may be used to better inform application evolution, to automate the process of identifying design anti-patterns, and to explore what-if scenario during testing of new features. Holistic methods to implement this vision need to be systematically investigated over the coming years.

2.11 Data Management

One of the key selling points of Cloud computing is the availability of affordable, reliable, and elastic storage that is collocated with the computational infrastructure. This offers a diverse suite of storage services to meet most common enterprise needs while leaving the management and hardware costs to the IaaS service provider. They also offer reliability and availability through multiple copies that are maintained transparently, along with disaster recovery with storage that can be replicated in different regions. A number of storage abstractions are also offered to suit a particular application's needs, with the ability to acquire just the necessary quantity and pay for it. *Object-based storage* (Amazon Simple Storage Service (S3), Azure File), *block storage services* (Azure Blob, Amazon Elastic Block Store) of a disk volume, and *logical* Hard Disk Drive *and* Solid-state Drive disks that can be attached to VMs are common ones. Besides these, higher-level data platforms such as NoSQL columnar databases, relational SQL databases and publish-subscribe message queues are available as well.

At the same time, there has been a proliferation of Big Data platforms [107] running on distributed VM's collocated with the data storage in the data centre. The initial focus has been on batch processing and NoSQL query platforms that can handle large data volumes from web and enterprise workloads, such as Apache Hadoop, Spark and HBase. However, fast data platforms for distributed stream processing such as Apache Storm, Heron, and Apex have grown to support data from sensors and Internet-connected devices. PaaS offerings such as Amazon ElasticMR, Kinesis, Azure HDInsight and Google Dataflow are available as well.

While there has been an explosion in the data availability over the past decade, and along with the ability to store and process them on Clouds, many challenges still remain. Services for data storage have not been adequately supported by services for managing their metadata that allows data to be located and used effectively [120]. Data security and privacy remain a concern (discussed further in Section 2.8), with regulatory compliance being increasingly imposed by various governments (such as the recent EU *General Data Protection Regulation (GDPR) and US CLOUD Act*), as well as leakages due to poor data protection by users. Data are increasingly being sourced from the edge of the network as IoT device deployment grows, and the latency of wide area networks inhibits their low-latency processing. Edge and Fog computing may hold promise in this respect [150].

Even within the data centre, network latencies and bandwidth between VMs, and from VM to storage can be variable, causing bottlenecks for latency-sensitive stream processing and bandwidth-sensitive batch processing platforms. Solutions such as Software Defined Networking (SDN) and Network Functions Virtualization (NFV), which can provide mechanisms required for allocating network capacity for certain data flows both within and across data centres with certain computing operations been performed in-network, are needed [110]. Better collocation guarantees of VMs and data storage may be required as well.

There is also increasing realisation that a lambda architecture that can process both data at rest and data at motion together is essential [105]. Big Data platforms such as Apache Flink and

Spark Streaming are starting to offer early solutions but further investigation is required [162]. Big Data platforms also have limited support for automated scaling out and in on elastic Clouds, and this feature is important for long-running streaming applications with dynamic workloads [106]. While the resource management approaches discussed above can help, these are yet to be actively integrated within Big Data platforms. Fine-grained per-minute and per-second billing along with faster VM acquisition time, possibly using containers, can help shape the resource acquisition better. In addition, composing applications using serverless computing such as AWS Lambda and Azure Functions has been growing rapidly [8]. These stateless functions can off-load the resource allocation and scaling to the Cloud platform provider while relying on external state by distributed object management services like Memcached or storage services like S3.

2.12 Networking

Cloud data centres are the backbone of Cloud services where application components reside and where service logic takes place for both internal and external users. Successful delivery of Cloud services requires many levels of communication happening within and across data centres. Ensuring that this communication occurs securely, seamlessly, efficiently and in a scalable manner is a vital role of the network that ties all the service components together.

During the past decade, there has been many network-based innovations and research that have explicitly explored Cloud networking. For example, technologies such as SDN and NFV intended to build agile, flexible, and programmable computer networks to reduce both capital and operational expenditure for Cloud providers. In online Appendix A.5 SDN and NFV are further discussed. Likewise, scaling limitations as well as the need for a flat address space and over subscription of servers also have prompted many recent advances in the network architecture such as VL2 [74], PortLand [123], and BCube [77] for the CDCs. Despite all these advances, there are still many networking challenges that need to be addressed.

One of the main concerns of today's CDCs is their high energy consumption. Nevertheless, the general practice in many data centres is to leave all networking devices always on [84]. In addition, unlike computing servers, the majority of network elements such as switches, hubs, and routers are not designed to be energy proportional and things such as, sleeping during no traffic and adaptation of link rate during low traffic periods, are not a native part of the hardware [113]. Therefore, the design and implementation of methodologies and technologies to reduce network energy consumption and make it proportional to the load remain as open challenges.

Another challenge with CDC networks is related to providing guaranteed QoS. The SLAs of today's Clouds are mostly centred on computation and storage [78]. No abstraction or mechanism enforcing the performance isolation and hence no SLAs beyond best effort is available to capture the network performance requirements such as delay and bandwidth guarantees. Within the data centre infrastructure, Guo et al. [78] propose a network abstraction layer called VDC that works based on a source routing technique to provide bandwidth guarantees for VMs. Yet, their method does not provide any network delays guarantee. This challenge becomes even more pressing, when network connectivity must be provided over geographically distributed resources, for example, deployment of a "virtual cluster" spanning resources on a hybrid Cloud environment. Even though the network connectivity problem involving resources in multiple sites can be addressed using network virtualization technologies, providing performance guarantees for such networks as it traverses over the public Internet raises many significant challenges that require special consideration [144]. The primary challenge in this regard is that cloud providers do not have privileged access to the core Internet equipment as they do in their own data centres. Therefore, cloud providers' flexibility regarding routing and traffic engineering is limited to a large extent. Moreover, the performance of public network such as the Internet is much more

unpredictable and changeable compared to the dedicate network of data centres that makes it more difficult to provide guaranteed performance requirements. Traditional WAN approaches such as Multi-Protocol Label Switching (MPLS) for traffic engineering in such networks are also inefficient in terms of bandwidth usage and handling latency-sensitive traffic due to lack of global view of the network [85]. This is one of the main reasons that companies such as Google invested on its own dedicated network infrastructures to connect its data centres across the globe [95].

In addition, Cloud networking is not a trivial task and modern CDCs face similar challenges to building the Internet due to their size [9]. The highly virtualized environment of a CDC is also posing issues that have always existed within network apart from new challenges of these multi-tenant platforms. For example in terms of scalability, VLANs (Virtual Local Area Network) are a simple example. At present, VLANs are theoretically limited to 4,096 segments. Thus, the scale is limited to approximately 4,000 tenants in a multitenant environment. VXLAN offers encapsulation methods to address the limited number of VLANs. However, it is limited in multicasting, and supports Layer 2 only within the logical network. IPv4 is another example, where some Cloud providers such as Microsoft Azure admitted that they ran out of addresses. To overcome this issue the transition to the impending IPv6 adoption must be accelerated. This requirement means that the need for network technologies offering high performance, robustness, reliability, flexibility, scalability, and security never ends [9].

2.13 Usability

The Human Computer Interface and Distributed Systems communities are still far from one another. Cloud computing, in particular, would benefit from a closer alignment of these two communities. Although much effort has happened on resource management and back-end-related issues, usability is a key aspect to reduce costs of organisations exploring Cloud services and infrastructure. This reduction is possible, mainly due to labour-related expenses as users can have better quality of service and enhance their productivity. The usability of Cloud [49] has already been identified as a key concern by NIST as described in their Cloud Usability Framework [142], which highlights five aspects: capable, personal, reliable, secure, and valuable. Capable is related to meeting Cloud consumers expectations with regard to Cloud service capabilities. Personal aims at allowing users and organizations to change the look and feel of user interfaces and to customise service functionalities. Reliable, secure, and valuable are aspects related to having a system that performs its functions under state conditions, safely/protected, and that returns value to users respectively. Coupa's white paper [34] on usability of Cloud applications also explores similar aspects, highlighting the importance of usability when offering services in the Internet.

For usability, current efforts in Cloud have mostly focused on encapsulating complex services into APIs to be easily consumed by users. One area where this is clearly visible is High Performance Computing (HPC) Cloud [122]. Researchers have been creating services to expose HPC applications to simplify their consumptions [32, 86]. These applications are not only encapsulated as services, but also receive Web portals to specify application parameters and manage input and output files.

Another direction related to usability of Cloud that got traction in the last years is DevOps [11, 130]. Its goal is to integrate development (Dev) and operations (Ops) thus aiding faster software delivery (as also discussed in Sections 2.10 and 4.10). DevOps has improved the productivity of developers and operators when creating and deploying solutions in Cloud environments. It is relevant not only to build new solutions in the Cloud but also to simplify the migration of legacy software from on-premise environments to multi-tenancy elastic Cloud services.

3 EMERGING TRENDS AND IMPACT AREAS

As Cloud computing and relevant research matured over the years, it led to several advancements in the underlying technologies such as containers and software defined networks. These developments in turn have led to several emerging trends in Cloud computing such as Fog computing, serverless computing, and software defined computing. In addition to them, other emerging trends in ICT such as Big Data, machine/deep learning, and blockchain technology also have started influencing the Cloud computing research and have offered wide opportunities to deal with the open issues in Cloud-related challenges. The emerging trends and impact areas relevant in the Cloud horizon are discussed in detail by the manifesto. However, due to the limitation in number of pages, the discussion is being produced as online Appendix to the article. The Appendix ponders the following topics:

A.1 - Containers: New type of virtualization technology with tiny memory footprint, lesser resource requirement and faster startup. *A.2 - Fog Computing*: Computing at the edge of the network and it envisions to make decisions as close as possible to the data source. *A.3 - Big Data*: Discusses rapid escalation in the generation of streaming data from IoT and social networking applications. *A.4 - Serverless Computing*: An emerging architectural pattern where the server is abstracted away and the resources are automatically managed for the user. *A.5 - Software-defined Cloud Computing*: Optimising and automating the Cloud configuration and adaptation by extending the virtualization to compute, storage, and networks. *A.6 - Blockchain*: Distributed immutable ledger deployed in a decentralised network that relies on cryptography to meet security constraints. *A.7 - Machine and Deep Learning*: Algorithms and models for optimised resource management and ML services offered from Clouds.

4 FUTURE RESEARCH DIRECTIONS

The Cloud computing paradigm, like the Web, the Internet, and the computer itself, has transformed the information technology landscape in its first decade of existence. However, the next decade will bring about significant new requirements, from large-scale heterogeneous IoT and sensor networks producing very large datastreams to store, manage, and analyse to energy- and cost-aware personalised computing services that must adapt to a plethora of hardware devices while optimising for multiple criteria including application-level QoS constraints and economic restrictions.

Significant research was already performed to address the Cloud computing technological and adoption challenges, and the state of the art along with their limitations is discussed thoroughly in Section 2. The future research in Cloud computing should focus at addressing these limitations along with the problems hurled and opportunities presented by the latest developments in the Cloud horizon. Thus the future R&D will greatly be influenced/driven by the emerging trends discussed in Section 3. Here the manifesto provides the key future directions for the Cloud computing research, for the coming decade.

4.1 Scalability and Elasticity

Scalability and elasticity research challenges for the next decade can be decomposed into hardware, middleware, and application-level.

At the Cloud computing hardware level, an interesting research direction is special-purpose Clouds for specific functions, such as deep learning—e.g., Convolutional Neural Networks, Multi-Layer Perceptrons, and Long Short-Term Memory—datastream analytics and image and video pattern recognition. While these functionalities may appear to be very narrow, they can be deployed for a spectrum of applications and their usage is increasingly growing. There are numerous

examples at control points at airports, social network mining, IoT sensor data analytics, smart transportation, and many other applications. Key Cloud providers are already offering accelerators and special-purpose hardware with increasing usage growth, e.g., Amazon is offering GPUs, Google has been deploying Tensor Processing Units (TPUs) [99] and Microsoft is deploying FPGAs in the Azure Cloud [129]. As new hardware addresses scalability, Clouds need to embrace non-traditional architectures, such as neuromorphic, quantum computing, adiabatic, nanocomputing, and many others (see Reference [90]). Research needed includes developing appropriate virtualization abstractions, as well as programming abstractions enabling just-in-time compilation and optimisation for special-purpose hardware. Appropriate economic models also need to be investigated for FaaS Cloud providers (e.g., offering image and video processing as composable micro-services).

At the Cloud computing middleware level, research is required to further increase reuse of existing infrastructure, to improve speed of deployment and provisioning of hardware and networks for very large-scale deployments. This includes algorithms and software stacks for reliable execution of applications with failovers to geographically remote private or hybrid Cloud sites. Research is also needed on InterClouds that will seamlessly enable computations to run on multiple public Cloud providers simultaneously. To support HPC applications, it will be critical to guarantee consistent performance across multiple runs even in the presence of additional Cloud users. New deployment and scheduling algorithms need to be developed to carefully match HPC applications with those that would not introduce noise in parallel execution or, if not possible, to use dedicated clusters for HPC [79, 122].

To be able to address large-scale communication-intensive applications, further Cloud provider investments are required to support high-throughput and low-latency networks [122]. The environment of these applications necessitates sophisticated mechanisms for handling multiple clients and for providing sustainable and profitable business provision. Moreover, Big Data applications are leveraging HPC capabilities and IoT, providing support for many modern applications such as smart cities [132] or industrial IoT [16]. These applications have demanding requirements in terms of (near-)real time processing of large scale of data, its intelligent analysis and then closing the loops of control.

4.2 Resource Management and Scheduling

The evolution of the Cloud in the upcoming years will lead to a new generation of research solutions for resource management and scheduling. Technology trends such as Fog will increase the level of decentralisation of the computation, leading to increased heterogeneity in the resources and platforms and also to more variability in the processed workloads. Technology trends, such as serverless computing and Edge computing, will also offer novel opportunities to reason on the tradeoffs of offloading part of the application logic far from the system core, posing new questions on optimal management and scheduling. Conversely, trends such as software-defined computing and Big Data will come to maturity, expanding the enactment mechanisms and reasoning techniques available for resource management and scheduling, thus offering many outlets for novel research.

Challenges arising from decentralisation are inherently illustrated in the Fog computing domain, edge analytics (discussed further in Section 4.7) being one interesting research direction. In edge analytics, the stream-based or event-driven sensor data will be processed across the complete hierarchy of Fog topology. This will require cooperative resource management between centralised CDCs and distributed Edge computing resources for real-time processing. Such management methods should be aware of the locations and resources available to edge devices for optimal resource allocation, and should take into account device mobility, highly dynamic network topology, and

privacy and security protection constraints at scale. The design of multiple co-existing control loops spanning from CDCs to the Edge is, by itself, a broad research challenge from the point of design, analysis and verification, implementation and testing. The adoption of container technology in these applications will be useful due to its small footprint and fast deployment [124].

Novel research challenges in the area of scheduling will also arise in these decentralised and heterogeneous environments. Recently proposed concepts such as multi-resource fairness [72] as well as non-conventional game theoretic methods [133], which today are primarily applied to small to medium-scale computing clusters or to define optimal economic models for the Cloud, need to be generalised and applied to large-scale heterogeneous settings comprising both CDCs and Edge. For example, mean-field games may help in addressing inherent scalability problems by helping to reason about the interaction of a large number of resources, devices and user types [133].

Serverless computing is an example of emerging research challenges in management and scheduling, such as offloading the computation far from the application core components that implement the business logic. From the end user standpoint, FaaS raises the expectation that functions will be executed within a specific time, which is challenging given that current performance is quite erratic [54] and network latency can visibly affect function response time. Moreover, given that function cost is per access, this will require novel resource management policies to decide when and to which extent rely on FaaS instead of microservices that run locally to the application.

From the FaaS provider perspective, allocation of resources needs to be optimal (neither excessive nor insufficient), and, from a user perspective, a desirable level of QoS needs to be achieved when functions are executed, determining suitable tradeoffs with execution requirements, network latency, privacy and security requirements. Given that a single application backed by FaaS can lead to hundreds of hits to the Cloud in a second, an important challenge for serverless platform providers will be to optimise allocation of resources for each class of service so that revenue is optimised, while all the user FaaS QoS expectations are met. This research will require to take into consideration soft constraints on execution time of functions and proactive FaaS provisioning to avoid high latency of resource start-up to affect the performance of backed applications. Moreover, providers and consumers, both for FaaS and regular Cloud services, often have different goals and constraints, calling for novel game-theoretic approaches and market-oriented models for resource allocation and regulation of the supply and demand within the Cloud platform.

The emerging SDN paradigm exemplifies a novel trend that will extend the range of control mechanisms available for holistic management of resources. By logically centralising the network control plane, SDNs provide opportunities for more efficient management of resources located in a single administrative domain such as a CDC. SDN also facilitates joint VM and traffic consolidation, a difficult task to do in traditional data centre networks, to optimise energy consumption and SLA satisfaction, thus opening new research outlets [36]. Service Function Chaining (SFC) is an automated process to set up the chain of virtual network functions (VNFs), e.g., network address translation (NAT), firewalls, intrusion detection systems (IDS) in an NFV environment using instantiation of software-only services. Leveraging SDN together with NFV technologies allows for efficient and on-demand placement of service chains [30]. However, optimal service chain placement requires novel heuristics and resource management policies. The virtualized nature of VNFs also makes their orchestration and consolidation easier and dynamic deployment of network services possible [108, 127], calling for novel algorithms that can exploit these capabilities.

In addition, it is foreseeable that the ongoing interest for ML, deep learning, and AI applications will help in dealing with the complexity, heterogeneity, and scale, in addition to spawn novel research in established data centre resource management problems such as VM provisioning, consolidation, and load balancing. It is, however, important to recognise that potential loss of control

and determinism may arise by adopting these techniques. Research in explainable AI may provide a suitable direction for novel research to facilitate the adoption of AI methods in Cloud management solutions within the industry [46].

For example, in scientific workflows the focus so far has been on efficiently managing the execution of platform-agnostic scientific applications. As the amount of data processed increases and extreme-scale workflows begin to emerge, it is important to consider key concerns such as fault tolerance, performance modelling, efficient data management, and efficient resource usage. For this purpose, Big Data analytics will become a crucial tool [42]. For instance, monitoring and analysing resource consumption data may enable workflow management systems to detect performance anomalies and potentially predict failures, leveraging technologies such as serverless computing to manage the execution of complex workflows that are reusable and can be shared across multiple stakeholders. Although today there exist the technical possibility to define solutions of this kind, there is still a shortage of applications of serverless functions to HPC and scientific computing use cases, calling for further research in this space.

4.3 Reliability

One of the most challenging areas in Cloud computing systems is reliability as it has a great impact on the QoS as well as on the long term reputation of the service providers. Currently, all the Cloud services are provided based on the cost and performance of the services. The key challenge faced by Cloud service providers is how to deliver a competitive service that meets end users' expectations for performance, reliability, and QoS in the face of various types of independent as well as temporal and spatial correlated failures. So the future of research in this area will be focused on innovative Cloud services that provide reliability and resilience with assured service performance; which is called Reliability as a Service (RaaS). The main challenge is to develop a hierarchical and service-oriented cloud service reliability model based on advanced mathematical and statistical models [126]. This requires new modules to be included in the existing Cloud systems such as failure model and workload model to be adapted for resource provisioning policies and provide flexible reliability services to a wide range of applications.

One of the future directions in RaaS will be using deep and machine learning for failure prediction. This will be based on failure characterisation and development of a model from massive amount of failure datasets. Having a comprehensive failure prediction model will lead to a failure-aware resource provisioning that can guarantee the level of reliability and performance for the user's applications. This concept can be extended as another research direction for the Fog computing where there are several components on the edge. While fault-tolerant techniques such as replication could be a solution in this case, more efficient and intelligent approaches will be required to improve the reliability of new type of applications such as IoT applications. This needs to be incorporated with the power efficiency of such systems and solving this tradeoff will be a complex research challenge to tackle [118].

Another research direction in reliability will be about Cloud storage systems that are now mature enough to handle Big Data applications. However, failures are inevitable in Cloud storage systems as they are composed of large-scale hardware components. Improving fault tolerance in Cloud storage systems for Big Data applications is a significant challenge. Replication and Erasure coding are the most important data reliability techniques employed in Cloud storage systems [121]. Both techniques have their own tradeoffs in various parameters such as durability, availability, storage overhead, network bandwidth and traffic, energy consumption and recovery performance. Future research should include the challenges involved in employing both techniques in Cloud storage systems for Big Data applications with respect to the aforementioned parameters [121]. This hybrid technique applies proactive dynamic data replication of erasure coded data based on node

failure prediction, which significantly reduces network traffic and improves the performance of Big Data applications with less storage overhead. So, the main research challenge would be solving a multivariable optimisation problem to take into account several metrics to meet users and providers requirements.

4.4 Sustainability

Sustainability of ICT systems is emerging as a major consideration [61] due to the energy consumption of ICT systems. Of course, sustainability also covers issues regarding the pollution and decontamination of the manufacturing and decommissioning of computer and network equipment, but this aspect is not covered in the present article.

In response to the concern for sustainability, viewed primarily through the lens of energy consumption and energy awareness, increasingly large CDCs are being established, with up to 1000 MW of potential power consumption, in or close to areas where there are plentiful sources of renewable energy [15], such as hydro-electricity in northern Norway, and where natural cooling can be available as in areas close to the Arctic Circle. This actually requires new and innovative system architectures that can distribute data centres and Cloud computing, geographically. To address this, algorithms have been proposed, which rely on geographically distributed data coordination, resource provisioning and energy-aware and carbon footprint-aware provisioning in data centres [47, 81, 104]. In addition, geographical load balancing can provide an effective approach for optimising both performance and energy usage. With careful pricing, electricity providers can motivate Cloud service providers to "follow the renewables" and serve requests through CDCs located in areas where green energy is available [111]. However, the smart grid focuses on controlling the flow of energy in the electric grid with the help of computer systems and networks, and there seems to be little if any work on the energy consumption by the ICT components in the smart grid, perhaps because the amount would be small as compared to the overall energy consumption of a country or region. Interestingly enough, there has been recent work on dynamically coupling the flow of energy to computing and communication resources, and the flow of energy to the components of such computer/communication systems [62] to satisfy QoS and SLAs for jobs while minimising the energy consumption, but much more work will be needed.

However, placing data centres far away from most of the end users places a further burden on the energy consumption and QoS of the networks that connect the end users to the CDCs. Indeed, it is important to note that moving CDCs away from users will increase the energy consumed in networks, so that some remote solutions that are based on renewable energy may substantially increase the energy consumption of networks that are powered through conventional electrical supplies. Another challenge relates to the very short end-to-end delay that certain operations, such as financial transactions, require; thus data centres for financial services often need to be located in proximity to the actual human users and financial organisations (such as banks) that are designing, maintaining and modifying the financial decision making algorithms, as well as to the commodity trading data bases whose state must accurately reflect current prices, since users need to buy and sell stock or other commodities at up-to-date prices that may automatically change within less than a second. Another factor is the proprietary nature of the data that is being used, and the legal and security requirements that can often only be verified and complied within national boundaries or within the EU. Thus if the data remains local, the CDCs that process it also have to be local. Thus in many cases, the Cloud cannot rely on renewable energy to operate effectively simply because renewal energy is not available locally and because some renewable energy sources (e.g., wind and photovoltaic) tend to be intermittent. At the other end, the power needs of CDCs and the Cloud are also growing due to the ever-increasing amount of data that need to be stored and processed. Thus running the Cloud and CDCs in an energy efficient manner remains a major priority.

Unfortunately, high performance and more data processing has always gone hand-in-hand with greater energy consumption. Thus QoS, SLAs, and energy consumption have to be considered simultaneously and need to be managed online [63]. Since all the fast-changing online behaviours cannot be predicted in advance or modelled in a complete manner, adaptive self-aware techniques are needed to face this challenge [154]. Some progress has been recently made in this direction [155] but further work will be needed. The actual algorithms that may be used will include machine learning techniques such as those described in Yin et al. [159], which exploits constant online measurement of system parameters that can lead to online decision making that will optimise sustainability while respecting QoS considerations and SLAs.

The Fog can also substantially increase energy consumption because of the greater difficulty of efficient energy management for smaller and highly diverse systems [62, 64]. At the same time, the reduced access distance and network size from the end users to the Fog servers can create energy savings in networks. Therefore, the interesting tradeoff between the increased energy consumption from many disparate and distribute Fog servers, and the reduced network energy consumption when the Fog servers are installed in close proximity to the end user, requires much further work [65]. Such research should include the improvements in network QoS that may be experienced by end users, when they access locally distributed Fog servers and their traffic traverses a smaller number of network nodes. There have been attempts to conduct experimental research in this direction with the help of machine learning based techniques [154].

Some approaches for improving sustainability and reducing energy consumption in the Cloud, primarily focus on the VM consolidation for minimising the energy consumption of the servers, which has been shown to be quite effective [13], while the Cloud cannot be accessed without the help of networks. However, reducing energy consumption in networks is also a complex problem [56, 67]. Saving energy for networking elements often disturbs other aspects such as reliability, scalability, and performance of the network [69]. Proposals have been made and tested regarding the design of smart energy-aware routing algorithms [66], but this area in general has received less attention compared to energy consumption and power efficiency of computing elements. With the advent of SDN, the global network awareness and centralised decision-making offered by SDN may provide a better opportunity for creating sustainable networks for Clouds [55]. This is perhaps one of the areas that will draw substantially more research efforts and innovation in the next decade.

4.5 Heterogeneity

Heterogeneity on the Cloud was introduced in the last decade, but awaits widespread adoption. As highlighted in Section 2.5, there are currently at least two significant gaps that hinder heterogeneity from being fully exploited on the Cloud. The first gap is between unified management platforms and heterogeneity. Existing research that targets resource and workload management in heterogeneous Cloud environments is fragmented. This translates into the lack of availability of a unified environment for efficiently exploiting VM level, vendor level and hardware architecture level heterogeneity while executing Cloud applications. The manifesto therefore proposes for the next decade an umbrella platform that accounts for heterogeneity at all three levels. This can be achieved by integrating a portfolio of workload and resource management techniques from which optimal strategies are selected based on the requirement of an application. For this, heterogeneous memory management will be required. Current solutions for memory management rely mainly on hypervisors, which limits the benefits from heterogeneity. Alternate solutions recently proposed rely on making guest operating systems heterogeneity-aware [102].

The second gap is between abstraction and heterogeneity. Current programming models for using hardware accelerators require accelerator specific languages and low-level programming

efforts. Moreover, these models are conducive for developing scientific applications. This restricts the wider adoption of heterogeneity for service oriented and user-driven applications on the Cloud. One meaningful direction to pursue will be to initiate a community-wide effort for developing an open-source high-level programming language that can satisfy core Cloud principles, such as abstraction and elasticity, which are suited for modern and innovative Cloud applications in a heterogeneous environment. This will also be a useful tool as the Fog ecosystem emerges and applications migrate to incorporate both Cloud and Fog resources.

Recent research in this area has highlighted the limitation of current programming languages, such as OpenCL [28]. The interaction between CPUs and the hardware accelerator need to be explicitly programmed, which limits the automatic transformation of source code in efficient ways. To this end, fine-grained task partitioning needs to be automated for general purpose applications. Additionally, the automated conversion from coarse-grained to fine-grained task partitioning is required. In the context of OpenCL programming, there is limited performance portability, which is to be addressed. However, currently available high-level programming languages, such as TANGRAM [29] provide performance portability across different accelerators, but need to incorporate performance models and adaptive runtimes for finding optimal strategies for interaction between the CPU and the hardware accelerator.

Although the Cloud as a utility is a more recent offering, a number of the underlying technologies for supporting different levels of heterogeneity (memory, processors etc.) in the Cloud came into inception a few decades ago. For example, the Multiplexed Information and Computing Service (Multics) offered single-level memory, which was the foundation of virtual memory for heterogeneous systems. Similarly, IBM developed CP-67, which was one of the first attempts in virtualizing mainframe operating systems to implement time-sharing. Later on VMWare used this technology for virtualizing x86 servers. The earlier technology was able to even provide I/O virtualization, and meaningful ways of addressing some of the challenges raised by modern heterogeneity may find inspiration in earlier technologies when the Cloud was not known.

Recently there is also a significant discussion about disaggregated data centres. Traditionally data centres are built using servers and racks with each server contributing the resources such as CPU, memory and storage, required for the computational tasks. With the disaggregated data centre each of these resources is built as a stand-alone resource "*blade*," where these blades are interconnected through a high-speed network fabric. The trend has come into existence as there is significant gap in the pace at which each of these resource technologies individually advanced. Even though most prototypes are proprietary and in their early stages of development, a successful deployment at the data centre level would have significant impact on the way the traditional IaaS are provided. However, this needs significant development in the network fabric as well [58].

4.6 Interconnected Clouds

As the grid computing and web service histories have shown, interoperability and portability across Cloud systems is a highly complicated area and it is clear at this time that pure standardisation is not sufficient to address this problem. The use of application containers and configuration management tools for portability, and the use of software adapters and libraries for interoperability are widely used as practical methods for achieving interoperation across Cloud services and products. However, there are a number of challenges [23], and thus potential research directions, that have been around since the early days of Cloud computing and, due to their complexity, have not been satisfactorily addressed so far.

One of such challenges is how to promote Cloud interconnection without forcing the adoption of the minimum common set of functionalities among services: if users want, they should be able to integrate complex functionalities even if they are offered only by one provider. Other research

directions include how to enable Cloud interoperation middleware that can mimic complex services offered by one provider by composing simple services offered by one or more providers - so that the choice about the complex service or the composition of simpler services were solely dependent on the user constraints—cost, response time, data sovereignty, and so on.

The above raises another important future research direction: how to enable middleware operating at the user-level (InterCloud and hybrid Clouds) to identify candidate services for a composition without support from Cloud providers? Given that providers have economic motivation to try to retain all the functionalities offered to their customers (i.e., they do not have motivation to facilitate that only some of the services in a composition are their own), one cannot expect that an approach that requires Cloud providers cooperation might succeed.

Therefore, the middleware enabling composition of services has to solve challenges in its two interfaces: in the interface with Cloud users, it needs to seamlessly deliver the service, in a level where how the functionality is delivered is not relevant for users: it could be obtained in all from a single provider (perhaps invoking a SaaS able to provide the functionality) or it could be obtained by composing different services from different providers. In the provider interface, it enables such more complex functions to be obtained, regardless of particular collaboration from providers: provided that an API exists, the middleware would be in charge of understanding what information/service the API can provide (and how to access such service) and thus decide by itself if it has all the required input necessary to access the API, and even the output is sufficient to enable the composition. This discussion makes clear the complexity of such middleware and the difficulty of the questions that need to be addressed to enable such vision.

Nevertheless, ubiquitously interconnected Clouds (achieved via Cloud Federation) can truly be achieved only when Cloud vendors are convinced that the Cloud interoperability adoption brings them financial and economic benefits. This requires novel approaches for billing and accounting, novel interconnected Cloud suitable pricing methods, along with formation of InterCloud marketplaces [144].

Finally, the emergence of SDNs and the capability to shape and optimise network traffic has the potential to influence research in Cloud interoperation. Google reports that one of the first uses of SDNs in the company was for optimisation of wide-area network traffic connecting their data centres [145]. In the same direction, investigation is needed on the feasibility and benefits of SDN and NFV to address some of the challenges above. For example, SDN and NFV can enable better security and QoS for services built as compositions of services from multiple providers (or from geographically distributed services from the same provider) by enforcing prioritization of service traffic across providers/data centres and specific security requirements [87].

4.7 Empowering Resource-Constrained Devices

Regarding future directions for empowering resource-constrained devices, in the mobile Cloud domain, we already have identified that, while task delegation is a reality, code offloading still has adaptability issues. It is also observed that, *"as the device capabilities are increasing, the applications that can benefit from the code offloading are becoming limited"* [140]. This is evident, as the capabilities of smartphones are increasing, to match or benefit from offloading, the applications are to be offloaded to Cloud instances with much higher capacity. This incurs higher cost per offloading. To address this, the future research in this domain should focus at better models for multi-tenancy in Mobile Cloud applications, to share the costs among multiple mobile users. The problem further gets complex due to the heterogeneity of both the mobile devices and Cloud resources.

We also foresee the need for incentive mechanisms for heterogeneous mobile Cloud offloading to encourage mobile users to participate and get appropriate rewards in return. This should encourage in adapting the mobile Cloud pattern to the social networking domain as well, in

designing ideal scenarios. In addition, the scope and benefits offered by the emerging technologies such as serverless computing, CaaS and Fog computing, to the mobile Cloud domain, are not yet fully explored.

The incentive mechanisms are also relevant for the IoT and Fog domains. Recently there is significant discussion about the establishment of Fog closer to the *things*, by infrastructure offered by independent Fog providers [27]. These architectures follow the consumer-as-provider (CaP) model. A relevant CaP example in the Cloud computing domain is the MQL5 Cloud Network [1], which utilises consumer's devices and desktops for performing various distributed computing tasks. Adaptation of such Peer-to-Peer (P2P) and CaP models would require ideal incentive mechanisms. Further discussion about the economic models for such Micro Data centres is provided in Section 4.9.

The container technology also brings several opportunities to this challenge. With the rise of Fog and Edge computing, it can be predicted that the container technology, as a kind of lightweight running environment and convenient packing tools for applications, will be widely deployed in edge servers. For example, the customised containers, such as Cloud Android Container [157], aimed at Edge computing and offloading features will be more and more popular. They provide efficient server runtime and inspire innovative applications in IoT, AI, and other promising fields.

Edge analytics in domains such as real-time streaming data analytics would be another interesting research direction for the resource constrained devices. The things in IoT primarily deal with sensor data and the Cloud-centric IoT (CIoT) model extracts this data and pushes it to the Cloud for processing. Primarily, Fog/Edge computing came to existence to reduce the network latencies in this model. In edge analytics, the sensor data will be processed across the complete hierarchy of Fog topology, i.e., at the edge devices, intermediate Fog nodes and Cloud. The intermediary processing tasks include filtering, consolidation, error detection and so on. Frameworks that support edge analytics (e.g., Apache Edgent [5]) should be studied considering both the QoS and QoE (Quality of Experience) aspects. Preliminary solutions related to scheduling and placement of the edge analytics tasks and applications across the Fog topology are already appearing in the literature [114, 138]. Further research is required to deal with cost-effective multi-layer Fog deployment for multi-stage data analytics and dataflow applications.

4.8 Security and Privacy

Due to the limitation in number of pages, the discussion is being produced as online Appendix B.1.

4.9 Economics of Cloud Computing

Due to the limitation in number of pages, the discussion is being produced as online Appendix B.2.

4.10 Application Development and Delivery

Due to the limitation in number of pages, the discussion is being produced as online Appendix B.3.

4.11 Data Management

Due to the limitation in number of pages, the discussion is being produced as online Appendix B.4.

4.12 Networking

Due to the limitation in number of pages, the discussion is being produced as online Appendix B.5.

4.13 Usability

There are several opportunities to enhance usability in Cloud environments. For instance, it is still hard for users to know how much they will spend renting resources due to workload/resource

fluctuations or characteristics. Tools to have better estimations would definitely improve user experience and satisfaction. Due to recent demands from Big Data community, new visualization technologies could be further explored on the different layers of Cloud environment to better understand infrastructure and application behaviour and highlight insights to end users. Easier API management methodologies, tools, and standards are also necessary to handle users with different levels of expertise and interests. User experience when handling data-intensive applications also needs further studies considering their expected QoS.

In addition, users are still overloaded with resource and service types available to run their applications. Examples of resources and services are CPUs, GPUs, network, storage, operating system flavour, and all services available in the PaaS. Advisory systems to help these users would greatly enhance their experience consuming Cloud resources and services. Advisory systems to also recommend how users should use Cloud more efficiently would certainly be beneficial. Advices such as whether data should be transferred or visualized remotely, whether resources should be allocated or deleted, whether baremetal machines should replace virtual ones are examples of hints users could receive to make Cloud easier to use and more cost-effective.

The main difficulty in this area lies on evaluation. Traditionally, Cloud computing researchers and practitioners mostly perform quantitative experiments, whereas researchers working closer to users have deep knowledge on qualitative experiments. This second type of experiments depends on selecting groups of users with different profiles and investigating how they use technology. As Cloud has a very heterogeneous community of users with different needs and skills and work in different Cloud layers (IaaS, PaaS, and SaaS), such experiments are not trivial to be designed and executed at scale. Apart from understanding user behaviour, it is relevant to develop mechanisms to facilitate or automatically reconfigure Cloud technologies to adapt to the user needs and preferences, and not assume all users have the same needs or have the same level of skills.

4.14 Discussion

As can be observed from the emerging trends and proposed future research directions (summarised in the outer ring of Figure 3), there will be significant developments across all the service models (IaaS, PaaS, and IaaS) of Cloud computing.

In the IaaS there is scope for heterogeneous hardware such as CPUs and accelerators (e.g., GPUs and TPUs) and special purpose Clouds for specific applications (e.g., HPC and deep learning). The future generation Clouds should also be ready to embrace the non-traditional architectures, such as neuromorphic, quantum computing, adiabatic, nanocomputing, and so on. Moreover, emerging trends such as containerisation, SDN and Fog/Edge computing are going to expand the research scope of IaaS by leaps and bounds. Solutions for addressing sustainability of CDC through utilisation of renewable energy and IoT-enabled cooling systems are also discussed. There is also scope for emerging trends in IaaS, such as disaggregated data centres where resources required for the computational tasks such as CPU, memory and storage, will be built as stand-alone resource blades, which will allow faster and ideal resource provisioning to satisfy different QoS requirements of Cloud based applications. The future research directions proposed for addressing the scalability, resource management and scheduling, heterogeneity, interconnected Clouds and networking challenges, should enable realising such comprehensive IaaS offered by the Clouds.

Similarly, PaaS should see significant advancements through future research directions in resource management and scheduling. The need for programming abstractions, models, languages and systems supporting scalable elastic computing and seamless use of heterogeneous resources are proposed leading to energy-efficiency, minimised application engineering cost, better portability and guaranteed level of reliability and performance. It is also foreseeable that the ongoing interest for ML, deep learning, and AI applications will help in dealing with the complexity,

Fig. 3. Future research directions in the Cloud computing horizon.

heterogeneity, scale and load balancing applications developed through PaaS. Serverless computing is an emerging trend in PaaS, which is a promising area to be explored with significant practical and economic impact. Interesting future directions are proposed such as function-level QoS management and economics for serverless computing. In addition, future research directions for data management and analytics are also discussed in detail along with security, leading to interesting applications with platform support such as edge analytics for real-time stream data processing, from the IoT and smart cities domains.

SaaS should mainly see advances from the application development and delivery, and usability of Cloud services. Translucent programming models, languages, and APIs will be needed to enable tackling the complexity of application development while permitting control of application delivery to future-generation Clouds. A variety of agile delivery tools and Cloud standards (e.g., TOSCA) are increasingly being adopted during Cloud application development. The future research should focus at how to continuously monitor and iteratively evolve the design and quality

of Cloud applications. It is also suggested to extend DevOps methods and define novel programming abstractions to include within existing software development and delivery methodologies, a support for IoT, Edge computing, Big Data, and serverless computing. Focus should also be at developing effective Cloud design patterns and development of formalisms to describe the workloads and workflows that the application processes, and their requirements in terms of performance, reliability, and security are strongly encouraged. It is also interesting to see that even though the technologies have matured, certain domains such as mobile Cloud, still have adaptability issues. Novel incentive mechanisms are required for mobile Cloud adaptability as well as for designing Fog architectures.

Future research should thus explore Cloud architectures and market models that embrace uncertainties and provide continuous "win-win" resolutions, for all the participants including providers, users and intermediaries, both from the Return On Investment (ROI) and satisfying SLA perspectives.

5 SUMMARY AND CONCLUSIONS

The Cloud computing paradigm has revolutionised the computer science horizon during the past decade and enabled emergence of computing as the fifth utility. It has emerged as the backbone of modern economy by offering subscription-based services anytime, anywhere following a pay-as-you-go model. Thus, Cloud computing has enabled new businesses to be established in a shorter amount of time, has facilitated the expansion of enterprises across the globe, has accelerated the pace of scientific progress, and has led to the creation of various models of computation for pervasive and ubiquitous applications, among other benefits.

However, the next decade will bring about significant new requirements, from large-scale heterogeneous IoT and sensor networks producing very large datastreams to store, manage, and analyse, to energy- and cost-aware personalised computing services that must adapt to a plethora of hardware devices while optimising for multiple criteria including application-level QoS constraints and economic restrictions. These requirements will be posing several new challenges in Cloud computing and will be creating the need for new approaches and research strategies, and force us to re-evaluate the models that were already developed to address the issues such as scalability, resource provisioning, and security.

This comprehensive manifesto brought the advancements together and proposed the challenges still to be addressed in realising the future generation Cloud computing. In the process, the manifesto identified the current major challenges in Cloud computing domain and summarised the state of the art along with the limitations. The manifesto also discussed the emerging trends and impact areas that further drive these Cloud computing challenges. Having identified these open issues, the manifesto then offered comprehensive future research directions in the Cloud computing horizon for the next decade. The discussed research directions show a promising and exciting future for the Cloud computing field both technically and economically, and the manifesto calls the community for action in addressing them.

ACKNOWLEDGMENTS

We thank anonymous reviewers, Sartaj Sahni (Editor-in-Chief) and Antonio Corradi (Associate Editor) for their constructive suggestions and guidance on improving the content and quality of this article. We also thank Adam Wierman (California Institute of Technology), Shigeru Imai (Rensselaer Polytechnic Institute), and Arash Shaghaghi (University of New South Wales, Sydney) for their comments and suggestions for improving the article. Regarding funding, G. Casale has been supported by the Horizon 2020 project DICE (644869).

REFERENCES

[1] 2017. MQL5 Cloud Network. Retrieved May 18, 2018 from https://cloud.mql5.com/.

[2] Rakesh Agrawal, Jerry Kiernan, Ramakrishnan Srikant, and Yirong Xu. 2004. Order preserving encryption for numeric data. In *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*. ACM, 563–574.

[3] Alain Andrieux, Karl Czajkowski, Asit Dan, Kate Keahey, Heiko Ludwig, Toshiyuki Nakata, Jim Pruyne, John Rofrano, Steve Tuecke, and Ming Xu. 2007. Web services agreement specification (WS-agreement). In *Open Grid Forum*, Vol. 128. 216.

[4] Jonatha Anselmi, Danilo Ardagna, John Lui, Adam Wierman, Yunjian Xu, and Zichao Yang. 2017. The economics of the cloud. *ACM Trans. Model. Perf. Eval. Comput. Syst.* 2, 4 (2017), 18.

[5] Apache Software Foundation. 2018. Apache Edgent—A Community for Accelerating Analytics at the Edge. Retrieved May 18, 2018 from http://edgent.apache.org/.

[6] Arvind Arasu, Spyros Blanas, Ken Eguro, Raghav Kaushik, Donald Kossmann, Ravishankar Ramamurthy, and Ramarathnam Venkatesan. 2013. Orthogonal security with cipherbase. In *Proceedings of the Conference on Innovative Data Systems Research (CIDR'13)*.

[7] Danilo Ardagna, Giuliano Casale, Michele Ciavotta, Juan F. Pérez, and Weikun Wang. 2014. Quality-of-service in cloud computing: Modeling techniques and their applications. *J. Internet Serv. Appl.* 5, 1 (2014), 11.

[8] Matt Asay. 2018. AWS Won Serverless—Now All Your Software Kinda Belong to Them. Retrieved May 18, 2018 from https://www.theregister.co.uk/2018/05/11/lambda_means_game_over_for_serverless/.

[9] Siamak Azodolmolky, Philipp Wieder, and Ramin Yahyapour. 2013. Cloud computing networking: Challenges and opportunities for innovations. *IEEE Commun. Mag.* 51, 7 (2013), 54–62.

[10] Enrico Bacis, Sabrina De Capitani di Vimercati, Sara Foresti, Stefano Paraboschi, Marco Rosa, and Pierangela Samarati. 2016. Mix&slice: Efficient access revocation in the cloud. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. 217–228.

[11] Armin Balalaie, Abbas Heydarnoori, and Pooyan Jamshidi. 2016. Microservices architecture enables DevOps: Migration to a cloud-native architecture. *IEEE Softw.* 33, 3 (2016), 42–52.

[12] Len Bass, Ingo Weber, and Liming Zhu. 2015. *DevOps: A Software Architect's Perspective*. Addison-Wesley Professional.

[13] Andreas Berl, Erol Gelenbe, Marco Di Girolamo, Giovanni Giuliani, Hermann De Meer, Minh Quan Dang, and Kostas Pentikousis. 2010. Energy-efficient cloud computing. *Comput. J.* 53, 7 (2010), 1045–1051.

[14] David Bernstein, Erik Ludvigson, Krishna Sankar, Steve Diamond, and Monique Morrow. 2009. Blueprint for the intercloud-protocols and formats for cloud computing interoperability. In *Proceedings of the International Conference on Internet and Web Applications and Services (ICIW'09)*. IEEE, 328–336.

[15] Josep L. Berral, Íñigo Goiri, Thu D. Nguyen, Ricard Gavalda, Jordi Torres, and Ricardo Bianchini. 2014. Building green cloud services at low cost. In *Proceedings of the IEEE 34th International Conference on Distributed Computing Systems (ICDCS'14)*. IEEE, 449–460.

[16] Flavio Bonomi, Rodolfo Milito, Jiang Zhu, and Sateesh Addepalli. 2012. Fog computing and its role in the internet of things. In *Proceedings of the 1st Edition of the MCC Workshop on Mobile Cloud Computing*. ACM, 13–16.

[17] Nicolas Bonvin, Thanasis G. Papaioannou, and Karl Aberer. 2011. Autonomic SLA-driven provisioning for cloud applications. In *Proceedings of the IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*. IEEE Computer Society, 434–443.

[18] Z. Brakerski and V. Vaikuntanathan. 2011. Efficient fully homomorphic encryption from (standard) LWE. In *Proceedings of the IEEE Symposium on Foundations of Computer Science (FOCS'11)*.

[19] Ross Brewer. 2014. Advanced persistent threats: Minimising the damage. *Netw. Secur.* 2014, 4 (2014), 5–9.

[20] Rajkumar Buyya and Diana Barreto. 2015. Multi-cloud resource provisioning with aneka: A unified and integrated utilisation of microsoft azure and amazon EC2 instances. In *Proceedings of the 2015 International Conference on Computing and Network Communications (CoCoNet'15)*. IEEE, 216–229.

[21] Rajkumar Buyya, Anton Beloglazov, and Jemal Abawajy. 2010. Energy-efficient management of data center resources for cloud computing: A vision, architectural elements, and open challenges. In *Proceedings of the 2010 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'10)*. CSREA Press.

[22] Rajkumar Buyya, Saurabh Kumar Garg, and Rodrigo N. Calheiros. 2011. SLA-oriented resource provisioning for cloud computing: Challenges, architecture, and solutions. In *Proceedings of the 2011 International Conference on Cloud and Service Computing (CSC'11)*. IEEE, 1–10.

[23] Rajkumar Buyya, Rajiv Ranjan, and Rodrigo N. Calheiros. 2010. Intercloud: Utility-oriented federation of cloud computing environments for scaling of application services. In *Proceedings of the International Conference on Algorithms and Architectures for Parallel Processing*. Springer, 13–31.

[24] Rajkumar Buyya, Chee Shin Yeo, Srikumar Venugopal, James Broberg, and Ivona Brandic. 2009. Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Fut. Gen. Comput. Syst.* 25, 6 (2009), 599–616.
[25] Emiliano Casalicchio and Luca Silvestri. 2013. Mechanisms for SLA provisioning in cloud-based service providers. *Comput. Netw.* 57, 3 (2013), 795–810.
[26] Israel Casas, Javid Taheri, Rajiv Ranjan, and Albert Y. Zomaya. 2017. PSO-DS: A scheduling engine for scientific workflow managers. *J. Supercomput.* 73, 9 (2017), 3924–3947.
[27] Chii Chang, Satish Narayana Srirama, and Rajkumar Buyya. 2017. Indie fog: An efficient fog-computing infrastructure for the internet of things. *IEEE Comput.* 50, 9 (2017), 92–98.
[28] Li-Wen Chang, Juan Gómez-Luna, Izzat El Hajj, Sitao Huang, Deming Chen, and Wen-mei Hwu. 2017. Collaborative computing for heterogeneous integrated systems. In *Proceedings of the 8th ACM/SPEC on International Conference on Performance Engineering (ICPE'17)*. 385–388.
[29] L. W. Chang, I. E. Hajj, C. Rodrigues, J. Gómez-Luna, and W. M. Hwu. 2016. Efficient kernel synthesis for performance portable programming. In *Proceedings of the 2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'16)*. 1–13.
[30] Daewoong Cho, Javid Taheri, Albert Y. Zomaya, and Pascal Bouvry. 2017. Real-time virtual network function (VNF) migration toward low network latency in cloud environments. In *Proceedings of the 2017 IEEE 10th International Conference on Cloud Computing (CLOUD'17)*. IEEE, 798–801.
[31] Byung-Gon Chun, Sunghwan Ihm, Petros Maniatis, Mayur Naik, and Ashwin Patti. 2011. Clonecloud: Elastic execution between mobile device and cloud. In *Proceedings of the 6th Conference on Computer Systems*. ACM, 301–314.
[32] Philip Church, Andrzej Goscinski, and Christophe Lefèvre. 2015. Exposing HPC and sequential applications as services through the development and deployment of a SaaS cloud. *Fut. Gen. Comput. Syst.* 43–44 (2015), 24–37.
[33] Valentina Ciriani, Sabrina De Capitani Di Vimercati, Sara Foresti, Sushil Jajodia, Stefano Paraboschi, and Pierangela Samarati. 2010. Combining fragmentation and encryption to protect privacy in data storage. *ACM Trans. Inf. Syst. Secur.* 13, 3 (2010), 22.
[34] Coupa Software. 2012. *Usability in Enterprise Cloud Applications*. Technical Report. Coupa Software.
[35] Steve Crago, Kyle Dunn, Patrick Eads, Lorin Hochstein, Dong-In Kang, Mikyung Kang, Devendra Modium, Karandeep Singh, Jinwoo Suh, and John Paul Walters. 2011. Heterogeneous cloud computing. In *Proceedings of the 2011 IEEE International Conference on Cluster Computing (CLUSTER'11)*. IEEE, 378–385.
[36] Richard Cziva, Simon Jouët, David Stapleton, Fung Po Tso, and Dimitrios P. Pezaros. 2016. SDN-based virtual machine management for cloud data centers. *IEEE Trans. Netw. Serv. Manage.* 13, 2 (2016), 212–225.
[37] Ernesto Damiani, S. D. C. D. Vimercati, Sushil Jajodia, Stefano Paraboschi, and Pierangela Samarati. 2003. Balancing confidentiality and efficiency in untrusted relational DBMSs. In *Proceedings of the 10th ACM Conference on Computer and Communications Security*. ACM, 93–102.
[38] Amir Vahid Dastjerdi and Rajkumar Buyya. 2014. Compatibility-aware cloud service composition under fuzzy preferences of users. *IEEE Trans. Cloud Comput.* 2, 1 (2014), 1–13.
[39] Sabrina De Capitani di Vimercati, Sara Foresti, Sushil Jajodia, Stefano Paraboschi, and Pierangela Samarati. 2016. Efficient integrity checks for join queries in the cloud. *J. Comput. Secur.* 24, 3 (2016), 347–378.
[40] S. De Capitani di Vimercati, Giovanni Livraga, Vincenzo Piuri, Pierangela Samarati, and Gerson A. Soares. 2016. Supporting application requirements in cloud-based iot information processing. In *Proceedings of the International Conference on Internet of Things and Big Data (IoTBD'16)*. Scitepress, 65–72.
[41] Jeffrey Dean. 2009. Large-scale distributed systems at google: Current systems and future directions. In *Proceedings of the 3rd ACM SIGOPS International Workshop on Large Scale Distributed Systems and Middleware (LADIS'09) Tutorial*.
[42] Ewa Deelman, Christopher Carothers, Anirban Mandal, Brian Tierney, Jeffrey S. Vetter, Ilya Baldin, Claris Castillo, Gideon Juve, et al. 2017. PANORAMA: An approach to performance modeling and diagnosis of extreme-scale workflows. *Int. J. High Perf. Comput. Appl.* 31, 1 (2017), 4–18.
[43] Travis Desell, Malik Magdon-Ismail, Boleslaw Szymanski, Carlos Varela, Heidi Newberg, and Nathan Cole. 2009. Robust asynchronous optimization for volunteer computing grids. In *Proceedings of the 5th IEEE International Conference on e-Science, 2009 (e-Science'09)*. IEEE, 263–270.
[44] Sabrina De Capitani di Vimercati, Sara Foresti, Riccardo Moretti, Stefano Paraboschi, Gerardo Pelosi, and Pierangela Samarati. 2016. A dynamic tree-based data structure for access privacy in the cloud. In *Proceedings of the 2016 IEEE International Conference on Cloud Computing Technology and Science (CloudCom'16)*. IEEE, 391–398.
[45] Hoang T. Dinh, Chonho Lee, Dusit Niyato, and Ping Wang. 2013. A survey of mobile cloud computing: Architecture, applications, and approaches. *Wireless Commun. Mob. Comput.* 13, 18 (2013), 1587–1611.
[46] Derek Doran, Sarah Schulz, and Tarek R. Besold. 2017. What does explainable AI really mean? A new conceptualization of perspectives. https://arxiv.org/abs/1710.00794.
[47] Hancong Duan, Chao Chen, Geyong Min, and Yu Wu. 2017. Energy-aware scheduling of virtual machines in heterogeneous cloud computing systems. *Fut. Gen. Comput. Syst.* 74 (2017), 142–150.

[48] Dave Evans. 2011. The internet of things: How the next evolution of the internet is changing everything. *CISCO White Paper* 1, 2011 (2011), 1–11.
[49] Chaudhry Muhammad Nadeem Faisal. 2011. Issues in cloud computing: Usability evaluation of cloud based application. LAMBERT Academic Publishing.
[50] Funmilade Faniyi and Rami Bahsoon. 2016. A systematic review of service level management in the cloud. *ACM Comput. Surv.* 48, 3 (2016), 43.
[51] W. Felter, A. Ferreira, R. Rajamony, and J. Rubio. 2015. An updated performance comparison of virtual machines and Linux containers. In *Proceedings of the 2015 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS'15)*. DOI: https://doi.org/10.1109/ISPASS.2015.7095802
[52] Huber Flores, Pan Hui, Sasu Tarkoma, Yong Li, Satish Srirama, and Rajkumar Buyya. 2015. Mobile code offloading: From concept to practice and beyond. *IEEE Commun. Mag.* 53, 3 (2015), 80–88.
[53] Huber Flores and Satish Narayana Srirama. 2014. Mobile cloud middleware. *J. Syst. Softw.* 92 (2014), 82–94.
[54] Geoffrey C. Fox, Vatche Ishakian, Vinod Muthusamy, and Aleksander Slominski. 2017. Status of serverless computing and function-as-a-service(FaaS) in industry and research. *CoRR* abs/1708.08028 (2017). http://arxiv.org/abs/1708.08028
[55] Frederic Francois and Erol Gelenbe. 2016. Towards a cognitive routing engine for software defined networks. In *Proceedings of the IEEE International Conference on Communications*. IEEE. DOI: https://doi.org/10.1109/ICC.2016.7511138
[56] F. Francois, N. Wang, K. Moessner, S. Georgoulas, and R. de Oliveira-Schmidt. 2014. Leveraging MPLS backup paths for distributed energy-aware traffic engineering. *IEEE Trans. Netw. Serv. Manage.* 11, 2 (2014), 235–249.
[57] Ivo Friedberg, Florian Skopik, Giuseppe Settanni, and Roman Fiedler. 2015. Combating advanced persistent threats: From network event correlation to incident detection. *Comput. Secur.* 48 (2015), 35–57.
[58] Peter Xiang Gao, Akshay Narayan, Sagar Karandikar, Joao Carreira, Sangjin Han, Rachit Agarwal, Sylvia Ratnasamy, and Scott Shenker. 2016. Network requirements for resource disaggregation. In *Proceedings of the USENIX Symposium on Operating Systems Design and Implementation (OSDI'16)*. 249–264.
[59] Saurabh Kumar Garg, Steve Versteeg, and Rajkumar Buyya. 2013. A framework for ranking of cloud computing services. *Fut. Gen. Comput. Syst.* 29, 4 (2013), 1012–1023.
[60] Erol Gelenbe. 2014. Adaptive management of energy packets. In *Proceedings of the 2014 IEEE 38th International Computer Software and Applications Conference Workshops (COMPSACW'14)*. IEEE, 1–6.
[61] Erol Gelenbe and Yves Caseau. 2015. The impact of information technology on energy consumption and carbon emissions. *Ubiquity* 2015, Article 1 (Jun. 2015), 1.
[62] Erol Gelenbe and Elif Tugce Ceran. 2016. Energy packet networks with energy harvesting. *IEEE Access* 4 (2016), 1321–1331.
[63] Erol Gelenbe and Ricardo Lent. 2012. Optimising server energy consumption and response time. *Theor. Appl. Inform.* 24, 4 (2012), 257–270.
[64] Erol Gelenbe and Ricardo Lent. 2013. Energy-qos trade-offs in mobile service selection. *Fut. Internet* 5, 2 (2013), 128–139.
[65] Erol Gelenbe, Ricardo Lent, and Markos Douratsos. 2012. Choosing a local or remote cloud. In *Proceedings of the 2012 Second Symposium on Network Cloud Computing and Applications (NCCA'12)*. IEEE, 25–30.
[66] Erol Gelenbe and Toktam Mahmoodi. 2011. Energy-aware routing in the cognitive packet network. In *Proceedings of the First International Conference on Smart Grids, Green Communications and IT Energy-aware Technologies (ENERGY'11)*. 7–12
[67] Erol Gelenbe and Christina Morfopoulou. 2011. A framework for energy-aware routing in packet networks. *Comput. J.* 54, 6 (2011), 850–859.
[68] Erol Gelenbe and Christina Morfopoulou. 2012. Power savings in packet networks via optimised routing. *Mobile Netw. Appl.* 17, 1 (2012), 152–159.
[69] Erol Gelenbe and Simone Silvestri. 2009. Reducing power consumption in wired networks. In *Proceedings of the 24th International Symposium on Computer and Information Sciences (ISCIS'09)*. IEEE, 292–297.
[70] C. Gentry. 2009. Fully homomorphic encryption using ideal lattices. In *Proceedings of the ACM Symposium on Theory of Computing (STOC'09)*.
[71] C. Gentry, A. Sahai, and B. Waters. 2013. Homomorphic encryption from learning with errors: Conceptually-simpler, asymptotically-faster, attribute-based. In *Proceedings of Annual International Cryptology Conference (CRYPTO'13)*. Santa Barbara, CA, USA.
[72] Ali Ghodsi, Matei Zaharia, Benjamin Hindman, Andy Konwinski, Scott Shenker, and Ion Stoica. 2011. Dominant resource fairness: Fair allocation of multiple resource types. In *Proceedings of the 8th USENIX Conference on Networked Systems Design and Implementation (NSDI'11)*. Vol. 11. 24–24.

[73] Rahul Ghosh, Kishor S. Trivedi, Vijay K. Naik, and Dong Seong Kim. 2010. End-to-end performability analysis for infrastructure-as-a-service cloud: An interacting stochastic models approach. In *Proceedings of the 2010 IEEE 16th Pacific Rim International Symposium on Dependable Computing (PRDC'10)*. IEEE, 125–132.

[74] Albert Greenberg, James R. Hamilton, Navendu Jain, Srikanth Kandula, Changhoon Kim, Parantap Lahiri, David A. Maltz, Parveen Patel, and Sudipta Sengupta. 2009. VL2: A scalable and flexible data center network. *SIGCOMM Comput. Commun. Rev.* 39, 4 (Aug. 2009), 51–62. DOI: https://doi.org/10.1145/1594977.1592576

[75] Jayavardhana Gubbi, Rajkumar Buyya, Slaven Marusic, and Marimuthu Palaniswami. 2013. Internet of things (IoT): A vision, architectural elements, and future directions. *Fut. Gen. Comput. Syst.* 29, 7 (2013), 1645–1660.

[76] Haryadi S. Gunawi, Thanh Do, Joseph M. Hellerstein, Ion Stoica, Dhruba Borthakur, and Jesse Robbins. 2011. *Failure as a service (faas): A cloud service for large-scale, online failure drills.* Technical Report UCB/EECS-2011-87. University of California, Berkeley.

[77] Chuanxiong Guo, Guohan Lu, Dan Li, Haitao Wu, Xuan Zhang, Yunfeng Shi, Chen Tian, Yongguang Zhang, and Songwu Lu. 2009. BCube: A high performance, server-centric network architecture for modular data centers. *SIGCOMM Comput. Commun. Rev.* 39, 4 (Aug. 2009), 63–74. DOI: https://doi.org/10.1145/1594977.1592577

[78] Chuanxiong Guo, Guohan Lu, Helen J. Wang, Shuang Yang, Chao Kong, Peng Sun, Wenfei Wu, and Yongguang Zhang. 2010. Secondnet: A data center network virtualization architecture with bandwidth guarantees. In *Proceedings of the 6th International Conference on Emerging Networking Experiments and Technologies (CoNEXT'10)*. ACM, 15.

[79] Abhishek Gupta, Paolo Faraboschi, Filippo Gioachin, Laxmikant V. Kale, Richard Kaufmann, Bu-Sung Lee, Verdi March, Dejan Milojicic, and Chun Hui Suen. 2016. Evaluating and improving the performance and scheduling of HPC applications in cloud. *IEEE Trans. Cloud Comput.* 4, 3 (2016), 307–321.

[80] Hakan Hacigümüş, Bala Iyer, Chen Li, and Sharad Mehrotra. 2002. Executing SQL over encrypted data in the database-service-provider model. In *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data*. ACM, 216–227.

[81] Abdul Hameed, Alireza Khoshkbarforoushha, Rajiv Ranjan, Prem Prakash Jayaraman, Joanna Kolodziej, Pavan Balaji, Sherali Zeadally, Qutaibah Marwan Malluhi, Nikos Tziritas, Abhinav Vishnu, Samee U. Khan, and Albert Zomaya. 2016. A survey and taxonomy on energy efficient resource allocation techniques for cloud computing systems. *Computing* 98, 7 (Jul. 2016), 751–774.

[82] Yi Han, Tansu Alpcan, Jeffrey Chan, Christopher Leckie, and Benjamin I. P. Rubinstein. 2016. A game theoretical approach to defend against co-resident attacks in cloud computing: Preventing co-residence using semi-supervised learning. *IEEE Trans. Inf. Forens. Secur.* 11, 3 (2016), 556–570.

[83] Yi Han, Jeffrey Chan, Tansu Alpcan, and Christopher Leckie. 2017. Using virtual machine allocation policies to defend against co-resident attacks in cloud computing. *IEEE Trans. Depend. Secure Comput.* 14, 1 (2017), 95–108.

[84] Brandon Heller, Srinivasan Seetharaman, Priya Mahadevan, Yiannis Yiakoumis, Puneet Sharma, Sujata Banerjee, and Nick McKeown. 2010. ElasticTree: Saving energy in data center networks. In *Proceedings of the 7th USENIX Conference on Networked Systems Design and Implementation (NSDI'10)*, Vol. 10. 249–264.

[85] Chi-Yao Hong, Srikanth Kandula, Ratul Mahajan, Ming Zhang, Vijay Gill, Mohan Nanduri, and Roger Wattenhofer. 2013. Achieving high utilization with software-driven WAN. In *Proceedings of the ACM SIGCOMM 2013 Conference (SIGCOMM'13)*. ACM, New York, NY, 15–26. DOI: https://doi.org/10.1145/2486001.2486012

[86] Qian Huang. 2014. Development of a SaaS application probe to the physical properties of the earth's interior: An attempt at moving HPC to the cloud. *Comput. Geosci.* 70 (2014), 147–153.

[87] Eduardo Huedo, Rubén S. Montero, Rafael Moreno, Ignacio M. Llorente, Anna Levin, and Philippe Massonet. 2017. Interoperable federated cloud networking. *IEEE Internet Comput.* 21, 5 (2017), 54–59.

[88] IDC. 2017. Worldwide Semiannual Big Data and Analytics Spending Guide. Retrieved May 18, 2018 from http://www.idc.com/getdoc.jsp?containerId=prUS42321417.

[89] IDG Enterprise. 2016. 2016 IDG Enterprise Cloud Computing Survey. Retrieved May 18, 2018 from https://www.idgenterprise.com/resource/research/2016-idg-enterprise-cloud-computing-survey/.

[90] IEEE. 2017. IEEE Rebooting Computing. Retrieved May 18, 2018 from https://rebootingcomputing.ieee.org/.

[91] Shigeru Imai, Thomas Chestna, and Carlos A. Varela. 2013. Accurate resource prediction for hybrid iaas clouds using workload-tailored elastic compute units. In *Proceedings of the 2013 IEEE/ACM 6th International Conference on Utility and Cloud Computing (UCC'13)*. IEEE, 171–178.

[92] Shigeru Imai, Pratik Patel, and Carlos A. Varela. 2016. Developing elastic software for the cloud. In *Encyclopedia on Cloud Computing* (2016).

[93] Shigeru Imai, Stacy Patterson, and Carlos A. Varela. 2017. Maximum sustainable throughput prediction for data stream processing over public clouds. In *Proceedings of the 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*. IEEE Press, 504–513.

[94] Shigeru Imai, Stacy Patterson, and Carlos A. Varela. 2018. Uncertainty-aware elastic virtual machine scheduling for stream processing systems. In *Proceedings of the 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid'18)*.

[95] Sushant Jain, Alok Kumar, Subhasree Mandal, Joon Ong, Leon Poutievski, Arjun Singh, Subbaiah Venkata, Jim Wanderer, Junlan Zhou, Min Zhu, et al. 2013. B4: Experience with a globally-deployed software defined WAN. *ACM SIGCOMM Comput. Commun. Rev.* 43, 4 (2013), 3–14.

[96] Bahman Javadi, Jemal Abawajy, and Rajkumar Buyya. 2012. Failure-aware resource provisioning for hybrid Cloud infrastructure. *J. Parallel Distrib. Comput.* 72, 10 (2012), 1318–1331.

[97] Barkha Javed, Peter Bloodsworth, Raihan Ur Rasool, Kamran Munir, and Omer Rana. 2016. Cloud market maker: An automated dynamic pricing marketplace for cloud users. *Fut. Gen. Comput. Syst.* 54 (2016), 52–67.

[98] Brendan Jennings and Rolf Stadler. 2015. Resource management in clouds: Survey and research challenges. *J. Netw. Syst. Manage.* 23, 3 (2015), 567–619.

[99] Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates et al. 2017. In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA'17)*. IEEE, 1–12.

[100] Christoforos Kachris, Dimitrios Soudris, Georgi Gaydadjiev, Huy-Nam Nguyen, Dimitrios S. Nikolopoulos, Angelos Bilas, Neil Morgan, Christos Strydis, et al. 2016. The VINEYARD approach: Versatile, integrated, accelerator-based, heterogeneous data centres. In *Proceedings of the International Symposium on Applied Reconfigurable Computing*. Springer, 3–13.

[101] Yiping Kang, Johann Hauswald, Cao Gao, Austin Rovinski, Trevor Mudge, Jason Mars, and Lingjia Tang. 2017. Neurosurgeon: Collaborative intelligence between the cloud and mobile edge. In *Proceedings of the 22nd International Conference on Architectural Support for Programming Languages and Operating Systems*. ACM, 615–629.

[102] S. Kannan, A. Gavrilovska, V. Gupta, and K. Schwan. 2017. HeteroOS - OS design for heterogeneous memory management in datacenter. In *Proceedings of the ACM/IEEE 44th Annual International Symposium on Computer Architecture*. 521–534.

[103] James M. Kaplan, William Forrest, and Noah Kindler. 2008. *Revolutionizing Data Center Energy Efficiency*. Technical Report. McKinsey & Company.

[104] Atefeh Khosravi and Rajkumar Buyya. 2017. Energy and carbon footprint-aware management of geo-distributed cloud data centers: A taxonomy, state of the art. *Sustainable Development: Concepts, Methodologies, Tools, and Applications*. 1456–1475.

[105] Mariam Kiran, Peter Murphy, Inder Monga, Jon Dugan, and Sartaj Singh Baveja. 2015. Lambda architecture for cost-effective batch and speed big data processing. In *Proceedings of the IEEE International Conference on Big Data*. IEEE, 2785–2792.

[106] Alok Gautam Kumbhare, Yogesh Simmhan, Marc Frincu, and Viktor K. Prasanna. 2015. Reactive resource provisioning heuristics for dynamic dataflows on cloud infrastructure. *IEEE Trans. Cloud Comput.* 3, 2 (2015), 105–118.

[107] Raghavendra Kune, Pramod Kumar Konugurthi, Arun Agarwal, Raghavendra Rao Chillarige, and Rajkumar Buyya. 2016. The anatomy of big data computing. *Softw. Pract. Exper.* 46, 1 (2016), 79–105.

[108] Tung-Wei Kuo, Bang-Heng Liou, Kate Ching-Ju Lin, and Ming-Jer Tsai. 2016. Deploying chains of virtual network functions: On the relation between link and server usage. In *Proceedings of the IEEE 35th Annual IEEE International Conference on Computer Communications (INFOCOM'16)*. IEEE, 1–9.

[109] Horacio Andrés Lagar-Cavilla, Joseph Andrew Whitney, Adin Matthew Scannell, Philip Patchin, Stephen M. Rumble, Eyal De Lara, Michael Brudno, and Mahadev Satyanarayanan. 2009. SnowFlock: Rapid virtual machine cloning for cloud computing. In *Proceedings of the 4th ACM European Conference on Computer Systems*. ACM, 1–12.

[110] Guyue Liu and Timothy Wood. 2015. Cloud-scale application performance monitoring with SDN and NFV. In *Proceedings of the 2015 IEEE International Conference on Cloud Engineering (IC2E'15)*. IEEE, 440–445.

[111] Zhenhua Liu, Minghong Lin, Adam Wierman, Steven Low, and Lachlan L. H. Andrew. 2015. Greening geographical load balancing. *IEEE/ACM Trans. Netw.* 23, 2 (2015), 657–671.

[112] Raquel V. Lopes and Daniel Menascé. 2016. A taxonomy of job scheduling on distributed computing systems. *IEEE Trans. Parallel Distrib. Syst.* 27, 12 (2016), 3412–3428.

[113] Priya Mahadevan, Puneet Sharma, Sujata Banerjee, and Parthasarathy Ranganathan. 2009. A power benchmarking framework for network devices. *Networking 2009* (2009), 795–808.

[114] Redowan Mahmud, Satish Narayana Srirama, Kotagiri Ramamohanarao, and Rajkumar Buyya. 2018. Quality of experience (QoE)-aware placement of applications in fog computing environments. *J. Parallel Distrib. Comput.* (2018). https://doi.org/10.1016/j.jpdc.2018.03.004

[115] Maciej Malawski, Gideon Juve, Ewa Deelman, and Jarek Nabrzyski. 2015. Algorithms for cost-and deadline-constrained provisioning for scientific workflow ensembles in IaaS clouds. *Fut. Gen. Comput. Syst.* 45 (2015), 1–18.

[116] Zoltán Ádám Mann. 2015. Allocation of virtual machines in cloud data centers-a survey of problem models and optimization algorithms. *ACM Comput. Surv.* 48, 1 (2015), 11.

[117] Sunilkumar S. Manvi and Gopal Krishna Shyam. 2014. Resource management for infrastructure as a Service (IaaS) in cloud computing: A survey. *J. Netw. Comput. Appl.* 41 (2014), 424–440.

[118] Farahd Mehdipour, Bahman Javadi, and Aniket Mahanti. 2016. FOG-engine: Towards big data analytics in the fog. In *Proceedings of the 2016 IEEE 14th International Conference on Dependable, Autonomic and Secure Computing, the 14th International Conference on Pervasive Intelligence and Computing, and the 2nd International Conference on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech'16).* IEEE, 640–646.

[119] Rafael Moreno-Vozmediano, Rubén S. Montero, and Ignacio M. Llorente. 2012. Iaas cloud architecture: From virtualized datacenters to federated cloud infrastructures. *Computer* 45, 12 (2012), 65–72.

[120] Kiran-Kumar Muniswamy-Reddy and Margo Seltzer. 2010. Provenance as first class cloud data. *ACM SIGOPS Operat. Syst. Rev.* 43, 4 (2010), 11–16.

[121] Rekha Nachiappan, Bahman Javadi, Rodrigo Calherios, and Kenan Matawie. 2017. Cloud storage reliability for big data applications: A state of the art survey. *J. Netw. Comput. Appl.* 97 (2017), 35–47.

[122] Marco A. S. Netto, Rodrigo N. Calheiros, Eduardo R. Rodrigues, Renato L. F. Cunha, and Rajkumar Buyya. 2018. HPC cloud for scientific and business applications: Taxonomy, vision, and research challenges. *Comput. Surv.* 51, 1, Article 8 (Jan. 2018), 29 pages.

[123] Radhika Niranjan Mysore, Andreas Pamboris, Nathan Farrington, Nelson Huang, Pardis Miri, Sivasankar Radhakrishnan, Vikram Subramanya, and Amin Vahdat. 2009. PortLand: A scalable fault-tolerant layer 2 data center network fabric. *SIGCOMM Comput. Commun. Rev.* 39, 4 (Aug. 2009), 39–50. DOI: https://doi.org/10.1145/1594977.1592575

[124] Claus Pahl and Brian Lee. 2015. Containers and clusters for edge cloud architectures–A technology review. In *Proceedings of the 2015 3rd International Conference on Future Internet of Things and Cloud (FiCloud'15).* IEEE, 379–386.

[125] Barbara Pernici, Marco Aiello, Jan vom Brocke, Brian Donnellan, Erol Gelenbe, and Mike Kretsis. 2012. What IS can do for environmental sustainability: A report from CAiSE'11 panel on green and sustainable IS. *Commun. Assoc. Inf. Syst.* 30 (2012), 18.

[126] Jorge E. Pezoa and Majeed M. Hayat. 2012. Performance and reliability of non-markovian heterogeneous distributed computing systems. *IEEE Trans. Parallel Distrib. Systems* 23, 7 (2012), 1288–1301.

[127] Chuan Pham, Nguyen H. Tran, Shaolei Ren, Walid Saad, and Choong Seon Hong. 2017. Traffic-aware and energy-efficient vNF placement for service chaining: Joint sampling and matching approach. *IEEE Trans. Serv. Comput.* (2017). DOI: 10.1109/TSC.2017.2671867

[128] Raluca Ada Popa, Catherine Redfield, Nickolai Zeldovich, and Hari Balakrishnan. 2011. CryptDB: Protecting confidentiality with encrypted query processing. In *Proceedings of the 23rd ACM Symposium on Operating Systems Principles.* ACM, 85–100.

[129] Andrew Putnam, Adrian M. Caulfield, Eric S. Chung, Derek Chiou, Kypros Constantinides, John Demme, Hadi Esmaeilzadeh, Jeremy Fowers, Gopi Prashanth Gopal, Jan Gray, et al. 2014. A reconfigurable fabric for accelerating large-scale datacenter services. In *Proceedings of the 2014 ACM/IEEE 41st International Symposium onComputer Architecture (ISCA'14).* IEEE, 13–24.

[130] M. Rajkumar, Anil Kumar Pole, Vittalraya Shenoy Adige, and Prabal Mahanta. 2016. DevOps culture and its impact on cloud delivery and software development. In *Proceedings of the International Conference on Advances in Computing, Communication, & Automation (ICACCA'16).* IEEE.

[131] Benny Rochwerger, David Breitgand, Eliezer Levy, Alex Galis, Kenneth Nagin, Ignacio Martín Llorente, Rubén Montero, Yaron Wolfsthal, Erik Elmroth, Juan Caceres, et al. 2009. The reservoir model and architecture for open federated cloud computing. *IBM J. Res. Dev.* 53, 4 (2009), 4–1.

[132] Eduardo Felipe Zambom Santana, Ana Paula Chaves, Marco Aurelio Gerosa, Fabio Kon, and Dejan S. Milojicic. 2017. Software platforms for smart cities: Concepts, requirements, challenges, and a unified reference architecture. *ACM Computing Surveys* 50, 6 (2017), 78.

[133] Prabodini Semasinghe, Setareh Maghsudi, and Ekram Hossain. 2017. Game theoretic mechanisms for resource management in massive wireless IoT systems. *IEEE Commun. Mag.* 55, 2 (2017), 121–127.

[134] Yogesh Sharma, Bahman Javadi, Weisheng Si, and Daniel Sun. 2016. Reliability and energy efficiency in cloud computing systems: Survey and taxonomy. *J. Netw. Comput. Appl.* 74 (2016), 66–85.

[135] Junaid Shuja, Raja Wasim Ahmad, Abdullah Gani, Abdelmuttlib Ibrahim Abdalla Ahmed, Aisha Siddiqa, Kashif Nisar, Samee U. Khan, and Albert Y. Zomaya. 2017. Greening emerging IT technologies: Techniques and practices. *J. Internet Serv. Appl.* 8, 1 (2017), 9.

[136] Sukhpal Singh and Inderveer Chana. 2016. QoS-aware autonomic resource management in cloud computing: A systematic review. *ACM Comput. Surv.* 48, 3 (2016), 42.

[137] Mukesh Singhal, Santosh Chandrasekhar, Tingjian Ge, Ravi Sandhu, Ram Krishnan, Gail-Joon Ahn, and Elisa Bertino. 2013. Collaboration in multicloud computing environments: Framework and security issues. *Computer* 46, 2 (2013), 76–84.

[138] Sander Soo, Chii Chang, Seng W. Loke, and Satish Narayana Srirama. 2017. Proactive mobile fog computing using work stealing: Data processing at the edge. *Int. J. Mobile Comput. Multimedia Commun.* 8, 4 (2017), 1–19.

[139] Borja Sotomayor, Rubén S. Montero, Ignacio M. Llorente, and Ian Foster. 2009. Virtual infrastructure management in private and hybrid clouds. *IEEE Internet Comput.* 13, 5 (2009).

[140] Satish Narayana Srirama. 2017. Mobile web and cloud services enabling internet of things. *CSI Trans. ICT* 5, 1 (2017), 109–117.

[141] Satish Narayana Srirama and Alireza Ostovar. 2014. Optimal resource provisioning for scaling enterprise applications on the cloud. In *Proceedings of the 6th International Conference on Cloud Computing Technology and Science (CloudCom'14)*. IEEE, 262–271.

[142] Brian Stanton, Mary Theofanos, and Karuna P. Joshi. 2015. Framework for cloud usability. In *Proceedings of the International Conference on Human Aspects of Information Security, Privacy, and Trust*. Springer, 664–671.

[143] Zahir Tari, Xun Yi, Uthpala S. Premarathne, Peter Bertok, and Ibrahim Khalil. 2015. Security and privacy in cloud computing: Vision, trends, and challenges. *IEEE Cloud Comput.* 2, 2 (2015), 30–38.

[144] Adel Nadjaran Toosi, Rodrigo N. Calheiros, and Rajkumar Buyya. 2014. Interconnected cloud computing environments: Challenges, taxonomy, and survey. *ACM Comput. Surv.* 47, 1 (2014), 7.

[145] Amin Vahdat, David Clark, and Jennifer Rexford. 2015. A purpose-built global network: Google's move to SDN. *Queue* 13, 8 (2015), 100.

[146] Carlos Varela and Gul Agha. 2001. Programming dynamically reconfigurable open systems with SALSA. *ACM SIGPLAN Not.* 36, 12 (2001), 20–34.

[147] Carlos A. Varela. 2013. *Programming Distributed Computing Systems: A Foundational Approach*. MIT Press.

[148] Blesson Varghese, Ozgur Akgun, Ian Miguel, Long Thai, and Adam Barker. 2016. Cloud benchmarking for maximising performance of scientific applications. *IEEE Trans. Cloud Comput.* (2016). DOI: 10.1109/TCC.2016.2603476

[149] Blesson Varghese and Rajkumar Buyya. 2018. Next generation cloud computing: New trends and research directions. *Fut. Gen. Comput. Syst.* 79, 3 (2018), 849–861.

[150] Prateeksha Varshney and Yogesh Simmhan. 2017. Demystifying fog computing: Characterizing architectures, applications and abstractions. In *Proceedings of the International Conference on Fog and Edge Computing (ICFEC'17)*.

[151] Sabrina De Capitani Di Vimercati, Sara Foresti, Sushil Jajodia, Stefano Paraboschi, and Pierangela Samarati. 2010. Encryption policies for regulating access to outsourced data. *ACM Trans. Database Syst.* 35, 2 (2010), 12.

[152] Kashi Venkatesh Vishwanath and Nachiappan Nagappan. 2010. Characterizing cloud computing hardware reliability. In *Proceedings of the 1st ACM Symposium on Cloud Computing*. ACM, 193–204.

[153] H. Wang and Laks V. S. Lakshmanan. 2006. Efficient secure query evaluation over encrypted XML databases. In *Proceedings of the International Conference on Very Large Data Bases (VLDB'06)*. Seoul, Korea.

[154] Lan Wang, Olivier Brun, and Erol Gelenbe. 2016. Adaptive workload distribution for local and remote clouds. In *Proceedings of the 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC'16)*. IEEE, 003984–003988.

[155] Lan Wang and Erol Gelenbe. 2018. Adaptive dispatching of tasks in the cloud. *IEEE Trans. Cloud Comput.* 6, 1 (2018), 33–45. DOI: https://doi.org/10.1109/TCC.2015.2474406

[156] Kim Weins. 2015. Cloud Computing Trends: 2015 State of the Cloud Survey. Retrieved May 18, 2018 from https://www.rightscale.com/blog/cloud-industry-insights/cloud-computing-trends-2015-state-cloud-survey.

[157] Song Wu, Chao Niu, Jia Rao, Hai Jin, and Xiaohai Dai. 2017. Container-based cloud platform for mobile computation offloading. In *Proceedings of the 2017 IEEE International Parallel and Distributed Processing Symposium (IPDPS'17)*. IEEE, 123–132.

[158] Liang Xiao, Dongjin Xu, Caixia Xie, Narayan B. Mandayam, and H. Vincent Poor. 2017. Cloud storage defense against advanced persistent threats: A prospect theoretic study. *IEEE J. Select. Areas Commun.* 35, 3 (2017), 534–544.

[159] Yonghua Yin, Lan Wang, and Erol Gelenbe. 2017. Multi-layer neural networks for quality of service oriented server-state classification in cloud servers. In *Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN'17)*. IEEE, 1623–1627.

[160] A. J. Younge, J. P. Walters, S. Crago, and G. C. Fox. 2014. Evaluating GPU passthrough in xen for high performance cloud computing. In *Proceedings of the 2014 IEEE International Parallel Distributed Processing Symposium Workshops*. 852–859. DOI: https://doi.org/10.1109/IPDPSW.2014.97

[161] Bowen Zhou, Amir Vahid Dastjerdi, Rodrigo Calheiros, Satish Srirama, and Rajkumar Buyya. 2017. mCloud: A context-aware offloading framework for heterogeneous mobile cloud. *IEEE Trans. Serv. Comput.* 10, 5 (2017), 797–810.

[162] Qunzhi Zhou, Yogesh Simmhan, and Viktor Prasanna. 2017. Knowledge-infused and consistent complex event processing over real-time and persistent streams. *Fut. Gen. Comput. Syst.* 76 (2017), 391–406.
[163] Tianqing Zhu, Gang Li, Wanlei Zhou, and S. Yu Philip. 2017. Differentially private data publishing and analysis: A survey. *IEEE Trans. Knowl. Data Eng.* 29, 8 (2017), 1619–1638.

Received November 2017; revised June 2018; accepted July 2018