Next Generation Cloud Computing: New Trends and Research Directions

Blesson Varghese*

School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, UK

Rajkumar Buyya

Cloud Computing and Distributed Systems (CLOUDS) Laboratory
School of Computing and Information Systems, The University of Melbourne, Australia

Abstract

The landscape of cloud computing has significantly changed over the last decade. Not only have more providers and service offerings crowded the space, but also cloud infrastructure that was traditionally limited to single provider data centers is now evolving. In this paper, we firstly discuss the changing cloud infrastructure and consider the use of infrastructure from multiple providers and the benefit of decentralising computing away from data centers. These trends have resulted in the need for a variety of new computing architectures that will be offered by future cloud infrastructure. These architectures are anticipated to impact areas, such as connecting people and devices, data-intensive computing, the service space and self-learning systems. Finally, we lay out a roadmap of challenges that will need to be addressed for realising the potential of next generation cloud systems.

Keywords: cloud computing, fog computing, cloudlet, multi-cloud, serverless computing, cloud security

1. Introduction

Resources and services offered on the cloud have rapidly changed in the last decade. These changes were underpinned by industry and academia led efforts towards realising computing as a utility [1]. This vision has been achieved, but there are continuing changes in the cloud computing landscape which this paper aims to present.

Applications now aim to leverage cloud infrastructure by making use of heterogeneous resources from multiple providers. This is in contrast to how resources from a single cloud provider or data center were used traditionally. Consequently, new computing architectures are emerging. This change is impacting a number of societal and scientific areas. In this discussion paper, we consider *'what future cloud computing looks like'* by charting out trends and directions for pursuing meaningful research in developing next generation computing systems as shown in Figure 1.

The remainder of this paper is organised as follows. Section 2 presents a discussion of the evolving infrastructure on the cloud. Section 3 highlights the emerging computing architectures and

*Corresponding author
Email addresses: varghese@qub.ac.uk (Blesson Varghese), rbuyya@unimelb.edu.au (Rajkumar Buyya)
URL: www.blessonv.com (Blesson Varghese), www.buyya.com (Rajkumar Buyya)

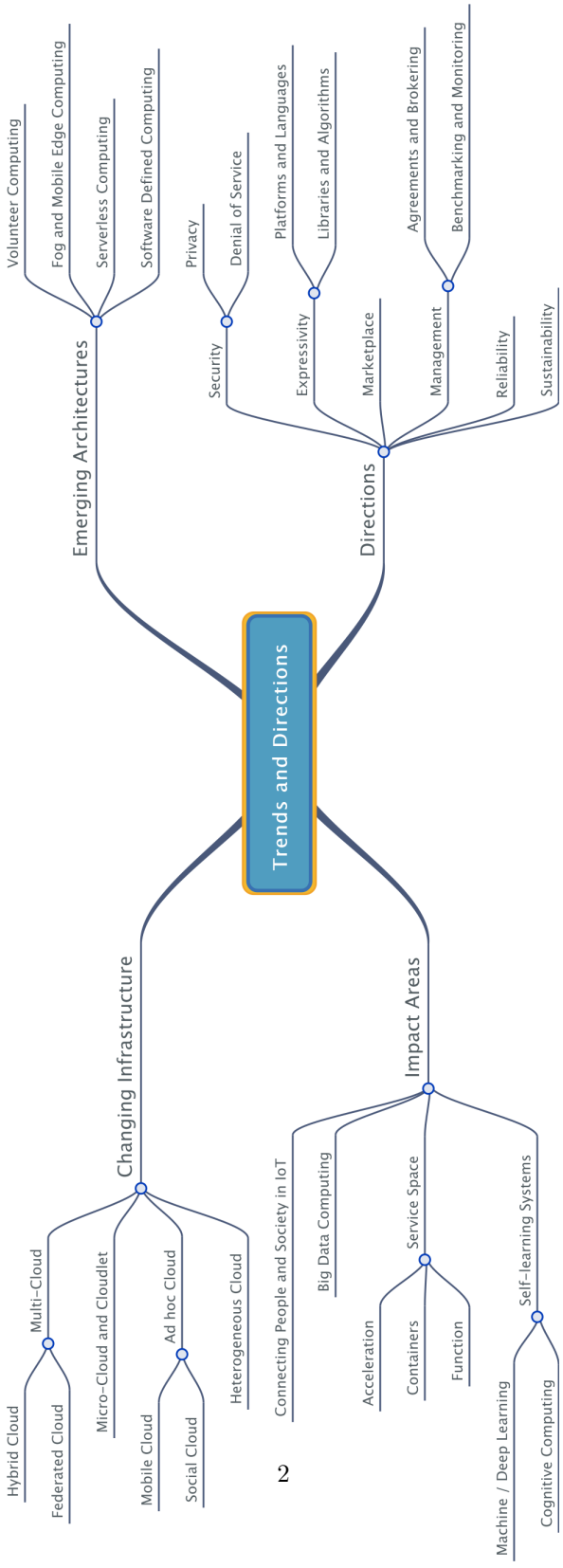

Figure 1. A snapshot of trends and directions in next generation cloud computing

their advantages. Section 4 considers a number of areas that future clouds will impact. Section 5 sets out a number of challenges that will need to be addressed for developing next generation cloud systems. Section 6 concludes this paper.

2. Changing Infrastructure

The majority of existing infrastructure hosting cloud services comprises dedicated compute and storage resources located in data centers. Hosting cloud applications on data centers of a single provider is easy and provides obvious advantages. However, using a single provider and a data center model poses a number of challenges. A lot of energy is consumed by a large data center to kee it operational. Moreover, centralised cloud data centers like any other centralised computing model is susceptible to single point failures. Additionally, data centers may be geographically distant from its users, thereby requiring data to be transferred from its source to resources that can process it in the data center. This would mean that applications using or generating sensitive or personal data may have to be stored in a different country than where it originated.

Alternate models of using cloud infrastructure instead of using data centers of a single provider have been proposed in recent years [2]. In this paper, we consider the multi-cloud, micro cloud and cloudlet, ad hoc cloud and heterogeneous cloud to demonstrate the trends in changing infrastructure of the cloud. The feasibility of these have been reported in literature and will find real deployment of workloads in next generation cloud computing.

2.1. Multi-cloud

The traditional notion of multi-cloud was leveraging resources from multiple data centers of a provider. Then applications were hosted to utilise resources from multiple providers [3, 4]. Rightscale estimates that current businesses use an average of six separate clouds[1].

The use of multi-clouds are increasing, but there are hurdles that will need to be overcome. For example, common APIs to facilitate multi-cloud need to account for different types of resources offered by multiple providers. This is not an easy given that more resources are rapidly added to the cloud marketplace and there are no unified catalogues that report a complete set of resources available on the cloud. Further, the abstractions, including network and storage architectures differ across providers, which makes the adoption of multi-cloud bespoke to each application rather than using a generic platform or service. Along with the different resources, hypervisors, and software suites employed, the pricing and billing models are significantly different across providers, all of which results in significant programming effort required for developing a multi-cloud application. All management tasks, such as fault tolerance, load balancing, resource management and accounting need to be programmed manually since there are no unifying environments that make these possible. Examples of APIs that alleviate some of these challenges include Libcloud[2] and jClouds[3]. However, further research is required for enabling adoption of clouds across multiple providers.

Hybrid Cloud: A multi-cloud can take the form of a hybrid cloud - a combination of public and private clouds or a combination of public and private IT infrastructure [5, 6]. These clouds cater for bursty demands or resource demands known beforehand. The benefit of using hybrid

[1] `http://www.forbes.com/sites/joemckendrick/2016/02/09/typical-enterprise-uses-six-cloud-computing-services-survey-shows/#e2207a47be31`

[2] `https://libcloud.apache.org/`

[3] `https://jclouds.apache.org/`

clouds for handling sensitive data is known [7]. It is estimated that 63% of organisations using the cloud have adopted a hybrid cloud approach[4] with use-cases reported in healthcare[5] and energy sectors[6]. The key challenge in setting up a hybrid cloud is network related. Bandwidth, latency and network topologies will need to be considered for accessing a public cloud from a private cloud [8]. Network limitations can result in an ineffective hybrid cloud. Dedicated networking between clouds may enable more effective infrastructure, but requires additional management of private resources, which can be a cumbersome task.

Federated Cloud: There are a number of benefits in bringing together different cloud providers under a single umbrella resulting in a federated cloud [9, 10]. This can provide a catalogue of services and resources available as well as makes applications interoperable and portable. The EU based EGI Federated Cloud is an example of this and brings together over 20 cloud providers and 300 data centers[7]. Federated clouds can address the vendor lock-in problem in that applications and data can be migrated from one cloud to another. This is not easy given different abstractions, resource types, networks and images as well as variable vendor specific costs for migrating large volumes of data. More recently, there are ongoing efforts to federate resources that are located outside cloud data centers, which is considered next.

2.2. Micro cloud and Cloudlet

Data centers occupy large amounts of land and consume lots of electricity to provide a centralised computing infrastructure. This is a less sustainable trend, and alternate low power and low cost solutions are proposed. There are recent efforts to decentralise computing towards the edge of the network for making computing possible closer to where user data is generated [11]. Small sized, low cost and low power computing processors co-located with routers and switches or located in dedicated spaces closer to user devices, referred to as micro clouds are now developed for this purpose [12, 13, 14]. However, there are no public deployments given the challenges in networking micro cloud installations over multiple sites. In the UK, there are efforts to connect micro clouds for general purpose computing[8].

Micro clouds lend themselves in reducing latency of applications and minimising the frequency of communication between user devices and data centers. Odroid boards[9] and Raspberry Pis[10] for example are used to develop micro clouds. However, integration of micro clouds to the existing computing ecosystem is challenging and efforts are being made in this direction [15]. One of the key challenges is scheduling applications during run time to make use of micro clouds along with a data center. This includes partitioning an application and its data across both high end and low power processors to improve the overall performance measured by user-defined objectives. In a decentralised cloud computing approach, application tasks will need to be offloaded both from data centers and user devices on to micro clouds. The challenge here is in using micro clouds (that may

[4] http://www.cloudpro.co.uk/cloud-essentials/hybrid-cloud/6445/63-of-organisations-embracing-hybrid-environments

[5] https://usa.healthcare.siemens.com/medical-imaging-it/image-sharing/image-sharing-archiving-isa

[6] http://w3.siemens.com/smartgrid/global/en/products-systems-solutions/smart-metering/emeter/pages/emeter-cloud.aspx

[7] https://www.egi.eu/

[8] http://gow.epsrc.ac.uk/NGBOViewGrant.aspx?GrantRef=EP/P004024/1

[9] http://www.hardkernel.com

[10] https://www.raspberrypi.org/

or may not be always available) with network management abstraction between the cloud and the edge without depending on the underlying hardware.

The aim of a cloudlet is similar to a micro cloud in extending cloud infrastructure towards the edge of the network [16, 17], but is used in literature in the context of mobile computing. It is used for improving the latency and overall Quality of Service (QoS) of mobile applications. Next generation computing systems will integrate computing on the cloudlet to service local traffic and reduce network traffic towards cloud data centers beyond the first hop in the network. The Elijah[11] project is an example of advances in the cloudlet arena.

2.3. Ad hoc cloud

The use of micro clouds and cloudlets will need to leverage on the concept of ad hoc computing that has existed from the grid era. For example, SETI@home[12] was a popular project that aimed to create a computing environment by harnessing spare resources from desktops using BOINC[13]. The concept of ad hoc clouds is based on the premise of ad hoc computing in that underutilised resources, such as servers owned by organisations can be harnessed to create an elastic infrastructure [18, 19]. This is in contrast to existing cloud infrastructure which is largely data center based and in which the resources available are known beforehand.

However, the context of an ad hoc cloud is changing with increasing connectivity of a large variety of resources to the cloud [20]. This is becoming popular for smaller mobile devices, such as smartphones [21, 22], which on an average have less than a 25% per hour of usage [23, 24]. The spare resources of smartphones can contribute to creating an ad hoc infrastructure (such as cloudlets) that supports low latency computing for non-critical applications in public spaces and transportation systems. The assumption here is that one device is surrounded by a large number of devices that will complement computing for the former device. Although such an infrastructure is not reliable, it may be used in conjunction with existing data centers to enhance connectivity. Such ad hoc clouds may be an enabler for deployments of cloudlets that improve the QoS of applications.

2.4. Heterogeneous Cloud

Heterogeneity in cloud computing can be considered in at least two ways. The first is in the context of multi-clouds, in which platforms that offer and manage infrastructure and services of multiple cloud providers are considered to be a heterogeneous cloud. Heterogeneity arises from using hypervisors and software suites from multiple vendors.

The second is related to low-level heterogeneity at the infrastructure level, in which different types of processors are combined to offer VMs with heterogeneous compute resources. In this paper, the latter is referred to as heterogeneous clouds. While supercomputers have become more heterogeneous by employing accelerators, such as NVIDIA GPUs or Intel Xeon Phis, cloud data centers mostly employ homogeneous architectures [25]. More recently heterogeneous cloud data center architectures have been proposed[14]. In the vendor arena, Amazon along with other providers offer GPU-based VMs, but accelerators are not yet fully integrated into the computing ecosystem. This is because it is not yet possible for a programmer to fully develop and execute code oblivious

[11] `http://elijah.cs.cmu.edu/`
[12] `http://setiathome.berkeley.edu/`
[13] `http://boinc.berkeley.edu/`
[14] `http://www.harness-project.eu/`

to the underlying hardware. There are a number of efforts in this direction, but the key challenge is achieving a high-level abstraction that can be employed across multiple architectures, such as GPUs, FPGAs and Phis [26, 27, 28, 29]. Further applications that already execute on the cloud cannot be scheduled onto heterogeneous resources. Efforts in this direction are made by the CloudLightning[15] project. The concept of a heterogeneous cloud may extend beyond the data center. For example, ad hoc clouds or microclouds could be heterogeneous cloud platforms.

3. Emerging Computing Architectures

Conventional cloud computing requires applications to simply follow a two tier architecture in which front end nodes use a service offered by the cloud. With the increasing number of sensor rich devices, such as smartphones, tables and wearables, large volumes of data is generated. Gartner forecasts that by 2020 over 20 billion devices will be connected to the internet consequently generating 43 trillion gigabytes of data[16]. This poses significant networking and computing challenges that will degrade the Quality-of-Service (QoS) and Experience (QoE) that cannot be met by existing infrastructure. Adding more centralised cloud data centers or eliminating them from the computing system will not address the problem. Instead a fundamentally different approach extending the computing ecosystem beyond cloud data centers towards the user will pave the way forward. This will include resources at the edge of the network or resources voluntarily contributed by owners, which is typically not considered in conventional cloud computing.

The cloud computing infrastructure is evolving and requires new computing models to satisfy large-scale applications. In this paper, we consider five computing models, namely volunteer computing, fog and mobile edge computing, serverless computing, resilient computing and software-defined computing that will set trends in future clouds.

3.1. Volunteer Computing

Ad hoc clouds and cloudlets are emerging to accommodate more innovative user-driven and mobile applications that can benefit from computing closer to user devices. The availability of compute resources is not guaranteed in an ad hoc cloud or cloudlet as in a conventional data center and therefore a pay-as-you-go or an upfront payment for reserving compute, storage or network resources will not be suitable. Instead, a crowd funded approach in which spare resources from user computers or devices are volunteered for creating an ad hoc cloud. Such a computing model may be used to support applications that have a societal or scientific focus.

Volunteer cloud computing [30, 31, 32] can take different forms. For example, users of a social network may share their heterogeneous computing resources in the form of an ad hoc cloud. This is referred to as 'social cloud computing' [33, 34]. More reliable owners are rewarded through a reputation marker within the social network. The Cloud@Home project rewards volunteers by payment for their resource donations [35]. Gamification is also reported as an incentive [36]. Similar research is also reported in the name of 'peer-to-peer cloud computing' [37].

The challenges that need to be overcome to fully benefit from volunteer cloud computing will be firstly in minimising the overheads for setting up a highly virtualised environment given that the underlying hardware will be heterogeneous and ad hoc. Moreover, there are security and privacy

[15] `http://cloudlightning.eu/`
[16] `http://www.gartner.com/newsroom/id/3165317`

concerns that will need to be addressed to boost confidence in the public to more readily become volunteers for setting up ad hoc clouds. Furthermore, a consistent platform that can integrate social networks with cloud management will need to emerge.

3.2. Fog and Mobile Edge Computing

The premise of fog computing is to leverage the existing compute resources on edge nodes, such as mobile base stations, routers and switches, or add computing capability to such nodes along the entire data path between user devices and a cloud data center. This will become possible if general purpose computing can be facilitated on existing edge nodes or additional infrastructure, such as micro clouds or cloudlets are deployed. Preliminary research reports the applicability of fog computing for use-cases, such as in online games and face recognition [38].

One characteristic of using fog computing is that applications can be vertically scaled across different computing tiers. This will enable only essential data traffic beyond the data source. Workloads can be offloaded from cloud data centers on to edge nodes [39, 40] or from user devices on to edge nodes [41, 42]. Additionally, an aggregating model [43, 44], in which data is aggregated from multiple devices or sensors will be possible. Application level and operating system containers may substitute the more heavyweight virtual machines for deploying workloads.

The term 'Mobile Edge Computing (MEC)' is used in literature [45, 46], which is similar to fog computing in that the edge of the network is employed. However, it is limited to the mobile cellular network and does not harness computing along the entire path taken by data in the network. In this computing model the radio access network may be shared with the aim to reduce network congestion. Application areas that benefit include low latency content delivery, data analytics and computational offloading [47, 48] for improving response time. Intel has reported the real life use of MEC[17] and industry led proof-of-concept models that support MEC have been developed[18]. It is anticipated that MEC will be adopted in 4G/5G networks[19].

To realise fog computing and MEC at least two challenges will need to be addressed. Firstly, complex management issues related to multi-party service level agreements [49, 50], articulation of responsibilities and obtaining a unified platform for management given that different parties may own edge nodes. Secondly, enhancing security and addressing privacy issues when multiple nodes interact between a user device and a cloud data center [51, 52]. The Open Fog consortium[20] is making a first step in this direction.

3.3. Serverless Computing

Conventional computing on the cloud requires an application to be hosted on a Virtual Machine (VM) that in turn offers a service to the user. If a web server is hosted on a cloud VM, for example then the service owner pays for the entire time the server application is hosted (regardless of whether the service was used). The metrics against which the performance of an application is generally benchmarked include latency, scalability, and elasticity. Therefore, development efforts on the cloud focus on these metrics. The cost model followed is 'per VM per hour' and does not take idle time into account (the VM was provisioned, but the server was idle since there were no requests or the

[17] `https://builders.intel.com/docs/networkbuilders/Real-world-impact-of-mobile-edge-computing-MEC.pdf`

[18] `http://mecwiki.etsi.org/index.php?title=Main_Page`

[19] `http://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp11_mec_a_key_technology_towards_5g.pdf`

[20] `https://www.openfogconsortium.org/`

application was not running). This is because the VM on which the server is running requires to be provisioned. However, with decentralised data center infrastructure that may have relatively less processing power, it will not be ideal to continually host servers that will remain idle for a prolonged period of time. Instead an application in a fog or MEC environment may be modularised with respect to the time taken to execute a module or the memory used by the application. This will require a different cost model that accounts for the memory consumed by the application code for the period it was executed and the number of requests processed

As the name implies 'serverless' does not mean that computing will be facilitated without servers. In this context, it simply means that a server is not rented as a conventional cloud server and developers do not think of the server and the residency of applications on a cloud VM[21]. From a developers perspective challenges such as the deployment of an application on a VM, over/under provisioning of resources for the application, scalability and fault tolerance do not need to be dealt with. The infrastructure, including the server is abstracted away from the user and instead properties, such as control, cost and flexibility are considered.

In this novel approach, functions (modules) of the application will be executed when necessary without requiring the application to be running all the time. Sometimes this is also referred to as Function-as-a-Service or event-based programming. An event may trigger the execution of a function or a number of functions in parallel. Examples of platforms that currently support this architecture includes AWS Lambda[22], IBM OpenWhisk[23] and Google Cloud Functions[24]. Forbes predicts that the use of serverless computing will increase given that billions of devices will need to be connected to the edge of the network and data centers[25]. It will not be feasible to have idle servers in resource constrained environments. The challenges that will hinder the widespread adoption of serverless computing will be the radical shift in the properties of an application that a programmer will need to focus on; not latency, scalability and elasticity, but those that relate to the modularity of an application, such as control and flexibility. Another challenge is developing programming models that will allow for high-level abstractions to facilitate serverless computing. The effect and trade-offs of using traditional external services along with serverless computing services will need to be investigated in orchestrating future cloud-based systems.

3.4. Software-Defined Computing

There is a large amount of traffic that traditionally did not exist in a two-tier cloud architecture. This is due to the ever increasing number of devices that are catered for by the Internet. Consequently, there is an increasing volume of data that needs to be transferred from one location to another to support applications that rely on multiple cloud services. To efficiently manage this, the networking technology needs to support a dynamic architecture. *Software Defined Networking (SDN)* is an approach of isolating the underlying hardware in the network from the components that control data traffic [53, 54]. This abstraction allows for programming the control components of the network to obtain a dynamic network architecture.

[21] https://d0.awsstatic.com/whitepapers/AWS_Serverless_Multi-Tier_Architectures.pdf

[22] https://aws.amazon.com/lambda/

[23] https://developer.ibm.com/openwhisk/

[24] https://cloud.google.com/functions/

[25] http://www.forbes.com/sites/ibm/2016/11/17/three-ways-that-serverless-computing-will-transform-app-development-in-2017/?cm_mc_uid=24538571706014848428726&cm_mc_sid_50200000=1484842872\#4f04f02565a3

In the context of future clouds, there are a number of challenges and opportunities relevant to developing SDN. Firstly, there are challenges in developing hybrid SDNs in lieu of centralised or distributed SDNs [55]. Research is required to facilitate physically distributed protocols while logically centralised control tasks can be supported. The second challenge is in developing techniques to capture Quality-of-Service by taking both the network and cloud infrastructure into account [55]. This is required for capturing end-to-end QoS and improving user experience in both virtualised network and hardware environments. Thirdly, the interoperability of Information-Centric Networking (ICN) will need to be facilitated as cloud networks adopt ICN over SDN [55]. Fourthly, developing mechanisms for facilitating network virtualisation for different granularities, for example per-job or per-task granularity [56].

With emerging distributed cloud computing architectures, 'software defined' could be applied not only to networking, but also to storage[26] and compute as well as resources beyond data centers for delivering effective cloud environments [57, 58]. This concept when applied to compute, storage and networks of a data center and resources beyond is referred to as Software Define Computing (SDC). This will allow for easily reconfiguring and adapting physical resources to deliver the agreed QoS metrics. The complexity in configuring and operating the infrastructure is alleviated in this case.

4. Avenues of Impact

Next generation cloud computing systems are aimed at becoming more ambient, pervasive and ubiquitous given the emerging trends of distributed, heterogeneous and ad hoc cloud infrastructure and associated computing architectures. This will impact at least the following four areas considered in this paper.

4.1. Connecting People and Devices in the Internet-of-Things

Innovation in the cloud arena along with prolific growth of the sensors and gadgets is bringing people, devices and the associated computing closer. The concept of combining multiple sensor environments, including sensors embedded into infrastructure (transportation, communication, buildings, healthcare and utilities) and sensors on user devices, wearables and appliances has resulted in the upcoming *Internet of Things (IoT)* [59, 60]. The aim is to improve the accuracy and efficiency of an actuation process and reduces human intervention. The 'things' in the IoT context may vary from microchips, biometric sensors, sensors on mobile phones and electrical gadgets at home to sensors embedded on infrastructure monitoring pollution, temperature, light etc. Conventional cloud computing architectures will be limiting in connecting billions of things, but a combination of computing architectures presented in Section 3 will facilitate the IoT vision.

Among many, a key challenge will be related to end-to-end security in networks given that sensor networks, wireless networks, RFID devices, cloud data centers, edge nodes, public and private clouds will need to be integrated for achieving IoT systems. Current mechanisms in securing networks involves encryption and authentication, which will prevent outsider attacks. However, additional techniques are required against insider malicious attacks. This will require the development of secure reprogrammable protocols that allow the authentication of events that triggers a function in the network (as in serverless computing), thereby preventing malicious installations.

[26] `https://www.opensds.io/`

Additionally, innovation in sensing and decision-making is required. Traditionally, sensing involved physical sensors that are integrated in the environment, but future IoT systems will involve the integration of physical sensors, people-centric sensors (low cost sensing of the environment localised to the user) and human sensors (people provide data about their environment) [61, 62, 63]. The key challenge here is that data is unstructured and platforms that account for this are required and currently not available.

4.2. Big Data Computing

The consequence of emerging computing models is that they generate large volumes of data, referred to as 'Big Data'. Data generated by organisations or users are transferred to a data store on the cloud (this is a result of employing a centralised 'application to data center' architecture). The data that is stored may never be used again and is often referred to as dark data. It is usually expensive to move data out of the store and perform any analytics. The opportunity to process data is before it is stored in the cloud.

As the cloud infrastructure becomes decentralised, more opportunities unveil to facilitate processing closer to where it is generated before storing it. For example, edge nodes may be used for processing image or video data before it is stored. However, existing research in big data usually considers centralised cloud architectures or multiple data centers. To leverage distributed cloud architectures there are a number of challenges that will need to be addressed.

Firstly, data processing and resource management on distributed cloud nodes[27] [64]. Whether they be ad hoc clouds, heterogeneous clouds or distributed clouds, there needs to be platforms that can take into account the ad hoc nature of nodes that may process data in a distributed cloud setting, the heterogeneity of processors and platforms that scale from low power processors to high-end processors without significant programming efforts.

Secondly, building models for analytics that scale both horizontally and vertically. Current models typically scale horizontally across multiple nodes in a data center or across nodes in multiple data centers. In the future, models that scale vertically from low end processors to data center nodes will need to be developed.

Thirdly, software stacks for end-to-end processing [65]. This relates to both the first and second challenges. Currently, most big data solutions assume a centralised cloud as the compute resource, but integrating micro clouds, cloudlets or traffic routing edge nodes in the software stack will need to be addressed.

In the real-world, the volume of unstructured data as opposed to structured data is increasing. Often unstructured data is interconnected (for example, generated from social networks) and takes the form of natural language. One key challenge is achieving accurate and actionable knowledge from unstructured data. To address this challenge, one approach will be to transform unstructured data to structured networks and then to knowledge, referred to as data-to-network-to-knowledge (D2N2K)[28]. However, this is challenging since automated and distant supervision methods will need to be firstly developed [66, 67]. Then methods will be required to derive knowledge from structured networks represented as graphs. Similarly, there are a number of challenges when performing analytics on large graphs. They include the need for designing novel mechanisms for fast searching and

[27] `http://www.cloud-council.org/deliverables/CSCC-Deploying-Big-Data-Analytics-Applications-to-the-Cloud-Roadmap-for-Success.pdf`

[28] `https://www.nsf.gov/awardsearch/showAward?AWD_ID=1705169\&HistoricalAwards=false`

querying [68, 69] and secure searching and indexing underpinned by homomorphic algorithms [70]. These remain open areas of research [71, 72].

4.3. Service Space

The abstraction of infrastructure, platforms and software were initially offered as services (IaaS, PaaS and SaaS) on the cloud. However, the service space is becoming richer with a wide variety of services. For example, to offer acceleration provided by GPUs to applications *Acceleration-as-a-Service (AaaS)* has been proposed [73]. In the future, as more applications make use of hardware accelerators the AaaS space is expected to become more mature. Currently, GPU virtualisation technologies, such as rCUDA facilitate the use of GPU services [74, 75]. However, most AaaS services still require applications to be specifically written for an accelerator. Further, a wider variety of accelerators, such as coprocessors, FPGAs and ASICs (such as Tensor Processing Units (TPUs) need to be integrated in future clouds to enable computing in device rich environments, such as fog computing and IoT. There is ongoing research to migitate these challenges, for example the Anyscale Apps[29] project and the OpenACC initiative[30].

Another area in the service space that is gaining significant traction is *Container-as-a-Service (CaaS)* [76, 77]. The benefits of deploying containers have been investigated for a variety of applications (although they are not applicable for all workloads). Consequently, containers are starting to be adopted as an alternate virtualisation technology. CaaS offers the deployment and management of containers, which will be required for workload execution in ad hoc clouds and micro clouds for enabling volunteer computing and fog computing, respectively. Examples include Google Kubernetes[31], Docker Swarm[32] and Rackspace Carina[33]. However, avenues such as container monitoring and live migration will need to be developed [78]. Dealing with dependencies and the portability of containers remains an open issue. The security aspects of containers due to weak isolation relative to cloud VMs needs to be further understood.

With the adoption of event-based platforms for enabling serverless computing, more applications will make use of *Function-as-a-Service (FaaS)*. The aim will be to execute functions on the cloud platform that are initiated by events. This is in contrast to current execution models in which an application is constantly running on the server to furnish a client request and is billed even when the server application remains idle when it is not servicing requests[34].

4.4. Self-learning systems

Currently, a large volume of user generated data in the form of photo, audio and video and metadata, such as network and user activity information, are moved to the cloud. This is due to the availability of relatively cheaper data storage and back up on the cloud. There is ongoing research in applying machine learning to speech/audio recognition, text, image and video analysis, and language translation applications [79]. This research is branded under the general umbrella of *'Deep Learning'* [80]. Traditionally machine learning algorithms were restricted to execution on large clusters given the large computational requirements. However, APIs and software libraries

[29] `http://anyscale.org/`
[30] `http://www.openacc.org/`
[31] `https://kubernetes.io/`
[32] `https://www.docker.com/products/docker-swarm`
[33] `https://getcarina.com/`
[34] `https://serverless.com/`

are now available to perform complex learning tasks without incurring significant monetary costs. Examples include the Google TensorFlow[35] and Nervana Cloud[36]. The availability of hardware accelerators, such as GPUs, in cloud environments has reduced the computing time for machine learning algorithms on large volumes of data [81, 82, 83]. Interest from the industry in this area is due to the potential of deep learning in predictive analytics.

A closely related avenue in the context of future clouds is *Cognitive Computing*. In this visionary model, cognitive systems will rely on machine learning algorithms and the data that is generated to continually acquire knowledge, model problems and determine solutions. Examples include the use of IBM Watson for speech and facial recognition and sentiment analysis[37]. APIs and SaaS supporting Watson are currently available. The hardware employed in cognitive systems may rely on functions of the human brain and are inherently massively parallel (examples include the SyNAPSE [84] and the SpiNNaker [85] architectures). It is anticipated that these architectures will be integrated in next generation clouds.

5. Research Directions

In this section, we present a a few directions that academic cloud research can contribute to in light of the new trends considered in the previous sections.

5.1. Guaranteeing Enhanced Security

The key to widespread adoption of computing remotely is security that needs to be guaranteed by a provider [86, 87, 88]. In the traditional cloud, there are significant security risks related to data storage and hosting multiple users, which are mitigated by robust mechanisms to guarantee user and user data isolation. However, this becomes more complex, for example in the fog computing ecosystem, the above risks are of greater concern, since a wide range of nodes are accessible to users (for example, the security of traffic routed through nodes, such as routers [89, 90]). For example, a hacker could deploy malicious applications on an edge node, which in turn may exploit a vulnerability that may degrade the QoS of the router. Such threats may have a significant negative impact. Moreover, if user specific data needs to be temporarily stored on multiple edge locations to facilitate computing on the edge, then privacy issues along with security challenges will need to be addressed. We recommend the design and development of methods to characterise and detect malwares at large scale [91].

Vulnerability studies that can affect security and privacy of a user when an application is scaled across both vertical (data centers, edge nodes, user devices) and horizontal (across multiple edge nodes and user devices) hierarchy will need to be considered. These studies will need to consider privacy concerns that are both inherited from traditional cloud systems and emerge from integrating sensors in the internet [92]. One open area that will need to be considered for distributed clouds is the authentication of distributed (edge, peer, ad hoc, micro cloud) nodes. Another area is developing suitable encryption-decryption mechanisms that are less resource hungry and energy consuming that scale on resource deprived nodes of distributed clouds. Also, methods to detect intrusion, such as anomaly detection, will need to be designed for real-time resource and bandwidth limited environments for upcoming IoT workloads.

[35] `https://www.tensorflow.org/`
[36] `https://www.nervanasys.com/cloud/`
[37] `https://www.ibm.com/watson/`

Distributed Denial of Service (DDoS) has become a well known security threat in the cloud arena and is a potential threat with greater negative impact on distributed clouds [93, 94]. Malicious users and hackers tend to exploit vulnerabilities of cloud services that are underpinned by virtualisation, auto-scaling mechanisms and multi-tenancy. Typically, the attack deprives other users on the cloud of resources and bandwidth, thereby making them incur more monetary cost for less optimised performance. This in turn negatively affects customer trust of the cloud provider and impacts cloud adoption. Often large-scale attacks that are reported in popular media have political and business motives. Three broad mechanisms, namely prevention, detection and mitigation are generally proposed to address such attacks. However, the development and adoption of concrete methods on the cloud are still in their infancy. The points of vulnerabilities increase as distributed cloud architectures are adopted and as more users and devices are connected to the cloud. Attack prevention, detection and mitigation methods will need to be further developed in conjunction with foundation technologies that enable the cloud.

5.2. Achieving Expressivity of Applications for Future Clouds

Expressing distributed applications using emerging computing models on changing infrastructure is an important research direction.

Platforms and Languages: in addition to popular programming languages, there is a wide variety of services to deploy applications on the cloud. However, with the increasing emphasis on distributed cloud architectures there will be the need for developing platforms and toolkits that account for the integration and management of edge nodes. Given that distributed cloud applications will find its use-cases in user-driven applications, existing platforms cannot be used to easily program an application, such as a distributed workflow. The programming model that aims to exploit edge nodes will need to execute workloads on multiple hierarchical levels. Languages that support the programming model will need to take the heterogeneity of hardware and the capacity of resources in the workflow into account. If edge nodes available are more vendor specific, then the platforms supporting the workflow will need to account for it. This is more complex than existing models that make the cloud accessible.

Platforms that facilitate serverless computing will need to be responsive in executing functions or microservices with limited start up latencies. Current delays in invoking functions are due to creating containers for each execution of a function. Although containers are faster than VMs existing container technology cannot be the unit of deployment. Alternate lean environments will be required to be integrated with platforms that facilitate Function-as-a-Service.

Libraries and Algorithms: unlike large servers distributed cloud architectures will not support heavyweight software due to hardware constraints. For example, if a small cell base station with a 4-core ARM-based CPU and limited memory is employed as an edge node in a distributed cloud model, then there is limited resources for executing complex data processing tools such as Apache Spark[38] that requires at least 8 cores CPU and 8 gigabyte memory for good performance. Here lightweight algorithms that can do reasonable machine learning or data processing tasks are required [95, 96]. Apache Quarks[39], for example, is a lightweight library that can be employed on small devices such as smart phones to enable real-time data analytics. However, Quarks supports basic data processing, such as filtering and windowed aggregates, which are not sufficient for advanced analytical tasks

[38] `http://spark.apache.org`
[39] `http://quarks.incubator.apache.org`

(e.g. context-aware recommendations). Machine learning libraries that consume less memory would benefit data analytics for edge nodes.

Current research is mostly targeted at developing platforms, libraries and languages for individual requirements of the emerging computing architectures. For example, individual software platforms are available for serverless computing or IoT. However, there are a number of common requirements for these emerging architectures and are not a design factor when developing platforms. Efforts towards developing a unified environment that can address the common requirements of emerging architectures to achieve interoperable and application independent environments will need to be a direction for research. Such unified environments can then be extended to suit individual requirements. We recommend the design and development of self-managing applications as a way forward for realising this [97].

5.3. Developing a Marketplace for Emerging Distributed Architectures

The public cloud marketplace is competitive and taking a variety of CPU, storage and communication metrics into account for billing [98, 99]. For example, Amazon's pricing of a VM is based on the number of virtual CPUs and memory allocated to the VM. Distributed cloud architectures will require the development of a similar yet a more complex marketplace and remains an open issue. This will need to be developed with industry-academic collaborations (for example, the Open Fog consortium is set up with industry and academic partners to achieve open standards in the fog computing architecture). The marketplace will need to take ownership, pricing models and customers into account.

Typically, public cloud data centers are owned by large businesses. If traffic routing nodes were to be used as edge nodes in distributed cloud architectures, then their ownership is likely to be telecommunication companies or governmental organisations that may have a global reach or are regional players (specific to the geographic location. For example, a local telecom operator). Distributed ownership will make it more challenging to obtain a unified marketplace operating on the same standards.

When distribution using the edge is considered, three possible levels of communication, which are between the user devices and the edge node, one edge node and another edge node, and an edge node and a cloud server, will need to be accounted in a pricing model. In addition, 'who pays what' towards the bill has to be articulated and a sustainable and transparent economic model will need to be derived. The priority of applications executing on these nodes will have to be considered. If a serverless computing model is developed, then monitoring tools at the fine-grain level of functions will need to be designed. These are open research areas.

Given that there are multiple levels of communication in emerging cloud architectures, there are potentially two customers. The first is an application owner running the service on the cloud who wants to improve the QoS for the application user. The second is the application user who could make use of a distributed architecture to improve the QoE when using a cloud service. For both the above, in addition to existing service agreements, there will be requirements to create agreements between the application owner, the nodes on to which an application is distributed and the user, which can be transparently monitored within the marketplace.

5.4. Offering Efficient Management Strategies in the Computing Ecosystem

On the cloud two key management tasks include (i) setting up agreements between parties involved and brokering services to optimise application performance, and (ii) benchmarking resources and monitoring services to ensure that high-level objectives are achieved. Typically, Service Level

Agreements (SLAs) are used to fulfil agreements between the provider and the user in the form of Service Level Agreements (SLAs) [49, 50]. This becomes complex in a multi-cloud environment [100, 101] and in distributed cloud environments (given that nodes closer to the user could also be made accessible through a marketplace). If a task were to be offloaded from a cloud server onto an edge node, for example, a mobile base station owned by a telecommunications company, then the cloud SLAs will need to take into account agreements with a third-party. Moreover, the implications to the user will need to be articulated. The legalities of SLAs binding both the provider and the user in cloud computing are continuing to be articulated. Nevertheless, the inclusion of a third party offering services and the risk of computing on a third party node will need to be articulated. Moreover, if computations span across multiple edge nodes, then keeping track of resources becomes a more challenging task.

Performance is measured on the cloud using a variety of techniques, such as benchmarking to facilitate the selection of resources that maximise performance of an application and periodic monitoring of the resources to ensure whether user-defined service level objectives are achieved [102, 103, 104, 105]. Existing techniques are suitable in the cloud context since they monitor nodes that are solely used for executing the workloads [106, 107]. On edge nodes however, monitoring will be more challenging, given the limited hardware availability. Benchmarking and monitoring will need to take into account the primary service, such as routing traffic, that cannot be compromised. Communication between the edge node and user devices and the edge node and the cloud and potential communication between different edge nodes will need to be considered. Vertical scaling along multiple hierarchical levels and heterogeneous devices will need to be considered. These may not be important considerations on the cloud, but becomes significantly important in the context of fog computing. The SLAs that are defined in future distributed clouds will need to implicitly account for security [108].

5.5. Ensuring Reliability of Cloud Systems

Reliability of the cloud continues to remain a concern while adopting the cloud for remote computing and storage. Cloud failures have been reported affecting a number of popular services, such as DropBox and Netflix[40,41,42]. It is also reported that a 49-minute outage suffered by Amazon.com in 2013 cost the company more than $4 million in lost sales[43]. As unplanned outages are inevitable, losses from outages will continue to escalate with the rapid growth of e-commerce businesses. Reliability becomes more challenging as the infrastructure becomes distributed. Recently, efforts are being made to design more reliable cloud data centers and services.

On the infrastructure level, to deal with hardware failures due to targeted attacks and natural disasters, VMs and data are rigorously replicated in multiple geographic locations [109, 110]. Proactive and reactive strategies so as to back up VMs taking network bandwidth and associated metrics are now inherent to designing cloud data centers. FailSafe is a Microsoft initiative for delivering disaster resilient cloud architectures which can be made use by a cloud application. However, incorporating resilient computing into distributed cloud applications remains challenging, still requires significant programming efforts and is an open area of research[44] [111]. Notwithstanding,

[40] https://www.techflier.com/2016/01/25/top-20-high-profile-cloud-failures-all-time/

[41] http://www.nytimes.com/2012/12/27/technology/latest-netflix-disruption-highlights-challenges-of-cloud-computing.html?_r=1

[42] http://spectrum.ieee.org/computing/networks/understanding-cloud-failures

[43] http://tinyurl.com/jjkn235

[44] https://docs.microsoft.com/en-us/azure/guidance/guidance-resiliency-overview

disaster recovery is an expensive operation, and is required as a service to minimise recovery time and costs after a failure has occurred [112]. Multi-cloud and multi-region architectures that scale both horizontally (geographically distributed) and vertically (not only in cloud data centers, but throughout the network) are recommended to avoid single points of failure [113].

5.6. Building Sustainable Infrastructure for the Future

In 2014, it was reported that the US data centers consumed about 70 billion kilowatt-hours of electricity. This is approximately 2% of the total energy consumed in the US[45]. Data centers are huge investments which have adverse environmental impact due to large carbon footprints. While it may not be possible to eliminate data centers from the computing ecosystem, innovative and novel system architectures that can geographically distribute data center computing are required for sustainability.

Useful contributions in this space can be achieved by developing algorithms that rely on geographically distributed data coordination, resource provisioning and carbon footprint-aware and energy-aware provisioning in data centers [114, 115, 116, 117]. These will in turn minimise energy consumption of the data center and maximise the use of green energy while meeting an application's QoS expectations. Incorporating energy efficiency as a QoS metric has been recently suggested [118]. This risks the violation of SLAs since VM management policies will become more rigorous aiming to optimise energy efficiency. However, there is a trade off between performance of the cloud resource and energy efficiency. This clearly is an open avenue for research. Intra and inter networking plays a key role in setting up efficient data centers. Virtualising network functions through software defined networking is an upcoming area to manage key services offered by the network. However, energy consumption is not a key metric that is considered in current implementations. An open area is the understanding of the trade off between energy consumption and network functions. Addressing this will provide insights into developing cloud infrastructure that are becoming more distributed.

Algorithms for application-aware management of power states of computing servers can be incorporated towards achieving more sustainable solutions in the long run. Moreover, methods that incorporate resilience in the event of outages and failures will be required.

Current Cloud systems primarily focus on consolidation of VMs to minimise energy consumption of servers. However, cooling systems (approximately 35% of energy) and networks consume significant energy. Emerging techniques will need to be developed that manage energy efficiency of servers, networks and cooling systems. These techniques can leverage the interplay between IoT-enabled cooling systems and data center managers that dynamically make decisions on which resources to switch on/off in both time and space dimensions based on workload forecasts.

6. Summary

So what does cloud computing in the next decade look like? The general trend seems to be towards making use of infrastructure from multiple providers and decentralising computing away from resources currently concentrated in data centers. This is in contrast to traditional cloud offerings from single providers. Consequently, new computing models to suit the demands of the market are emerging.

[45] `http://www.datacenterknowledge.com/archives/2016/06/27/heres-how-much-energy-all-us-data-centers-consume/`

In this paper, we considered computing models that are based on voluntarily providing resources to create ad hoc clouds and harnessing computing at the edge of the network both for mobile and online applications. A computing model which will replace the traditional notion of paying for a cloud VM even when a server executing on the VM is idle was presented. The concept of integrating resilience and software-defined into distributed cloud computing is another emerging computing model that was highlighted in this paper.

Both the changing cloud infrastructure and emerging computing architecture will impact a number of areas. They will play a vital role in improving connectivity between people and devices to facilitate the Internet-of-Things paradigm. The area of data intensive computing will find novel techniques to address challenges related to dealing with volume of data. New services, such as containers, acceleration and function, is anticipated become popular. A number of research areas will find convergence with next generation cloud systems to deliver self-learning systems.

These changes are being led both by the industry and academia, but there are a number of challenges that will need to be addressed in the future. In this paper we considered directions in enhancing security, expressing applications, managing efficiently and developing sustainable systems for next generation cloud computing.

References

[1] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, I. Brandic, Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing As the 5th Utility, Future Generation Computer Systems 25 (6) (2009) 599–616.

[2] N. Grozev, R. Buyya, Inter-Cloud Architectures and Application Brokering: Taxonomy and Survey, Software: Practice and Experience 44 (3) (2014) 369–390.

[3] D. Petcu, G. Macariu, S. Panica, C. Crciun, Portable Cloud Applications: From Theory to Practice, Future Generation Computer Systems 29 (6) (2013) 1417–1430.

[4] Z. Wu, H. V. Madhyastha, Understanding the Latency Benefits of Multi-cloud Webservice Deployments, SIGCOMM Computer Communications Review 43 (2) (2013) 13–20.

[5] D. Bernstein, E. Ludvigson, K. Sankar, S. Diamond, M. Morrow, Blueprint for the Intercloud - Protocols and Formats for Cloud Computing Interoperability, in: Proceedings of the 4th International Conference on Internet and Web Applications and Services, 2009, pp. 328–336.

[6] K. Zhang, X. Zhou, Y. Chen, X. Wang, Y. Ruan, Sedic: Privacy-aware Data Intensive Computing on Hybrid Clouds, in: Proceedings of the 18th ACM Conference on Computer and Communications Security, 2011, pp. 515–526.

[7] X. Xu, X. Zhao, A Framework for Privacy-Aware Computing on Hybrid Clouds with Mixed-Sensitivity Data, in: 17th IEEE International Conference on High Performance Computing and Communications, 7th IEEE International Symposium on Cyberspace Safety and Security, and 12th IEEE International Conference on Embedded Software and Systems, 2015, pp. 1344–1349.

[8] A. N. Toosi, R. O. Sinnott, R. Buyya, Resource Provisioning for Data-intensive Applications with Deadline Constraints on Hybrid Clouds Using Aneka, Future Generation Computer Systems.

[9] B. Rochwerger, C. Vzquez, D. Breitgand, D. Hadas, M. Villari, P. Massonet, E. Levy, A. Galis, I. M. Llorente, R. S. Montero, Y. Wolfsthal, K. Nagin, L. Larsson, F. Galn, An Architecture for Federated Cloud Computing, John Wiley & Sons, Inc., 2011, pp. 391–411.

[10] R. Buyya, R. Ranjan, R. N. Calheiros, InterCloud: Utility-oriented Federation of Cloud Computing Environments for Scaling of Application Services, in: Proceedings of the 10th International Conference on Algorithms and Architectures for Parallel Processing, 2010, pp. 13–31.

[11] B. Varghese, N. Wang, S. Barbhuiya, P. Kilpatrick, D. S. Nikolopoulos, Challenges and Opportunities in Edge Computing, in: IEEE International Conference on Smart Cloud, 2016, pp. 20–26.

[12] F. P. Tso, D. R. White, S. Jouet, J. Singer, D. P. Pezaros, The Glasgow Raspberry Pi Cloud: A Scale Model for Cloud Computing Infrastructures, in: 33rd IEEE International Conference on Distributed Computing Systems Workshops, 2013, pp. 108–112.

[13] A. Sathiaseelan, A. Lertsinsrubtavee, A. Jagan, P. Baskaran, J. Crowcroft, Cloudrone: Micro Clouds in the Sky, in: Proceedings of the 2nd Workshop on Micro Aerial Vehicle Networks, Systems, and Applications for Civilian Use, 2016, pp. 41–44.

[14] Y. S. S. A. Elkhatib, B. F. Porter, H. B. Ribeiro, M. F. Zhani, J. Qadir, E. Rivire, On Using Micro-Clouds to Deliver the Fog, IEEE Internet Computing.

[15] M. Villari, M. Fazio, S. Dustdar, O. Rana, R. Ranjan, Osmotic Computing: A New Paradigm for Edge/Cloud Integration, IEEE Cloud Computing 3 (6) (2016) 76–83.

[16] K. Gai, M. Qiu, H. Zhao, L. Tao, Z. Zong, Dynamic Energy-aware Cloudlet-based Mobile Cloud Computing Model for Green Computing, Journal of Network and Computer Applications 59 (C) (2016) 46–54.

[17] M. Satyanarayanan, P. Bahl, R. Caceres, N. Davies, The Case for VM-Based Cloudlets in Mobile Computing, IEEE Pervasive Computing 8 (4) (2009) 14–23.

[18] G. A. McGilvary, A. Barker, M. Atkinson, Ad Hoc Cloud Computing, in: Proceedings of the IEEE 8th International Conference on Cloud Computing, 2015, pp. 1063–1068.

[19] M. Hamdaqa, M. M. Sabri, A. Singh, L. Tahvildari, Adoop: MapReduce for Ad-hoc Cloud Computing, in: Proceedings of the 25th Annual International Conference on Computer Science and Software Engineering, 2015, pp. 26–34.

[20] I. Yaqoob, E. Ahmed, A. Gani, S. Mokhtar, M. Imran, S. Guizani, Mobile Ad Hoc Cloud: A Survey, Wireless Communications & Mobile Computing 16 (16) (2016) 2572–2589.

[21] D. Huang, X. Zhang, M. Kang, J. Luo, MobiCloud: Building Secure Cloud Framework for Mobile Computing and Communication, in: Proceedings of the 5th IEEE International Symposium on Service Oriented System Engineering, 2010, pp. 27–34.

[22] G. Huerta-Canepa, D. Lee, A Virtual Cloud Computing Provider for Mobile Devices, in: Proceedings of the 1st ACM Workshop on Mobile Cloud Computing & Services: Social Networks and Beyond, 2010, pp. 6:1–6:5.

[23] H. Falaki, D. Lymberopoulos, R. Mahajan, S. Kandula, D. Estrin, A First Look at Traffic on Smartphones, in: Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement, 2010, pp. 281–287.

[24] H. Falaki, R. Mahajan, S. Kandula, D. Lymberopoulos, R. Govindan, D. Estrin, Diversity in Smartphone Usage, in: Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services, 2010, pp. 179–194.

[25] S. P. Crago, J. P. Walters, Heterogeneous Cloud Computing: The Way Forward, Computer 48 (1) (2015) 59–61.

[26] J. Auerbach, D. F. Bacon, I. Burcea, P. Cheng, S. J. Fink, R. Rabbah, S. Shukla, A Compiler and Runtime for Heterogeneous Computing, in: Proceedings of the 49th Annual Design Automation Conference, 2012, pp. 271–276.

[27] C. J. Rossbach, Y. Yu, J. Currey, J.-P. Martin, D. Fetterly, Dandelion: A Compiler and Runtime for Heterogeneous Systems, in: Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles, 2013, pp. 49–68.

[28] K. J. Brown, A. K. Sujeeth, H. J. Lee, T. Rompf, H. Chafi, M. Odersky, K. Olukotun, A Heterogeneous Parallel Framework for Domain-Specific Languages, in: Proceedings of the 2011 International Conference on Parallel Architectures and Compilation Techniques, 2011, pp. 89–100.

[29] P. Harvey, K. Bakanov, I. Spence, D. Nikolopoulos, A Scalable Runtime for the ECOSCALE Heterogeneous Exascale Hardware Platform, 2016.

[30] A. Marosi, J. Kovács, P. Kacsuk, Towards a Volunteer Cloud System, Future Generation Computer Systems 29 (6) (2013) 1442 – 1451.

[31] V. D. Cunsolo, S. Distefano, A. Puliafito, M. Scarpa, Volunteer Computing and Desktop Cloud: The Cloud@Home Paradigm, in: 8th IEEE International Symposium on Network Computing and Applications, 2009, pp. 134–139.

[32] F. Costa, L. Silva, M. Dahlin, Volunteer Cloud Computing: MapReduce over the Internet, in: IEEE International Symposium on Parallel and Distributed Processing Workshops, 2011, pp. 1855–1862.

[33] S. Caton, K. Bubendorfer, K. Chard, O. F. Rana, Social Cloud Computing: A Vision for Socially Motivated Resource Sharing, IEEE Transactions on Services Computing 5 (2012) 551–563.

[34] K. Chard, S. Caton, O. Rana, K. Bubendorfer, Social Cloud: Cloud Computing in Social Networks, in: 3rd IEEE International Conference on Cloud Computing, 2010, pp. 99–106.

[35] S. distefano, A. Puliafito, Cloud@Home: Toward a Volunteer Cloud, IT Professional 14 (1) (2012) 27–31.

[36] A. Shahri, M. Hosseini, R. Ali, F. Dalpiaz, Gamification for Volunteer Cloud Computing, in: Proceedings of the 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing, 2014, pp. 616–617.

[37] R. Ranjan, L. Zhao, Peer-to-Peer Service Provisioning in Cloud Computing Environments, The Journal of Supercomputing 65 (1) (2013) 154–184.

[38] B. Zhou, A. V. Dastjerdi, R. Calheiros, S. Srirama, R. Buyya, mCloud: A Context-aware Offloading Framework for Heterogeneous Mobile Cloud, IEEE Transactions on Services Computing PP (99) (2016) 1–1.

[39] S. Deng, L. Huang, J. Taheri, A. Y. Zomaya, Computation Offloading for Service Workflow in Mobile Cloud Computing, IEEE Transactions on Parallel & Distributed Systems 26 (12) (2015) 3317–3329.

[40] S. Sardellitti, G. Scutari, S. Barbarossa, Joint Optimisation of Radio and Computational Resources for Multicell Mobile-Edge Computing, IEEE Transactions on Signal and Information Processing over Networks 1 (2) (2015) 89–103.

[41] K. Bhardwaj, P. Agrawal, A. Gavrilovska, K. Schwan, AppSachet: Distributed App Delivery from the Edge Cloud, in: 7th International Conference Mobile Computing, Applications, and Services, 2015, pp. 89–106.

[42] P. Simoens, Y. Xiao, P. Pillai, Z. Chen, K. Ha, M. Satyanarayanan, Scalable Crowd-Sourcing of Video from Mobile Devices, Tech. Rep. CMU-CS-12-147, School of Computer Science, Carnegie Mellon University (December 2012).
URL `http://elijah.cs.cmu.edu/DOCS/CMU-CS-12-147.pdf`

[43] O. Incel, A. Ghosh, B. Krishnamachari, K. Chintalapudi, Fast Data Collection in Tree-Based Wireless Sensor Networks, IEEE Transactions on Mobile Computing 11 (1) (2012) 86–99.

[44] H. A. B. F. de Oliveira, H. S. Ramos, A. Boukerche, L. A. Villas, R. B. de Araujo, A. A. F. Loureiro, DRINA: A Lightweight and Reliable Routing Approach for In-Network Aggregation in Wireless Sensor Networks, IEEE Transactions on Computers 62 (2013) 676–689.

[45] S. Wang, R. Urgaonkar, M. Zafer, T. He, K. Chan, K. K. Leung, Dynamic Service Migration in Mobile Edge-Clouds, in: IFIP Networking Conference, 2015, pp. 1–9.

[46] K. Habak, M. Ammar, K. A. Harras, E. Zegura, Femto clouds: Leveraging mobile devices to provide cloud service at the edge, in: Proceedings of the 8th IEEE International Conference on Cloud Computing, 2015, pp. 9–16.

[47] G. Orsini, D. Bade, W. Lamersdorf, Computing at the Mobile Edge: Designing Elastic Android Applications for Computation Offloading, in: 8th IFIP Wireless and Mobile Networking Conference, 2015, pp. 112–119.

[48] X. Chen, L. Jiao, W. Li, X. Fu, Efficient Multi-User Computation Offloading for Mobile-Edge Cloud Computing, IEEE/ACM Transactions on Networking 24 (5) (2016) 2795–2808.

[49] S. A. Baset, Cloud SLAs: Present and Future, ACM SIGOPS Operating Systems Review 46 (2) (2012) 57–66.

[50] R. Buyya, S. K. Garg, R. N. Calheiros, SLA-oriented Resource Provisioning for Cloud Computing: Challenges, Architecture, and Solutions, in: Proceedings of the International Conference on Cloud and Service Computing, 2011, pp. 1–10.

[51] I. Stojmenovic, S. Wen, X. Huang, H. Luan, An overview of Fog Computing and its Security Issues, Concurrency and Computation: Practice and Experience 28 (10) (2016) 2991–3005.

[52] Y. Wang, T. Uehara, R. Sasaki, Fog Computing: Issues and Challenges in Security and Forensics, in: IEEE 39th Annual Computer Software and Applications Conference, Vol. 3, 2015, pp. 53–59.

[53] D. Kreutz, F. M. V. Ramos, P. E. Verssimo, C. E. Rothenberg, S. Azodolmolky, S. Uhlig, Software-Defined Networking: A Comprehensive Survey, Proceedings of the IEEE 103 (1) (2015) 14–76.

[54] B. A. A. Nunes, M. Mendonca, X. N. Nguyen, K. Obraczka, T. Turletti, A Survey of Software-Defined Networking: Past, Present, and Future of Programmable Networks, IEEE Communications Surveys Tutorials 16 (3) (2014) 1617–1634.

[55] A. Hakiri, A. Gokhale, P. Berthou, D. C. Schmidt, T. Gayraud, Software-Defined Networking: Challenges and research opportunities for Future Internet, Computer Networks 75, Part A (2014) 453–471.

[56] Z. Zhang, B. Bockelman, D. W. Carder, T. Tannenbaum, Lark: An effective approach for software-defined networking in high throughput computing clusters, Future Generation Computer Systems 72 (2017) 105 – 117.

[57] Y. Jararweh, M. Al-Ayyoub, A. Darabseh, E. Benkhelifa, M. Vouk, A. Rindos, Software Defined Cloud, Future Generation Computer Systems 58 (C) (2016) 56–74.

[58] R. Buyya, R. N. Calheiros, J. Son, A. V. Dastjerdi, Y. Yoon, Software-Defined Cloud Computing: Architectural Elements and Open Challenges, in: International Conference on Advances in Computing, Communications and Informatics, 2014, pp. 1–12.

[59] F. Mattern, C. Floerkemeier, From Active Data Management to Event-based Systems and More, 2010, Ch. From the Internet of Computers to the Internet of Things, pp. 242–259.

[60] J. Gubbi, R. Buyya, S. Marusic, M. Palaniswami, Internet of things (iot): A vision, architectural elements, and future directions, Future Generation Computer Systems 29 (7) (2013) 1645–1660.

[61] D. He, S. Chan, M. Guizani, Privacy and Incentive Mechanisms in People-centric Sensing Networks, IEEE Communications Magazine 53 (10) (2015) 200–206.

[62] F. Delmastro, V. Arnaboldi, M. Conti, People-centric Computing and Communications in Smart Cities, IEEE Communications Magazine 54 (7) (2016) 122–128.

[63] J. P. J. Peixoto, D. G. Costa, Wireless Visual Sensor Networks for smart City Applications: A Relevance-based Approach for Multiple Sinks Mobility, Future Generation Computer Systems 76 (2017) 51 – 62.

[64] M. D. Assuno, R. N. Calheiros, S. Bianchi, M. A. Netto, R. Buyya, Big Data computing and clouds: Trends and future directions, Journal of Parallel and Distributed Computing 7980 (2015) 3–15.

[65] J. Horey, E. Begoli, R. Gunasekaran, S.-H. Lim, J. Nutaro, Big Data Platforms As a Service: Challenges and Approach, in: Proceedings of the 4th USENIX Conference on Hot Topics in Cloud Computing, 2012.

[66] J. Liu, J. Shang, C. Wang, X. Ren, J. Han, Mining Quality Phrases from Massive Text Corpora, in: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, 2015, pp. 1729–1744.

[67] C. Wang, Y. Song, H. Li, M. Zhang, J. Han, Text Classification with Heterogeneous Information Network Kernels, in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 2016, pp. 2130–2136.

[68] C. Nguyen, P. J. Rhodes, Accelerating Range Queries for Large-scale Unstructured Meshes, in: IEEE International Conference on Big Data, 2016, pp. 502–511.

[69] C. Nabti, H. Seba, Querying Massive Graph Data: A Compress and Search Approach, Future Generation Computer Systems 74 (2017) 63 – 75.

[70] S. Q. Ren, B. H. M. Tan, S. Sundaram, T. Wang, Y. Ng, V. Chang, K. M. M. Aung, Secure Searching on Cloud Storage Enhanced by Homomorphic Indexing, Future Generation Computer Systems 65 (2016) 102 – 110.

[71] J. Han, On the Power of Big Data: Mining Structures from Massive, Unstructured Text Data, in: IEEE International Conference on Big Data, 2016.

[72] S. Ma, J. Li, C. Hu, X. Lin, J. Huai, Big Graph Search: Challenges and Techniques, Frontiers of Computer Science 10 (3) (2016) 387–398.

[73] B. Varghese, J. Prades, C. Reao, F. Silla, Acceleration-as-a-Service: Exploiting Virtualised GPUs for a Financial Application, in: 11th IEEE International Conference on e-Science, 2015, pp. 47–56.

[74] A. J. Peña, C. Reaño, F. Silla, R. Mayo, E. S. Quintana-Ortí, J. Duato, A Complete and Efficient CUDA-sharing Solution for HPC Clusters, Parallel Computing 40 (10) (2014) 574–588.

[75] J. Prades, B. Varghese, C. Reao, F. Silla, Multi-tenant Virtual GPUs for Optimising Performance of a Financial Risk Application, Journal of Parallel and Distributed Computing (2016) –.

[76] C. Pahl, Containerisation and the PaaS Cloud, IEEE Cloud Computing 2 (3) (2015) 24–31.

[77] B. Arnold, S. A. Baset, P. Dettori, M. Kalantar, I. I. Mohomed, S. J. Nadgowda, M. Sabath, S. R. Seelam, M. Steinder, M. Spreitzer, A. S. Youssef, Building the IBM Containers cloud service, IBM Journal of Research and Development 60 (2-3) (2016) 9:1–9:12.

[78] U. Deshpande, K. Keahey, Traffic-sensitive Live Migration of Virtual Machines, Future Generation Computer Systems 72 (2017) 118 – 128.

[79] S. Rendle, D. Fetterly, E. J. Shekita, B.-y. Su, Robust Large-Scale Machine Learning in the Cloud, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1125–1134.

[80] J. Dean, G. S. Corrado, R. Monga, K. Chen, M. Devin, Q. V. Le, M. Z. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, A. Y. Ng, Large Scale Distributed Deep Networks, in: Proceedings of the 25th International Conference on Neural Information Processing Systems, 2012, pp. 1223–1231.

[81] A. Coates, B. Huval, T. Wang, D. Wu, B. Catanzaro, N. Andrew, Deep learning with COTS HPC systems, in: Proceedings of the 30th International Conference on Machine Learning, Vol. 28, 2013, pp. 1337–1345.

[82] H. Cui, H. Zhang, G. R. Ganger, P. B. Gibbons, E. P. Xing, GeePS: Scalable Deep Learning on Distributed GPUs with a GPU-specialized Parameter Server, in: Proceedings of the Eleventh European Conference on Computer Systems, 2016, pp. 4:1–4:16.

[83] R. Raina, A. Madhavan, A. Y. Ng, Large-scale deep unsupervised learning using graphics processors, in: Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 873–880.

[84] P. Merolla, J. Arthur, R. Alvarez-Icaza, A. Cassidy, J. Sawada, F. Akopyan, B. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. Esser, R. Appuswamy, B. Taba, A. Amir, M. Flickner, W. Risk, R. Manohar, D. Modha, A Million Spiking-Neuron Integrated Circuit With a Scalable Communication Network and Interface, Science (2014) 668–673.

[85] A. D. Brown, S. B. Furber, J. S. Reeve, J. D. Garside, K. J. Dugan, L. A. Plana, S. Temple, SpiNNaker - Programming Model, IEEE Transactions on Computers 64 (6) (2015) 1769–1782.

[86] K. Hashizume, D. G. Rosado, E. Fernandez-Medina, E. B. Fernandez, An Analysis of Security Issues for Cloud Computing, Journal of Internet Services and Applications 4 (1) (2013) 5.

[87] N. Gonzalez, C. Miers, F. Redígolo, M. Simplício, T. Carvalho, M. Näslund, M. Pourzandi, A Quantitative Analysis of Current Security Concerns and Solutions for Cloud Computing, Journal of Cloud Computing: Advances, Systems and Applications 1 (1) (2012) 11.

[88] B. K. Chejerla, S. K. Madria, Qos guaranteeing robust scheduling in attack resilient cloud integrated cyber physical system, Future Generation Computer Systems 75 (2017) 145 – 157.

[89] I. Stojmenovic, S. Wen, X. Huang, H. Luan, An overview of Fog Computing and its Security Issues, Concurrency and Computation: Practice and Experience 28 (10) (2016) 2991–3005.

[90] Y. Wang, T. Uehara, R. Sasaki, Fog Computing: Issues and Challenges in Security and Forensics, in: Computer Software and Applications Conference (COMPSAC), 2015 IEEE 39th Annual, Vol. 3, 2015, pp. 53–59.

[91] X. Wang, W. Wang, Y. He, J. Liu, Z. Han, X. Zhang, Characterizing Android Apps Behavior for Effective Detection of Malapps at Large Scale, Future Generation Computer Systems 75 (2017) 30 – 45.

[92] J. Lopez, R. Rios, F. Bao, G. Wang, Evolving privacy: From sensors to the internet of things, Future Generation Computer Systems 75 (2017) 46 – 57.

[93] O. Osanaiye, K.-K. R. Choo, M. Dlodlo, Distributed Denial of Service (DDoS) Resilience in Cloud, Journal of Network and Computer Applications 67 (C) (2016) 147–165.

[94] S. Yu, Y. Tian, S. Guo, D. O. Wu, Can We Beat DDoS Attacks in Clouds?, IEEE Transactions on Parallel and Distributed Systems 25 (9) (2014) 2245–2254.

[95] S. Kartakis, J. A. McCann, Real-time Edge Analytics for Cyber Physical Systems Using Compression Rates, in: Proceedings of the International Conference on Autonomic Computing, 2014, pp. 153–159.

[96] I. Santos, M. Tilly, B. Chandramouli, J. Goldstein, DiAl: Distributed Streaming Analytics Anywhere, Anytime, Proceedings of the VLDB Endowment 6 (12) (2013) 1386–1389.

[97] G. Toffetti, S. Brunner, M. Blchlinger, J. Spillner, T. M. Bohnert, Self-managing Cloud-native Applications: Design, Implementation, and Experience, Future Generation Computer Systems 72 (2017) 165 – 179.

[98] B. Sharma, R. K. Thulasiram, P. Thulasiraman, S. K. Garg, R. Buyya, Pricing Cloud Compute Commodities: A Novel Financial Economic Model, in: Proceedings of the 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, 2012, pp. 451–457.

[99] H. Xu, B. Li, A Study of Pricing for Cloud Resources, SIGMETRICS Performance Evaluation Review 40 (4) (2013) 3–12.

[100] A. J. Ferrer, F. HernáNdez, J. Tordsson, E. Elmroth, A. Ali-Eldin, C. Zsigri, R. Sirvent, J. Guitart, R. M. Badia, K. Djemame, W. Ziegler, T. Dimitrakos, S. K. Nair, G. Kousiouris, K. Konstanteli, T. Varvarigou, B. Hudzia, A. Kipp, S. Wesner, M. Corrales, N. Forgó, T. Sharif, C. Sheridan, OPTIMIS: A Holistic Approach to Cloud Service Provisioning, Future Generation Computer Systems 28 (1) (2012) 66–77.

[101] A. Barker, B. Varghese, L. Thai, Cloud Services Brokerage : A Survey and Research Roadmap, in: Proceedings of the 8th IEEE International Conference on Cloud Computing, 2015, pp. 1029–1032.

[102] B. Varghese, O. Akgun, I. Miguel, L. Thai, A. Barker, Cloud Benchmarking for Performance, in: Proceedings of the IEEE International Conference on Cloud Computing Technology and Science, 2014, pp. 535–540.

[103] W. Lloyd, S. Pallickara, O. David, M. Arabi, T. Wible, J. Ditty, Demystifying the Clouds: Harnessing Resource Utilization Models for Cost Effective Infrastructure Alternatives, IEEE Transactions on Cloud Computing PP (99) (2017) 14.

[104] B. F. Cooper, A. Silberstein, E. Tam, R. Ramakrishnan, R. Sears, Benchmarking Cloud Serving Systems with YCSB, in: Proceedings of the ACM Symposium on Cloud Computing, 2010, pp. 143–154.

[105] B. Varghese, O. Akgun, I. Miguel, L. Thai, A. Barker, Cloud Benchmarking For Maximising Performance of Scientific Applications, IEEE Transactions on Cloud Computing PP (99) (2016) 14.

[106] J. Povedano-Molina, J. M. Lopez-Vega, J. M. Lopez-Soler, A. Corradi, L. Foschini, DARGOS: A Highly Adaptable and Scalable Monitoring Architecture for Multi-tenant Clouds, Future Generation Computer Systems 29 (8) (2013) 2041–2056.

[107] J. Montes, A. Sánchez, B. Memishi, M. S. Pérez, G. Antoniu, GMonE: A Complete Approach to Cloud Monitoring, Future Generation Computer Systems 29 (8) (2013) 2026–2040.

[108] R. Trapero, J. Modic, M. Stopar, A. Taha, N. Suri, A Novel Approach to Manage Cloud Security SLA Incidents, Future Generation Computer Systems 72 (2017) 193 – 205.

[109] S. Ferdousi, F. Dikbiyik, M. F. Habib, M. Tornatore, B. Mukherjee, Disaster-aware Datacenter Placement and Dynamic Content Management in Cloud Networks, IEEE/OSA Journal of Optical Communications and Networking 7 (7) (2015) 681–694.

[110] R. D. S. Couto, S. Secci, M. E. M. Campista, L. H. M. K. Costa, Network Design Requirements for Disaster Resilience in IaaS Clouds, IEEE Communications Magazine 52 (10) (2014) 52–58.

[111] T. A. Nguyen, D. S. Kim, J. S. Park, Availability modeling and analysis of a data center for disaster tolerance, Future Generation Computer Systems 56 (2016) 27 – 50.

[112] T. Wood, E. Cecchet, K. K. Ramakrishnan, P. Shenoy, J. van der Merwe, A. Venkataramani, Disaster Recovery As a Cloud Service: Economic Benefits & Deployment Challenges, in: Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing, 2010.

[113] X. Yuan, G. Min, L. T. Yang, Y. Ding, Q. Fang, A Game Theory-based Dynamic Resource Allocation Strategy in Geo-distributed Datacenter Clouds, Future Generation Computer Systems 76 (2017) 63 – 72.

[114] A. Hameed, A. Khoshkbarforoushha, R. Ranjan, P. P. Jayaraman, J. Kolodziej, P. Balaji, S. Zeadally, Q. M. Malluhi, N. Tziritas, A. Vishnu, S. U. Khan, A. Zomaya, A Survey and Taxonomy on Energy Efficient Resource Allocation Techniques for Cloud Computing Systems, Computing 98 (7) (2016) 751–774.

[115] C.-W. Lee, K.-Y. Hsieh, S.-Y. Hsieh, , H.-C. Hsiao, A Dynamic Data Placement Strategy for Hadoop in Heterogeneous Environments, Big Data Research 1 (C) (2014) 14–22.

[116] A. J. Younge, G. von Laszewski, L. Wang, S. Lopez-Alarcon, W. Carithers, Efficient Resource Management for Cloud Computing Environments, in: Proceedings of the International Conference on Green Computing, 2010, pp. 357–364.

[117] H. Duan, C. Chen, G. Min, Y. Wu, Energy-aware scheduling of virtual machines in heterogeneous cloud computing systems, Future Generation Computer Systems 74 (2017) 142 – 150.

[118] S. Singh, I. Chana, M. Singh, The Journey of QoS-Aware Autonomic Cloud Computing, IT Professional 19 (2) (2017) 42–49.