Cloud Futurology

Blesson Varghese*[1], Philipp Leitner[2], Suprio Ray[3], Kyle Chard[4], Adam Barker[5], Yehia Elkhatib[6], Herry Herry[7], Cheol-Ho Hong[8], Jeremy Singer[9], Fung Po Tso[10], Eiko Yoneki[11], and Mohamed-Faten Zhani[12]

[1]Queen's University Belfast, UK; b.varghese@qub.ac.uk
[2]Chalmers University of Technology, Sweden; philipp.leitner@chalmers.se
[3]University of New Brunswick, Canada; sray@unb.ca
[4]University of Chicago, USA; chard@uchicago.edu
[5]University of St Andrews, UK; adam.barker@st-andrews.ac.uk
[6]Lancaster University, UK; y.elkhatib@lancaster.ac.uk
[7]University of Glasgow, UK; h@herry.co
[8]Chung-Ang University, South Korea; cheolhohong@cau.ac.kr
[9]University of Glasgow, UK; jeremy.singer@glasgow.ac.uk
[10]Loughborough University, UK; p.tso@lboro.ac.uk
[11]University of Cambridge, UK; eiko.yoneki@cl.cam.ac.uk
[12]École de Technologie Supérieure, Canada; mohamed-faten.zhani@etsmtl.ca

The Cloud has become integral to most Internet-based applications and user gadgets. This article provides a brief history of the Cloud and presents a researcher's view of the prospects for innovating at the infrastructure, middleware, and application and delivery levels of the already crowded Cloud computing stack.

The global Cloud computing market exceeds $100 billion and research in this area has rapidly matured over the last decade. During this time many buzz words have come and gone. Consequently, traditional concepts and conventional definitions related to the Cloud are almost obsolete [1]. The current Cloud landscape looks very different from what may have been envisioned during its inception. It may seem that the area is saturated and there is little innovative research and development to be done in the Cloud, which naturally raises the question - *'What is the future of the Cloud?'*

This article first examines the multiple generations of innovation that the Cloud has undergone in the last decade, it then presents a researcher's view of the prospects and opportunities for innovation in this area across the entire Cloud stack - highlighting the infrastructure, middleware, and application and delivery levels.

*Corresponding Author

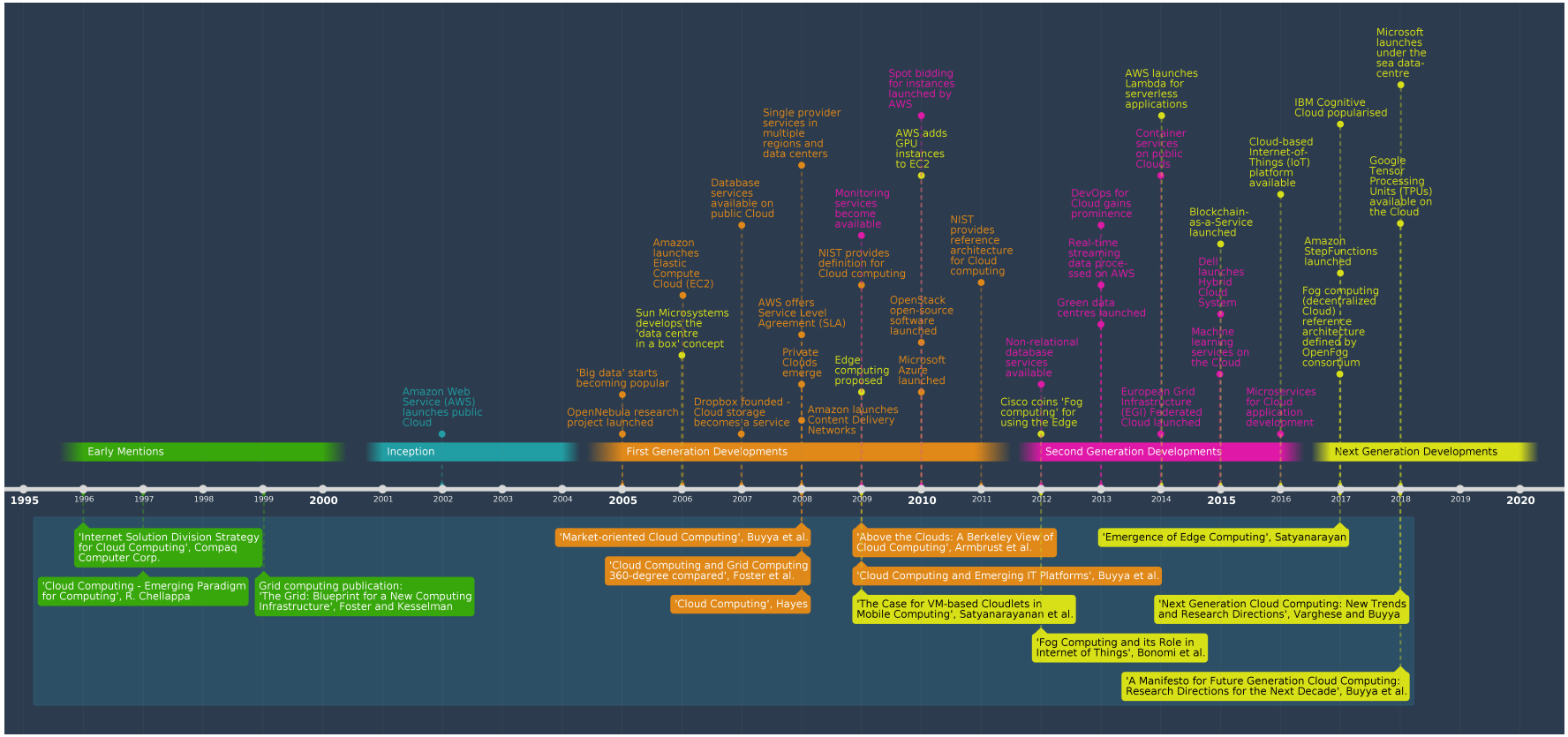

Figure 1: A timeline of the Cloud computing landscape. Early mentions of the Cloud in literature are described in the green block. The period of inception is highlighted in turquoise. This is followed by two generations of developments on the Cloud - the orange block highlights first generation and the purple colour presents second generation developments. Upcoming developments and trends are presented as next generation developments in the yellow block.

The Cloud Landscape

> The first known suggestion of Cloud-like computing was by Professor John McCarthy at MIT's centennial celebration in 1961 *"Computing may someday be organized as a public utility just as the telephone system is a public utility ... Each subscriber needs to pay only for the capacity he actually uses, but he has access to all programming languages characteristic of a very large system ... Certain subscribers might offer service to other subscribers ... The computer utility could become the basis of a new and important industry"*

Two fundamental technologies required for developing the Cloud, namely virtualization and networking, were first developed in the 60's. In 1967, IBM virtualized operating systems to allow multiple users to share the same computer and in 1969, the US Department of Defense launched the Advanced Research Projects Agency Network (ARPANET), which defined network protocols that led to the development of the Internet. Although the earliest mentions of Cloud computing in literature appear in the 90s as shown in Figure 1, it was Grid computing [2] that laid the foundation for offering computing resources as a service to users in the 90's and early 21st century. The inception of the Cloud as a utility service was realized when Amazon launched its commercial public Cloud in 2002.

Significant advances over the last decade can be divided into two generations as seen in Figure 1. The first generation focuses on development at the infrastructure level, for example creating data centres, which are centralized infrastructure that host significant processing and storage resources across different geographic regions. Across other layers of the Cloud stack a range of user-facing services emerged, some of which were available only in specific geographic regions. Software developed by OpenNebula (`http://www.opennebula.org`) and OpenStack (`https://www.openstack.org`) allowed organizations to own private Clouds and set up their own data centres.

As big data started to gain popularity in 2005, the Cloud was a natural first choice to tackle big data challenges. This led to the popularity of storage services such as Dropbox that relied on the Cloud. Relational databases hosted in the Cloud to support enterprise applications emerged.

Since inception in early 2000 and despite significant research and development efforts spanning over half a decade, a reference architecture for the Cloud was not defined until 2011 (`https://ws680.nist.gov/publication/get_pdf.cfm?pub_id=909505`). Since the Cloud was a new technology it may have taken a few years before definitions were articulated, circulated, and widely accepted.

Second generation developments focused on enriching the variety of services and their quality. In particular, management services and modular applications emerged. Monitoring services of compute, network and storage resources offering aggregate and fine metrics became available to application owners, allowing them to maximize performance. More flexible pricing strategies and service level agreements (SLAs), in addition to the posted price, pay-as-you-go model, such as spot bidding and preemptible virtual machine instances emerged in 2010.

Furthermore, the move from immutable virtual machines to smaller, loosely coupled execution units in the form of microservices and containers was a game changer for decomposing applications within and across different data centres. Combining public and private (on-premise) Clouds of different scale (a.k.a *cross-cloud or hybrid Cloud computing* [3]) gained prominence in order to alleviate concerns related to privacy and vendor lock-in.

An important step in the evolution of the Cloud was the development of Content Delivery Networks (CDNs). Compute and storage resources were geographically distributed for improving the overall quality of a variety of services, including streaming and caching. CDNs are the basis of upcoming trends in decentralizing Cloud resources towards the edge of the network, which will be considered in this article. Amazon launched their CDN in 2008.

Containers are namespaces with ring-fenced resources. For example, Docker [4] is a popular container technology for creating and managing self-sufficient execution units. Containers offer the prospect of seamless application development, testing, and delivery over heterogeneous environments. The microservice architectural style focuses on how the application logic is implemented rather than how it is hosted by dividing services into atomic functions in order to tame operational complexity. The emphasis is on developing small, replaceable service units instead of maintaining monolith services [5].

Innovation for the Next Generation

Tangential innovations in the first and second generation developments have made way for another generation of Cloud development that will focus on decentralization of resources. Although there has been a decade long explosion of Cloud research and development, there is significant innovation yet to come in the infrastructure, middleware, and application and delivery areas.

As shown in the 'Next Generation Development' block of Figure 1 in the next five years, computing as a utility will be miniaturized and available outside large data centres. Referred to as 'Cloud-in-a-Box', Sun Microsystems first demonstrated these ideas in 2006 and paved the way for Fog/Edge computing. The use of hardware accelerators, such as Graphics Processing Units (GPUs), in the Cloud began in 2010 is now leading to inclusion of even more specialized accelerators that are, for example, customized for modern machine learning or artificial intelligence workloads. Google provides Tensor Processing Units (TPUs) that are customized for such workloads with the aim to deliver new hardware and software stacks that extend machine learning and artificial intelligence capabilities both within and outside the cloud.

Infrastructure

A range of hardware accelerators, such as Graphics Processing Units (GPUs), Field-Programmable Gate Arrays (FPGAs), and more specialized Tensor Processing Units (TPUs) are now available for improving the performance of applications on the Cloud. Typically, these accelerators are not shared between applications and therefore result in an expensive data centre setup given the large amount of underutilized hardware.

Accelerator virtualization is the underlying technology that allows multiple applications to share the same hardware accelerator [6]. All existing virtualization solutions have performance limitations and are bespoke to each type of accelerator. Given that this is a relatively new area of study, we are yet to see a robust production-level solution that can easily incorporate and virtualize different types of accelerators with minimal overheads.

Currently, applications leverage Cloud resources from geographically distant data centres. Consequently, a user device, such as a smartphone must transfer data to a remote Cloud for processing data. However, this will simply not be possible when billions of devices are connected to the Internet, as frequent communication and communication latencies will affect the overall service quality and experience of the user. These latencies can be reduced by bringing compute resources to the network edge in a model often referred to as ***Fog/Edge computing*** [7, 8].

Edge computing is challenging - Edge resources need to be publicly and securely available. However, the risks, security concerns, vulnerabilities and pricing models are not articulated, or even fully known. Implementations of standardized Edge architectures and a unified marketplace is likely to emerge. Furthermore, there are no robust solutions to deploy an application across the Cloud and the Edge. Toolkits for deploying and managing applications on the Edge will materialize given the community led efforts by the European Telecommunications Standards Institute, the OpenFog consortium, and OpenStack.

Fog/Edge computing [7,8] refers to the use of resources at the Cloud to the network edge continuum. Data from user devices can be processed at the edge instead of remote data centers. Gartner and Forester define Edge computing as an enabling and strategic technology for realizing IoT and for building future Cloud applications. The Edge computing market is estimated to be worth US $6.72 billion within the next five years with a compound annual growth rate of over 35%.

The definition of Cloud data centres is also changing - what was conventionally 'dozens of data centres with millions of cores,' is now moving towards 'millions of data centres with dozens of cores.' These ***micro data centres*** are novelty architectures and miniatures of commercial cloud infrastructure (see Figure 2). Such deployments have become possible by networking several single board computers, such as the popular Raspberry Pi [9]. Consequently, the overall capital cost, physical footprint and power consumption is only a small fraction when compared to the commercial macro data centres [10].

Figure 2: A micro data centre comprising a cluster of Raspberry Pis that was assembled at the University of Glasgow, UK. These data centres in contrast to large Cloud data centres are low cost and low power consuming.

Micro data centres are a compelling infrastructure to support education, where students can be exposed to near-real data centre scenarios in a sandbox. They are widely used in educational contexts and more convincing use-cases will emerge. Global community engagement will facilitate the further adoption of micro data centres. International competitions and league tables for micro data centre deployment would catalyze this process.

The challenges to be addressed in making micro data centres operational include reducing overheads of the software infrastructure. For example, machine learning libraries must be optimized for use on micro data centre hardware. Additionally, management tools must run on devices that have only limited network access, perhaps due to the presence of firewalls, or intermittent connectivity as commonly found on edge networks. This is different from traditional data centres, which operate on the assumption that managed nodes are directly and permanently reachable.

Data centers are now one of the largest consumers of energy (`http://www.climatechangenews.com/2017/12/11/tsunami-data-consume-one-fifth-global-electricity-2025/`). This is in part due to the end of Moore's law and the fact that increasing processor speed no longer offsets energy consumption. It is also due to the rapid growth of Internet-of-Things (IoT) - with estimates of up to 80 billion devices online by 2025 - and increasing use in developing countries. As such, Cloud providers are facing both economic and legislative pressures to reduce energy consumption. It is unlikely that new power plants will be sufficient to meet these growing energy needs and thus there will be widespread use of renewable energy sources and 'stranded power' [11].

Middleware

Cloud customers are spoiled for choice, to the extent that it has become overwhelming for many. This is because of the incredible rate at which the Cloud resource and service portfolio has expanded. As such, there is now a need not only for middleware technologies that abstract differences between Clouds and services, but also for decision support systems to aid

customer deployment of applications. For example, to help guide selection of the best Cloud providers, resources and services, and configuration.

Cross-cloud challenges, such as identifying optimal deployment resources from across the vast array of options from different providers, seamlessly moving workloads between providers, and building systems that work equally well across the services of different providers will need to be surmounted for establishing a viable ***Cloud federation***.

Such systems, naturally, cannot be one-size-fits-all, but they must be tailored to the needs of the customer and to follow any changes in the Cloud provisioning market. ***Resource brokers*** are likely to become necessary to fill this void and provide a means of exploiting differences between Cloud providers, and identifying the real performance profiles of different Cloud services before matching them to the customer needs. Currently however, no practical brokering solutions are available.

Brokers can also be used to automatically configure the parameters of applications for a set of selected resources so as to maximize the performance of applications. Typically, the configuration parameters are manually tuned, which is cumbersome given the plethora of resources. A generic ***auto-tuner*** that operates in near real-time and measures performance cost-effectively is ideal. However, measuring performance in the Cloud is time consuming and therefore expensive [12,13]. In a transient environment, such as the Cloud, the performance metrics will be obsolete if they cannot be gathered in real time. The complexity increases when resources from multiple providers are considered. A starting point may be to develop bespoke auto-tuners based on domain specific knowledge that a system engineer may possess using a combination of machine learning techniques, such as deep neural networks and reinforcement learning [14].

While resource brokerage allows customers to select and combine services to their liking, Cloud data centre operators can also choose a variety of Network Functions (NFs) to create in-network services and safeguard networks to improve an application's performance. The process of orderly combining different combination or subsets of NFs, such as firewalls, network monitors, and load balancers, is referred to as ***service chaining*** [15].

Service chaining will be particularly useful for Edge computing to, for instance, improve data privacy. Recent proposals of intent-driven networking [16] allow operators and end-users to define *what* they want from the network, not *how*. This enables the on demand composition of bespoke network logic, allowing much more refined application control and dynamicity. A research challenge here is to manage services across Cloud and Edge networks and resources that have different ownership and operating objectives.

> An interesting starting point for implementing service chaining will be creating personalized network services across Cloud and Edge environments. This will provide, for example, a user with a personalized security profile that is adaptive to different environments, such as home or office. Network functions will need to be miniaturized for Edge resources to facilitate chaining.

Application and Delivery

The Cloud will continue to be an attractive proposition for ***big data applications*** to meet the volume, velocity, and variety challenges [17]. Apache Spark, Hadoop MapReduce and its variants have been extensively used to process volumes of data in the last five years. However, for many users these frameworks remain inaccessible due to their steep learning curve and lack of interactivity.

Further, while frameworks such as Storm and Kafka address some of the challenges associated with data velocity (e.g., as seen in real-time streaming required by social media, IoT and high frequency trading applications), scaling resources on the Cloud to meet strict response time requirements is still an active research area. This requires complex stream processing with low latency, scalability with self-load balancing capabilities and high availability. Many applications that stream data may have intermittent connectivity to Cloud back-end services for data processing and it would be impossible to process all data at the edge of the network. Thus, efficient stream processing frameworks that can replicate stream operators over multiple nodes and dynamically route stream data to increase potential paths for data ex-

change and communication will be desirable.

Innovation will be seen in taming traditional data challenges. For example, emergence of tools that alleviate the burden on the user, that support elasticity - dynamic provisioning and fair load balancing, and that cleverly move data. As data are increasingly large and distributed, the cost of moving data can now exceed the cost of processing it. Thus, there will be increasing interest in moving computation to data and constructing federated registries to manage data across the Cloud.

A new class of distributed data management systems, called NewSQL [18], are emerging. These systems aim to offer similar scalable performance to NoSQL while supporting Atomicity, Consistency, Isolation, and Durability (ACID) properties for transactional workloads that are typical of traditional relational databases. However, providing ACID guarantees across database instances that are distributed across multiple data centers is challenging. This is simply because it is not easy to keep data replicas consistent in multiple data locations. More mature approaches that will allow for data consistency will emerge to support future modularization of applications on the Cloud.

A variety of existing and upcoming applications, such as smart homes, autonomous trading algorithms and self-driving cars, are expected to generate and process vast amounts of time-series data. These applications make decisions based on inputs that change rapidly over time. Conventional database solutions are not designed for dealing with scale and easy-use of time-series data. Therefore, time-series databases are expected to become more common.

> Gartner predicts that by 2022 nearly 50% of enterprise-generated data will be processed outside the traditional Cloud data centre. Processing data outside the Cloud is the premise of Edge computing. It is likely that novel methods and tools that perform complex event processing across the Cloud and Edge will emerge.

A new class of applications are starting to emerge on the Cloud, namely ***serverless applications***. They are exemplified by the Function-as-a-Service (FaaS) Cloud, such as AWS Lambda or Azure Cloud Functions. FaaS Clouds are implemented on top of upcoming containerization technology, but provide convenient developer abstractions on top of, for instance, Docker. Contrary to traditional Cloud applications that are billed for the complete hour or the minute, serverless applications are billed by the millisecond [19]. FaaS has not yet seen widespread adoption for business-critical Cloud applications. This is because FaaS services and tooling are still immature and sometimes unreliable Furthermore, since FaaS Clouds rely on containers significant overheads are incurred for on-demand boot up. This may be problematic for an end-user facing use cases. FaaS Clouds have limited support for reuse, abstraction and modularization, which are usually taken for granted when using distributed programming models. Another practical challenge is that current-day FaaS services are entirely stateless. All application state needs to be handled in external storage services, such as Redis. FaaS providers are currently working towards stateful storage solutions, but it is as of yet unclear what these solutions will look like in practice. Consequently, significant developer effort is currently required to take advantage of FaaS services.

> FaaS Clouds have obvious advantages and we will witness innovation at the virtualization front either to reduce the overheads of containers or in the development of entirely new lightweight virtualization technologies. More powerful programming abstractions for composing and reusing Cloud functions will emerge.

Delivery of cloud services via economic models is transforming computing into a commodity, inline with McCarthy's vision of computing utility. These economic models, and benefits from economies of scale, have underpinned much of the success of the Cloud. The Cloud now utilizes a range of economic models, from posted price models, through to dynamic, spot markets for delivering infrastructure resources and a suite of higher-level platform and software services [20]. AWS even allows users to directly exchange reserved resources with one another via a reseller market. As the Cloud moves towards further decoupled resources (e.g., as seen in serverless computing) new economic models are needed to address granular resource bundles, short-duration leases, and

flexible markets for balancing supply and demand. Furthermore, granular, differentiated service levels are likely to become common, enabling greater user flexability with respect to price and service quality. This increasingly diverse range of economic models will further enable flexibility; however, it will require greater expertise to understand trade-offs and effectively participate in the market. Cloud automation tools will emerge to alleviate this burden by enabling users to directly quantify and manage inherent trade-offs such as cost, execution time, and solution accuracy.

> Current Cloud providers operate as independent silos, with little to no ability for users to move resources between providers. From a technology standpoint, creating federated Clouds via standardized abstractions is a solution, but general markets in which Cloud offerings can be compared and delivered are expected to appear soon. Users may be potentially offered many interchangeable alternatives, and therefore new economic models will need to cater for competition between providers rather than consumers.

Opportunities and Outlook

This article highlights a researcher's view of the prospects at the infrastructure, middleware, and application and delivery level of the Cloud computing landscape.

Although there are ongoing efforts to tackle research challenges at the infrastructure level in four avenues, namely accelerator virtualization, Fog/Edge computing, micro data centres, and power and energy-aware solutions, the timeline to mass adoption will vary. rCUDA (http://rcuda.net/) and gVirtus (https://github.com/RapidProjectH2020/GVirtuS) are exemplars of accelerator virtualization solutions that have undergone a decade of innovation, but are yet to become available on production level Cloud systems. With the increasing number of accelerators used in the Cloud it is foreseeable that accelerator virtualization technologies are adopted within the next decade. Similarly, there are limited Fog/Edge computing test-beds and real deployments of such systems are yet to be seen. While the mechanisms required to adopt Fog/Edge systems are currently unknown, the growing emphasis on 5G communication will accelerate interest in the area in the short term. Most micro data centres are still prototypes (`https://news.microsoft.com/features/under-the-sea-microsoft-tests-a-datacenter-thats-quick-to-deploy-could-provide-internet-connectivity-for-years/`) and will become widespread as more compelling use-cases, for example in Fog/Edge computing emerge over the next five years.

At the middleware level Cloud federation, resource brokers and auto-tuners, and service chaining were considered. Cloud federation in its strictest sense, i.e. reaching operational arrangement between independent providers, has for long been discussed but never in fact emerged on the horizon. Nevertheless, efforts into tools to bridge services of different providers are expected to continue. Current resource brokers and auto-tuners focus on the challenge of abstraction, but very few tackle this along with delegation, i.e., fully adaptive life cycle management. With increasing use of machine learning, the mechanisms of enabling such smart brokerage and auto-tuning are becoming available. The related challenges of expressing and interpreting customer requirements is also becoming a significant trend towards a vision where the customer needs to know far less about the infrastructure than current DevOps. In a similar vein, capturing and satisfying what an application needs from the network as seen in service chaining is a prominent research challenge. However, network operators are not as advanced in the pay-as-you-go model as cloud providers are. We expect this to change over the next five years, opening up a wide range of resources for supporting applications across an end-to-end network.

The four key areas identified for innovation at the applications and delivery level are complex stream processing, big data applications and databases, serverless applications and economic models. The rapid increase in connected devices, IoT, and streaming data is driving the immediate development of new real-time stream processing capabilities. In the next five years, we expect to see a variety of new stream processing techniques, including those designed to

leverage edge devices. Big data applications and databases have been the focus of recent research and innovation, and thus much immediate focus is on the adoption and use of these efforts. In the next 2-5 years, there will be increasing effort focused on time-series databases (following on from the release of Amazon Timestream and Azure Time Series Insights). Serverless computing is still in its infancy and is likely to be the most disruptive of the changes at the applications and delivery level. Over the next ten years, serverless technologies will likely permeate IT infrastructure, driving enhancements in terms of performance, flexibility (for example, support for more general application types), and economic models. Economic models will evolve over the next 2-5 years due to maturing serverless and decoupled computing infrastructure. The underlying models will account for energy policies and the desire for more flexible SLAs, and will leverage new opportunities for intercloud federations.

We recommend research focus in designing and developing the following important areas:

- Novel lightweight virtualization technologies for workload specific accelerators, such as FPGAs and TPUs, that will proliferate the Cloud to Edge continuum and facilitate low overhead serverless computing.
- System software of micro data centres for remote management of the system stack and workload orchestration.
- New power-aware strategies at the middleware and application levels for reducing energy consumption of data centres.
- A unified marketplace for Fog/Edge computing that will cater for competition between providers rather than consumers, and techniques for adopting Fog/Edge computing rather than simply making devices and resource Fog/Edge enabled.
- Common standards for Cloud federation to allow users to freely move between providers and avoid vendor lock-in. This includes, developing interoperable systems that enforce user policies, manage the lifecycle of workloads, and negotiate Service Level Agreements (SLAs).
- Mechanisms for vertically chaining network functions across different Cloud and Edge networks, multiple operators and heterogeneous hardware resources.
- Novel techniques for spatio-temporal compression to reduce data sizes, adaptive indexing to managing indexes while processing queries, and developing 'NewSQL' systems to support both transactional and analytical workloads on the Cloud.

Although Cloud computing research has matured over the last decade, there are numerous opportunities to pursue meaningful and impactful research in this area.

References

[1] B. Varghese and R. Buyya, "Next Generation Cloud Computing: New Trends and Research Directions," *Future Generation Computer Systems*, vol. 79, pp. 849 – 861, 2018.

[2] I. Foster and C. Kesselman, Eds., *The Grid: Blueprint for a New Computing Infrastructure*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999.

[3] Y. Elkhatib, "Mapping Cross-Cloud Systems: Challenges and Opportunities," in *USENIX Conference on Hot Topics in Cloud Computing*. USENIX Association, Jun. 2016, pp. 77–83.

[4] D. Merkel, "Docker: Lightweight Linux containers for consistent development and deployment," *Linux Journal*, vol. 2014, no. 239, p. 2, 2014.

[5] I. Nadareishvili, R. Mitra, M. McLarty, and M. Amundsen, *Microservice Architecture: aligning principles, practices, and culture*. O'Reilly Media, Inc., 2016.

[6] C.-H. Hong, I. Spence, and D. S. Nikolopoulos, "GPU Virtualization and Scheduling Methods: A Comprehensive Survey," *ACM Computing Surveys*, vol. 50, no. 3, p. 35, 2017.

[7] M. Satyanarayanan, "The Emergence of Edge Computing," *Computer*, vol. 50, no. 1, pp. 30–39, 2017.

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

[9] S. J. Johnston, P. J. Basford, C. S. Perkins, H. Herry, F. P. Tso, D. Pezaros, R. D. Mullins, E. Yoneki, S. J. Cox, and J. Singer, "Commodity Single Board Computer Clusters and Their Applications," *Future Generation Computer Systems*, June 2018.

[10] Y. Elkhatib, B. F. Porter, H. B. Ribeiro, M. F. Zhani, J. Qadir, and E. Rivière, "On Using Micro-Clouds to Deliver the Fog," *Internet Computing*, vol. 21, no. 2, pp. 8–15, 2017.

[11] F. Yang and A. A. Chien, "ZCCloud: Exploring Wasted Green Power for High-Performance Computing," in *IEEE International Parallel and Distributed Processing Symposium*, 2016, pp. 1051–1060.

[12] J. Scheuner, P. Leitner, J. Cito, and H. Gall, "Cloud Work Bench - Infrastructure-as-Code Based Cloud Benchmarking," in *IEEE International Conference on Cloud Computing Technology and Science*, 2014, pp. 246–253.

[13] B. Varghese, O. Akgun, I. Miguel, L. Thai, and A. Barker, "Cloud Benchmarking For Maximising Performance of Scientific Applications," *IEEE Transactions on Cloud Computing*, 2018.

[14] V. Dalibard, M. Schaarschmidt, and E. Yoneki, "BOAT: Building Auto-Tuners with Structured Bayesian Optimization," in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 479–488.

[15] L. Cui, F. P. Tso, and W. Jia, "Enforcing Network Policy in Heterogeneous Network Function Box Environment," *Computer Networks*, vol. 138, pp. 108 – 118, 2018.

[16] A. Elhabbash, G. Blair, G. Tyson, and Y. Elkhatib, "Adaptive Service Deployment using In-Network Mediation," in *International Conference on Network and Service Management*, 2018.

[17] M. D. Assuno, R. N. Calheiros, S. Bianchi, M. A. Netto, and R. Buyya, "Big Data Computing and Clouds: Trends and Future Directions," *Journal of Parallel and Distributed Computing*, vol. 79-80, pp. 3 – 15, 2015.

[18] A. Pavlo and M. Aslett, "What's Really New with NewSQL?" *SIGMOD Record*, vol. 45, no. 2, pp. 45–55, Sep. 2016.

[19] S. Hendrickson, S. Sturdevant, T. Harter, V. Venkataramani, A. C. Arpaci-Dusseau, and R. H. Arpaci-Dusseau, "Serverless Computation with OpenLambda," in *8th USENIX Workshop on Hot Topics in Cloud Computing*, 2016.

[20] I. A. Kash and P. B. Key, "Pricing the Cloud," *IEEE Internet Computing*, vol. 20, no. 1, pp. 36–43, 2016.

Author Biography

Blesson Varghese is a lecturer in computer science at Queens University Belfast and an honorary lecturer at the University of St Andrews. His research interests are in tackling system oriented challenges of next-generation distributed systems that leverage the Cloud–Edge continuum. Varghese received a PhD in computer science from the University of Reading. Contact him at b.varghese@qub.ac.uk.

Philipp Leitner is an assistant professor of software engineering at Chalmers and the University of Gothenburg. His research interests are in software engineering for cloud- and web-based systems, with a special focus on software and service performance. Leitner received a PhD in business informatics from the Vienna University of Technology. Contact him at philipp.leitner@chalmers.se.

Suprio Ray is an assistant professor at the University of New Brunswick. His research interests include big data, database management systems and data management in the Cloud. Ray received a PhD in computer science from the university of Toronto. Contact him at sray@unb.ca.

Kyle Chard is a senior researcher and Fellow at the University of Chicago. His research interests include high performance computing, data-intensive computing, and cloud computing. Chard received a PhD in computer science from Victoria University of Wellington. Contact him at chard@uchicago.edu.

Adam Barker is a Professor in Computer Science at the University of St Andrews. His research interests are in scalable systems that address current and future large-scale data challenges. Barker received a PhD in Informatics from the University of Edinburgh. Contact him at adam.barker@st-andrews.ac.uk.

Yehia Elkhatib is an assistant professor at Lancaster University and a visiting professor at L'École de Technologie Supérieure, Montréal. He works to enable distributed applications to traverse infrastructural boundaries, enforcing high-level application requirements, and involving associated interoperability and decision support issues. Elkhatib received a PhD in computer science from the University of Lancaster and is the creator of the Cross-Cloud workshop series. Contact him at y.elkhatib@lancaster.ac.uk.

Herry Herry is a research associate at the University of Glasgow. His research interests include cloud computing, configuration management and AI. Herry received a PhD in cloud computing from the University of Edinburgh. Contact him at h@herry.co.

Cheol-Ho Hong is an assistant professor at Chung-Ang University. His research interests include operating systems and accelerator virtualization. Hong received a PhD in computer science from Korea University. Contact him at cheolhohong@cau.ac.kr.

Jeremy Singer is a senior lecturer at the University of Glasgow. His research interests are in complex systems engineering, focusing on the application of mathematical models to runtime system behaviour. Singer received a PhD in computer science from the University of Cambridge. Contact him at `jeremy.singer@glasgow.ac.uk`.

Fung Po Tso is a senior lecturer at Loughborough University, UK. His research interests include cloud networking, edge computing and network function chaining. Tso received a PhD in computer science from the City University of Hong Kong. Contact him at `p.tso@lboro.ac.uk`.

Eiko Yoneki is a research fellow at the University of Cambridge and a Turing fellow at the Alan Turing Institute. Her research interests include distributed systems, large-scale graph processing, and computer systems optimisation with machine learning methods. Yoneki received a PhD in data centric asynchronous communication from the University of Cambridge. Contact her at eiko.yoneki@cl.cam.ac.uk.

Mohamed Faten Zhani is an associate professor at École de Technologie Supérieure, University of Quebec. His research interests include cloud computing, network and service management and large-scale distributed systems. Zhani received a PhD in Computer Science from the University of Quebec. Contact him at mfzhani@etsmtl.ca.