Cloud Computing at the *Edges*

Luiz F. Bittencourt[1](✉), Omer Rana[2], and Ioan Petri[2]

[1] Institute of Computing, University of Campinas, Campinas, Brazil
bit@ic.unicamp.br
[2] School of Computer Science and Informatics, Cardiff University, Cardiff, UK
{ranaof,petrii}@cardiff.ac.uk

Abstract. Currently, most cloud computing deployments are generally supported through the use of large scale data centres. There is a common perception that by developing scalable computation, storage, network, and by energy-acquisition at preferential prices, data centres are able to provide more efficient, reliable and cost effective hosting environments for user applications. However, although the network capacity within and in the proximity of such a data centre may be high – the connectivity of a user to their first hop network may not be. Understanding how a distributed cloud can be provisioned, enabling capability to be made available "closer" to a user (geographically or based on network metrics, such as number of hops or latency), remains an important challenge – aiming to provide the same benefits as a centralised cloud, but with better Quality of Service for mobile users. With increasing proliferation of mobile devices and sensor-based deployments, understanding how data from such devices can be processed in closer proximity to the device (ranging from capability directly available on the device or through first-hop network nodes from the device) also forms an important requirement of such distributed clouds. We review a number of technologies that could be useful enablers of distributed clouds – outlining common themes across them and identifying potential business models.

Keywords: Distributed clouds · Mobile computing · Edge device integration

1 Introduction and Overview

There has been a recent increase in the diversity, type and number of devices used to access cloud services – with such devices expected to reach 24 billion by 2020 [1] and generally part of the increasing interest in *Internet of Things* (IoT). IoT comprises any kind of objects that are able to generate a minimal piece of data and transmit it through the network, ranging from small fixed sensors to mobile, smart devices. The amount of data that can be generated by these devices and that need to be processed and/or stored has no precedents. Although the now established cloud computing paradigm could be utilised to store and process data generated by IoT devices, the expected amount of data can make this inefficient or even unpractical.

M. Helfert et al. (Eds.): CLOSER 2015, CCIS 581, pp. 3–12, 2016.
DOI: 10.1007/978-3-319-29582-4_1

One drawback of using a centralised data centre alone to process and store IoT data is related to constraints with existing network capacity and latency. Devices constantly generating and transferring data to the cloud can result in poor network conditions, yielding congestion and service disruption for many applications. Moreover, much of the data generated by IoT devices do not need to be stored in its raw form. There is now significant interest in combining cloud computing, offered at large scale data centres, with services that have been made available at regional data centres. With interest in providing cloud computing capability across different types of data centres, this often implies that there needs to be suitable coordination between distributed data centres that are able to receive and process data from such devices, which may be located at different geographical areas and operating with varying reliability criteria. The extent of this distributed cloud model also encompasses recent interest in supporting multiple micro and nano data centres, which may be connected over network links with varying bandwidth, availability profiles and latency.

The distributed cloud deployment model enables a variety of different types of market players to also engage and provision services and infrastructure, from telecom operators who may use their existing network infrastructure to offer cloud services, to a variety of businesses (such as coffee shops, supermarkets etc.), who can host cloud services to enable a better Quality of Experience (QoE) for a user. The benefits of this model are many and include: (i) improved resilience of cloud services; (ii) location specific contextualisation of provisioned services; (iii) ability to integrate regionally provisioned services in a seamless manner; (iv) latency hiding through automated service "hand-off"; (v) better coupling between cloud services and wireless access networks.

The distributed cloud model shares similarities with a number of emerging technologies and approaches – in all cases attempting to move data and processing closer to the user, thereby moving cloud provisioning from centralised data centres to *edge* servers with varying capability and connectivity. We briefly outline some of these in Sect. 2, to demonstrate common themes and outline a generalised architecture that attempts to combine features from these. Although each of these approaches have their own specific use scenarios, and have been developed by different communities, we notice a significant overlap in the underlying concepts being used. We characterise these in Sect. 3.

2 Related Approaches

The maturity of the cloud computing paradigm has contributed to a large number of distributed network applications that take advantage of cloud capacity to overcome computing and data storage requirements of a user. This centralised data centre architecture allows access to a large (potential) computing pool with unbounded[1] capacity. *Elasticity* is often a key enabler in such applications, allowing dynamic scale up/down based on instantaneous resource requirements [2,3].

[1] Unbounded here refers to the user perception of endless on-demand capacity.

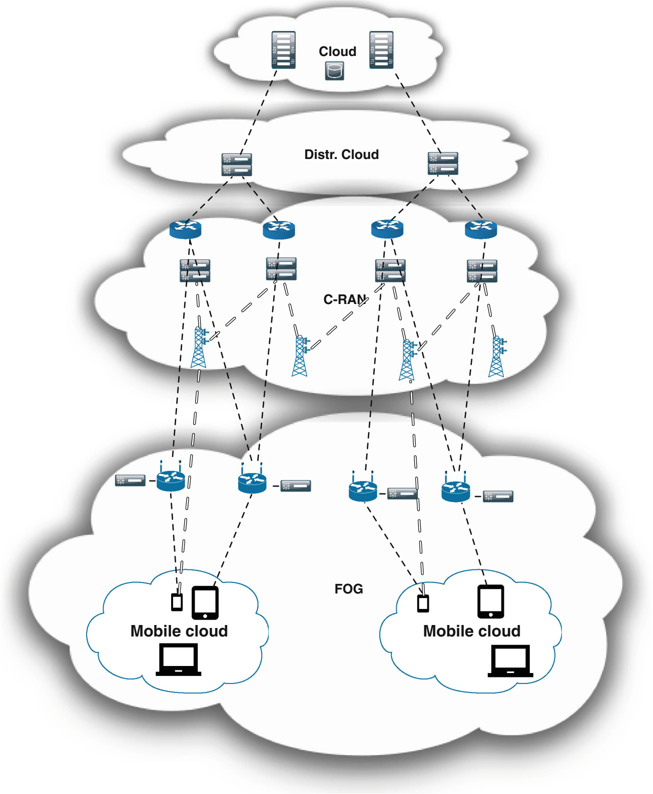

Fig. 1. Distributed Clouds – a conceptual perspective.

Data centres-based Cloud systems are able to fulfil many application requirements, needing limited upfront investment and easing the management of continuous change in requirements over time. Recently, however, there has been interest in providing support for "distributed clouds", which provide similar benefits but focus on cloud provisioning across multiple providers. We briefly describe each of the approaches to support distributed clouds outlined in Fig. 1, identifying their key characteristics and emphasising their similarities.

Distributed Data Centres: This approach involves the use of multiple types of linked data centres, each type offering differing capabilities. Two general types are often identified: (i) network data centres (NDC); (ii) cloud/enterprise data centres (CDC). The first of these are generally owned and managed by network operators and able to provide a limited programmatic interface to external users. Existing efforts towards network function virtualisation and software defined networks have enabled capability on network elements within an NDC, such as routers and switches, to be directly accessed by external users – enabling a variety of in-network operations to be made available (ranging from data encryption, data transcoding, etc.). Techniques such as MiddleBox technologies (often also referred to as network appliances) can be combined with such approaches to enable data streams arriving at NDCs to be aggregated in some way. Conversely, an enterprise data centre offers computational and data storage capability of a much greater capacity than an NDC, but often not situated at an intermediate point in the network. A CDC has much greater potential for infrastructure scalability and can be part of a much wider, global deployment (e.g. a CDC in Asia-Pacific, in Europe, etc.). A provider is able to use the combined capability of multiple CDCs to enable elastic provisioning to a user and to provide fault masking in one such CDC. There may be a variety of other types of data centres (some user owned) in addition to the two identified here, and significant recent research is attempting to identify potential types that could be use to support resource provisioning to a user [4,5].

Mobile Cloud and Cloud-Offloading: This approach involves moving computation (generally) and data stored on mobile devices to an enterprise/cloud data centre. The general motivation is to enable more computationally intensive processes to be carried out on large scale data centres (CDC) rather than on the device, enabling: (i) improved battery usage on the device; (ii) latency and network outage masking on the device – especially when a device user is moving across geographical areas with varying network coverage; (iii) handling wireless connectivity across highly heterogeneous networks (always-on connectivity, on-demand scalability and energy efficiency is a difficult problem across heterogeneous networks); (iv) improve (potential) availability when using a CDC – due to reduced capacity on radio/wireless networks. Approaches can range in complexity from providing a complete *copy* of a mobile device within a CDC, with periodic synchronisation of state between the processes on the device and the CDC-hosted copy (e.g. the CloneCloud system [6] to create a device clone on the Cloud and provide an application level Virtual Machine (VM) at the data centre). CloneCloud requires device to cloud connectivity for the clone to remain in sync. with the device. An alternative approach is to annotate program source code to identify what should run on the device and what should be cloud-hosted – e.g. the Maui system [7]. In this approach, two versions of a program are created, a local and a cloud version. The "reflection" technique from programming languages is then used to determine which part should run where and how the two copies should remain in sync. This approach generally requires source code annotation (and can tolerate disconnection from network). Other related approaches focus on annotating a program call graph (method calls) to determine which parts should be off loaded – making use of criteria such as data transfer costs and security/data privacy concerns (i.e. determining what should remain local to the device and what can be moved to a CDC) [8].

Cloudlets and Fog Computing: This approach considers that processing and storage can be performed on edge devices, as in the mobile cloud computing paradigm, whenever this brings optimisation to the system and better quality of service. Fog computing introduces the notion of *cloudlets* – "small clouds" which are geographically scattered across a network and acting as "small data centres" at the edge of the network [9,10]. Cloudlets aim to give support to IoT devices by providing increased processing and storage capacity as an extension of those devices, but without the need to move data/processing to a CDC. This leads to reduced communication delays and the overall size of data that needs to be migrated to a CDC. Data processing offered by cloudlets can employ a set of mechanisms to process data on behalf of the IoT device, effectively sending to the cloud only data that are aggregated results or that need data/processing that is not available at the cloudlet [11].

Cloud Radio Access Network (C-RAN): This approach provides an optimisation over an existing de-centralised Radio Access Network (RAN), due to a significant increase in mobile internet traffic over recent years and the cost associated with operating, building and upgrading such a network. The Cloud-RAN approach involves splitting the capability offered at a mobile base station into two: a Remote Radio Head (RRH) and a BaseBand Unit (BBU). In the C-RAN approach, the BBU is centralised and shared amongst multiple sites in a virtualised BBU pool – and often hosted at a data centre. This centralisation enables reduced operating costs, improves scalability and reduces potential energy consumption. As BBU's are virtualised and hosted on a single data centre, this enables multiple physical cells/sites to interact with lower delays leading additionally to increased spectral efficiency and throughput. The C-RAN approach also aligns well with recent interest in creating Heterogeneous and Small Cell networks (HetSNets), primarily leading to increased network capacity due to the additional cells now available. The C-RAN approach is particularly relevant in the context of distributed data centres as they enable improved handoff mechanisms for mobile users (due to the use of the same BBU hosting location) – being geographically closer the user and able to support partial processing [12].

3 Common Themes

There are conceptual similarities that arise in the paradigms listed in the previous section. In this section we discuss related concepts and general aspects on how these relate to cloud computing at network edges.

3.1 Architecture and Deployment

Enabling cloud computing at the edges involves, primarily, a decision on where processing/storage capacity should be placed in order to fulfil users' application requirements. This decision can depend on several factors, including how efficient and reliable the network is in connecting users to the edge processing/storage equipment, as well as connecting those equipment among themselves. Other criteria can also influence this decision – such as: (i) overall cost of undertaking computation; (ii) size of data that needs to be transferred from a local (proximity-based) device to a data centre; (iii) network reliability/availability, amongst others.

It is necessary to consider the trade off between the computational infrastructure needed to host services (such as cloudlet) and their proximity to the user. Locating a service closer to a user could potentially require a greater number of facilities to deploy such services. This incurs higher costs, but smaller latencies/delays for users accessing cloud data/applications. For example, a more geographically distributed architecture such as advocated in Fog Computing would be able to act as a real-time capacity extension for mobile devices, leading to a one-hop connection to processing/storage resources. On the other hand, deployment costs may require different business models to make it feasible.

The deployment of equipment to support such edge services leads to greater reliance on a dependable network. The straightforward approach is to let communication go through existing infrastructure, i.e., with traffic between distributed processing/storage equipment traversing the core network using ordinary TCP/IP communication – potentially leading to increased traffic in the core network. A second approach would be to provide a direct connection using a dedicated link (radio, fibre, or even ethernet), which increases cost but improves performance. This trade-off between cost and performance can be also a focus of study: distributed equipment "clusters" could be built using direct network connections in places where demand is significantly higher, preventing routing through the core network. Conversely, where communication requirement is lower (or sparse), the core network could be utilised.

An important aspect is a consideration of who would be responsible for deployment and maintenance of equipment when making use of distributed cloud computing resources. Feasible/ potential options include cloud providers, network (broadband) providers, mobile phone carriers, and/or local businesses. While cloud/broadband providers seem like the obvious choices, mobile phone carriers (especially in developing countries) and local businesses can utilise their intrinsic distributed presence to host equipment and provide computing services in addition to communication through 4G/LTE/5G and WiFi connections.

3.2 Virtualisation

Virtualisation enables sharing of infrastructure amongst users with software and, potentially, hardware isolation. The hypervisor (or virtual machine monitor – VMM) has the ability to replicate hardware interfaces and trap the necessary instructions in order to share the underlying hardware among multiple privileged tenants. Therefore, tenants generally have no knowledge they are running on a virtualised and shared hardware.

Efficient resource virtualisation is essential to enable various Quality of Service provisioning to be supported across a shared infrastructures – enabling different users (with varying service requirements and QoS needs) to be isolated from each other. In deploying cloud-based services, virtualisation is also important to ease management through the use of virtual machines (VMs), which can be migrated to different physical machines to fulfil an objective function, such as infrastructure cost reduction. What is virtualised can vary – for instance: (i) a physical machine or a mobile device can be virtualised (with CPU, memory and network interface); (ii) network function, e.g. routing and forwarding of packets can be virtualised; (iii) a base station capability (in C-RAN) can be virtualised, (iv) a physical sensor may be virtualised – enabling the same "virtual" sensor interface to communicate with different physical sensors at different times, or to enable data from multiple sensors to be aggregated and offered as a virtual sensor; (v) a firewall or security interface can be virtualised, etc. Over recent years, there has been interest in providing virtualisation at different levels of the computational infrastructure – with "enterprise" and "data centre" virtualisation enabling an aggregation of different levels of virtualisation to co-exist, leading to a much greater efficiency in how the physical infrastructure is used, providing isolation for users and enabling dynamic update of physical infrastructure that is accessed through a virtualised interface.

In a distributed cloud context, such virtualisation capability can now extend beyond a single data centre – along the different layers outlined above. Additionally, the isolation provided by virtualisation, the ability to replicate a user session across different VMs and support for VM migration can help in reducing latencies when the user moves from one geographical location to another. Services hosted within such a VM can be utilised to perform data/process migration along with user movement, aiming to reduce delays for specific applications. This could be specially interesting in the fog computing paradigm, where VMs can migrate among cloudlets to support users applications [13].

3.3 Data Migration and Management

When using a distributed cloud for mobility-based scenarios, support for efficient data migration is necessary, enabling data to be placed closer to the user (with a user location potentially changing several times during a single day). Nodes within a distributed cloud may be used for storing more "volatile" data that does not need to be kept for long periods of time, and such nodes can provide a pre-processing facility to reduce data transfer to the centralised cloud, where long-term data storage/processing can occur. To enable QoS-based provisioning, user data and applications should be placed as closest (in terms of number of hops or network latency) as possible to his/her device(s). The (potentially real-time) need for migration introduces new challenges in resource management. Data and processing should follow users, demanding mechanisms for mobility detection/prediction to anticipate migration and reduce the number of service disruptions seen by a user.

4 Business Models

Several business models may become relevant when considering virtualised distributed cloud environments. Nodes associated with a distributed cloud must be deployed and managed by an individual or organisation, and the costs of the infrastructure must be taken into account in the business model. Similar to current broad availability of WiFi access points and cell phone antennas, we envisage four general ways of funding cloud at the edge: (i) by cloud providers; (ii) by local businesses; (iii) by public funding; and (iv) by mobile carriers. Various trust models exist that may be associated with each of these four options.

Service Selection: in this model, the user would be able to choose a cloud at the edge provider on-the-go, according to his/her current activity or provider's availability and potential reputation within a market place. The use of a service-based approach enables loose coupling, enabling an eco-system of providers to co-exist. However, there is no guarantee that integrating externally provisioned services will lead to the fulfilment of the user objectives, since this would depend on providers' agreements to support data and process migration. Therefore, interoperability and trust issues are expected to dominate this selection decision.

Service Contracts: in this model, contracts are signed between the user and the provider, where criteria that adequately captures the circumstances that influence the performance of the externally provisioned services must be specified and pre-agreed. Contracts can be based on particular (monitorable) service-level objectives – where short-term contracts have proved to be more profitable options for service providers. Providers can also offer in-contract guarantees performance metrics (e.g. availability) to the customer, which is reflected in the associated price.

All-in-One Enterprise Cloud: this model is a more comprehensive approach, where a distributed node is actually hosted at a data centre. Therefore, large cloud providers could joining local businesses/ network providers in order to build a larger business ecosystem with greater financial stability, allowing users content/data/processing to freely travel across their boundaries.

Business models are important to make distributed clouds profitable, as well as to help users make informed decisions about providers. Each business model is associated with a set of cost models according to the provider's service strategies and business objectives, as for example:

- *Consumption-Based Cost Model*: clients only pay for the resources they use. For distributed clouds a user could be charged according to the size of his/her files or processing time utilised by applications that need edge computing.
- *Subscription-Cost Pricing Model*: clients pay a subscription charge for using a service for a period of time – typically on a monthly basis. This subscription cost typically provides unlimited usage (subject to some "fair use" constraints) during the subscription period. For example, local businesses can offer a subscription to their infrastructure that enables a user to have content/applications to be placed on that infrastructure.
- *Advertising-Based Cost Model*: clients get a no-charge or heavily-discounted service whereas the providers receive most of their revenue from advertisers. This model is quite common in cloud-based media services such as free TV providers (e.g. net2TV) and can also be adopted in distributed clouds.
- *Market-Based Cost Model*: clients are charged on a per-unit-time basis. When bringing computing to the edges, the user can have a configuration dashboard to establish the maximum usage quota/capacity and other relevant parameters, similarly to IaaS offerings such as Amazon EC2.
- *Group Buying Cost Model*: clients can acquire reduced cost services only if there are enough clients interested in a deal. This can be adapted for distributed clouds, enabling users to have access to a larger set of edge infrastructure but with limited concurrency among shared users, for example.

5 Application Scenarios

We describe two potential scenarios where the approach being proposed in this paper could be benefit:

- Crowd-sourced surveillance: this application would involve making use of user provisioned resources to capture local data, aggregated through the use of a Cloud-based platform. As increasing number of individuals posses mobile devices able to record (via photos, videos or text-based data) information about a scene locally, each of these devices could be used to record such information and tag this with the location of the user. Such information could then be submitted to a data centre for aggregation. While the information is in-transit from the capture source to the data centre, it could be aggregate enroute. Additional content related to crime rates within a geographical area, known crime reports within a particular time frame, etc. could be combined with such content to increase the potential veracity of information that is subsequently submitted to the data centre. The device owned by a user could connect to the nearest available "cloudlet" to offload some of the data recorded about the particular event being monitored. Cloudlets would interact with each other, based on the geographical proximity of other users to check if the same incident has been recorded by others.
- Real time streaming: this application would involve a user interacting with a real time information source, with a requirement tomaintain a persistent, high quality (low latency, high throughput) connection to the information source. In this scenario, the user would initially register their quality of service requirements to a cloudlet, and as the user moves from one region to another, there would need to be hand-off to other cloudlets. This hand-off could be supported through technologies such as C-RAN, where a common regional data centre may be used to host multiple cloudlets, with a potential predictive hand-off with user movement.

6 Conclusion

We describe a variety of emerging technologies that promote the integration of edge devices with Cloud computing, enabling both to be used in coordination. With increasing deployment and availability of sensing capability, there is a realisation that not all of this data needs to be migrated to a centralised data centre. Undertaking data processing and storage closer to a user allows masking of the *last mile* connectivity concerns that have been highlighted in Content Distribution Networks. Understanding how resources that have a more efficient (small number of hops or low latency) connection to a user, can be combined with large scale data centres remains an important challenge for many applications. This contribution attempts to highlight common issues that occur within multiple approaches addressing this concern.

Acknowledgements. We would like to acknowledge various individuals who have contributed to our views expressed in this article, these include: Manish Parashar, Javier Diaz-Montes, Mengsong Zou, Ali Reza Zemani (Rutgers University, USA), Rafael Tolosana-Calasanz, Jose Banares (University of Zaragoza, Spain), Congduc Pham (University of Pau, France), Yacine Rezgui, Tom Beach, Stuart Allen (Cardiff University).

References

1. Gubbi, J., Buyya, R., Marusic, S., Palaniswami, M.: Internet of things (IoT): a vision, architectural elements, and future directions. Future Gener. Comput. Syst. **29**, 1645–1660 (2013)
2. Bittencourt, L.F., Madeira, E.R.M., Da Fonseca, N.L.S.: Scheduling in hybrid clouds. IEEE Commun. Mag. **50**, 42–47 (2012)
3. Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., et al.: A view of cloud computing. Comm. of the ACM **53**, 50–58 (2010)
4. Mazmanov, D., Curescu, C., Olsson, H., Ton, A., Kempf, J.: Handling performance sensitive native cloud applications with distributed cloud computing and SLA management. In: 2013 IEEE/ACM 6th International Conference on Utility and Cloud Computing (UCC), pp. 470–475 (2013)
5. Nygren, E., Sitaraman, R.K., Sun, J.: The akamai network: a platform for high-performance internet applications. SIGOPS Oper. Syst. Rev. **44**, 2–19 (2010)
6. Chun, B.G., Ihm, S., Maniatis, P., Naik, M., Patti, A.: Clonecloud: elastic execution between mobile device and cloud. In: Proceedings of the Sixth Conference on Computer Systems, EuroSys 2011, pp. 301–314. ACM, New York (2011)
7. Cuervo, E., Balasubramanian, A., Cho, D.K., Wolman, A., Saroiu, S., Chandra, R., Bahl, P.: Maui: making smartphones last longer with code offload. In: Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services, MobiSys 2010, pp. 49–62. ACM, New York (2010)
8. Pedersen, M., Fitzek, F.: Mobile clouds: the new content distribution platform. Proc. IEEE **100**, 1400–1403 (2012)
9. Bonomi, F., Milito, R., Zhu, J., Addepalli, S.: Fog computing and its role in the internet of things. In: MCC Workshop on Mobile Cloud Computing, pp. 13–16. ACM (2012)
10. Bonomi, F., Milito, R., Natarajan, P., Zhu, J.: Fog computing: a platform for internet of things and analytics. In: Bessis, N., Dobre, C. (eds.) Big Data and Internet of Things: A Roadmap for Smart Environments. SCI, vol. 546, pp. 169–186. Springer, Heidelberg (2014)
11. Fesehaye, D., Gao, Y., Nahrstedt, K., Wang, G.: Impact of cloudlets on interactive mobile cloud applications. In: 2012 IEEE 16th International Enterprise Distributed Object Computing Conference (EDOC), pp. 123–132 (2012)
12. Checko, A., Christiansen, H., Yan, Y., Scolari, L., Kardaras, G., Berger, M., Dittmann, L.: Cloud ran for mobile networks - a technology overview. IEEE Commun. Surv. Tutorials **17**, 405–426 (2015)
13. Bittencourt, L.F., Lopes, M.M., Petri, I., Rana, O.F.: Towards virtual machine migration in fog computing. In: 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (2015)