EPiC Series in Computing

Volume 91, 2023, Pages 66–75

Proceedings of 38th International Conference on Computers and Their Applications

Edge Networks the Future of Cloud Computing

Sean McGrath[1†] Colin Flanagan[1*] Xiaoxiao Liu[1†] Liam Quinn[1] and Liaoyuan Zeng[2]

[1]University of Limerick, Ireland.

[2]UESTC, China.

sean.mcgrath@ul.ie, colin.flanagan@ul.ie, lbq954@gmail.com, xiaoxiaoliu53@foxmail.com, liaoyuan.zeng@gmail.com

Abstract

Computing architecture continues the pendulum swing between centralized and distributed models – driven by technological innovation in CPU/GPU architecture, memory/storage, I/O, networking, performance, and emerging use cases. For a number of years, the most recent architecture has been the cloud-based centralized computing model, which is now shifting to a distributed edge computing model. New businesses, technologies, usage models, and new applications are driving this change. The rapid growth of IoT across all segments of society is driven by 5G, Edge, low-cost sensors, embedded SoC controllers, and new enterprise applications. The impact is more data being generated by these smart sensors, increased demand for storage, data analytics, and network capacity to move this data to adjacent nodes or cloud models. These and other parameters are driving the emergence of the Edge computing model – and there are many Edge types. However, in terms of optimisation and efficiency, these approaches may not be the best solution.

Key words: Edge Networks, decentralized network, Cloud Computing

1 Introduction

Communication networks can be broadly classified into two categories – switched and broadcast communication networks. Examples of switched networks are circuit-switched networks, such as the public telephone network, and packet-switched networks, such as computer networks based on TCP/IP.

[*] Reviewed and revised the document.

[†] Masterminded and created the first stable version of this document.

Liam Quinn and Liaoyuan Zeng edited part of the contents.

Broadcast communication networks include packet radio networks, satellite networks, and multi-access local area networks such as Ethernet [1]. Circuit-switching and packet-switching are two different technologies that have evolved over time. During the high-level reviews of each of these network architectures, the evolution of networks over the past number of decades continues to reflect the similar changes between centralized and decentralized network topologies – taking advantage of new features, capabilities, and integration in silicon, applications, and the separation of data and control plane processing and routing.

2 Cloud Computing Architecture Model

Cloud computing is the ability to deliver on-demand computing services over the internet on a pay-as-you-go basis. This means, that rather than managing files on a local storage device, cloud computing makes it possible to save, store and access them over the internet, which supports the growing number of mobile and remote users. As with all changes in architecture, several models have been developed based on the deployment of cloud and the associated services and how they are consumer by the end-user based several factors, including cost, applications, IT resources, security, and scalability. Cloud optimization is the process of correctly selecting and assigning the right resources to a workload or application, aligned with the optimal deployment and services Table 1 and Table 2. All applications and workload needs are different, and the infrastructure requirements are unique, and evolve over time [2].

Public Cloud	Private Cloud	Hybrid Cloud
The cloud infrastructure is made available to the public over the Internet and is owned by a public cloud provider. The capex for the user is low, and they pay for the service and time used – so the cost is the lowest in this model. Providers of public clouds include AWS, Microsoft Azure, IBM's Blue Cloud and Sun Cloud.	This is where the end-user installs their own cloud infrastructure- the capex cost is much larger, and the operation and support costs are significant. The cloud infrastructure is exclusively operated by a single organization, and it can be managed by the operator or a third party and can be located on-premise or off-premise. Providers of private clouds include AWS and VmWare.	This takes the best features and capabilities of both the private and public cloud models, where the user can use the optimized services and features of both the public and private cloud models. In this model a customer may use the private cloud service for secure data and content, and the public cloud services for sharing non-sensitive data with partners or with the public.

Table 1: Deployment Models

Infrastructure-as-a-Service	Platform-as-a-Service	Software-as-a-Service
(IaaS) is a cloud service that provides basic computing infrastructure, on which the operating system and applications can be deployed. The services are available on a pay-for-use model and the end-user is typically an IT administrator. This service is typically used where the customer needs a virtual machine with local on-site expertise to install the software on the end-device(s). Providers of IaaS include AWS, Microsoft Azure, and Google.	(PaaS) provides a computing platform and runtime environment for developing, testing, and managing applications. The user can develop software using tools and libraries from the cloud provider and deploy the software onto cloud services. It allows software developers to deploy applications without requiring all the related infrastructure. Typical users are software developers, and providers of PaaS services include Microsoft Azure, AWS, and Google.	(SaaS) – this model is where the cloud providers host and manage the software applications on a payas-you-go pricing model. All software and hardware are provided and managed by the cloud provider, so users don't maintain any of the infrastructure, software, libraries, or applications (OPEX model). This is equivalent to getting a finished product or service where the software and data are hosted on the cloud and provided as a service. Google Gmail is an example of SaaS.

Table 2: Service Models

3 Edge Computing Architecture Model

Edge computing is a distributed IT architecture where client data is processed at the periphery of the network, as close as possible to the originating source. There are several technology trends and usage models that is driving the transition toward edge computing, including mobile computing, the decreasing cost of computer components, the growth and adoption of IoT, emergence of 5G and AI. Depending on the implementation, time-sensitive data in an edge computing architecture may be processed at the point of origin by an intelligent device or sent to an intermediary server located in close geographical proximity to the client. Data that is less time-sensitive is sent to the cloud for historical analysis, big data analytics and long-term storage. With cloud computing, data from a range of sources is sent to large data centres often geographically far away from the source of the data, while distributed edge computing is located closer to the data source through a network of edge devices – bringing (some) the advantages and benefits of the cloud closer to the data source. This means the data has less distance to travel and can be processed and acted on with less latency for mission critical applications – mission control, automation, smart devices, and utilities. The term 'edge' in edge computing is synonymous with networking, at which point traffic enters or exists the network. With Edge, more compute, storage, and AI based training is executed on data, with local decision capabilities to reduce the traffic moving to the cloud [3]. The Edge model supports improved optimization at each stage in the overall compute ecosystem, with local decisions on data and what data is stored or moved to the cloud. As with other IT architectures over the past number of decades, the 'pendulum' swing from central-distribute-central continues to play out, with Edge the latest iteration of the change from central compute with cloud to a more distributed architecture with Edge. There continues to exist the structure of store and forward with Edge and cloud – it is not the ideal solution, certainly the transition to edge is coming at the right time enabled by other technology trends in cost of silicon, feature integration and IoT in a future connected

world. The Edge architecture is a key architectural change to satisfy the requirements of 5G, due to the functions of services localization, local network breakout, caching, computational offloading, and network context information (deep packet inspection). The edge can decrease the end-to-end latency dramatically through service localization and caching, that are key enablers of 5G low latency requirements. These features also reduce the requirement for backhaul bandwidth, reducing the network operational and maintenance cost, and overall system performance in optimizing the cloud and edge architecture. The computational, AI and storage capabilities, offloads these functions from IoT and other end devices that prolong the battery life of resource constrained IoT devices. The Edge therefore bridges between the cloud and IoT/end-devices and applications, which have diverse requirements, ex low latency, high bandwidth, location aware, cost, power, performance, and reliability. For true optimization and time-based decision, the network is the future of computing – computing is not a location (cloud/Edge) – rather the network is the compute, where applications should be capable of accessing resources (HW/SW) as-a-service independent of where in the network/compute continuum those applications are running [4]. Edge computing is a modern architectural approach to IoT systems. By provisioning parts of the workload in the cloud, it leverages the maturity and advances of the Cloud, in terms of compute, storage and machine learning. Simultaneously, provisioning services directly on the edge, closer to the physical devices, enables the management of the edge-specific workloads more efficiently, and allows a better distribution of resources between the infrastructure layers. The emphasis in this paper is that for true optimization and time-based decision, the network is the future of computing – computing is not a location (cloud/Edge) – rather the network is the compute, where applications should be capable of accessing resources (HW/SW) as a service independent of where in the network/compute continuum those applications are running.

Decentralized systems, like the distributed architecture, have no single points of failure. Even if several nodes go down, the network is still up. There is no single entity control, so there is zero possibility of the network going down anytime unless all the nodes go down simultaneously which is a rare possibility when their nodes connected globally with the Internet. Decentralized and distributed is more in the context of control and where the control functions are located. Like a peer-to-peer network, decentralized means there is no central authority, every node acts as both a client and server [5]. Distributed is more in the context of scalability – the difference is very subtle. A distributed system can also be implemented in a centralized system to process computational tasks, supporting features like data replication, intelligent fault tolerant policies, high availability. This is more in the context of centralized cloud computing, where one of more centralized domains are distributed in multiple data centres – all supporting a controlled centralized compute set of functions to keep the overall system operational.

4 Compute/Networking AAS Model

Building on several of the computing architectures above, essentially a lot of the same models are repeated in the context of the pendulum swing from distributed-central-distributed. With each successive iteration, new features and capabilities are added, including the use and deployment of smart devices and improved networking with better bandwidth, lower latency, and virtualized functions and capabilities. Additionally, the compute architecture today has expanded significantly outside IT and the traditional data centre, that cross adjacent market segments such as smart cities, smart vehicles, and smart homes and enterprises. Virtually all segments of society and industry is highly dependent on computer automation, process control and a growing dependency on an always connected and available computer system. This leads to what is next and forms the basis of this thesis in proposing a networking computing model as-a-service, like a networked compute utility that provide seamless compute to an

end-user. A close metaphor is the electric utility, where power is available at the nearest power socket to deliver power to a device or appliance in a safe, secure, and seamless way. The basis of this proposal is the user simply connects to the 'network compute plug' and request access to data, content, and services from the network compute portal. The user or (IoT) device is then provided with the requested service is a timely, secure, and seamless way. With AI and voice-assisted applications, elements of this exist today with Google, and Amazon, where content is requested by voice and delivered to the user at the application layer and independent of the underlying compute and networking hardware. Extending the discussion and overview of the cloud and Edge architectures, more smart devices filling out the distributed architecture model. With networks becoming the computing platform of the future, the growing compute, storage, and intelligence in those networks provides the capability for the network to make decisions on better performance and optimization of the overall network. The addition of AI across integrated silicon and platforms, each compute node or edge device should have a deep awareness of the features and capabilities of the adjacent nodes. This translates into more of a mesh structure populated by intelligent n awareness of the adjacent nodes – aka nearest neighbour node. The use and deployment of AI and smart devices is rapidly evolving and will continue to grow and expand to where the underlying networking and compute become hidden in the background and the true innovation will be virtualizing the underlying the intelligent hardware resources. By abstracting the usage model and capabilities above the physical devices and networking layer, the outcome will be capable of delivering intelligent, scalable network compute on-demand. Network Compute-as-a-Service (NCaaS) uses a virtualized network infrastructure to provide network services to users. A virtualized network consists of the physical underlying network, which are all the physical switches, routers, compute, and media elements. To realize the vision of network compute as a service, each smart compute node and device should have an awareness of the adjacent smart nodes and devices around them – nearest neighbour details [6].

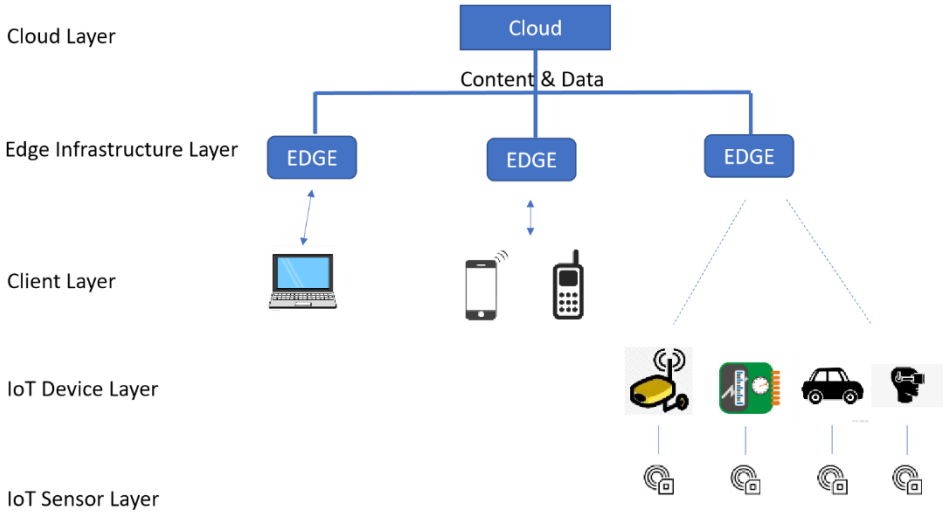

Figure 1: Edge Architecture

5 Network Evolution

As networks evolve away from cloud-based architectures, we are now seeing a dramatic shift to edge networks and edge processing. This has been partly brought about by the growth of the Internet of Things and the evolution of mobile technologies such as 5G. IoT systems are moving computation to the edge for reasons of cost, power-efficiency, latency, and timeliness. Systems have never experienced such large amounts of data being moved across networks between the edge and cloud. This presents a technological challenge in how we optimize decisions within these networks and where should these decisions take place. The optimal choices, if they exist, will vary with time and load. The competing "paradigms" are: do it all at the edge with the inherent problems with co-ordination across multiple domains; or use 5G to backhaul to the warehouse data centre, but this has problems of scalability, responsiveness, and power. The complexity of future networks will require the use of sophisticated AI algorithms to help achieve this process [7].

As we move to a society where information is processed and managed at the edge, networks are evolving in how they manage and move data. As network complexity increases, the question becomes what characteristics at the various parts of the network change and how should we make decisions because of such changes. The characteristics of various resources are changing as we move between the edge to cloud, such as bandwidth, processing power, latency, and trust. Traditionally the control of these resources resided within the cloud and the resource control dynamically allocates what services were necessary at the edge [8]. The model in Figure 1 consists of three elements the edge where processing takes place and normally are very resource-limited in terms of processing and power. It is important to clarify good in this paper what we mean by the edge, network, and the cloud. The edge network or edge processing edge computing is not new and in 2002 Microsoft published a technical report titled “Enabling rich content services on the edge”. European telecommunications standards Institute (ETSI) organized an industrial group for Mobile Edge Computing (MEC) [9]. The dawn of the Internet of Things allowed connection with every application to the network, as the development of smart sensors, embedded devices, and low-power short-range wireless technologies emerged. With the evolution of 5G it encourages transport to the data centre, which is an opposite trend to computation at the edge, but these data centre are with edge computing.

In this paper when discussing the network, we infer the middle section of the system which deals with transporting and routing up information between the edge and cloud. This network generally moves and routes information between the edge and cloud as efficiently as possible. However, these networks have always to compromise between the different applications of voice video and data in providing an optimized solution [10]. The routers themselves are designed to reduce processing time and switch data as efficiently as possible to reduce latency. The routers themselves do not make decisions optimizing the network from an end-to-end perspective. Finally, the cloud provided unlimited storage processing, and security but the downside being latency issues. The difficulty for networks has always been a scalability issue do you have a distributed system where each element has full autonomy in the decision-making process regarding the data. In a centralized client-server system, information is controlled and managed in a very efficient and secure manner. Trying to resolve the problem of distributed or centralized networks has always been a challenge and as technology develops to solve some of the problems that are presented. This three-tier system evolves the Edge with intermediate level datacentre, and "backend" warehouse-scale datacentre [11]. The decision then is how to place compute to achieve goals of efficiency, responsiveness, accuracy/timeliness and low cost. These networks can never be fully optimized as we are dealing with a variety of different applications which have very different constraints, voice, video, and data all of which have very different requirements from a service aspect.

6 Decision Making in Networks

If we consider an end-to-end system of three elements, the edge, the network, and the cloud where information flows from the edge through the network to the cloud and vice versa. The critical issue in networks is where the decisions take place and more precisely what are these decisions and what are the implications of where these decisions occur. Different services at the edge from temperature sensors, voice systems, and high-definition video transmission. Each of these services will have very different resource implications that need to be optimised. For example, can edge devices make an independent decision at a local level irrespective of the implications that that may have on the network. The whole area of software-defined networks was intended to provide network optimization from the physical to the application layer right across the network. From the other perspective it is necessary that all information generated by the sensor at the edge is required to be transmitted across the network for decisions to be made. The major implication of some of this information being transmitted is its time dependency and relevance. In Figure 2 we see a diagram showing the basic structure of this architecture. The question is where this control takes place, or the decisions occur. Now the trend is that the edge will make local decisions resulting in faster processing time and lower latency, the middle block traditionally the network doesn't have real intelligence but essentially just routes the information in an "efficient" manner, and the final block the cloud is not resource-limited concerning power. bandwidth etc but it's not very efficient and is extremely power hunger and inefficient.

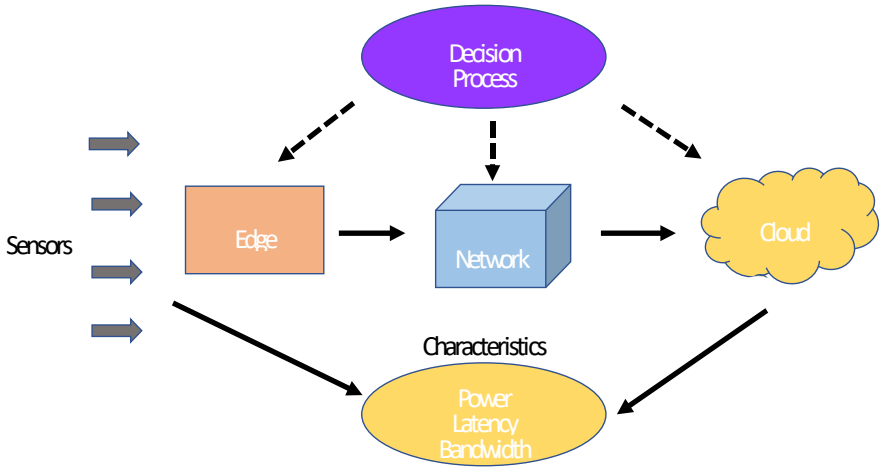

Figure 2: Edge to cloud network

The question is essentially one of network optimisation, but we need to define what parameters or characteristics we consider for optimisation to achieve the desired goal or objective. As an example, the classical vision of a data network transports data from the edge, where it is acquired, to some set of servers for processing, but with the growth in computing power in recent years, we need to ask whether it makes sense to distribute processing functions so they are closer to the edge, the principle being that computation is becoming cheaper (in terms of power) at an exponential rate, but communication remains expensive. So, a system that passively transports all data acquired from a set of edge sensors to processing on a server farm in the cloud does not make the best use of resources, and a more energy-efficient solution may be possible by using edge or intermediate stage processing. For example, Movidius (Intel) markets very efficient vision and inference engines that can perform low-power

processing at or near to the edge [12]. The vision is to distribute computational resources through the net and allocate them to processing tasks in an efficient way, so that the cloud becomes a true cloud, with computation happening where it makes most sense in terms of lowering power consumption, rather than today's pseudo-cloud of distributed sensors and large, power-hungry datacentres, with what is essentially passive communication between them. Of course, this will require that the decisions about what to process and where to do it must be made on an ongoing, real-time, basis. And ideally, we would like to distribute the decision-making functions across the computational resources of the enhanced cloud so that the system has no single point of failure. This is a big ask: the main consideration is that any network is a complex entity, in that it is distributed, time varying (in some but not all aspects), and *resource limited*. The application of decision theory to this problem of distributed resource allocation is research question to address. What we want is a framework or model that allows us to decide where in a network (edge, cloud, intermediate stages?), both processing and decisions about resource allocation are made. And we want to do this in a way that is responsive in near real-time.

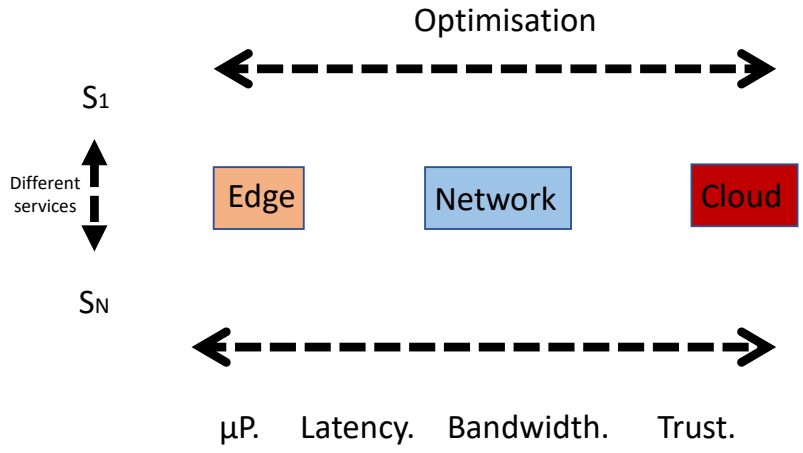

Figure 3: The network characteristics: BW, processing, power, latency, trust

Here we defined the system consisting of the edge, the network to the cloud. How do we optimize this system with so many variables, most of which are time-dependent? The question becomes what is to be optimized? Is it for the network? Is it for the user? is it for the service? The characteristics that we consider the network to be important are bandwidth, processing, power, latency, and trust. To model this process, we look at the diagram in Figure 3 which consists of three columns with inputs from different sensors S_1 two S_N each of these inputs will be processed at each stage going from left to right. As we process at each stage the characteristics that we previously mentioned of bandwidth, processing, etc. will vary within the network. Latency for example can be divided into processing time and transmission time. The physical transmission we can do very little about, however reducing latency in processing at the edge has its advantages but processing is resource limited unlike processing in the cloud.

For example, latency as we go from edge to cloud varies across the network. which we can model as a simple distribution. The traditional approach is to manage and optimize different characteristics across the network such as increasing dependence and reducing latency with varying degrees of success. It is important to consider number is the problem a scalability issue as data will continue to increase

and therefore networks cannot maintain the same architectures that evolved overtime. These systems by their very nature of evolution over time are not optimal and modifications and improvements efficiency and processing at various part of the networks provide ad hoc solutions to this optimization problem.

7 Conclusion

By comparing traditional fixed and mobile distributed computing systems. Ad-hoc, Edge, and cloud introduce several complex challenges due to the heterogeneous computing environment, heterogeneous and dynamic network environment, node mobility, and limited battery power. The real time requirements associated with internet of things and cyber physical system applications make the problem even more challenging. The existing limitations of edge cloud and ad hoc networks due to their evolution to their present form and their existing constraints and functionality. What is required is a possible future network/distributed-computing architecture, avoiding the problems outlined.

References

[1] Lam, Simon Sin-Sing. Packet switching in a multi-access broadcast channel with application to satellite communication in a computer network. University of California, Los Angeles, 1974.

[2] Chaisiri, Sivadon, Bu-Sung Lee, and Dusit Niyato. "Optimization of resource provisioning cost in cloud computing." IEEE transactions on services Computing 5.2 (2011): 164-177.

[3] Shi, Yuanming, et al. "Communication-efficient edge AI: Algorithms and systems." IEEE Communications Surveys & Tutorials 22.4 (2020): 2167-2191.

[4] Moustafa, Nour. "A new distributed architecture for evaluating AI-based security systems at the edge: Network TON_IoT datasets." Sustainable Cities and Society 72 (2021): 102994.

[5] McCann, Jeff, et al. "Towards the Distributed Edge–An IoT Review." 2018 12th International Conference on Sensing Technology (ICST). IEEE, 2018.

[6] Sun, Huo-Ching, and Yann-Chang Huang. "Optimization of power scheduling for energy management in smart homes." Procedia engineering 38 (2012): 1822-1827.

[7] Wang, Honggang, Mahmoud Daneshmand, and Hua Fang. "Artificial intelligence (AI) driven wireless body area networks: Challenges and directions." 2019 IEEE International Conference on Industrial Internet (ICII). IEEE, 2019.

[8] O'Callaghan, Maria, Neville Gawley, Michael Barry, and Sean McGrath. "Admission control for heterogeneous networks." 13th IST Mobile and Wireless Commun. Summit (2004).

[9] Mao, Yuyi, et al. "A survey on mobile edge computing: The communication perspective." IEEE communications surveys & tutorials 19.4 (2017): 2322-2358.

[10] Hogan, Bryan J., Michael Barry, and Sean McGrath. "Congestion avoidance in source routed ad hoc networks." 13th IST Mobile and Wireless Communications Summit, Lyon(2004).

[11] Crăciunescu, Mihai, et al. "IIoT gateway for edge Computing applications." International Workshop on Service Orientation in Holonic and Multi-Agent Manufacturing. Springer, Cham, 2019.

[12] Hochstetler, Jacob, et al. "Embedded deep learning for vehicular edge computing." 2018 IEEE/ACM Symposium on Edge Computing (SEC). IEEE, 2018.