Towards the Realization of Converged Cloud, Edge and Networking Infrastructures in Smart MegaCities

Panagiotis Kokkinos[1,2]
[1]Department of Digital Systems, University of Peloponnese, Greece
[2]Institute of Communication and Computer Systems (ICCS), National Technical University Athens, Greece
p.kokkinos@uop.gr

***Abstract*— The emergence of Internet of Things (IoT) and the anticipated 5G/6G applications lead to several challenges regarding the rapid and the efficient processing of massive amounts of data, which are generated, transferred and processed within a city boundaries. Towards this end, the convergence of computing, storage and networking infrastructures operating in a megacity environment is pivotal. In this work, we present several related research innovations regarding the service of user and application demands, the orchestration of cloud and edge resources and the realization of edge infrastructures.**

Keywords*— *Edge, Clouds, Networks, Infrastructures, Convergence, Megacities

I. INTRODUCTION

The United Nations projects that 60% of the global population will be urbanized by 2030, while the number of megacities - cities with population greater than 10 million - grew from 28 in 2014 to 33 in 2018, with the largest one, Tokyo, Japan, reaching 37.5 million population [1]. At the same time the requirements for computing, storage and network resources is constantly increasing through the use of popular on-line services, the Internet of Things (IoT) and smart-city applications.

To ensure high quality services, content distribution networks and service providers, tend to push more and more content, data and services as close as possible to the end-users/devices, to offer high quality digital services, minimizing the experienced latency and alleviated the overhead from other core infrastructures. Hence, a large percentage of traffic flows, start and end within the city boundaries, increasing regional traffic and the provisioned local capacity even with a faster pace than the core-capacity [2].

It is expected that smart megacities will become the main source of data, characterized by massive data growth and processing requirements. Today these requirements are served by a variety of optical and wireless networking and edge/fog/cloud computing technologies and infrastructures deployed in the cities and belonging to different providers, while smaller business owners (e.g., operators) are also deploying their own infrastructures (Figure 1). It is clear that the deployment and the management of these computing, networking and storage infrastructures pose numerous challenges. Today, these are actively investigated by researchers, research projects and the industry all over the world.

In this paper, we discuss some of these challenges and identify three main directions that will play an major role towards the realization and efficient operation of converged computing and networking infrastructures operating in a megacity: i) intent-driven user access, ii) autonomous operation, through cognitive orchestration and closed loop control, iii) mass edge resources deployment.

In Section 2, we present the basic technologies and characteristics of the computing and networking infrastructures operating in a city environment. In Section 3, we discuss about intent- and recommendation-driven cloud services. In Section 4, we describe the cognitive and close-loop operation of the converged infrastructures. Section 5 presents methods towards the realization of edge computing infrastructures. Finally, in Section 6 we conclude this paper.

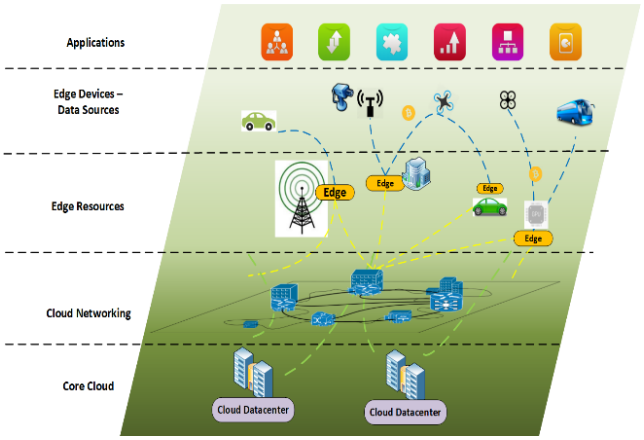

Figure 1: Smart megacities are served by a variety of networking and computing technologies and infrastructures, belonging to different providers.

II. INFRASTRUCTURE

A. Computing Resources

Centralized cloud computing infrastructures are currently handling the processing and storage workload of most applications' and services, rendering cloud computing a key component of modern economy. There is plethora of computing and storage resource offerings by multiple cloud providers like Google, Microsoft Azure and Amazon Web Services (AWS) and smaller ones like Vultr, UpCloud, Linode. The resources offered differ in terms of computing, networking, storage and memory capacity that target different use cases. These also differ in terms of cost, availability, security, region of operation and other parameters of interest. A number of companies also offer multi-cloud services, incorporating multiple resources from more than one cloud providers and offer to their customers the ability to deploy their workloads and store their data in a (semi-)transparent manner.

978-1-6654-9792-3/22/$31.00 ©2022 IEEE

Recently, edge computing has also emerged offering computation and storage at the very edge of the network where data is produced, in order to reduce latency and limit the load that is carried to higher layers of the infrastructure hierarchy. Edge devices range in size and capabilities (Figure 2): micro datacenters (mDC), modular data centers in shipping containers, specialized computing devices (FPGA, GPU) and IoT computing devices (e.g., Arduino, Rasberry Pi, NVIDIA Jetson). Edge together with the traditional cloud resources form an edge-cloud continuum [3] that offers better quality of services and lower monetary and energy costs.

In this way, the provided computing and storage capacity increases when moving from the edge to the cloud, however at the same time performance limitations in terms of processing latency and available bandwidth also appear. Tasks and data are assigned respectively: ephemeral storage and low-latency required computations on the edge, permanent storage and complex computations at the cloud.

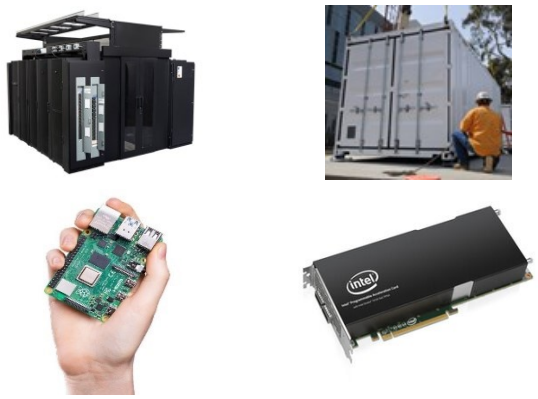

Figure 2: Edge devices range in size and capabilities.

B. *Networking Resources*

Metropolitan networks that interconnect the access networks with the core communication infrastructures are becoming the focus of attention. Both wired (optical) and wireless (5G, beyond 5G and 6G) can be employed in the various network segments.

An intra-city network is usually based on a hierarchical optical interconnection network (Figure 3) that aggregates traffic from access networks into the different metro layers, namely the (i) metro access, (ii) metro and (iii) metro core. Metro networks exploit a number of technologies such as Coarse Wavelength Division Multiplexing (WDM) and Dense WDM for lower and higher layers of the metro hierarchy (wavelengths-colored lines in Figure 3). Also, different switching approaches using filtered-based Reconfigurable Optical Add-Drop Multiplexer (ROADMs) and filterless nodes are being considered. ROADMs use optically reconfigurable and colored components to pass, add and drop wavelengths, while filterless nodes are based on passive splitters and combiners. As a result, the wavelengths dropped in filterless nodes are not filtered out, and the respective spectrum cannot be reused in the subsequent fiber-links. This introduces a trade-off between the nodes' reduced cost and the increased spectrum waste. Passive Optical Networks (PONs) is the most prominent technology in the access network segment. PON's topology structure enables operators to deliver in a flexible manner, high bandwidth connections from a central location (Optical Line Terminal - OLT) to multiple endpoints (Optical Network Unit - ONU) over long distances and low latency.

C. *Integrared Infrastructures*

Software Defined Networking (SDN) and in general the programmability of the infrastructures both wireless and wired is key for their efficient operation and the provision of advanced services. SDN promise to realize network slicing mechanisms through which dedicated computing, networking and storage resources become available, realizing the vision of tailored quality characteristics per service or application over a shared physical and integrated infrastructure. This is challenging considering the various infrastructure segments, technologies and providers involved.

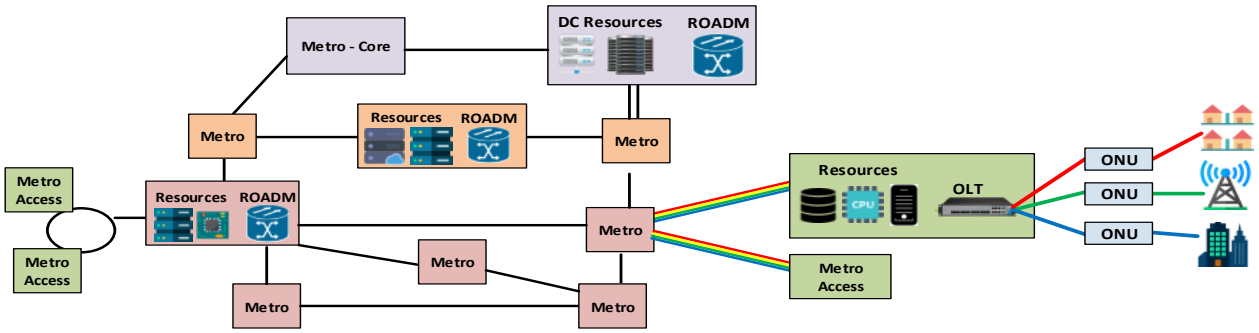

Figure 3: A multilayer converged computing and networking infrastructure.

Computing and storage resources can be placed in various locations or alongside networking components (Figure 3). For, example, selected nodes of the metro hierarchy can be equipped with storage and computational resources that range from general purpose equipment (e.g., servers, mini-racks) to special purpose hardware accelerators (e.g., FPGA, GPU). So, a hierarchical computing infrastructure is created and interconnected with high-speed optical connections through the metropolitan network [4]. This enables the offloading of computational operations close to the data origin, introducing intelligence at the edge and low latency computations for critical applications.

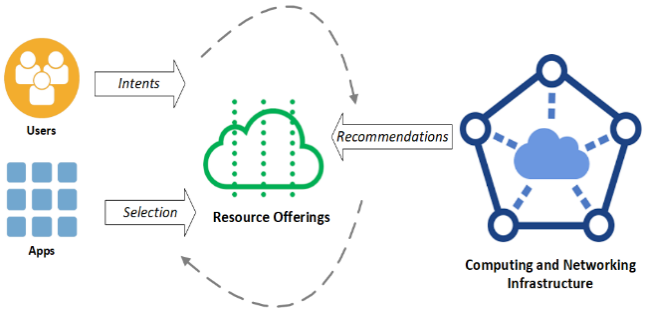

Figure 4: Intent-based requirements declaration and the service/offerings through recommendation mechanisms.

III. Intent Driven Operations

The seamless access of applications and users applications to converged computing and networking infrastructures is pivotal towards the realization of the smart cities vision, by leading to increased application development and deployment. Towards, this end, we suggest the use of intent-based requirements declaration and the offering of candidate services/resources through recommendation mechanisms (Figure 4).

Intent based operations have been proposed, relative recently, by various actors (standardization organizations, providers, academia) as a way to specify requirements on an infrastructure . Intent-driven operations have initially focused on networks [5][6] but recently their application in cloud and edge computing is also investigated [3]. Different actors may can have different perspectives for intent driven operations in the networking or the cloud area, e.g., who can use them and how these can be used. For example, one approach is to include technical details that require some level of expertise, while another approach is to shield from such details, enabling applications and users to express their high-level requirements in an infrastructure agnostic manner without requiring any expertise (e.g., in security). [5] identifies a number of principles that can be common among intents and the different usage scenario: i) intents should be declarative, ii) an easy-to-use interface should be provided for their definition, iii) intents should portable across similar systems.

Overall, intent-based operations enable transparent, adaptive and efficient access to heterogeneous processing and storage resources in the cloud and in the edge that may also belong to different providers (multi-cloud).

Also, in a multi-cloud and multi-edge infrastructure environment the service of a particular workload and storage request may be served with multiple ways by allocating different combinations of the available resources. Intent-based operations can also be coupled with recommendation systems that propose service/resource offerings for users and applications to choose from based on their past and present requirements. The level of user satisfaction, in the form of quality of experience metrics, submitted by the user or an application programmatically or through a user interface can also enable better recommendations.

IV. Cognitive Orchestration of the Edge-Cloud Continuum

The realization of an edge-cloud continuum requires the efficient orchestration of the available edge, cloud and other resources of the heterogeneous infrastructure. This will enable the efficient processing of massive data originating from IoT and other applications [7]. Currently, there is a transition from top-down-designed architectures that apply centralized resource control, towards federations of loosely coupled autonomous or semi-autonomous systems, that are self-organized in a distributed manner.

A. Hierarchical orchestration

The overall orchestration must be performed in a lean, automated, holistic and integrated manner, overcoming the complexity barriers stemming from the heterogeneity of the computing units. Such a hierarchical architecture that enables end-to-end cognitive resource orchestration and supports the transparent application deployment, should include multiple levels of orchestration. In its simplest form there is a high-level/central orchestrator and multiple local orchestrators (Figure 5) that handle the individual parts of the overall infrastructure in terms of the type of the resource (edge, cloud, HPC), the local orchestration software (e.g. Kubernetes, Docker Swarm) for edge and cloud resources and the hardware characteristics (generic server, FPGA, GPUs etc). The interaction of the central orchestrator with the local ones, will be based on the APIs and tools exposed by each local orchestration component.

The high-level orchestrator receives resource and performance requirements from the intent and recommendation driven services (Section III) and coordinates the allocation of resources and the initial deployment of workload in particular sites along with the necessary supplemental actions (e.g. transfer of required data). The hierarchical nature of the orchestration architecture leaves several degrees of freedom to each local orchestrator for serving in an optimal manner the "request", satisfying both the central orchestrator and the resource's objectives. The local orchestrators provide loose coupling between the high-level resource orchestrator and the underlying resources by expressing the general requests to explicit instructions for the local resource, while being responsible for the actual deployment. Also, through hierarchical and distributed orchestration, administrative and geographical limitations are

overcomed, responding locally and rapidly in case of unexpected events.

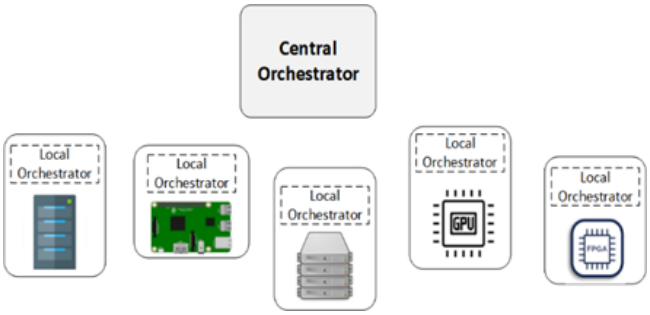

Figure 5: A high-level/central orchestrator and multiple local orchestrators for managing the computing and storage infrastructure.

The complexity and heterogeneity of distributed computing infrastructures pose significant challenges for management and service deployment. Heading to more complex environments, the development of intelligent orchestration algorithms is key to optimize resource utilization for present and future workloads. Towards, this end multi-objective optimization, graph theory, AI/ML techniques, and heuristics need to be exploited to design a set of algorithms aiming to provide different trade-offs between optimality and complexity.

B. Closed-Loop-Operation and Edge-Cloud Hierarchy

In a high dynamic environment where application requirements and resource characteristics can change rapidly it is necessary to couple orchestration together with a continuous closed-loop control (Figure 6), based on the principles of observe, decide and act that will run over an infinite [8]. This will support proactively and reactively re-optimization adjustments in the computing, networking and storage resources in cloud and edge, based on the applications' and the resources' current state. The goal is to automatically determine the most appropriate (computing, storage, networking) resources of the cloud continuum to be used by an application, and then transparently deploy and migrate workloads and coordinate data movement.

A key to this close-loop operation is data-driven hierarchical cloud and network telemetry mechanisms that collect and analyse telemetry data across the distributed edge/cloud infrastructure and the running applications. Advanced real-time telemetry and AI/ML is expected to play a key role in efficient and cognitive resource management, providing load predictions, performance degradation detection and recommendation actions, increasing performance, improving customer experience and reducing costs. Telemetry mechanism can operate in the resources to collect, pre-process and correlate telemetry information at local level, enabling local optimization and re-optimization decisions. These mechanisms will also aggregate and forwards monitoring data to the central orchestrator triggering global management actions. With such a hierarchical approach, the management complexity and the intervention to the infrastructure will be kept as low as possible, so as not to overload the central orchestration mechanisms.

In this way, an edge-cloud hierarchy is created, where processing tasks and data can move both horizontally, across the edge, and vertically, from the edge to the cloud, and vice versa, so as to dynamically adapt to the requirements of the users and applications and the characteristics of the resources in terms of processing latency, storage capacity, location, cost and other parameters of interest. This hierarchy will be powered with mechanisms enabling extreme-scale analytics on the input data, which are necessary for increasing the utilization efficiency of the edge resources, for reducing the volume of data transferred to the cloud and for increasing scalability and energy efficiency.

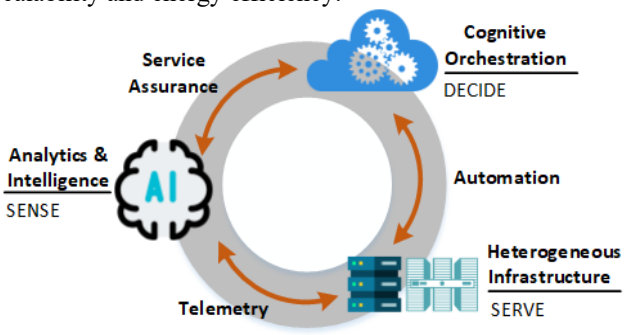

Figure 6: Couple orchestration with a continuous closed-loop control that runs over an infinite.

V. EDGE INFRASTRUCTURES

A. Marketable edge

The realization of any edge infrastructure depends highly on its deployability: deployment needs to be massive for the infrastructure to provide the anticipated edge services. Deployability, in turn, depends on the economic benefits expected and the desired Return on Investment (RoI). Today, on the one hand the existing edge resources are limited and isolated. On the other hand, no actual incentives exist, mainly in terms of uptake of investment, for deploying and supporting new edge infrastructures, while the heterogeneity of the infrastructures hinders application development in the edge.

We support the idea that resources operating in the edge (computing, networking and storage) should be viewed as a new marketable resource, where the available capacity becomes a commodity item able to be traded in this market. The idea of market-based operation of edge infrastructure requires research and development at multiple levels that will lead to the formation of a new kind of cloud-edge services, both infrastructure and platform ones, based on collaborative scenarios and innovative service execution approaches.

Towards this end, we envisage a sharing model for the edge resources, where their computing and storage capacity is used primary for serving the local (resource owners') needs, while the remaining capacity is allocated to the market. In this way, owners of edge resource become prosumers (i.e., producers and consumers simultaneously) of processing and storage capacity, in the same way small energy producers, use solar panels or wind turbines, to cover their energy needs and sell (or buy) capacity to (from) the large energy providers or to their community (micro- grids). On the other hand, cloud and network operators act like large energy producers in the energy grids arena. In this way, edge resource owners will have some clear benefits that will propel them to first buy and then support and upgrade as needed the computing and storage equipment, making it "self-maintained/sustained". These edge resources can also be co-located with local renewable energy sources, for covering their energy needs.

Of course, a key point in any system involving sharing is trust. For this reason, security and trustworthiness mechanisms, operating in multiple layers and providing resiliency against security failures or breaches are necessary. For example, a high secure and distributed blockchain-based architecture [9], can support users sell and buy, resource capacity on the move, while a market operator entity can be responsible for clearing bid/ask procedures in the market, deciding what is the current cost of an edge resource, fulfilling the role of a demand and supply operator, handling inbalances in the market, and other.

B. Resource Aggregation and White Boxes

Generally, edge resources are underutilized and cannot provide the required scalability and resiliency. Through the aggregation of resources (Figure 7) into dynamic set/cluster of resources [10], edge resource owners that are otherwise too weak in the cloud business and too small and below a critical size, are able to participate in the edge market, to trade and obtain better-negotiating power in "selling" their free resources. Higher dependability, statistical multiplexing gains and the ability to provide service level agreements (SLAs) are also obtained, since a reduction in the availability of one resource (due to increase of local use) can be counterbalanced by an increase in the availability of another resource or by the aggregation of more edge resources. The aggregation also abstracts the complexity and the dynamicity of the underlying infrastructures. Also, application/service providers will see their applications and services efficiency improved, even reducing their costs in using central cloud resources, while users will also benefit from low cost applications/services and improved performance (e.g., in terms of latency). Cloud and network providers will be also benefited from the computational and storage offloading to the edge, removing the unnecessary load put in their infrastructures and reducing the related costs. The deployed edge-based infrastructures can also be used complementary by cloud providers, when for some reason exhibit difficulties in serving their customers etc.

Another important parameter is the development of computing and storage edge disaggregated white boxes, with open specifications, based on modular software and on off-the-shelf hardware (e.g., accelerators and SmartNICs, P4 switches). Corresponding white boxes have been proposed for the networking industry [11]. Edge-based which boxes can lead to large cost savings and deployment flexibility to scale to significant levels. Commoditization of edge hardware, starting with disaggregation, has the potential to create an open and competitive market for interchangeable parts that will help cloud-edge achieve economies of scale.

VI. Conclusions

In this work, we present a number of research innovations on progress, for the realization of converged cloud, edge and networking infrastructures. It is clear that the size of current and future megacities will result in data produced, transferred and consumed within the city boundaries. Intent- and recommendation-driven cloud services can enable seamless access to the available infrastructures. The creation of the envisaged edge-cloud continuum requires hierarchical orchestration and continuous adaptation through a close-loop operation that involves monitoring, analysis and optimization. In the end, the degree to which edge resources will be deployed and operate in the cities, will be the key. Efforts for the marketization of edge computing and the disaggregation of involved hardware and software components can provide the boost required.

Figure 7: Aggregated and shared resources utilized by IoT and other applications.

Acknowledgment

This work is supported by the EU research project SERRANO (101017168).

References

[1] Statista, Megacities - Statistics & Facts, 2019

[2] Cisco Visual Networking Index, 2017–2022, 2018.

[3] A. Kretsis, et al. "SERRANO: Transparent Application Deployment in a Secure, Accelerated and Cognitive Cloud Continuum", IEEE MeditCom, 2021

[4] P. Soumplis, et al., "Network Slicing and Workload Placement in Megacities", IEEE ICTON, 2020.

[5] L. Pang, et al., "A survey on intent-driven networks", IEEE Access, 2020.

[6] Intent Classification: https://datatracker.ietf.org/doc/html/draft-draft-li-intent-classification-00

[7] Y. Wu, "Cloud-edge orchestration for the internet-of-things: Architecture and ai-powered data processing", IEEE Internet of Things Journal, 2020.

[8] K. Christodoulopoulos, et al., Observe-decide-act: Experimental demonstration of a self-healing network, OFC, 2018.

[9] V. K. Rathi, et al. "A blockchain-enabled multi domain edge computing orchestrator." IEEE Internet of Things Magazine, Vol. 3, No. 2 pp. 30-36, 2020.

[10] P. Kokkinos, et al.. "Virtual resource consolidation in the edge for 5G networks", IEEE PIMRC, 2018.

[11] E. Riccardi, et al., "An operator view on the introduction of white boxes into optical networks", Journal of Lightwave Technology, 36(15), 3062-3072, 2018.