Edge Intelligence

Schahram Dustdar
Distributed Systems Group
TU Wien, Austria
dustdar@dsg.tuwien.ac.at

***Abstract*—In this talk we discuss the challenges ahead when researching the confluence of Internet of Things, Edge Computing, Fog Computing, and Cloud Computing. In particular we discuss the topics related to research issues in the area of AI and Edge Computing.**

Keywords—Edge Computing, Edge AI, Edge Intelligence

I. Introduction

Today's systems infrastructure is composed out of three building blocks: People, Software Services, and Things. There are two fundamental approaches to discuss such infrastrucures. The first one is the cloud centric perspective. This basically views the Cloud as the center of the world, and everything else is connected to the cloud. Some people call it the brain. In this case, everything is centralized and the brain is the most important thing. In this view, IoT is always connected to the Cloud as all machine learning and decision making is done on the Cloud, nothing goes unnoticed by the Cloud. The second perspective is the *Internet-centric* view. Here, decision making, learning, model building etc. is also done on the edge of a network; partially consolidated and transferred in a federated fashion to Cloud systems, if needed. In this talk we propose to look at the whole compute continuum and utilize IoT, Edge, Fog, and Cloud in all our systems development and engineering efforts. 5G or 6G base stations in the future, they are typically general-purpose computing infrastructures, sometimes they are telecom operator-controlled pieces of equipment, sometimes they are belonging to organizations or even to individuals. Currently, the fog infrastructure base stations, for example, is one example for that. The Cloud has essentially unlimited compute and storage resources, with the full spectrum of cloud services with high availability and lower costs is another important part of the compute continuum. Now, we know that there is a new family of applications, which require extremely low latency and different levels of privacy and security that actually cannot work only with the Cloud; autonomous driving is a good example for that. One needs to have extremely low latencies. Another example is telemedicine in real time surgery. Hence, this means that you have to make sense out of the whole spectrum from the IoT via the Edge, Fog, to the Cloud.

II. Compute Continuum

In this talk I suggest a software-intensive Edge systems focus. We completely rethink the design and the operation of such an environment. Why is that necessary? The main reason is that we have fundamentally conflicting factors concerning the system requirements, which need to be resolved. So on the one hand side, we have latency. There is this an inherent traditional division of the Cloud and IoT, which has different time factors and performance factors, which we can manage better when we look at it from a software intensive side. Secondly, we have computation as an edge resource, which basically means that we need to use edge infrastructures similarly to cloud infrastructures to perform complex infrastructure tasks, such as safety and security. And thirdly, we have the question of locality and mobility, where we can introduce novel solutions to privacy software configuration and system evolution. So the question is, which characteristics of edge computing systems then should be abstracted as first class citizens to the underpinning model?

III. Elastic Diffusion

In this talk we will first understand this from a hypothetical perspective. Is it possible to move the computation and decision making the model creation, etc, closer to where the data is actually being created? In other words, to take proximity, context or capability and energy more into account. That is definitely some an important area of research that people are working on. In this talk, we will focus on what I call the main principle, which is *elastic diffusion*. We will break it down into essentially two points. The first one is elasticity. Elasticity is a property that we know from physics, and it is basically a property, which says something about the resilience. We know elasticity from physics, it's a property of returning to initial form or a state following some deformation. In other words, you put force on a material, it changes its shape. When you take away that force, it goes back to its initial form. That is the principle of elasticity. In, in science in neuroscience for the brain, they call it plasticity. So you learn something and something changes its shape, so to speak. The second principle we discuss is Osmotic Computing. Osmosis is a principle from Chemistry. Molecules flow from higher to lower concentration. Similarly, we aim at mimicking the flow of microservices (functionality) from Cloud, Fog, and Edge devices from and to each other.

In this talk we will discuss what Elasticity and Osmosis mean for the domain of Edge Intelligence and what the fundamental research question entail.

978-1-6654-1681-8/21/$31.00 ©2021 IEEE
DOI 10.1109/ICWS53863.2021.00097