Towards next generation digital infrastructures to contribute to a more resilient and sovereign European economy

For several years now, we have been witnessing the irreversible generalization of a "cloudification" of most economic sectors in Europe. The use of digital infrastructures (covering the cloud to IoT continuum) is now essential to ensure the competitiveness of European service and manufacturing industries in the increasingly global world market. However, this trend brings new threats as these infrastructures are dependent on a resilient and secure energy and network infrastructures to let users to access remote services in a reliable and trusted way. Although these assumptions were taken granted the last twenty years, the today and probably forthcoming geopolitical threats make these assumptions invalid. The excessive centralization of today's infrastructures represents an Achilles' heel for the European economy. The threat is all the greater for Europe because it often depends on infrastructures that are not located on its own territory. Recent attacks on energy infrastructures, as well as massive cyberattacks, by unidentified organizations are clear illustrations on the kind of dangers the European economy might face in the future.

It is therefore a paramount importance to rethink the way we build and deploy computing and data storage infrastructures and their associated software services by promoting a more decentralized approach to avoid as much as possible single point of failures. Edge computing is a first step to change the way future Cloud platforms will be built. Europe has already invested in several R&I projects dealing with Edge computing, but mainly with an IoT perspective. Such an effort must be pursued in a broader perspective by addressing a larger spectrum of service-based applications. Making the deployment of services near to their users' location or processing data as close as possible as their production location will generate several advantages such as reducing the latency, preserving privacy, ensuring resilience, and even more lowering the energy consumption. However, although this concept has been in vogue for several years, Edge Computing is still in its infancy and many issues require cutting-edge research.

Resource management

Compared to a classical centralized cloud infrastructure for which is it easy to have a global view of the resource availability, edge computing is based on massively geo-distributed ICT infrastructures. In that context, the management of geo-dispersed computing and data storage resources is a very complex issue due to high variability of network latencies bandwidth, network failures, heterogeneity of these resources, churn rates which is the percentage of resources that leave the federated infrastructure for a given period, etc… It is thus necessary to propose new models to estimate/predict resource availabilities able to serve the need of hundreds of thousands of users. A promising direction is to consider this computing continuum (from the centralized data centers to the edge devices) as a set of complex but autonomous systems capable of satisfying local requests even in case of network partitioning. Machine learning techniques and advanced probabilistic models are good candidates to achieve this goal. Adaptation of traditional approaches to supervision itself, considering the plasticity and great dynamics of the global system while taking into account the specific needs of the various tenants. At this level of scale and complexity, resources will have to be dynamically managed, monitored and automatically reconfigured thanks to distributed probes. New

methods will have to be proposed to monitor both data exchanges, most of which are encrypted, at different levels of the software stack, and the behaviours of cloud tenants and/or administrators.

Software engineering

The management of the life cycle of service-based applications is usually carried out with a low level of abstraction. This leads to a lack of optimization because the management is done by the applications themselves, which often settle for simple solutions due to a lack of expertise. Elasticity management mechanisms in Clouds are an example of how a higher level of abstraction (e.g. by providing rules for adding and/or removing VMs instances to host services) simplifies resource management while allowing better optimization of their use. One of the challenges consists in abstracting the description of the whole application structure to be able to globally optimize the resources used with respect to multi-criteria objectives (price, deadline, performance, energy, etc.). Given the complexity of the choice, it seems important to decouple as much as possible the description of the application structure from the infrastructure in order to use external services to adapt the application. This also provides a framework to address the challenges associated with the reconfiguration of applications so that they automatically adapt to the use of resources. This requires defining novel models and associated languages to describe applications, their objective functions, placement and scheduling algorithms supporting system and application-level criteria, etc. The encapsulation of applications and their environment in light and mobile containers will allow for greater migration possibilities in a more decentralized cloud infrastructure and facilitates the use of distributed resources. In addition to being efficient, such automated systems must also be secure and frugal, especially when it comes to managing failures.

Energy consumption

Energy consumption of large-scale cloud infrastructures (i.e. data centers) is known to be an important issue due to its environmental impacts. Data centers hosting services still rely heavily on oversized approaches to achieve high quality of service. The decentralization will give new opportunities to lower down the energy consumption as mini-data centers will be installed in urban areas within which there exists opportunities to use wasted energy (cooling of the infrastructure) to the benefit of people who live in this urban area. However, to ensure the resilience of the services deployed in such infrastructures, it will be necessary to deploy multiple instances of a given service to several data centers to tolerate failures. A trade-off will have to be found between resilience and energy consumption. To do so, there is an urgent need to address new challenges to guide future designs of digital infrastructures and service deployments having digital sobriety in mind:

- End-to-end analysis and energy management of large-scale hierarchical infrastructures considering the processing, networking and storage aspects.
- Monitoring and profiling of virtual resources (software containers, virtual machines) on heterogeneous infrastructures (CPU, GPU, ...).
- Trade-off between energy efficiency and other performance metrics in virtualized infrastructures, placement of tasks between Cloud, Fog and Edge infrastructures while ensuring resilience and lowering energy consumption.
- Eco-design of applications and digital services: frugal AI, blockchain, streaming... exposure of the energy characteristics of the different services to offer choice to orchestrators.

Cybersecurity, privacy, trust

Cloud infrastructures, like any information system, requires the establishment of specific mechanisms to ensure the confidentiality, integrity and availability of data, applications, and services. Centralization has one main benefit: a

strong team of cybersecurity experts oversees the infrastructure to maintain the required security level to handle threats and manage attacks. However, decentralization will increase the complexity of the infrastructure and will make this task much more difficult. It thus important to think about new solutions to ensure the security:

- Adaptation to the constraints induced by the regulations of the different locations of the machines constituting the Cloud. Failing to repatriate the data in a trusted location, it is necessary to be able to process this data in encrypted form as allowed by homomorphic and functional encryption mechanisms. Research work is still needed to achieve the required efficiency in terms of processing time, ciphertext volume and key management.
- Integrity of the data can benefit of a geo-distributed infrastructure thanks to their replication in several locations. The protection of personal data (confidentiality) may result from their location in certain places, or even among the people themselves (personal Cloud) who thus regain sovereignty over their data. Therefore, distributed algorithmic work is necessary to allow the exploitation of these data.
- Strengthening data protection also requires innovation in systems, languages and hardware. In recent years, processor manufacturers have been offering new instruction sets to encrypt data on the fly (e.g., Intel SGX or ARM Trustzone or ARM CCA - Confidential Compute Architecture). These instruction sets ensure that the data can only be accessed by the codes deployed by the data owner, which prevents private data from being exfiltrated by hackers or cloud operators. Unfortunately, applications hardly ever use these instruction sets because cloud software layers and programming languages are incapable of exposing these abstractions in a simple way. It is therefore necessary to revisit programming languages to offer simple abstractions that can be easily used by cloud application developers.
- Traditional security mechanisms must be adapted to deal with specific threats for the different types of cloud paradigms (SaaS, PaaS, IaaS) such as flaws in sandboxing technologies (VMs, hypervisors, containers, etc.) and orchestrators, virtual network technologies (SDN, NFV), programming or access interfaces: (1) Adapting traditional prevention approaches to these specific threats with end-to-end encryption and application security at the language level; (2) Adaptation of security policies to a more complex environment. Real challenges arise to ensure overall consistency, to deal with inconsistencies/incompatibilities between tenants or between administrators/managers.