H-CLOUD

Cloud Computing in Europe

Appendix 12

Infrastructure and Technology Landscape Analysis

15 July 2020

h-cloud.eu

Title	Cloud Computing in Europe. Appendix 12: Infrastructure and Technology Landscape Analysis
Lead Editor	Mark Dietrich
Contributors (in alphabetical order)	Carla Arend, Federico Facca, Mark Dietrich
Version	0.7
Date	15 July 2020
Confidentiality Notice	**Confidential** - The information contained in this document and any attachments are confidential. It is governed according to the terms of the project consortium agreement

Important Notice: Working Document

This briefing is an annex to the Green Paper v 0.7. The Green Paper is an outcome of the H-CLOUD project, a Coordination and Support Action that has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement n. 871920.

CONTENTS

1 Services being offered

1.1 Different meanings of “cloud”

"Cloud" is also used to mean "infrastructure hosted outside of the organization." Most commercial cloud suppliers are providing “cloudy” service using their own off-site infrastructure, so the term “cloud” sometimes signals the move away from on-premise infrastructure.

The business model of public (commercial) cloud suppliers is perceived as a "pay as you go" revenue model, so the term “cloud” can also refer to a shift from CapEx+OpEx to a more pure OpEx cost model. (This shift is not unique to IT service providers, e.g. General Electric can supply jet engines in a similar “propulsion as a service” business model.) In practice, many cloud services are sold with year-long or multiyear contracts (for example reserved instances), which makes the commercial cloud models less flexible than what “pay as you go” suggests.

Public (commercial) cloud suppliers achieve economies of scale by hosting all clients on a shared infrastructure, rather than a dedicated infrastructure hosted in a commercial data centre or on-premise. Essentially cloud suppliers are “reselling” the same physical infrastructure to multiple clients over time -- for example, by assigning free capacity to a client when they request it, and releasing that capacity when the client releases it. This applies to all three cloud service models: from Infrastructure as a Service to Software as a Service. In SaaS services, the service itself has been designed in a “multi-tenant” model, storing all client data in the same database and giving all clients access to the same running software, but separating the clients’ experiences and possible actions by restricting clients to their own data and related processing steps. (Again this business model is not unique to the cloud. IaaS resources are analogous to airplane seats or hotel rooms. Financial institutions “resell” currency deposits through multiple loans.)

"Cloud" is also used to refer to the various software technologies used to deploy IT workloads in a “cloudy” way (typically using virtual machines or containers, and eventually microservices and functions). Turning again to NIST, “*The cloud infrastructure can be viewed as containing both a physical layer and an abstraction layer. The physical layer consists of the hardware resources that are necessary to support the cloud services being provided, and typically includes server, storage and network components. The abstraction layer consists of the software deployed across the physical layer, which manifests the essential cloud characteristics. Conceptually the abstraction layer sits above the physical layer.*” For our purposes, we refer to the technologies involved in this abstraction layer as “cloud-style” -- yielding “cloud style deployment” and “cloud style technologies.” etc..

S-T Challenge 1: The term “cloud” has a wide range of explicit and implicit meanings.

1.2 Cloud Service Models

The many kinds of cloud infrastructure -- public, community and private -- can be analyzed through the common lens of the services they provide to clients, organized in the three models defined by NIST in 2011.

1.2.1 Infrastructure as a Service (IaaS)

Services closely tied to the physical technologies themselves, notably various types of compute, storage and file systems, and networking, typically exposed as virtual machines (“instances”) and managed with virtualization tools such as OpenStack. In the IaaS model, the user (client organization) is responsible for the infrastructure, even if it is virtual -- that is, additional compute “instances” must be requested (provisioned) when demand increases, additional storage configured, etc. In the cloud all of these steps can be automated -- client staff don't have to purchase and install more physical servers or hard drives, but the virtual infrastructure still has to be managed.

In IaaS, the client also needs to assemble and manage the "stack" of software components needed on each virtual machine, including database software, library and package dependencies, communications interfaces, etc.

In Europe all the major public (commercial) cloud suppliers (including US based) offer IaaS, from a range of standard compute instances and storage solutions, to more specialized solutions, including database tools, ingest/storage of streaming data, edge device management, etc.

1.2.2 Platform as a Service (PaaS)

PaaS is the developer-focused layer, where cloud service providers offer building blocks that developers can use, for example database services, developer environments, natural language interfaces, digital assistants, etc. At the PaaS layer, developers can consume the services in their applications and don't need to manage the underlying infrastructure. In PaaS, cloud-based business functions are developed in the selected PaaS environment and can be configured to scale up or down automatically -- the client will be charged for this but the client doesn't have to track utilization and adjust configurations directly.

The PaaS market includes all of the revenue of IT capability in the application development and deployment primary software market when it is composed and delivered as a cloud service. PaaS provides integrated (i.e., made up of multiple discrete software functions) services organized around the tasks of application development and life-cycle management; application deployment; code testing, quality, and application life cycle; data management; and integration when they are provided as a service delivered through public cloud or specifically designed to be included in a private cloud implementation.

When PaaS solutions are designed and offered as private cloud–ready solutions, IT assets are typically owned and managed by the customer (there are models available for premise-based private cloud remote management by professional services firms) and dedicated to a single customer. Virtualization and dynamic scalability can help optimize resource utilization but do not change any of the underlying assignments of key roles and responsibilities. When PaaS is offered as a public cloud, customers use shared runtime platform assets, ownership, and management of the platform shifts from the customer to the service provider, and the use of platform capabilities is presumed to be shared.

Public cloud PaaS is packaged as configurable, turnkey offerings sold directly from intellectual property (IP) owners/providers, cloud OEM partners/service providers, systems integrators, and a variety of other mechanisms. When offered with underlying infrastructure, PaaS frequently includes access to system infrastructure capability such as workload automation, scheduling, change and configuration management, storage management, security, and network management.

Whether designed for public or private cloud, PaaS exhibits eight basic cloud characteristics:

- Solution packaged
- Shared/standard services
- Elastic resource scaling
- Self-service
- Elastic, term-based pricing (no perpetual license)
- Ubiquitous (authorized) network access
- Standard UI technologies
- Published service interface/API.

Two major PaaS offerings are Google App Engine and RedHat OpenShift. App Engine allows clients to develop and deploy applications in a range of programming languages (e.g. Python, Ruby, PHP, Java) and integrate with a number of common tools such as SQL/NoSQL databases and authentication systems.

PaaS services are very similar to “Container as a Service” (CaaS) services which can directly execute containers (such as Docker) that have already been developed by clients.

In Europe all the major public cloud suppliers (including US based) offer container-focussed services, as well as PaaS. IaaS vendors are adding PaaS capabilities to their offerings, as PaaS services create a higher level of customer “stickiness” compared to Iaas services. At the same time, we also see SaaS providers offer PaaS platforms to give customers the possibility to customise the SaaS services and to create extensions to the SaaS services or new SaaS services using the platform’s building blocks.

1.2.3 Software as a Service (SaaS):

Software-as-a-service (SaaS) application services are based on a service composition and delivery model made up of a utility computing environment in which unrelated customers share a common application and infrastructure that is managed by an independent software vendor (ISV) or a third-party service provider. The code or intellectual property of the service is typically owned by the software-as-a-service ISV. There are many emerging models for SaaS providers to leverage third-party infrastructure, business services, and other providers as hosting, selling, fulfillment, or support partners, and many new models are forming far beyond the comparatively well-understood direct versus tiered distribution models of packaged software. These new models provide customers with access to and consumption of software and application functionality built specifically for network delivery and hosted, provisioned, and accessed by users over the Internet.

This delivery model goes well beyond prior online delivery approaches — combining efficient use of multitenant (shared) resources, radically simplified "solution" packaging, self-service provisioning, highly elastic and granular scaling, flexible pricing, and broad leverage of internet-standard technologies — to make offerings dramatically easier and generally cheaper to consume.

Fully integrated software suites typically centre around specific activities. Major examples include customer relationship management (CRM), enterprise resource planning (ERP) and financial accounting, as well as productivity tools like G Suite by Google, Microsoft Azure’s Office 365 and DropBox. Microsoft’s Azure offerings are largely SaaS, supported by flexible IaaS capabilities that are also available to clients. The dividing line between SaaS and PaaS is blurred by products like MS Dynamics, which is a full-featured CRM system (SaaS) that allows customization and programming as if it were a PaaS product. Salesforce.com’s SaaS CRM tool is similar, allowing customization and programming through its Force.com PaaS capability.

Except for SAP and Visma, the top ten commercial SaaS vendors are US based. There is a long list of small European SaaS vendors, as the market is very fragmented. SaaS tools for training and deploying AI models are being offered in specific domain areas (for example optimizing marketing campaigns).

Many of the services provided in research clouds could be classed as SaaS, including science gateways and research platforms, and data/artefact management tools (archiving and replication, search and discovery, policy management).

1.2.4 Consulting and other human-based Services:

Effective use of cloud services depends, ultimately, on personnel trained to select, configure and operate those services. Services can range from initial outreach, advice and consultation, training, helpdesk to extended support and implementation and operation offerings.

The major public (commercial) cloud suppliers vary in their approach to service. Amazon and Google provide extensive documentation and knowledge base support to clients in their use of their respective services, as well as training and conferences for cloud professionals, but do not provide significant consultation, help desk or direct operational assistance. Microsoft offers more assistance than Amazon or Google. Smaller cloud suppliers (by market share) often distinguish themselves by offering more support, guidance, and direct operational assistance (e.g. IBM).

Major consulting and IT service firms (e.g. Deloitte, Accenture, Tata, Wipro, Atos, T-Systems, TietoEvry, Capgemini, Computacenter as well as cloud-native services firms like Cloudreach, Nordcloud, Reply, etc) have built significant practices around assisting clients to implement and manage their cloud-based IT infrastructure. There are also a significant number of SMEs in Europe offering cloud-focussed IT services.

1.2.5 Cloud Management Services

In addition to these user-facing services, other services are available to help users (and cloud managers) manage the cloud services themselves. Some of these are unique to each cloud supplier, while some generic tools (e.g. based on OpenStack) can be used to manage services from different suppliers (e.g. as long all the suppliers complied with the OpenStack standards), and new services are being developed to manage the orchestration of services among multiple suppliers in a "multi-cloud" environment. The primary categories here are:

- Virtualization Tools: OpenStack and VMWare
- Containerization: Open Container (from Docker), LXC, (Singularity and Shifter for HPC environments)
- Resource Management/Orchestration Tools: Kubernetes, Apache Mesos.
- Scheduling Tools (for queue-based resource management): SLURM, Torque, PBS Pro.

Finally, there are a range of services and functions that cloud suppliers themselves need to manage the cloud services. For integrated cloud suppliers (e.g. Amazon AWS) these services are not visible to the public. However in federated cloud environments, these services and functions are required by federation managers to ensure smooth interoperation of the services being provided by different suppliers, for example configuration management databases, IM (an infrastructure management tool developed by INDIGO-Datacloud).

1.3 Cloud Data Protection Security and Privacy

Shared responsibility models between the cloud service provider and the customers are clearly defined by the cloud service provider, typically in cloud services agreements or data processing agreements. However this division of responsibility is often not clearly understood by the customer. IaaS cloud service providers ensure that the infrastructure, up to the virtual machine or container layer, is secure, but any applications or data that the customer runs on this infrastructure, is the customer's responsibility, as well as the processes that the customer supports with the cloud service. Ultimately, the customer has the responsibility to handle the data correctly.

Such shared responsibility models raise questions about how data is kept secure in any IT system that contains "cloudy" services or technologies. Physically hosting infrastructure outside the boundaries of an organization raises other questions.

These questions can be organized as follows:

- **"Broad network access"** implies that most "cloudy" systems are visible to the public (e.g. from the web)[1], and will require authentication and authorization systems to control access. Data being transferred to/from the user, e.g. populating a web page with customer information, must also be protected from interception and misuse.

- **Public Cloud IaaS/PaaS services**: A public cloud supplier's physical infrastructure will be virtualized for a specific client's exclusive use while assigned to that client, but the physical infrastructure is still controlled by the cloud supplier and can theoretically be accessed by its

[1] Some private clouds are accessible only through virtual private networks (VPNs) which enforce their own authentication and authorization process.

employees, making it possible for unauthorized access to the client's virtualized software and data. In some situations clients may want to allow such access to enable the cloud supplier to provide technical support.

There are also concerns about access to software and data that may persist in physical infrastructure after it is released by the client for re-use by others.

Encryption is critical to mitigate against unauthorised access. Customers need to bring their own encryption keys and manage them to ensure that their data is not accessed without permission. That is also true for data that is being subpoenaed under the US CLOUD Act to be used for prosecution in the US. Delivering data to the US courts under the CLOUD Act will put European organizations in violation of GDPR, as these two regulations have not yet been harmonized.

Even without unauthorized access (intentional or not) by cloud supplier employees, client organizations can fail to configure their virtualized resources to fully prevent other types of unauthorized access. Since the client organization is responsible for developing its own applications for deployment to the cloud, it must take direct responsibility for their security -- the cloud supplier is not responsible for any vulnerabilities the client itself may create. As a trivial example, clients must ensure that all of the software components used within their virtual machine images or containers have had all the latest security patches applied. At the same time, clients can implement additional tools to protect data, such as using encrypted data storage and managing the related "keys" in a robust way.

In PaaS deployments, clients have less control over these choices. The cloud supplier will provide access to various "secure" tools, but the client is responsible for using those tools in a way that protects data against unauthorized disclosure or use. It is the client's responsibility to create correct access policies, as well as building their entire software stack in a secure way. There could be vulnerabilities in the PaaS layer (the CSP's responsibility), the client's layer(s), as well as the interaction between the two (hard to define responsibility).

  - As one example, if a client would like to enforce encryption of data at rest in their solution, the client needs to choose a PaaS provider that offers this at the outset of the development process -- since building encryption on top of an existing PaaS environment is less secure.

- **Public Cloud SaaS services**: Multi-tenant services rely on consistent implementation of pervasive security controls throughout their application(s). Clients cannot use techniques like data encryption for data protection, and must rely on the supplier's confirmation of security, normally through their own compliance with relevant security certification standards and maintenance of corresponding certifications. For example, Salesforce.com maintains compliance against roughly 25 certifications, standards and regulations. Large SaaS suppliers actively scan for security vulnerabilities in their own applications, in integrations with other applications, and in any software dependencies (including underlying IaaS/PaaS software components). Where clients can "program" their own customizations in a PaaS environment (as in both MS Dynamics and Force.com) it is possible that vulnerabilities may be created that will not be tested/examined/corrected by the SaaS vendor.

- **Private Cloud deployments** are intended for use only by the client organization. Clients can contract with public IaaS cloud suppliers to create such environments ("managed private clouds"), with the vulnerabilities described above. Alternatively, clients can construct their own private cloud infrastructure with purchased hardware combined with virtualization tools (such as OpenStack), but the client must take direct responsibility for security configuration and maintenance.

- **Multi-cloud and Hybrid Cloud deployments** raise additional concerns through inter-application communication, and the need to protect data in transit between these applications.

A number of EU R&I projects have addressed important topics related to data protection, security and privacy ("DPSP"), and these projects formed a DPSP cluster to consolidate their perspectives and expertise. Three reviews in particular attempted, respectively, to identify synergies among proposed solutions, to identify common research and innovation challenges that must still be addressed, and to illustrate future technology options in support of the free flow of data:

Synergies: (See Research and Innovation Challenges in Data Protection, Security and Privacy in the Cloud: Map of synergies of the clustered projects) Despite the title, this report highlights the lack of synergies among the clustered projects, suggesting that additional work is needed to evaluate the different solutions, and either pick the most promising or motivate integration of synergistic solutions.

Future Challenges: (See Challenges for trustworthy (multi-)Cloud- based services in the Digital Single Market) The research challenges were identified by 16 projects in the DPSP cluster as research gaps preventing fulfillment of Initiative #14: *Initiatives on data ownership, free flow of data (e.g. between cloud providers) and on a European Cloud* of the Digital Single Market initiative. It is unknown whether these challenges have been taken up by any H2020 projects.

Future Directions: (See Cloud technology options towards Free Flow of Data) This report consolidates the learning of the DPSP projects in the context of roughly 50 use cases from industry, government, research and the technology sector. Several technologies were proposed that hold promise for better "DPSP", including:

- Secure Web Containers (SPECS project).
- Federation-as-a-Service (SUNFISH)

Potentially an open-source effort to create solutions that could be employed by EU firms to deploy GDPR compliant applications to the cloud (any cloud). The Gaia-X initiative is also trying to define a secure data layer that will conform with European values and regulations, and will develop a certification scheme for "Gaia-X compliant nodes".

As described in Appendix 10, there are several efforts to create cloud "codes of conduct" governing the behaviour and security approach of public cloud providers. Given the "shared responsibility model" described above, selecting a cloud provider that is governed by such a code of conduct still does not ensure that client implementations ON that cloud provider will be kept secure, or that any related data is being managed consistent with the GDPR.

Finally a new category of cloud software is emerging that claims to provide clients with a "data layer" that is GDPR compliant (see Section 2.4 below).

1.4 Cloud Adoption and Migration Dynamics

Other sections of this report characterize the extent of cloud adoption -- generally concluding that adoption has been slower in Europe than, e.g. in the US. It is useful to think through what "cloud adoption" actually means for different organizations -- their different circumstances highlight different obstacles to cloud adoption which could merit different responses from the EC to increase overall adoption.

Client organizations are developing and updating their IT infrastructure plans based on available cloud services and their internal experience with on-premise solutions. Clients can turn to cloud suppliers to meet some or all of their needs, and increasingly they consider a mix of on-premise and cloud-based services to balance a number of criteria (CapEx/OpEx, predictability, reliability, security, privacy, agility, flexibility, access to innovative services, access to new features and functions, evergreen implementations, support for agile / devops working processes and methodologies, demands by developers and digital business units etc.). Some workloads will stay on-premise (not private cloud) because they are legacy applications (perhaps mainframe based for very large organizations) that will require effort and time to convert to cloud-style deployment formats. Other workloads may have to stay on-premise (strictly controlled by the client) for regulatory, privacy or confidentiality reasons, but sometimes these applications are migrated to cloud-style formats to simplify workload management, even if they will still be deployed to a private cloud.

One criteria affecting cloud adoption is a client organization's ability to architect, develop and deploy software in the cloud-style format required, and to sustain the maintenance and upgrading of that software in that format over the planned lifetime of the application. This often requires a change in methodology for the user organization (e.g. moving from SDLC to agile development, refactoring monolithic corporate applications into smaller software components that interact with one another, integrating "off-the-shelf" with internally-developed functions).

In the SaaS market, client organizations often combine SaaS services with other functions developed and managed by the corporate IT department, but deployed on IaaS or PaaS cloud infrastructure, creating a "multi-cloud" solution in which communications between the "packaged service" (such as CRM) and corporate-developed applications are managed by mechanisms such as APIs or message queues. Client organizations often need to integrate multiple SaaS functions, running on different physical cloud infrastructure, and need to tie this multi-cloud structure together with integrating tools such as an enterprise service bus (ESB). (Many other integration architectures are used.)

1.4.1 How Large Clients Adopt the Cloud

Large organizations may find that a mixed cloud structure ("hybrid cloud") offers the best way to balance flexibility and user-centricity with compliance and control. For large organizations, even those with internal IT capabilities, clients can contract with outside consultants to guide the redesign process, guide the change in development methodology, train staff on the new processes and methodology, and leave the client able to support the new systems on their own.

- When large clients convert their workloads to cloud-style formats (such as containers or virtual machine images), it then becomes possible for them to choose the "best" hosting location (private cloud, one or more public cloud providers) for each workload according to workload-specific criteria. There are no standard tools to automate this decision process, but several H2020 projects have developed prototype solutions to this problem.

  Large organizations use contractual structures to manage relationships with their cloud suppliers, possibly creating multi-party agreements where different cloud suppliers must interoperate to achieve the client's business objectives. Specific suppliers may be chosen because they adhere to one or more standards that support interoperation, but ultimately the client (or its IT consultants) is responsible for successful integration of the different components.

1.4.2 How Smaller Clients Adopt the Cloud

By contrast, smaller organizations may not have the scale to make hybrid cloud solutions feasible, and will have to choose between public cloud and private cloud, possibly mixed with cloud-based SaaS functions if the economics make sense. Each of these options has obstacles:

- Public cloud adoption would require the organization to migrate its applications to the cloud, which in turn may trigger a need to re-architect significant portions of their IT capability, potentially leveraging off-the-shelf or SaaS components instead of migrating previous customized solutions. Typically in house resources are not available to deal with a large project of this nature, and outside consulting advice could be too expensive.

  - The data protection, security and privacy concerns described above also hinder the adoption of public cloud by smaller organizations -- clients must clearly take responsibility for these issues themselves in a public cloud environment, increasing the obstacles to adoption.

- Private cloud might simplify adoption for a smaller organization, or where security concerns are more significant (e.g. a clinic or physician practice). For example, organizations may implement a private cloud "stack" on their own hardware within an outer security firewall. The private cloud deployment approach allows the client to take advantage of new cloud-compatible software solutions, as well as modernizing its development approach and improving retention of IT personnel. The outer firewall provides a first line of defence against intrusion and data loss, and reduces risks associated with keeping the full "internal" software stack completely impenetrable against attack.

- Integrating SaaS solutions into a "multi-cloud" exposes clients both to potential weaknesses in the individual SaaS solutions as well as in application integration. The EU Code of Conduct helps to manage this risk but does not protect against vulnerabilities in application integration points.

Comments received in H-CLOUD's webinar of experts on supply side challenges indicated that cloud adoption may also be hindered by the perceived cost of moving data to and from the cloud, which also creates penalties for switching cloud providers. Other experts noted the need for tools to manage multi-cloud implementations and the value of a shared marketplace in which different stakeholders can collaborate rather than compete.

S-T Challenge 2: Organizations are hesitant to adopt cloud technologies because of the risk and costs associated with complying with EU privacy and security regulations, including GDPR.

S-T Recommendation 2: A "GDPR compliant" cloud abstraction layer for cloud deployments (that sits above the physical infrastructure) might be useful for small organizations looking to deploy cloud technology.

1.4.3 How Collaborating Clients Adopt the Cloud

Organizations that need to collaborate with other organizations face additional challenges implementing secure interoperation of multiple IT capabilities. Different organizational circumstances can work for and against this collaboration:

- Benefits from (or need to) share data, constrained by laws/regulations/policy preventing that sharing.
- A desire for standards, but worries about losing control to another organization.
- Benefits from (or need to) share resources, but needing to retain control over critical assets.

The Sunfish project (see Appendix 17) has identified several use cases where secure data access and sharing are needed, but the stakeholders want to work in a "peer-to-peer" fashion to achieve this, rather than creating a new entity (to hold data through a shared service) or losing autonomy to an existing entity (to dictate a solution to stakeholders). Sunfish has developed a "federation-as-a-service" approach to address these kinds of problems. This could be promising in a variety of demand scenarios, including research, healthcare, public administration, as well as scenarios where future dynamic edge capabilities will be needed, and those capabilities are not necessarily owned/controlled by a single entity.

S-T Challenge 3: Various groups of organizations need to share sensitive data with the group, but do not have the tools or frameworks to do so while complying with EU privacy and security regulations, including GDPR.

S-T Recommendation 3: Support development of "GDPR compliant" tools and/or frameworks that enable secure access to and sharing of distributed data. These tools might function through peer-to-peer software components that are certified to be GDPR compliant, or through participation in coordinated structures such as federation.

2 Technology Landscape

2.1 IT Hardware Technologies

The main EU suppliers of the hardware to build cloud datacenters or private clouds are the same as operate throughout the globe. Original device manufacturers include Lenovo, Dell Technologies, HPE (including Cray and SGI), IBM, Lenovo, Cisco, ATOS+Bull, Fujitsu, Oracle, Huawei, Hitachi, NetApp, Arista and Juniper. There has been consolidation and increased competition among major hardware manufacturers. Major IaaS providers (Google, AWS) also design and develop their own compute servers to meet their specific needs.

Even at the chip level there has been increased competition: Intel continues to dominate but is under attack. AMD and Nvidia GPUs are gaining market share. ARM-based chips are making inroads (Cavium). FPGAs/ASICs are playing increased roles in specialized applications. With increased demand for AI performance, many new AI-focussed chips have been developed and deployed, from Google's TPU, to Graphcore [UK], Cerebras, etc. The fragmentation of this competitive landscape raises questions about how the European Processor Initiative (EPI) will break into this ecosystem.

Data centre interconnect technologies (networking within the data centre, connecting servers and storage) also present a rapidly shifting landscape: Nvidia bought Mellanox (Infiniband, historically the dominant interconnect for high performance computing). Cray (part of HPE) is now winning leadership class supercomputing projects with its proprietary Slingshot interconnect. 10-100Gigabit ethernet interconnect is a commodity that dominates in cloud data centres, but increasingly customers are looking for enhanced ethernet (with proprietary software features such as flow/congestion control). New interconnect paths (NVlink, enhanced PCIe) are being introduced to physically link nearby servers/nodes, creating mesh networks as an alternative to traditional "hub and spoke" topologies, in order to provide additional performance improvements.

There is continued evolution of both chip and node architectures, both with the explosion of chip architectures noted above (CPUs [Intel, AMD, ARM], GPUs [NVidia, Intel Xeon, AMD], DSPs, FPGAs [Xilinx], specialist AI chips), as well as new designs for "nodes" (what used to be called the motherboard), varying traditional decisions about where to place memory, how is memory shared, how pieces are connected, where are the paths to other nodes, are there new interconnect paths for communication with neighbouring nodes. The scale of many cloud providers allows them to define/design their own chips and nodes (Google and AWS) or packaging (OVH in Europe).

Data storage is moving away from spinning disks to solid state, which in turn is enabling higher storage densities, different performance ranges, and increased need for "software defined storage" to create smart storage capabilities. File systems sit above actual storage, encompassing traditional network attached storage, high performance block-oriented file systems (Lustre, GPFS, etc.), object-oriented storage (Ceph, S3 (AWS) interfaces), Hadoop File System (HDFS) and similar systems optimized for data analytics, and generally separating metadata from actual data. There is a growing need for hierarchical storage management, allowing files to be placed on the "right" storage system based on access, latency, preservation requirements, and balancing the use of fast but expensive solid state disk, slower but cheaper spinning hard drives, slower and cheaper still tape systems.

2.2 Networking

Physically, the Cloud depends, in the simplest terms, upon computers connected together by (data) networks. Networks are engineered to deliver a satisfactory quality of service to the users of the network.

Network design is a fundamental part of the user experience of Cloud computing; when the network 'misbehaves' people have a poor Cloud experience. The engineering design of networks takes into account current and anticipated traffic volumes in each sector. The client server reality of data networks means that, regardless of logical topology, the physical reality of any network is a tree structure with the highest capacity circuits deployed in the core and the lowest capacity circuits closest to the end-points. Note that, "lowest capacity" is a relative measure, and low capacity circuits are, by no means, equal: this feature is part of the network engineering process.

Cloud services have two important aspects: a development and deployment aspect and a delivery aspect. The cloud SRIA will inevitably focus mainly on the development and deployment face. However, to ensure that society gains benefit from future developments in Cloud computing, it is important to include the delivery side in the analysis.

In terms of user experience, the "last leg" low capacity connection of the data network is one of the most crucially important sectors in network engineering[2]. It can range in capacity from dedicated synchronous T1, T3, or SONNET/SDH[3] lines for an office complex to shared asynchronous fibre or copper lines for domestic services. The current engineering designs of existing networks address these needs simply, on a geographical basis (offices, homes and network cables are located in the real world) by using well understood consumer models. However, as cloud services develop these models will be forced to change in order to avoid unintended consequences and unplanned outages. We have evidence of what happens when these usage models cannot quickly adapt to changing consumer needs. The current COVID-19 pandemic has forced people to rapidly change their daily routines, with some people working from home while others are at home because they are unable to work. This is placing unusual demand on the domestic lines due to increased levels of:

- home working services used by those working remotely.
    - Four of the UK access networks went down on 18 Mar.
- entertainment services used by the isolating and quarantined.
    - Netflix announced on 19 Mar that it was reducing the resolution of all European services due to traffic congestion.

Networks are clearly not engineered for these kinds of rapid shifts in traffic activity and volume. They need to be. The success of future cloud deployment depends just as much on the delivery face as it does on development and deployment face.

The consequence of this reality is that, from the user perspective, the future of cloud computing depends equally upon the ability to deploy adaptive and responsive networks to support service delivery as much as upon the cloud service itself.

Given that the Digital Single Market and the EU research community will rely exclusively on cloud services, there could be (significant) short and long-term (economic) costs associated with getting this part of the picture wrong. We must ensure that the future cloud experience does not depend upon circumstances or location. To achieve this, the SRIA needs to make room for the networking aspects of Cloud services.

S-T Challenge 5: To ensure that society benefits from future developments in cloud computing and other related new technologies, Europe needs to develop networking techniques that are able to

[2] In terms of its impact on user experience only. In reality, the entire network design is crucial. There is a commonly held belief that "the internet" is both robust and resilient. This is not the case, sectors fail routinely with little impact on users because of the overall engineering design. The last time there was a complete internet failure was in the early-mid 1990s, just before the web really took off.

[3] https://en.wikipedia.org/wiki/Comparison_of_T-carrier_and_E-carrier_systems

accommodate rapid shifts in large-scale user behaviour and location. It is unclear if 5G technology can deliver this adaptability across the entire region.

S-T Recommendation 5: Support creation of networking and service delivery capabilities that can adapt both to new patterns of demand, but also new patterns of infrastructure investment and location. [Deployment]

2.3 The Growth of Open Source Software

Open source software is a critical element in the cloud stack. In particular, in between the infrastructure and the platform layer there are many capabilities, from virtualization, to cloud management, to containerization, to data pipelining and artificial intelligence, where the uptake of open source is vast. Open source offers the opportunity for the European tech SMEs to be part of a vibrant innovation ecosystem, without spending too much in terms of license and maintenance. But they also need to be empowered to contribute to those communities through upskilling.

Open source provides the benefits of open standards, compatibility, interoperability, community support, democracy, meritocracy, transparency, and no lock in. However, if you want to use open source tools in an enterprise context and in IT operations, you would most likely use a supported version from the likes of Red Hat, which means that you have deviated from the "open" version and have accepted some level of lock in and proprietary code, which is necessary to create the enterprise grade user experience, as opposed to the developer user experience.

Conversely using pure open source software requires a higher level of skill, as well as a commitment of time and resources to enable participation in the relevant communities.

The downsides of using open source software are:

- not enough people have the skills to use open source tools,
- they are not familiar with the business model and the service model,
- they are not comfortable with using these sometimes raw building blocks.

The degree of challenge depends on the maturity of the open source components. Kubernetes for example has been adopted almost as a *de facto* standard by all industry players, and supported versions are available from many different providers, so it is easy to use and well accepted. On the other hand, there are many unsupported or semi-supported versions and building blocks, which require understanding of how the open source world works before you can use them with success.

Openstack is a good example. It required a very high level of sophistication in the beginning, upgrades from one release to the next were very difficult if not impossible, there were only a few supported version out there (from Red Hat, Suse, etc), and even those were only usable because the vendor was adding a lot of proprietary code and hardening to it.

Open source tools are typically designed not for IT operators, but for developers, which can make them awkward to use in day to day operations.

- As an example of how open source tools are evolving, Kubernetes is being "adopted" by RedHat as OpenShift. I.e. following the same path RedHat took with Linux (RHEL = RedHat Enterprise Linux).
- Many open source tools (for example the high performance file system, Lustre), rely on developer communities to keep up to date, and often an adopting organization needs to commit -- not to a maintenance license, but to a person on staff to stay up to date with the community and potentially make contributions. This obviously is not an appropriate model for small organizations or for commercial enterprises that are used to a different model, and want a "throat to choke" when things go wrong!
- There are also gaps in open source functionality -- OpenStack has been continuously evolving/expanding in scope to fill gaps, and many commercial users have rejected it in favour

of other solutions (such as VMWare) because of these gaps. There are therefore several variations on OpenStack offered by various vendors (e.g. Huawei has a huge effort in this area, but it represents a vendor lock in).

- While these tools are definitely seen as "toys" for the developers (and therefore not ideal for an enterprise environment), there is also a move toward DevOps (as a broader trend in IT development, deployment and operation), and DevOps professionals may be able to work with open source tools in the stack as effectively as with more stable commercial components.

Many of the larger cloud providers are bringing open source software into production-ready status on their offering. There is some concern about whether these providers are contributing enough back to the original open source projects or not.

Once a client has chosen one “flavour” of open source tool, it may be difficult to shift to another flavor. For example, AWS only reluctantly provided their own Kubernetes engine, they would much rather have customers use lambda functions.

2.4 Data Privacy Tools and Management Software

As described above under “Shared Responsibility” (section 1.2), a growing category of cloud software is emerging that provides clients with tools to protect and secure sensitive data, including private data, as well as administrative workflows for managing privacy-related activities. These products appear to provide solutions for single organizations only -- and it is unclear the extent to which any of these products provide consistent management of data stored on multiple platforms (i.e. on premise, private cloud, public cloud, hybrid and multi-cloud combinations). While many of these products come out of the US, some specifically help clients comply with GDPR (as well as the similar California Consumer Privacy Act -- CCPA[4]).

The strategies of these products vary:

- Several (OneTrust, the market leader, as well as BigID and TrustArc, all US-based) focus on inventorying and classifying sensitive data, and then using these inventories to manage policies in support of certification.
- UK-based Privitar provides additional tools for encryption, anonymization as well as “watermarking” sensitive data to help track unauthorized access and use.
- US-based InCountry takes the most aggressive approach, segregating data according to applicable regulations and restricting the storage of affected data to geographically appropriate facilities (operated by InCountry as a value-added service on top of existing IaaS providers, including AWS and Microsoft).

2.5 Data Centres

The essential enabler of cloud technology and cloud service providers has been the ability to build and operate highly efficient data centres at very large scale -- hence the colloquial term “hyperscaler”. Despite the phrase “in the cloud,” cloud-based solutions operate from physical data centers, constructed where real estate and power costs are low, network connectivity is high, and environmental factors (temperature and humidity) allow the heat generated by installed computer equipment to be efficiently removed from the data center.

Data centres are requiring higher power levels (increasingly greater than 1 MegaWatt). Commercial data centres require high availability which means redundant power supplies, ideally from different suppliers, plus backup systems (UPS + generators). Research computing is usually configured for lower reliability, and redundant power is not needed (UPS + generators as required, for graceful shutdown

[4] https://en.wikipedia.org/wiki/California_Consumer_Privacy_Act

after power outage). Ideally there is access to renewable power, but from a commercial standpoint it is more common for data center operators to arrange for “renewable energy credits”.

Physically, data centers are requiring higher density floor configurations (computer racks are getting heavier). Often this means a move away from the classic “raised floor” designs of the past.

Data centers also require increased heat rejection capability. Most commercial data centres are air cooled, with racks arranged in cold and hot aisles and careful attention to air flow. Air is cooled by heavy duty air conditioners, which in turn require external chillers, or links to cold water loops (e.g. district/campus cooling, or access to bodies of water). However these approaches are increasingly not adequate or too disruptive to accommodate the power density of newer compute technologies, and operators are shifting to liquid cooled technologies. These include rear door heat exchangers (air cooling within the rack, cooled by a liquid system in each door, liquid typically exchanges heat with a secondary loop); direct liquid to chip systems (liquid delivered by plumbing to each component); and immersion systems (full systems immersed in dielectric liquid bath). With the trend toward higher power densities, chip manufacturers are designing chips with Increased tolerance for higher ambient temperatures, allowing free air or local water supplies to be used for heat exchange. Ideally heat exchanged away from computer equipment can be used productively, for example through district/campus heating found in some parts of Europe.

3 Edge and Fog Technologies

Two interrelated paradigms have been defined: fog and edge computing. According to NIST:

- Fog computing is a computing paradigm where data processing and storage services are located between cloud data centres and end-user devices,
- edge computing fosters the processing and storage of data at the edge of the network including smart devices and their end-users.

The advantages of processing and storing data closer to their source are essentially: the reduced latency and network usage, the increased data security and governance.

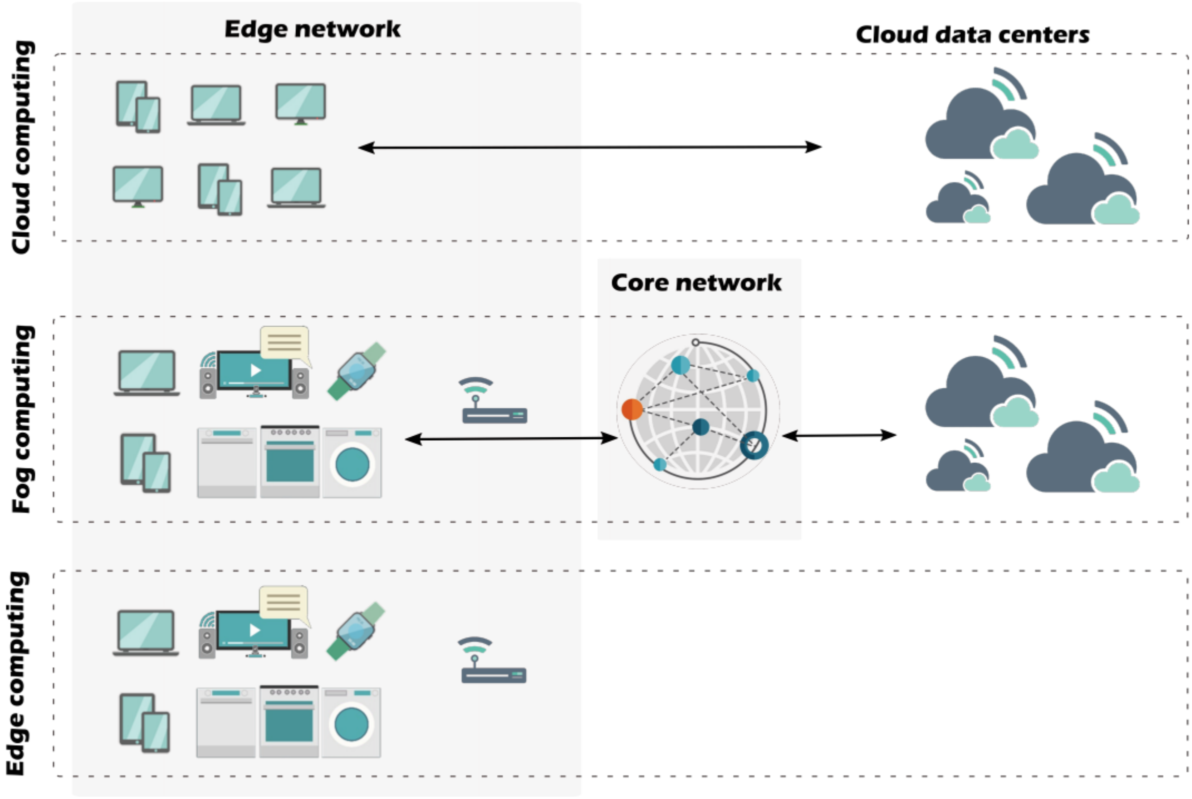


Figure 1. Cloud, fog and edge domains (based on Mahmud et al. [8]).

FIGURE 1. FROM CLOUD COMPUTING TO EDGE COMPUTING[5]

3.1 Edge and fog is a continuum

Edge computing is not simply the pure functionality of supporting computation at the network edge. Rather it is supporting computation through the full cloud to edge continuum. That means, from the cloud data centres, via intermediary edges, till the devices at the network edge. (This is in the context of our analysis and research, in line with current market trends evidenced by IDC)

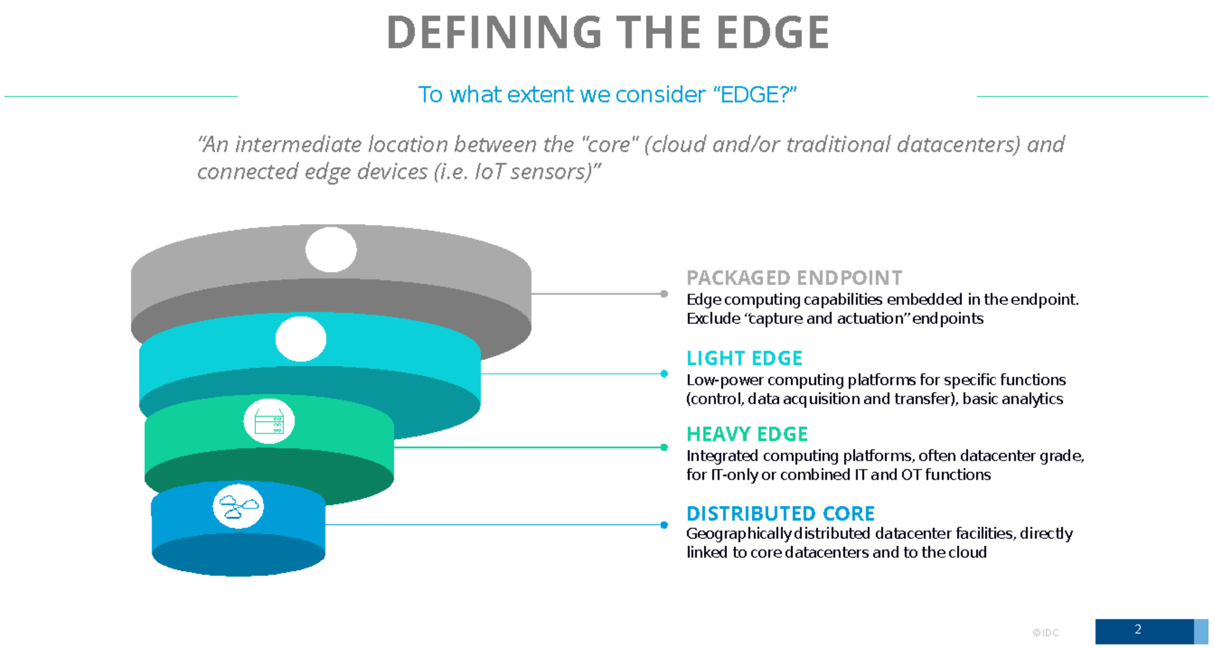
Generally, the continuum from the cloud data centres (core) to the network edge (packaged endpoint) corresponds to:

a) decreasing computational capacity, power consumption and latency.
b) Increasing dispersion and distribution of the infrastructure and complexity of management of the infrastructure (also related to reduced network availability).

3.2 The distinction between light and heavy edge

While core and heavy edge are fixed infrastructures, light and packaged edges may be mobile. (see figure 2)

[5] Svorobej, S.; Takako Endo, P.; Bendechache, M.; Filelis-Papadopoulos, C.; Giannoutakis, K. M.; Gravvanis, G. A.; Tzovaras, D.; Byrne, J.; Lynn, T.. Simulating Fog and Edge Computing Scenarios: An Overview and Research Challenges.



DEFINING THE EDGE

To what extent we consider "EDGE?"

"An intermediate location between the "core" (cloud and/or traditional datacenters) and connected edge devices (i.e. IoT sensors)"

PACKAGED ENDPOINT
Edge computing capabilities embedded in the endpoint. Exclude "capture and actuation" endpoints

LIGHT EDGE
Low-power computing platforms for specific functions (control, data acquisition and transfer), basic analytics

HEAVY EDGE
Integrated computing platforms, often datacenter grade, for IT-only or combined IT and OT functions

DISTRIBUTED CORE
Geographically distributed datacenter facilities, directly linked to core datacenters and to the cloud

© IDC

2

Figure 2. Defining the EDGE.

3.3 Different applications will use different edge approaches

Different application scenarios may make use of different "edges" in the cloud-edge continuum. For instance, Telco operators mostly exploit heavy and light edges (in some cases) whereas Industry 4.0 exploits mostly light and endpoint edges.

3.4 Public, private and hybrid edge can still apply

The distinctions between "public", "private" and "hybrid" cloud computing, can be applicable as well to edge computing. Thus, we can define:

- private edge as an edge computing services offered either over the Internet or a private internal network and only to select users instead of the general public;
- public edge as an edge computing services offered by third-party providers over the public Internet. (Public edges are still limited on the market.)
- hybrid edge as an aggregation of edge resources including at least one private edge and one public edge.

4 Technology Trends

Compute: The end of both Moore's Law (first presented in 1965) and Dennard Scaling (first presented in 1974) means that compute price/performance will no longer improve significantly each year. At the same time compute requirements are growing exponentially with global digital transformation (including growth in training and distributed deployment of AI models, where compute requirements for training alone are growing 10x per year[6]).

S-T Challenge 4.1: Significant growth in compute spending. This will create pressure to extend economic lifetimes for compute investment and reduce pressure to refresh as rapidly as in the past. This may also limit the future appeal of cloud-based solutions on a "total cost of ownership" basis.

[6] https://openai.com/blog/ai-and-compute/

Storage: Data volumes are growing 27% per year[7], but demand is growing more rapidly in some sectors (e.g. research[8]). Storage costs are falling, but costs are generally not keeping pace with increased demand. For the most price-competitive hard drive storage, cost/unit capacity has been falling 20-30% per year historically[9], but is projected to fall by only 15% per year going forward. Solid state drive costs have fallen more rapidly but command higher prices because of higher performance -- and future costs are projected to fall by only 15% per year as well. Growing focus on archival data storage will further increase absolute data volumes by an estimated 10-20%.

S-T Challenge 4.2: Significant growth in storage spending. Storage costs are becoming a more significant component of ICT budgets, even as compute spending increases more rapidly than seen in decades.

Networking: The rate of networking bandwidth growth is significant but still lags growth of compute and storage demands. Dataset sizes are projected to increase faster than network capacities. Transferring meaningful amounts of data will take more time in the future (even with network upgrades).

S-T Challenge 4.3: Increased data gravity. It will be increasingly important to process data where it is stored.

A growing proportion of newly created data will be generated "in the field" -- by IoT, edge devices, sensors, autonomous vehicles, etc. It is unclear how the following activities will be balanced:

- Local storage/transmission of raw data to the next layer of devices (edge, fog, cloud, etc.)
- Local processing of raw data
- Transmission of processed data to the next layers
- Intermediate processing/aggregation/storage
- Central storage of data.

Managing the interplay of these activities will require new tools/services. These activities will also be constrained by policy and regulatory factors (GDPR, etc.).

S-T Recommendation 4.1: Bring compute to the data. Increasingly the data required for analysis will be distributed, should not require transfer to a "central" location for processing, and instead the processing should be applied to the data where it is stored. For "big science" projects, it may not even be feasible to collect data for processing in one place, since the size of that data may require extreme investments in storage and processing or create unacceptable delays associated with data transfer. Moreover, the environmental cost of reproducing, transferring and then storing this "big data" is becoming more and more significant.

S-T Recommendation 4.2: Analyze data where it is generated. Today, data generated at the edge is a special case, but will increasingly become the dominant case. Data processing (including AI training and inference) at the edge should be beneficial compared to the investments in intermediate networking and centralized storage and processing required to support centralized collection/concentration/processing of that data.

5 R&I Project Analysis: ICT-06-2016 - Cloud Computing (H2020)

[Based on review of select projects in "Cordis Project Summary" tab of Google Sheet: https://docs.google.com/spreadsheets/d/1_s8N6QZg0b1L9jIcT6BXRpfdIesag7PD-Ko7pSBLcZI/edit?usp=sharing]

The 12 most recently completed set of R&I projects (ICT-06-2016 - Cloud Computing) address:

[7] IDC 2018

[8] In 2015, Compute Canada estimated Canadian research data volumes would grow 50% per year.

[9] https://indico.cern.ch/event/713888/contributions/3122779/attachments/1719287/2774787/storage_tech_market_BPS_Sep2018_v6.pdf

“Recent trends in cloud computing go towards the development of new paradigms (heterogeneous, federated, distributed clouds) as opposed to the current centralised model, with tight interactions between the computing and networking infrastructures. The challenge is to address, from the research and experimentation perspectives, the necessary evolution in cloud architectures, cloud networking, deployment practices and run-time management as well as the associated security and privacy needs. The ambition is to increase the uptake of cloud technology by providing the robustness, trustworthiness, and performance required for applications currently considered too critical to be deployed on existing clouds. From the innovation side, the challenge is in fostering the provision and adoption of competitive, innovative, secure and reliable cloud computing services by SMEs and public sector organisations across Europe. “ (from Cordis)

Of these projects, two promised results of particular relevance:

> **MELODIC**: “MELODIC will enable data-intensive applications to run within defined security, cost, and performance boundaries seamlessly on geographically distributed and federated cloud infrastructures. Serving the user’s needs and constraints, MELODIC will realise the potential of Cloud computing for big data and data-intensive applications by transparently taking advantage of distinct characteristics of available private and public clouds, dynamically optimise resource utilisation, consider data locality, conform to the user’s privacy needs and service requirements, and counter vendor lock-in.” (Cordis)
>
> The project builds on 3 earlier R&I projects (PaaSage, PaaSword, CACTOS). The project’s primary “product” is Cloudiator v2, which implements Hadoop over arbitrarily distributed cloud resources.
>
> The MELODIC team also offers insightful guidance in “Future Cloud Systems Design: Challenges and Research Directions”
>
> **LightKone**: “The goal of LightKone is to develop a scientifically sound and industrially validated model for doing general-purpose computation on edge networks. … However, today’s state of the art, the gossip and peer-to-peer models, give no solution for defining general-purpose computations on edge networks, i.e., computation with shared mutable state. LightKone will solve this problem by combining two recent advances in distributed computing, namely synchronisation-free programming and hybrid gossip algorithms, both of which are successfully used separately in industry. ” (from Cordis).
>
> The project has developed several interesting products, which in combination allow the configuration and operation of edge computing networks, with local storage and processing, without assuming that data must be transferred to the centre.
>
> The LightKone team proposed a “reference architecture” (LiRA) that highlights some of the shortcomings of current edge computing paradigms.