NEXUS FORUM.EU

RESEARCH & INNOVATION ROADMAP

Cognitive Computing Continuum

WHITE PAPER

First Version - September 2024

www.eucloudedgeiot.eu

EXECUTIVE SUMMARY

This is the first public draft version of the Cognitive Computing Continuum Research and Innovation Roadmap for Europe. The roadmap will be released annually in two main versions: one after December 2024, and one after December 2025.

This roadmap identifies several technology priorities for the European Cognitive Computing Continuum, to support the EU cloud-edge industry with urgently needed R&D. It complements the industrial roadmaps, innovations, and vision from the EU Cloud Alliance and the new IPCEI-CIS, with a view towards R&D needs in the medium to long term.

The roadmap is organised into four main topics, each with its own main section in this document, that are aligned with the four IPCEI-CIS Workstreams:

- **AI for Cloud-Edge**: Orchestration and managing a multi-provider continuum.
- **Cloud-Edge for AI**: Enabling and facilitating AI applications across the continuum.
- **Telco Cloud-Edge**: Telco as one of the main tenants and infrastructure providers.
- **Cloud-Edge Use Cases**: Requirements from next-generation use cases and applications.

These four main topics have been further divided into several subtopics, and each subtopic is presented with some context and challenges, along with R&D priorities and potential impact. The main subtopics identified can be summarised as follows:

AI for Cloud-Edge

- "Cross-layer" optimization in a multi-provider cloud-edge continuum.
- Large Language Models and AI for debugging and root cause analysis.
- AI for an energy-efficient, carbon- and energy-grid-aware cloud-edge continuum.
- Efficient resource allocation for "continuum-native" workloads.
- Stateful serverless computing for the cloud-edge continuum.

Cloud-Edge for AI

- Integrating (Euro)HPC in the cloud-edge continuum.
- Integrating RISC-V hardware in the cloud-edge continuum.
- Open compiler infrastructure and lightweight virtualization/container technologies for heterogeneous hardware.
- Federated computation for Foundation Models.
- Confidential computing and data privacy & security.
- LLMOps and GenAIOps.

Telco Cloud-Edge

- Open radio access networks (O-RAN).
- Resilient on-prem 5G/6G edge-cloud for Industry 4.0 and beyond.
- AI and Generative AI to optimize network resources and operational efficiency.

Cloud-Edge Use Cases

- Software-defined vehicles.
- Decentralized spatial computing for AR/VR technologies.
- Satellite communication for edge applications.
- Cyberspace and physical space fusion.

More details can be found in the respective sections of this document. The roadmap additionally identifies several areas that might be interesting to explore in the context of the Cognitive Computing Continuum and further develop in future versions of the roadmap:

- Neuromorphic computing.
- Quantum computing.
- Space edge.
- Hyper-decentralized edge.

Four key transversal topics that are aligned with focus areas under the Cloud-Edge WG of the EU Cloud Alliance, have been taken into consideration in the roadmap, and will be covered in dedicated sections in future versions of the roadmap:

- Sovereignty & Open Source.
- Sustainability.
- Interoperability.
- Cybersecurity.

Dedicated Working Groups will be set up as part of NexusForum.EU, corresponding to the main topics and the transversal topics, to further develop and validate this roadmap and engage the broader community in its development. More details about this can be found in the deliverable D4.1.

TABLE OF CONTENTS

1 INTRODUCTION

The goal of the Cognitive Computing Continuum Research & Innovation Roadmap for Europe is to deliver a strategic outlook on the future of the Cognitive Computing Continuum, effectively combining the vision, priorities, and advancements by the Horizon Europe programme, the new IPCEI-CIS, and the *European Alliance for Industrial Data, Edge, and Cloud*. The Cognitive Computing Continuum Research & Innovation Roadmap for Europe incorporates: 1) an EU Research & Technological Capabilities Analysis; 2) an International Landscape Analysis; 3) and a bold long-term outlook and vision for potential research and innovation topics addressing the gaps in EU work programmes and incorporating relevant future breakthroughs.

The present version of the roadmap is divided into four main sections, based on the priorities identified by the European Commission, the IPCEI-CIS, and the *European Alliance for Industrial Data, Edge, and Cloud*:

- **AI for Cloud-Edge** (more automation and AI for managing an increasingly complex computing continuum infrastructure).
- **Cloud-Edge for AI** (how the computing continuum can support and enable the development and deployment of AI technologies in Europe).
- **Telco Cloud-Edge** (how telecom operators can become a more integrated part of the computing continuum, both as a tenant and as a service provider).
- **Cloud-Edge Use Cases** (the needs, requirements, and challenges for different use cases to use cloud and edge computing technologies).

In addition to those four main sections, the roadmap already identifies a number of interesting technology areas that could have a disruptive impact on the European Cognitive Computing Continuum in the longer term:

- Neuromorphic computing.
- Space edge.
- Quantum computing.

The roadmap will be released annually in two main versions: one after M12 (December 2024), and one after M24 (December 2025). A draft version of the roadmap is released to the public after M6 (June 2024), as part of this deliverable (D2.1).

1.1 European market context and strategy for the Cognitive Computing Continuum

The European Union (EU) acknowledges the significance of cloud computing, edge computing, and the Internet of Things (IoT) in shaping its digital future toward a European Computing

Continuum Infrastructure.1,2,3 The strategy to safeguard digital sovereignty must ensure that Europe possesses the necessary capabilities to act and make independent decisions within the global context concerning edge and cloud services, systems, and technologies. Achieving European technological sovereignty requires addressing several challenges within the specific socio-economic context of Europe.

Dependency on non-EU Technologies. Europe's heavy reliance on technologies developed by non-EU vendors, especially those from the United States and China, poses challenges to its ability to control critical infrastructure and data processing services. Meanwhile, non-EU hyperscalers are rapidly advancing in the race to dominate the future cloud-edge computing value chain. Although European cloud offerings span a wide spectrum of services, customers often need to collaborate with multiple providers to match the quality and breadth of services offered by leading global cloud providers. European providers differentiate themselves by focusing on specific niches where they excel. European cloud providers (including telcos) are in a unique position to be able to leverage proximity, resource optimization, QoS evaluation, and regulatory compliance to enhance edge computing performance and address latency challenges.

Securing Europe's strategic autonomy in the digital world is not only a technical hurdle; it also presents a geopolitical puzzle. Europe contends with other global regions in the race for becoming technological leaders. Navigating the delicate balance between economic interests, security imperatives, and strategic partnerships demands thoughtful management and governance. The impact of this process goes beyond individual choices by end-users and businesses, as illustrated by Ursula von der Leyen's February 2020 op-ed, where she described tech sovereignty as "*the capability that Europe must have to make its own choices, based on its own values, respecting its own rules*."

For instance, **ensuring data privacy and security** is crucial for maintaining sovereignty. European citizens and businesses rightfully anticipate that their data will be treated with care and shielded from unauthorized access. Balancing the imperative of data sharing for innovation with the need to protect privacy remains a formidable challenge. While regulations like the General Data Protection Regulation (GDPR) aim to tackle this issue, their implementation across the EU members states can be intricate and geopolitical threats are also an issue to be addressed.

Furthermore, Europe consists of multiple countries with diverse regulations, languages, cultural contexts and markets. This fragmentation can hinder the development of unified technological solutions and the emergence of a **fully integrated European Digital Single Market**. Therefore, harmonizing standards and creating interoperable systems across borders is essential for technological sovereignty. Despite significant initiatives at EU level to harmonize regulatory standards,[4] the continent's digital market remains fragmented into local realms, individually lacking the critical mass for players to scale and compete with their American and Asian counterparts.

On the networking and communications dimension**, the connectivity infrastructure across the EU** is not yet ready to address the current and future challenges of the data-driven society

---

[1] *European Alliance for Industrial Data, Edge and Cloud | Shaping Europe's digital future (europa.eu)*.
[2] *European data strategy - European Commission (europa.eu)*.
[3] *Europe's digital decade: 2030 targets | European Commission (europa.eu)*.
[4] *Relevant examples include the GDPR and the Free Flow of Non-Personal Data Regulation.*

and economy and the future needs of all end-users.[5] The EU's lag in fibre coverage and the deployment of standalone 5G networks, when compared to regions like Japan, South Korea and China, poses a significant vulnerability for Europe's economy, also reducing the collaboration possibilities. The delivery of advanced data services and AI-based applications relies heavily on these infrastructure developments.

The deployment of public edge computing infrastructure is equally relevant. It serves as a critical enabler for time-critical applications and computing capabilities, especially in real-time data-intensive use cases and related to the Internet of Things (IoT). Notably, there exists a strong correlation between the adoption of capable digital networks and the uptake of modern technologies, which are currently not developing at large scale.

Within this context and with the willingness to **decentralize data processing**, allowing it to happen closer to the user, the European Union is aiming at the creation of a **greener, more resilient and more secure ecosystem of cloud edge computing**. Indeed, by 2030, the bloc wants to deploy 10,000 climate neutral highly secure edge nodes across its territory, thus enabling a 75% increase of European enterprises using advanced cloud computing services.

At the end of 2023, the EC approved the **Important Project of Common European Interest on Next Generation Cloud Infrastructure and Services (IPCEI-CIS)**, the first ever IPCEI focused on cloud technologies; 19 companies from 7 Member States are taking part in it as Direct Participants, with a total of around 100 organizations currently active in this initiative.[6] The aim of this innovative approach is to foster a **multi-provider cloud edge continuum** in Europe, that can create an open environment where data is processed based on a network of interconnected cloud and edge infrastructures.[7] This kind of cloud and edge ecosystem is expected, in return, to contribute to the development of data processing technologies that are federated, energy-efficient and trustworthy, thus touching upon European businesses and citizens through the advancement of the EU's transition towards **a green, digital, resilient, secure and sovereign future**.[8]

Furthermore, within the respect of competition law and the monitoring of its implementation, IPCEIs are an instrument to fairly tackle the technological and financial risks that joint innovation involves. Creating a cloud-edge continuum ensures a **European alternative**, pushing for stronger competitiveness in these advanced technologies and **increasing welfare for its business and citizens**. The spill-over effects of such a sovereign cloud-edge continuum will also be felt by companies not directly involved in the IPCEI, as results and knowledge of the initiative will be shared, to enable collaborative innovation to unfold.

On a similar note, EU businesses and Member States have come together under the sponsorship of DG CONNECT to form the ***European Alliance for industrial Data, Edge and Cloud***,[9] an initiative targeting the development and deployment of next generation edge and cloud technologies, willing to **empower the EU industry's position in the field**, while fostering effective and secure processing of sensitive data across businesses and public administrations. The EU Cloud Alliance produces roadmaps to identify priorities for co-investment in cloud and edge technologies, providing recommendations on how to

[5] *European Commission "How to master Europe's digital infrastructure needs?" (europa.eu)*.
[6] *Cloud - European Commission (europa.eu)*.
[7] *SA_102517_707E5C8E-0000-C216-8C1C-3081176554C2_287_1.pdf (europa.eu)*.
[8] *IP_23_6246_EN.pdf (europa.eu)*.
[9] European Alliance for Industrial Data, Edge and Cloud | Shaping Europe's digital future (europa.eu).

implement such investments and finally, it advises legal and commercial conditions for cloud services. Industry stakeholders willing to join the EU Cloud Alliance must be legally based in the EU and must comply with the overall objectives defined by the EC.[10] The work of the EU Cloud Alliance underlines the importance for more technological sovereignty as a crucial approach for the development of a competitive digital European single market.

Open technologies, including standards, open-source software (OSS) and hardware (OSH), contribute to economic growth and innovation, and they hold significant potential in strengthening European digital sovereignty.[11] Recognising that foundational cloud-edge technologies are currently, to a large extent, being developed and maintained within opensource foundations like the Linux Foundation and the Eclipse Foundation, the roadmap of the *European Alliance for Industrial Data, Edge and Cloud* outlines several priorities aimed at improving European capabilities and autonomy in the global open-source sector, including the need to "*strengthen existing (and create new) open source software communities, led by European organisations, to manage and maintain important technologies in the long term as well as to lead open source reference implementations*".[12]

While the critical importance of OSS for cloud computing is well established, interest in OSH has developed more recently and is centred on the role that it can play in lowering barriers to entry and thereby mitigating strategic dependencies in the fields of processors and, more broadly, semiconductors. Such dependencies were made salient by disrupted supply chains during the COVID-19 and increased geopolitical volatility in recent years. Currently, the growing role of AI across various sectors is further heightening the strategic importance of semiconductors and the need for the EU to address particular areas of strategic weakness, including chip design.[13]

Of particular interest in this context is the surge in open innovation, catalysed by the development and adoption of RISC-V, a free and open instruction set architecture (ISA) standard. The European Chips Act,[14] along with initiatives like the Chips for Europe initiative, is providing an impetus for increased collaboration within Europe, emphasising the importance of RISC-V and open-source. Building on open and shared standards, open-source licenses facilitate seamless collaboration between researchers and industry, helping to accelerate innovation and counteract dependencies. However, to fully harness the potential of open technologies in achieving digital sovereignty, a strategic approach is required to build and strengthen competencies in key areas, ensuring Europe has the capacity to fork projects and maintain the code it relies on, thereby reducing dependencies.[15]

---

[10] EUSurvey - Survey (europa.eu).
[11] Blind et al. (2021). The impact of Open Source Software and Hardware on technological independence, competitiveness and innovation in the EU economy. Final Study Report.
[12] European Alliance for Industrial Data, Edge and Cloud presents its first deliverables (europa.eu).
[13] Kleinhans (2021). The lack of semiconductor manufacturing in Europe. Policy Brief. SNV Berlin.
[14] EC (2022). European Chips Act: Communication, Regulation, Joint Undertaking and Recommendation.
[15] EC Working Group on OSH and OSS (2022). Recommendations and roadmap for European sovereignty on open source hardware, software and RISC-V Technologies.

1.2 Cognitive Computing Continuum Research & Innovation Roadmap for Europe

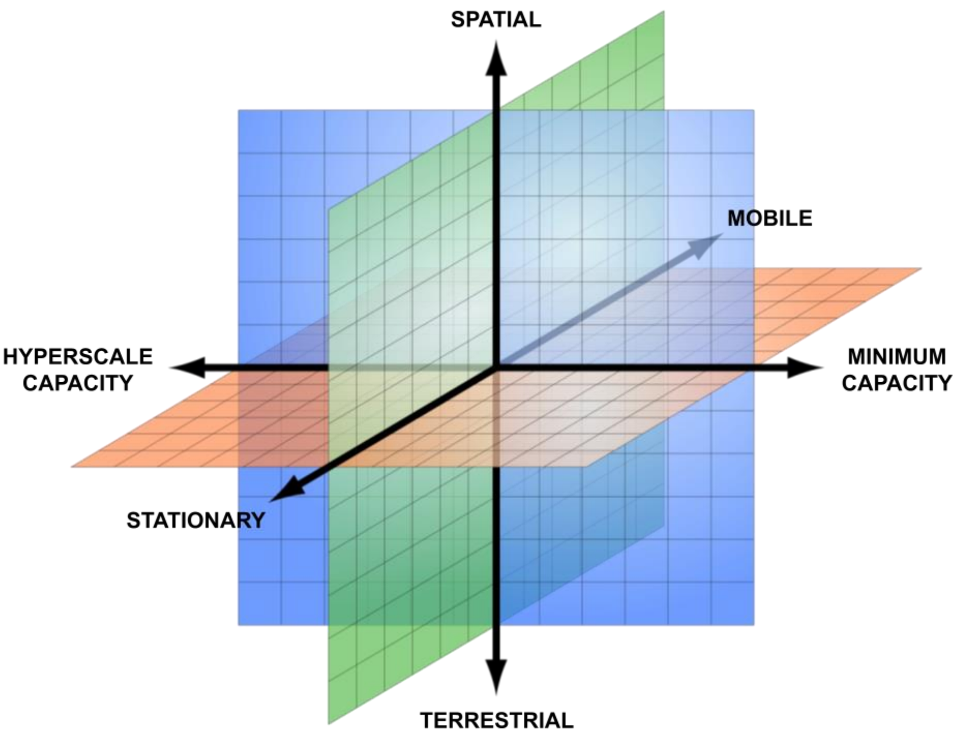
EU-level policies on Cloud, Edge and Internet of Things are far-reaching, touching upon domains ranging from energy efficiency to economic security, to standardisation and interoperability.

The commitment of the European Union towards the objectives of the priority 'A Europe fit for the digital age' is clear, with efforts also dedicated to the 'European Green Deal', 'An economy that works for people' and 'European way of life' Commission priorities. This translates to a policymaking exercise aiming to strike a balance between innovation, competitiveness, economic security, environmental sustainability (energy efficiency in particular) and social sustainability and inclusion.

This is all but an easy challenge to take on, as testified by the breadth of the efforts of EU institutions to produce a coherent yet wide range of strategic and legislative documents addressing the priorities it has set for the 2020-2024 Commission.

Figure 1: Three main dimensions of the emerging Cloud-Edge Computing Continuum infrastructure.

With reference to this research and innovation roadmap, some key policy domains have been identified.



This roadmap identifies the role of **energy efficiency** and **environmental footprint** as cross-cutting priorities for cloud, edge and IoT. This aligns with both the high-level Commission's priority 'the European Green Deal' and Directive (EU) 2023/1791 on energy efficiency, and Communication COM(2020) 67 final 'Shaping Europe's digital future', which push towards increased energy efficiency in data centres and in the wider ICT supply chain.

Standardisation, open-source, and competition are also key cross-cutting domains when examining the four main topics covered by this roadmap and are closely connected to Regulation (EU) No 1025/2012 on European standardisation and responds to legislation on interoperability, including flagship Regulation (EU) 2024/903 (Interoperable Europe Act), the Digital Markets Act (EU) 2022/1925, and the upcoming Cyber Resilience Act (COM(2022) 454 final). Additionally, the roadmap identifies open-source solutions as valuable tools in the push for increased standardisation, in coherence with policy actions aimed at fostering the role of the open-source ecosystem throughout Europe, including the EC's *Open Source Software Strategy 2020-2023*. In terms of competition and access to market, the relevance for the cloud and edge service providers of the Digital Markets Act (EU) 2022/1925 might be expected to grow in the coming years as the regulation is enforced across Member States. might be expected to grow in the coming years as the regulation is enforced across Member States.

This also links to a specific focus on the deployment of solutions aimed at increasing **interoperability** across a diverse range of systems: this is increasingly important in a changing context in which, for example, ARM and RISC-V architectures are increasingly adopted in addition to x86-64. From a policy perspective, this connects once more with the Interoperable Europe Act (Regulation (EU)2024/903) and with the Chips Act (Regulation (EU) 2023/1781).

Some of the topics outlined in the roadmap are relevant to European policies area focusing on **economic security**. In particular, the Joint Communication on European Economic Security Strategy (JOIN(2023) 20 final) mentions Artificial Intelligence, 6G, quantum computing and advanced semiconductors and microchips among the technologies underpinning European economic security in a changing global context. 6G has also been identified as an important technology for European digital sovereignty in the 2020 State of the European Union Speech. 6G is likely to be a game changer in the IoT part of the Cloud-Edge-IoT continuum particularly, due to its facilitation of higher bandwidth communications for mobile devices other than smartphones.

Data privacy and **cybersecurity** are also pivotal in the Cloud, Edge and IoT domains, with particular emphasis on privacy and security as related to Artificial Intelligence. This resonates with the AI Act (COM(2021) 206 final), the Cybersecurity Act (Regulation (EU) No 526/2013), and the upcoming Cyber Resilience Act (COM(2022) 454 final) and responds to the principles of the General Data Protection Regulation ((EU) 016/679), the Declaration on the Future of Internet[1], the European Declaration on Digital Rights and Principles and the wider Digital Decade Policy Programme 2030 (Decision (EU) 2022/2481).

The following sections of the document present the corresponding sections of the Cognitive Computing Continuum Research & Innovation Roadmap for Europe. For each specific topic, a brief context is presented, along with the associated research challenges; then, some identified R&D priorities are outlined, before concluding with the envisaged potential impact.

2 AI for Cloud-Edge

Creating a federated, energy-efficient, and trustworthy European data processing ecosystem, spanning a multi-provider cloud to edge continuum, will require the development of new automated management technologies for the cloud-edge stack and new mechanisms for interoperation between different providers. The management of the Cognitive Computing Continuum should enable real-time and low-latency services by distributing workloads in the cloud-edge continuum and reduce transmission of large volumes of data to centralised cloud data centres. At the same time, the computing continuum needs to be energy-efficient and sustainable, with as low carbon footprint as possible. Below, we list several research and development challenges that need to be solved to enable the operation and management of such a distributed data processing ecosystem.

2.1 Cross-layer optimization of the continuum infrastructure

Context and challenges

The different layers of the continuum must be operated in a coordinated fashion. For example, in a data centre the distribution of workloads should be coordinated to optimize resource utilization and energy use of the cooling system. In a multi-provider cloud-edge continuum, this becomes a distributed optimization problem in which each entity has partial information and possibilities to adjust operational parameters. Trying to optimize each layer separately will lead to suboptimal performance – this is the problem of "local" vs "global" optima.

R&D priorities

- Open "cross-layer" protocols and mechanisms to coordinate the optimization across different levels: 1) the cooling and resource utilization of the data centre infrastructure and individual servers; 2) meta-orchestration and workflow execution across machines and clusters; and 3) optimizations at the application and service levels. It is important that these coordinate network, data, and compute resources.
- New collaborative business models may be needed to fully address these challenges, including payment solutions, federation models, and pricing models-market mechanisms.
- Establishing a set of "standard" services and APIs to facilitate the optimization across the cloud-edge continuum, service discovery, etc.
- Moving towards "digital shadows" and "digital twins" of the continuum, here focusing on the data flows and the integrations needed to optimize the management of network, data, and compute resources across the cloud-edge continuum.

Potential impact

Hyperscalers have a big advantage, since they control their entire cloud stack from facilities to servers/hardware to service offerings. New AI-based solutions for managing the emerging cloud-edge continuum in a holistic way will improve resource utilization and reduce energy use across the multi-provider continuum. This will help cut down the costs of operation and provide a more viable and competitive EU alternative.

2.2 Efficient resource allocation for AI workloads in the HPC-integrated continuum

Context and challenges

AI workloads can be resource-intensive and require careful orchestration to ensure efficient use of cloud resources. EuroHPC provides high-performance computing (HPC) resources for AI development, with plans to expand the capacity, but traditional HPC is sometimes considered too different or too complicated for some companies. For example, file management, such as moving files back and forth from the cloud to the supercomputer, can be cumbersome in AI model training. They might not see the value in using HPC – they're fine with cloud but might not realize that running at scale is cheaper using HPC.

R&D priorities

- Develop strategies for dynamic resource allocation based on workload requirements to optimize resource usage in the cloud-edge continuum for AI.
- Simpler user interfaces and better portability of workflows would make it easier for companies to adopt HPC and integrate it with their cloud deployments. Tools like ColonyOS can be beneficial as they create a level of abstraction, simplifying file management.
- Embrace hardware heterogeneity to make the software stack and skills more portable and robust, and avoiding high reliance on vendor-specific hardware-optimized software, like cuDNN.

Potential impact

If implemented, this solution would lead to more efficient use of resources, potentially reducing financial and energy costs and improving the performance of AI workloads across the computing continuum.

See also Cloud-Edge for AI subtopics 1 (Integrating HPC in the Cognitive Computing Continuum) and 6 (Operationalization of future AI systems in the Computing Continuum (GenAIOps)).

2.3 Sustainable, energy-efficient, and energy-grid-aware cloud-edge continuum

Context and challenges

As the global share of energy used by data centres increases, the carbon footprint of the computing continuum will play an increasingly important role in achieving our ambitious climate goals. The management and operation of the computing continuum needs to become not only more energy-efficient, but also carbon-aware. However, the supply of renewable energy is not unlimited, and the energy grid is facing several challenges, such as the variability of renewable energy, and the limited capacity of transmission and distribution networks, mainly affecting urban and highly populated areas. These challenges cause a number of issues, including grid congestion, and excess renewable energy supply that is typically dealt with either through energy curtailment or storing the energy in batteries. The highly distributed nature of the cloud-edge continuum, and the relatively instantaneous reaction time of computing systems, makes it a promising candidate for providing energy flexibility services and taking advantage of regional excess of renewable energy resources.

Cooling systems account for a substantial portion of the energy consumption of data centres. Therefore, the energy-efficiency and sustainability of this sector is intricately linked to advancing holistic optimization strategies, prominently through the adoption of digital twins and comprehensive data centre management systems. The availability of widespread data in this area is hampered by the natural urge of companies operating in this area to regard their power usage and heating/cooling data as key confidential business data. Similar issues in the electricity generation market have been hampering developments in that area as well. The Open Energy Modelling Initiative (https://openmod-initiative.org/) has been seeking to help solve this problem via a multi-pronged strategy. Similar approaches in the area of energy use in cloud computing need consideration.

R&D priorities

- *Carbon- and energy grid-aware scheduling and orchestration* can allow workloads to be shifted in time and space to react to local fluctuations in renewable energy supply, as well as take advantage of excess energy supply.
- *Beyond energy-efficiency and towards energy grid- and carbon-aware*: Incorporate advances in open energy data, carbon-aware grids, and forecasting of local energy mixes to optimize carbon emissions rather than just energy efficiency. The recovery of waste heat needs to be further investigated and implemented.
- *Establish a standardized set of sustainability metrics and benchmarks*: While metrics like Power Usage Effectiveness (PUE) are valuable, they fall short in capturing the efficiency of processing per unit of energy, especially in the context of a highly heterogeneous continuum.
- *Digital twins using a mix of physical models, CFD and AI models* (e.g., scientific/physics-based ML) to control cooling, management and optimization of the digital infrastructure for the cloud-edge continuum. By enabling a detailed simulation and understanding of energy flows within data centres and their integration into electricity and thermal grids, digital twins facilitate a nuanced analysis of operational efficiencies and inefficiencies. This deep insight allows for the identification and implementation of strategies that can significantly enhance energy efficiency. For instance, by closely monitoring critical operational variables such as CPU temperatures and dynamically adjusting data centre loads, it becomes possible to optimize the balance between IT performance and cooling energy consumption.
- *Reference designs* for different environments and use cases.
- *Whole-system approach*: It's vital to recognize that power consumption isn't confined to data centres alone but extends to the network and specific applications, necessitating benchmarking considerations based on workflows. Innovation should also target energy-efficient software for IoT devices and energy-restricted devices to minimize power consumption.
- *New cooling solutions*, e.g., different liquid or even submersion cooling, to enable next-generation high-performance hardware.

- *Hardware and software co-design* for specific use case applications, e.g., highly energy-efficient AI inference or AI training at the edge.

Potential impact

Moving towards a carbon- and energy grid-aware computing continuum that minimizes the carbon emissions of the energy it uses. The availability of renewable energy does not always match the demand and transmission capacity limits the possibilities to transfer electricity between regions. Storing renewable energy in batteries incurs lifecycle costs that negatively impact the sustainability of the energy system. Taking advantage of excess energy when it is available, and where it is available, improves the overall sustainability of the energy networks.

The role of a holistic data centre management system, overseeing both hardware and software components, ensures that every aspect of the data centres' operation is aligned with efficiency goals. Traditionally, data centres have operated under the paradigm of maintaining the lowest possible temperature to safeguard hardware integrity. However, with real-time insights provided by digital twins into the actual cooling needs based on operational loads and external conditions, data centres can afford to safely increase set temperatures. This adjustment not only reduces the energy consumed in cooling but also extends the lifespan of cooling equipment, further contributing to environmental sustainability.

Green Computing (including carbon-neutral computation) for all aspects of cognitive computing continuum is a key interest for government and industrial players in Japan and South Korea.

2.4 Large Language Models (LLMs) and AI-assisted debugging and root cause analysis

Context and challenges

Components will fail in the computing continuum due to infrastructure or network failures, or an error during application deployment or operation. It is generally hard to track down the relevant information in the system logs and extract insights from them, and there is a large gap between human-operator insights, for example in network operations centres, and system-generated logs. An example would be a ticket where a system (such as telecom network) is known to have failed and a DevOps engineer trying to find the component that caused the failure associated to the ticket. Understanding and identifying the cause of failures in the continuum will be even more complicated, due to the highly distributed and complex multi-provider environment. New solutions and tools are needed both for collecting relevant metrics and logging information across the lifetime of an application in the continuum, and for extracting relevant insights from those logs and metrics, will be needed.

R&D priorities

- *Developing and applying AI technologies for debugging and root cause analysis*: State-of-the-art AI technologies, such as large language models (LLMs) and multi-modal models, have shown an impressive ability to understand context and extract insights from semi-structured data. They have the potential to synthesize large quantities of data, including system logs, while providing a flexible human-machine-interface to query the system for relevant insights.

Potential impact

Bugs in complex systems cause enormous disruptions. For example see the Air Traffic system breakdown in the UK, where some 45 million passengers were affected. This was due to the NAT system being down. Essentially the error was 2 airports having the same 3 letter code, e.g. LHR. LLM and AI tools would make root cause analysis and debugging complex systems easier, as the continuum *will* be much more complex than current cloud systems. Improvements will save time, frustration, and money.

It can also be useful from a cybersecurity perspective for Security Operations Centre (SOC) teams to assess the degree of visibility they have over security incidents in the continuum for different Tactics, Techniques and Procedures (TTPs) from real-world threat scenarios. This approach could be very useful to: (i) carry out the identification of the different data sources in the cloud environment; (ii) to train an LLM to classify which information can be extracted and verify its completeness; and (iii) align the results with the techniques from well-established security frameworks (i.e., MITRE ATT&CK) to pinpoint how the threat monitoring capabilities of a SOC can be extended.

See also Cloud-*Edge for AI subtopic 6 (Operationalization of future AI systems in the Computing Continuum (GenAIOps)).*

2.5 Stateful serverless for the Cloud-Edge

Context and challenges

Data is increasingly generated by devices and systems at the edges of the network. Continuous analytics of such data streams will require data management and analytics solutions that work in highly distributed environments. In use cases where these analytics workflows depend on previous data, or where the processing is distributed between edge sites, there is a need for efficient state synchronization mechanisms. This should also take into consideration mobility between edge sites (which could also be triggered by exogenous circumstances).

R&D priorities:

- *State Synchronization*: Develop mechanisms for seamless state synchronization between cloud and edge, ensuring consistency and reliability.
- *Programming Models*: Create programming models that simplify the development of stateful serverless applications for edge computing.
- *Resource Utilization*: Optimize resource utilization on edge devices to balance workload distribution and energy efficiency.
- *Scalability and Flexibility*: Enhance scalability and flexibility of applications through dynamic execution of workloads at the edge.
- *Extreme data logs consumption*: facilitating the decision-making and the automation of operational and cyber-incident handling.

Potential impact

Enable future stateful continuous analytics at the edge, which include various advanced IoT and digital twin use cases and applications. This would greatly reduce network traffic and latency for such use cases.

2.6 Distributed AI for safety-critical applications in the Cognitive Computing Continuum

Context and challenges

We can expect an increase in the number of dedicated critical software functions and complexity of the underlying hardware, operating systems, and communication networks. This complexity creates the need for an orchestration layer that ensures the functioning of the overall system, especially when it comes to safety-critical operations. At the same time, it imposes new challenges in system design practices and, consequently, in gaining acceptance from the public. Due to limited compute resources on devices, it is necessary to integrate compute resources from cloud data centres and edge. It is also necessary to address privacy concerns, particularly for distributed learning models.

R&D priorities:

- *Seamless cloud-device integration*: Develop platform features that can seamlessly integrate cloud resources in onboard applications. Such integration with cloud infrastructure also enables the use of big data applications (in the cloud). Thus, instead of being restricted to the data produced only within one owned system, other data sources become instantly usable using a high-availability database.
- *Security requirements*: For safety-critical systems, the orchestration layer should be able to provide the hosting process with the capability to communicate with other processes through secure and real-time communication channels. Due to this security requirement, the operating system must ensure a safe environment such that the processes interact only through authorised channels.
- *Orchestration*: Orchestration of computing resources and services is the key to enabling smart and compute-intensive services in the cloud-edge-IoT continuum. Many orchestration frameworks have been proposed in the recent past for cloud applications. However, the current orchestrators mainly focus on resource allocation and load balancing of small-scale hardware and additionally do not take real-time requirements into account. In this context, we should not only study how to reconfigure the node resources but also the impact of the reconfiguration on functional safety and the real-time requirements.

Potential impact

Next-generation intelligent systems are distributed by design (multiple spatially distributed sensors) with high safety requirements and have to cope with a limited amount of energy and often limited cooling capabilities. However, their demand for integration of state-of-the-art AI components (e.g., used for perception or classification tasks) typically require high-performance hardware. High-availability cloud-edge solutions with real-time considerations are key to bridging these constraints. To achieve this, we can use an AI-powered orchestrated approach, with a reduced response time relying on performance metrics as well as on non-functional ones such as energy consumption and security, among others.

Figure 2: Are we ready for millions of privately-owned edge nodes as part of a future hyper-decentralized continuum?

2.7 Towards a hyper-decentralized cloud-edge continuum

Context and challenges

It will be difficult for centrally governed continuum federation models to deal with scenarios in which many new providers join the edge computing market, or in cases in which the market becomes hyper-fragmented. For example, in the future it may be possible for individuals or small entities (e.g., under a cooperative model) to purchase and set up their own edge computing servers powered by green energy, and to make them available for hosting locally-produced data under a principle of sovereignty or even to be commercialized under a pay-per-use business model to third-parties (including local cloud service providers (CSPs) requiring to occasionally increase the capacity of their data centres with additional resources in order to deal with unexpected peaks of demand). This will challenge our current models of service discovery, trust, and management of the application lifecycle in the continuum, as well as the scalability of many of the platforms and services that initiatives like the IPCEI-CIS are expected to produce in coming years.

R&D priorities:

- Decentralized service discovery, for example using distributed hash tables, and payment methods in highly dynamic multi-provider scenarios.
- New decentralized security and trust mechanisms.
- New approaches for managing the continuum and application lifecycles in such a hyper-decentralized computing continuum.

- Mechanisms for smaller operators (even for individuals) to offer available compute resources when available.
- Trustworthy certification of program execution, proof of execution.

Potential impact

Greater flexibility for new players to join the cloud-edge market. It could enable a more sophisticated market that could take advantage of market dynamics, such as smaller players providing edge nodes that can take advantage of local availability of distributed renewables. Opening up for smaller entities to offer up their spare compute capacity would lead to more efficient use of existing compute resources. The quick response of starting/stopping computing workloads, hyper-distributed edge nodes could help take advantage of sudden bursts of local availability of distributed renewables, that otherwise would lead to grid instability and energy curtailment, to increase the sustainability and resilience of the overall energy system.

3 Cloud-Edge for AI

The Coordinated Plan on Artificial Intelligence[16] outlines the EU's strategy and approach to AI built on excellence and trust, putting people first and building strategic leadership in high-impact sectors, as elaborated in the "*White Paper on Artificial Intelligence: a European approach to excellence and trust*".[17] The importance of securing critical computing capacity in the EU can be seen in several actions that have been implemented towards this plan, for example:

- The European Chips Act and Chips Joint Undertaking (Chips JU)[18] which aim to address semiconductor shortages and enhance the European semiconductor industry.
- The European High Performance Computing Joint Undertaking (EuroHPC JU)[19] initiative to develop a world class supercomputing ecosystem in Europe.
- The Testing and Experimentation Facilities (TEFs)[20] to support technology development of components and systems for edge AI.
- The Important Projects of Common European Interest on Next Generation Cloud Infrastructure and Services (IPCEI-CIS). [21]

RISC-V in particular has been identified as a key next-generation technology, and alternative to proprietary processor and accelerator solutions for the European computing continuum and High-Performance Computing (HPC) ecosystem.[22] The recent amendment of the EuroHPC JU Regulation to include the development and operation of AI factories,[23] further highlights the critical role of HPC in strengthening and supporting the European AI ecosystem. Below, we list several research and development needs for the medium to long term that can be identified from these challenges and developments.

3.1 Integrating HPC in the Cognitive Computing Continuum

Context and challenges

Industries may find it difficult to adopt EuroHPC JU supercomputers due to various challenges such as integrating HPC and Cloud services, optimizing software for different architectures like RISC-V, and implementing multi-cloud federations for edge computing. Furthermore, HPC systems can have compatibility issues with standard Machine Learning (ML) libraries due to differences in their software stacks, which can hinder the development and deployment of AI services in the cloud. For instance, the EuroHPC system LUMI, which uses AMD processors, is more open-source than NVIDIA-based systems. AMD uses the ROCm open-source software stack, while NVIDIA has its proprietary CUDA stack. However, machine learning libraries are

[16] https://digital-strategy.ec.europa.eu/en/policies/plan-ai.
[17] https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.
[18] https://digital-strategy.ec.europa.eu/en/policies/european-chips-act, https://www.chips-ju.europa.eu/.
[19] https://eurohpc-ju.europa.eu/index_en.
[20] https://digital-strategy.ec.europa.eu/en/activities/testing-and-experimentation-facilities.
[21] https://www.bmwk.de/Redaktion/EN/Artikel/Industry/ipcei-cis.html.
[22] https://eurohpc-ju.europa.eu/new-call-developing-hpc-ecosystem-based-risc-v-2023-02-01_en.
[23] https://eurohpc-ju.europa.eu/ai-factories-amendment-eurohpc-ju-regulation-enters-force-2024-07-09_en.

not as mature on AMD GPUs compared to NVIDIA. Continuous development and improvement of these open-source stacks can help bridge the gap between different HPC systems.

R&D priorities

- Provide concrete scenarios and guidelines to help industries understand and overcome these challenges.
- Promote the development of an open-source ecosystem to facilitate the integration of HPC and Cloud services, and encourage the use of open-source software stacks to enable AI.
- Bridge the gap between EuroHPC supercomputers and EU clouds, embracing a hybrid cloud paradigm.

Potential impact

Foster the industrial uptake of EuroHPC and AI. If implemented, this solution would make it easier for European startups and industries to adopt HPC and leverage the EuroHPC initiative, potentially leading to more efficient use of resources, significant cost savings, improved competitiveness of EU industry in AI technologies, and advancement towards the Computing Continuum. These challenges are also relevant for cloud computing and the introduction of new hardware architectures.

3.2 Open compiler infrastructure and lightweight virtualization for AI applications

Context and challenges

The cloud-edge continuum will almost certainly be a heterogeneous set of hardware and software systems. Therefore, running applications on different architectures, such as ARM, Intel (x86), RISC-V, Arduino and differing operating systems makes it hard to find a one size fits all for software developers coding in the continuum. In large organizations even specialised teams (e.g., IOS, Android) are needed. Tools applying the Infrastructure as Code (IaC) approach, such as Ansible, may help, but are still often too error prone.

There are several ongoing open initiatives in this direction, including the open W3C standard called WebAssembly[24] (WASM), and the open LLVM[25] project—a collection of modular and reusable compiler and toolchain technologies—with its associated Multi-Level Intermediate Representation, or MLIR,[26] project. These initiatives open up producing code from (m)any high-level languages. The expression “write once, run everywhere” typically formulates this principle.

[24] https://www.w3.org/groups/wg/wasm/, https://webassembly.org/.
[25] https://llvm.org/.
[26] https://mlir.llvm.org/.

R&D priorities:

- *A modern middleware, and support for lightweight virtualization/container technologies*: Use compiler technology to produce machine-like code for any architecture. Support for lightweight virtualization technologies on devices, such as WASM or unikernels, is needed.
- *Heterogeneous Computing Support*: Improve support for heterogeneous computing environments to ensure efficient AI workload execution across different hardware architectures, and in particular interoperability with RISC-V.
- *Open and community-driven development*: Promote a community-driven approach to foster innovation and contributions to open ecosystems, such as the LLVM and MLIR ecosystems which are gaining important traction.
- *Cross-platform efficiency*: Address the need for cross-platform compilation and execution efficiency to support diverse AI applications.
- *Standardization and collaboration*: Encourage standardization and collaboration among industry stakeholders to accelerate the development of open compiler infrastructures across EU institutes and industry.

Potential impact

Harmonize the compute continuum, using open standards, according to the principle "write once, run everywhere". Factory 4.0 and 5.0 as well as systems with many IoT devices are ideal candidates.

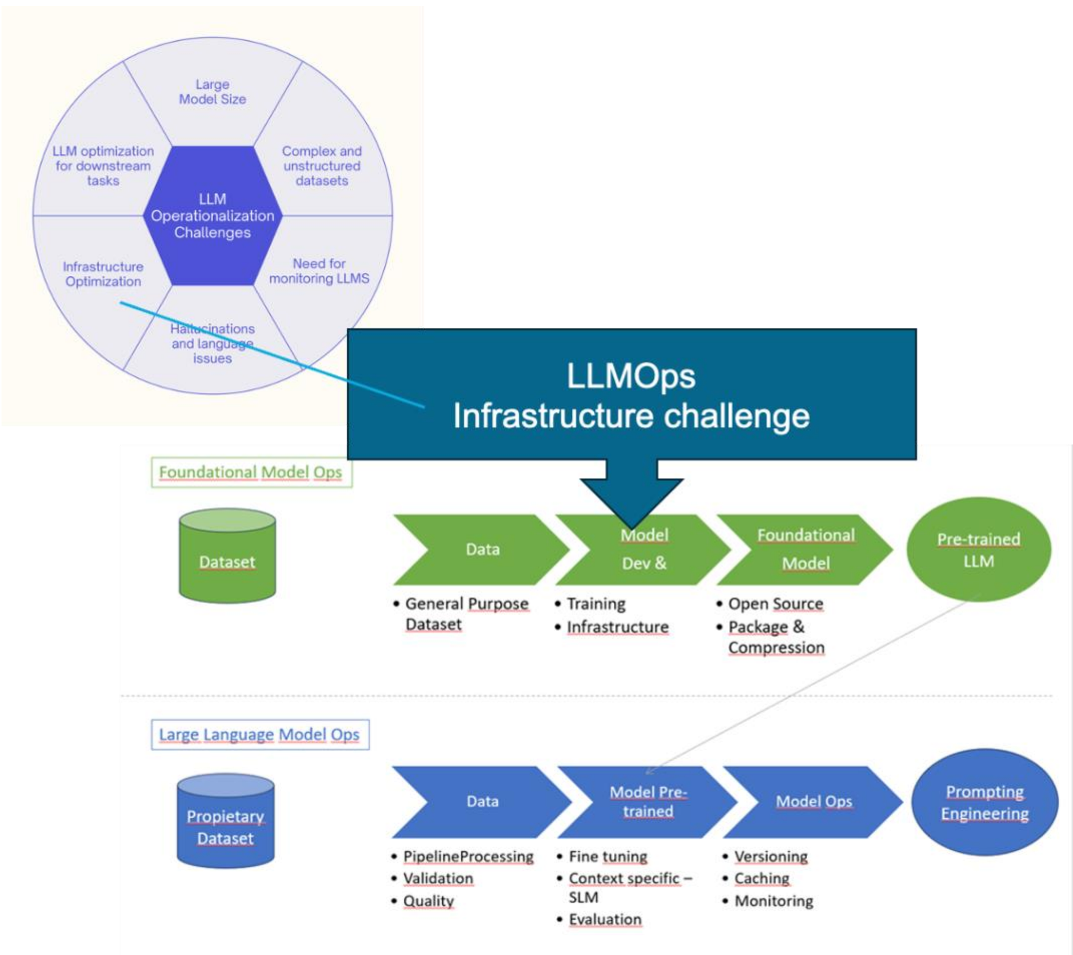


Figure 3: LLMOps.

3.3 Federated computation for Foundational Models

Context and challenges

Most LLMs are highly demanding and mainly use HPC (High Performance Computing). The training of the foundational models (FMs) can take advantage of the recent advances on GPUs and use the computation offered by the Open continuum in what is called "Federated Computation". Implementing federated computation in LLMs is a complex task that requires careful consideration of technical, ethical, and legal aspects. It's a multi-disciplinary effort that involves advancements in machine learning, data privacy, and distributed systems.

R&D priorities

Implementing federated computation for large language models (LLMs) requires several key considerations:

- Establishing a distributed network infrastructure with reliable communication protocols and robust data security measures is essential.
- Prioritizing data privacy through techniques like differential privacy or secure multi-party computation is crucial.
- Adapting LLM designs to support federated computation may involve modifying them for decentralized data sources and partial updates.
- Developing efficient federated computation algorithms capable of handling LLM complexity is imperative for effective model aggregation. Additionally, optimizing resources ensures scalability and efficiency, while rigorous evaluation and testing across diverse data distributions and network conditions ensure performance.
- Encouraging user participation through transparency and trust-building measures and continuous monitoring for issues like data drift are also vital.
- Fostering collaboration among academia, industry, and regulatory bodies is essential for standardizing federated computation methodologies and advancing LLMs.

Potential impact

Federated computation could play a significant role in the future of AI training by enabling more efficient and privacy-preserving methods to train foundational models. The creation of foundation models, which are a subset of large language models, has changed the methodology of how AI can be created. These models can be adapted for a wide variety of use cases with much higher productivity than before. Federated computation could further enhance this by allowing for distributed training across various devices and data centres, potentially leading to more robust and widely applicable AI models.

In summary, the radical shift of Federated computation for Foundational Models training could lead to more specialized and efficient AI architectures that are capable of handling the complex demands of training large-scale neural networks while preserving privacy and leveraging distributed data sources. This shift is part of the broader trend towards more powerful and versatile AI systems that can be adapted for a multitude of business and consumer applications.

3.4 Large-scale testbeds for AI services

Context and challenges

There is a need for large-scale testbeds that can simulate real-world conditions for testing AI services, pipelines, and workflows in the computing continuum. Without such test environments, it's challenging to validate and optimize these services for use in operational conditions. There is demand for an International Testbed for "Cross-border Data flows", for example with Japan and South Korea, considering the "Cross-border Data flow deal" signed between EU and Japan.

R&D priorities

- Develop large-scale testbeds that can accurately simulate real-world conditions, enabling thorough testing and optimization of AI services, e.g. AI Factories.
- Real life 5G/6G-integrated cloud-edge continuum testbeds for projects developing applications using the low latency introduced by edge computing in 5G networks, and for projects optimizing the digital infrastructure, software architectures, and operations and management of such networks of computation and communication.
- Managing cross-border data flows.

Potential impact

Implementing this solution would allow for more robust testing of AI services, leading to improved performance and reliability in operational conditions. They would provide a "testing ground" for gathering operational data to drive further advances in AI-enabled cloud-edge management.

3.5 Data privacy and security in AI services

Context and challenges

Data-driven insights can contribute to vital decisions in many domains (e.g. crisis management, predictive maintenance, mobility, public safety, and cybersecurity). However, obtaining the trust of decision makers to exploit data-driven insight is still a pending issue due to (i) the data and their fluctuating quality and volumetry and (ii) the finality of big-data processing not necessarily suited to decision-maker comprehension.

R&D priorities

- Implement robust data protection measures and comply with regulations like the AI Act to ensure data privacy and security in AI services:
- The development of advanced privacy toolkits. Solutions that can focus on the encryption of data records by exploring attribute-based encryption and proxy re-encryption, symmetric encryption, and searchable encryption. Additionally, could consider the combination of Searchable Encryption and Distributed Ledger Technology (DLT) platforms.

- End-to-end secure privacy-preserving AI pipelines leveraging confidential computing technologies.
- Ensure privacy in the information exchange and certification processes. The application of Zero Knowledge Proof technology and Layer 2 solutions like ZK-EVM (Zero Knowledge Ethereum Virtual Machine) could help too as a part of the solution.

Potential impact

AI services often require access to sensitive data, making data privacy and security a significant challenge. This is especially true in multi-provider cloud federations where data might be stored and processed in different jurisdictions with varying data protection laws. Implementing this solution would increase trust in AI services and could lead to wider adoption, particularly in sectors handling sensitive data.

3.6 Operationalization of future AI systems in the Computing Continuum (GenAIOps)

Context and challenges

Deploying emerging AI learning technologies in the continuum introduces new challenges. Specifically, Large Language Models (LLMs) and Foundation Models (FMs) are very large and costly in terms of resources such as power, time, and storage, and therefore challenges the lower/smaller end of the continuum. Other issues for the continuum are heterogeneity and full-stack dependencies. Infrastructure needs for FM training and inference, i.e., pre-trained on vast amounts of unstructured data to learn complex concepts, need to be addressed.

R&D priorities

Dynamic Orchestration: Implement comprehensive OS solutions to simplify the deployment of future AI systems in the continuum addressing the specific needs of future AI models, such as FMs and LLMs plus the combination of classic ML models. A computational aspect is to use compression techniques to reduce computational cost, memory footprint and energy consumption. An example is to enable edge-based deployment such as using TinyMLOps.

Potential impact

European competitive advantage and independence for upcoming AI solutions deployed in the continuum.

Reference: https://dl.acm.org/doi/full/10.1145/3625289, https://arxiv.org/abs/2205.01423.

3.7 Confidential computing

Context and challenges

Secure and confidential computing supported all the way down to the hardware level could enable new applications with stricter data privacy requirements. Parts of this solution exist today, such as secure enclaves and Trusted Execution Environments, in a commercial context, for example solutions from Intel (SGX and TDX), AMD (SEV-SNP) and IBM. However, the various CPUs and solutions require different toolchains, as compiler support is needed to mark code parts that should execute in the enclave. There isn't really a turnkey confidential computing functionality not dominated by a large, single player. This fragmentation and lack of standardization is similar with the HPC case above.

R&D priorities

Open architectures for the hardware part, see enclaves for RISC-V. Increased support for different hardware solutions in virtualization/container technologies. Multi-tenancy in data centres allows access to parts of datasets, often Kerberos based. Securing and trusting the end-points, think "TLS++" as a concept, in which one can trust the end-points of the connection, the keys, and the end architecture.

Potential impact

Confidential computing is enormous and as yet unresolved. People, companies, organisations, even countries could gain value by sharing data, or parts thereof, but don't or can't.

3.8 Data spaces for AI

Context and challenges

Artificial Intelligence (AI) systems require robust and diverse datasets to learn, adapt, and provide accurate outputs. These datasets can be produced by different stakeholders or companies, and stored along the cloud continuum from edge devices, to fog or cloud environments. Each data producer can have different strategies for data sharing, including monetization of data, constraints on the companies allowed to use the data and the expected use of the data. Furthermore, trust, data sovereignty and data quality are paramount to create a data governance framework able to build AI solutions using data spread all along the cloud continuum. Data Spaces are crucial in this context as they provide the infrastructure for secure data sharing and management, which is essential for the development and deployment of AI. Data space technology still lacks the maturity level needed for seamless deployment in a cloud continuum environment. However, current initiatives in the EU, like Gaia-X, IDS and the SIMPL project among others are designing and developing the standards, tools and frameworks needed to facilitate the creation and operation of data spaces.

R&D priorities

A data-space-based solution for data sharing allows the data producers to share data with other stakeholders without losing control of their data. Here's how Data Spaces could support AI:

- *Data spaces distributed in the cloud-edge-IoT continuum*: Data may be produced and processed at multiple levels of the computing continuum. Additionally considering that data may be owned by different entities with different data policies, it is necessary to create data space solutions that work across the computing continuum.

- *Data Availability*: AI models need access to large volumes of data to train effectively. Data Spaces facilitate the aggregation of data from various sources, making it readily available for AI processing.
- *Data Governance*: Ensuring that data is used responsibly and ethically is paramount. Data Spaces offer governance frameworks that help in managing data access, usage rights, and compliance with regulations such as GDPR.
- *Data Quality and Integrity*: AI outputs are only as good as the data fed into them. Data Spaces help maintain high-quality data standards, ensuring that the data used to train AI models is accurate, complete, and reliable.
- *Interoperability*: AI systems often need to work with data from different domains and sources. Data Spaces support interoperability standards, allowing AI systems to integrate and analyse data across various platforms and environments.
- *Security and Privacy*: Protecting sensitive data is critical, especially when used in AI applications. Data Spaces provide secure environments that protect data privacy and ensure that AI systems are not exposed to data breaches or unauthorized access.
- *Scalability*: As AI applications grow, the need for more data and more complex data processing increases. Data Spaces are designed to scale, supporting the growth of AI systems without compromising performance.

Potential impact

Once a data space is deployed, it is possible to access and manage data from heterogeneous data producers in a transparent, secure, trusted, and fair way, no matter where the data is produced and stored in the cloud continuum, making possible a complete new kind of AI systems and applications.

4 Telco Cloud-Edge

Europe has enjoyed a long-standing history as a global leader in the telco industry, but there are potential risks to this position. As the telco industry moves toward an increasingly cloud-native and open paradigm, such as O-RAN,[27] the traditional telco industry in Europe faces competition from non-EU cloud and tech industry. Furthermore, Europe's economy faces a considerable risk due to the EU's slower progress in expanding fibre coverage and rolling out standalone 5G networks, particularly in contrast to the advancements seen in regions such as South Korea and China. This not only hampers the industry's ability to remain competitive but also reduces the possibilities for collaboration, underscoring the urgency for Europe to accelerate its cloud and network infrastructure developments to support the capabilities of the telco industry.

These challenges, and the convergence of connectivity and computing, are outlined and discussed in the white paper "**How to master Europe's digital infrastructure needs?**",[28] which proposes the creation of the "Connected Collaborative Computing Network", or "3C Network". Several research and development needs for the medium to long term can be identified from these challenges and developments.

4.1 Seamless data connectivity across different networks

Context and challenges

In terms of networking and communications, the European Union's connectivity infrastructure is still unprepared to meet the demands of a society and economy increasingly driven by data. For example, handover between common and future networks (4G, 5G, 6G, LoRaWAN, Wifi7, SatCom, StarLink) is not totally seamless, causing disruptions in ongoing continuous data flows when moving between them.

R&D priorities

A set of measures that will enable a more seamless handover between networks in the Cloud-Edge Continuum:

- Secure base accessibility to all networks.
- Sensing and QoS detection of all physical network channels.
- AI prediction of near future channel QoS.
- Predictive and fast hardware switch, accompanied by channel switch software protocol.
- Develop hardware and software stack that enables easy integration of SatCom connectivity, utilizing multiple services in the background (Inmarsat, Iridium, Starlink, etc.).
- Buffer and recovery mechanism on the edge devices to bridge switch over phase.

---

[27] https://www.o-ran.org.
[28] https://digital-strategy.ec.europa.eu/en/library/white-paper-how-master-europes-digital-infrastructure-needs.

- APIs and support for application design that can handle network disruption with graceful performance degradation and quick recovery on reestablishment of connectivity.

Potential impact

This enables a connected society, as IoT/edge devices or humans on the move continuously become disconnected from the cloud continuum.

4.2 Generative telco cloud

Context and challenges

The telecommunications industry is facing growth and revenue challenges. To enhance operational efficiency, it must embrace automation and foster flexibility to reduce costs. Concurrently, there is a pressing need to explore new revenue streams to sustain its financial health. There is an urgency for Europe to accelerate its cloud and network infrastructure developments to support the capabilities of the telco industry.

R&D priorities

Implement cloud computing to scale infrastructure dynamically based on demand, combined with Generative AI tools to automate processes, optimize network resources, and enhance operational efficiency. To achieve this, an ecosystem of service providers and network operators is required. Networks need to be designed horizontally and federate open network APIs to integrate mobile operators and cloud services.

Potential impact

- *Efficiency and Scalability*: Telcos will have flexibility, cost-effectiveness, and the ability to rapidly deploy and manage services, ensuring efficient infrastructure utilization and capacity management.
- *Customer Experience Enhancement*: AI-powered chatbots and virtual assistants will handle customer queries, troubleshoot issues, and personalize interactions. Predictive analytics will anticipate customer needs, leading to proactive service offerings.
- *Revenue Growth and Cost Reduction*: they will be able to monetize AI by offering data analytics services to other industries. Insights derived from telco data can drive business decisions across sectors. AI-driven content recommendations can boost revenue by increasing user engagement.
- *Predictive Maintenance and Network Optimization*: Telcos will be able to enhance network reliability by identifying potential issues before they impact services as well as automate processes, optimize network resources, and enhance operational efficiency.
- *Predictive Protection Actions*: User's exposure to specific threats, leveraging the information of threat frameworks to facilitate the adoption of preventive actions. Behaviour based analytics through artificial intelligence techniques will assess if the behaviour of a group of users is akin to the behaviour of a known threat changing Network level security posture dynamically.

References

- The GSMA Open Gateway initiative, launched in MWC Barcelona 2023 and currently has 47 operator groups already on board.
- During MWC Barcelona 2024, Huawei Cloud showcased how it is building an AI-ready cloud infrastructure.
- Cloud Native Elite Club (CNEC) stated at MWC Barcelona 2024 that looking ahead to 2024, CNEC will prioritise the strategic integration of cloud native and AI technologies, an effort to accelerate the global adoption and advancement of AI.

4.3 Open radio access network (O-RAN)

Context and challenges

The radio access network (RAN) is mostly closed in terms of an "open development space" except for large OEM telecommunications companies. Off-the-shelf hardware, open software, existing licenses *and* licensing, standards access, protective patents, as well as willing customers prohibit entry into RAN & its development. Unlike cloud native, where 3rd parties can develop microservices, advertise and sell/license, the RAN is a HW+SW monolith which is a hurdle to enter and offer services within. Antennas, base stations, MSCs, 3GPP + IETF protocol development + plus global testing are huge undertakings.

An open RAN is the vision of some, who are out of the traditional OEMs group, through alliances such as the O-RAN Alliance[29] and the Telecom Infra Project (TIP)[30] which consolidate interested parties, identify missing parts and provide a base on which to build projects. O-RAN is very much discussed, and as of 2024 fall into 2 categories: i) large players 'promising' to develop an open solution for the benefit of all, e.g., the Ericsson - AT&T $14 billion deal,[31] or ii) smaller component-like development, a la Cloud native. Either way it remains to be seen whether either will deliver a truly open, competitive RAN.

R&D priorities

Identify parts outside the existing alliances that could be promoted as topics for future EU calls with OEM partnerships. Delve into the 6G and 7G generations that are still open before they are subsumed by the 4 major telecom players.

Potential impact

Promote open and fair competition in the telecoms market. Lower entry cost for new operators resulting in a more competitive market. Make the possibility of European lead in telecoms whilst breaking the dominance of US hyperscalers in the Cloud market. Much more open-source in telecommunications solutions in the future, à la SIP servers, Asterisk (Open PABXs).

[29] https://www.o-ran.org.
[30] https://telecominfraproject.com.
[31] https://www.ericsson.com/en/news/2023/12/ericsson-and-att-in-major-future-network-of-the-future-deal.

4.4 Resilient on-prem 5G/6G edge-cloud for Industry 4.0 and beyond

Context and challenges

The seamless delivery of advanced data services and AI-based applications is integral to Industry 4.0, necessitating robust on-premises 5G/6G edge cloud infrastructure to ensure resilience and reliability amidst the challenges of integrating legacy automation systems with cutting-edge cloud technologies. In many industrial applications resilience and reliability are critical. The automation systems in industry often rely on old computer systems, making it difficult to seamlessly integrate them with modern cloud technologies and keep the systems up to date. Most of the systems need to be on-prem very close to the production line and it is therefore not possible to use public telecom and cloud resources that might suffer from short disconnects, making it necessary to have private on-premise 5G/6G edge cloud deployments.

R&D priorities

- Portability between public telco operator environments and private 5G/6G environments, compatibility with APIs, and using cloud-native technologies.
- Seamless management of IoT devices that do not use 5G/6G.

Potential impact

Easier to seamlessly port applications and services between private on-prem environments and public cloud-edge.

References

- The telco global API alliance: https://camaraproject.org.
- Open source 5G Core network: https://magmacore.org.

5 Cloud-Edge Use Cases

The cloud-edge computing continuum will enable both new use cases and applications across various sectors, and transform existing ones. The specific requirements that these pose on the cloud-edge infrastructure, and how it is operated and managed, will differ between use cases.

5.1 Software-defined vehicles

Context and challenges

The industry is now moving towards Software Defined Vehicle (SDV) architectures where features are updated continuously over the air, and connectivity and cloud services offer new possibilities, as well as risks. The SDV provides a use case spanning the whole continuum from the Edge, to IoT, to Cloud Computing, and integrates directly to the smart city and neighbouring environments, especially with the rise of autonomous driving. It is costly for each manufacturer to develop their own solutions. Therefore, consolidation in the market at top-level is needed, e.g., Tacton (Scania, VW), Stellantis (Opel, Vauxhall, Peugeot). There has been previous disharmony with competing standards such as 5G (3GPP-based) and G5 (IEEE-based), 5G versus IEEE 802.11p. Cooperative Intelligent Transport Systems (C-ITS) for vehicle-to-infrastructure communication could become fractured.

R&D priorities

- Better standardisation through EU organisations. Leverage open-source, as in IT as de-facto solutions. Proposals from the community. Ref: Report.
- *Promote Open Standards*: ensure that any standards promoted are implementable in open-source software (OSS).
- *Establish Functional Safety Norms*: Collaborate with industry consortia and stakeholders to develop clear guidelines and standards for functional safety certification that take into account and leverage the benefits of OSS characteristics.
- *Facilitate Collaboration*: Cultivate environments that promote collaboration among OEMs, suppliers, and OSS communities to drive innovation.
- *Educate on OSS*: Implement educational initiatives about the benefits and best practices of OSS for businesses, particularly in sectors like automotive.
- *Consider OSS as a Geopolitical Concern*: Take proactive steps to understand the geopolitical landscape of software and its implications for the industry. Adopt and support open technologies, including OSS, open standards, and open hardware, both in industry and government, to enhance regional sovereignty and global competitiveness.
- *Engage with Global OSS Sustainability Efforts*: Governments should engage in and encourage industry participation in global OSS sustainability efforts to ensure the longevity and support of OSS. Support may include diverse options such as funding critical projects, promoting OSS in public procurement, raising awareness and knowledge of the risks related to the global supply chain, and encouraging organisations to sustain their own upstream dependencies.

Potential impact

Open-source software (OSS) is pivotal to enabling the transition towards the SDV but is today mainly limited to non-safety critical areas such as infotainment. The industry is slowly finding its way in moving towards a networked collaboration structure and function-safety compliant development processes and defining the non-differentiating arenas where actors can collaborate with standardized interfaces that allow for modular and flexible configurations of the future SDV.

5.2 Decentralized spatial computing for AR/VR technologies

Context and challenges

Extended reality (XR) technologies, such as augmented reality (AR) and virtual reality (VR), are expected to transform everyday life fundamentally, and enable new use cases and applications in industry. Due to the high data throughput and low latencies required to deliver a seamless VR/AR/XR experience, these technologies are among the main use cases for 5G and edge computing technologies.

R&D priorities

- *Consistency in Virtual Environments*: Develop algorithms to ensure seamless and consistent user experiences across decentralized networks.
- *Latency Reduction*: Deploy and reinforce telecom and network infrastructure to minimize latency for real-time interaction within VR environments.
- *AI Integration*: Integrate AI to dynamically adapt and optimize VR experiences based on user interaction and environmental changes.
- *Interoperability Standards*: Establish standards to ensure interoperability among diverse VR platforms and decentralized computing resources.
- *Secure Data Exchange*: Create secure protocols for spatial data exchange in decentralized VR applications to protect user privacy and data integrity.
- *Operational Datasets*: Develop open operational datasets for training and evaluating AI models within VR scenarios.

Potential impact

VR/AR/XR are enabling new use cases and applications, with the potential of bringing great value to industry and enterprises.

5.3 Cyberspace and physical space fusion

Context and challenges

Originating from Japan, the concept of "Society 5.0" envisions a connected cyberspace where AI surpasses human capabilities, feeding optimal outcomes back into the physical realm.

Ensuring data privacy and security on the Cloud is paramount as AI systems process vast amounts of data. High latency can hinder performance, particularly in critical applications such as healthcare or autonomous vehicles, where real-time processing is essential. Additionally, AI systems must efficiently scale with increasing data volumes and user demand to maintain optimal performance. The computational power required for AI can be costly, especially with the utilization of large neural networks, thus affecting overall expenses. Interoperability is crucial as AI systems often need to integrate with various other technologies and data sources, necessitating seamless compatibility. Moreover, regulatory compliance, particularly regarding data protection regulations like GDPR in Europe, is vital for AI systems operating on the Cloud, emphasizing the importance of cross-border data flows and specialized legal expertise. Lastly, addressing ethical considerations surrounding AI's integration into society, alongside ensuring data quality and standardization, remains imperative to harness its full potential while mitigating risks and limitations.

R&D priorities

- *Data Privacy and Security*: There are risks associated with data breaches and unauthorized access, which necessitate robust encryption and security protocols.
- *Latency and Performanc*e: Cloud infrastructure must be optimized to minimize latency and provide the necessary computational power.
- *Scalability*: Cloud platforms need to provide flexible and scalable resources to accommodate the growth of AI applications without compromising performance.
- *Cost*: Organizations must manage the cost of Cloud resources effectively to make AI integration economically viable.
- *Interoperability*: Ensuring interoperability between different Cloud services and AI models requires standardized protocols and interfaces.
- *Regulatory Complianc*e: AI Governance platforms: Automate the identification of regulatory changes and translation into enforceable policies, Risk management and lifecycle governance.
- *Technical Expertise*: Training and recruiting talent are necessary to bridge this gap and drive integration forward.
- *Ethical Considerations*: Ensure that AI systems are transparent, explainable, and aligned with human values.
- *Data Quality and Standardization*: Ensure that shared data is of high quality and standardized for interoperability.

Potential impact

The convergence of cyberspace and physical space envisages a seamless integration, facilitating smart cities and environments where data exchange between sensors and cyberspace optimizes resources and enhances quality of life. Society 5.0 embraces AI-driven analysis, wherein AI not only processes data but also offers feedback and solutions, augmenting decision-making. This feedback loop, empowered by AI, is anticipated to spawn new value across sectors, fostering economic growth, job creation, and societal well-being.

6 Interesting topics with potentially disruptive impact

In addition to the sections and topics above, we have identified several topics of interest that could have a potentially disruptive impact on the cognitive computing continuum in the longer term.

6.1 Integration of neuromorphic systems

Context and challenges

Traditional digital computers are struggling to keep up with the demands of advanced technologies and applications, such as AI and Edge computing, as well as the optimization of intricate systems. This is evident in simulations used in chemistry and pharmaceuticals to speed up the development of new medications and vaccines, in the process optimization of complex production lines, as well as in the area of traffic and freight transport. Neuromorphic computing, which aims to emulate the self-organizing and self-learning nature of the brain, offers a promising solution to this challenge. Despite its potential, neuromorphic hardware has not found its way into commercial AI data centres. One reason is due to insufficient opportunities for application-oriented testing of the hardware developments required for the highly complex computing technologies, as well as for a rapid implementation of the results in prototypes and small series. Another is the lack of standardized neuron models and common training techniques.

As an example, the EU-funded HYBRAIN[32] project aims to develop a new computing system that is inspired by the human brain. This system will be based on ‘ultra-fast response’ technologies, combining a number of highly innovative solutions based on integrating complementary technology platforms. Companies like SynSense and Innatera are designing neuromorphic hardware specifically for integration into consumer devices, for sensor-adjacent processing. They have deployed several audio inference use cases to Xylo Audio, and a bunch of vision inference use cases to Speck. Moreover, the first EU exascale supercomputer JUPITER will include a neuromorphic module.

R&D priorities

- *Hardware-software codesign*: Seamless integration of neuromorphic computing capabilities in the cloud-edge continuum. This will require adapting existing software stacks and architecture patterns commonly used today and creating novel interfaces between the neuromorphic and digital computing domains.
- Event-based paradigm: Neuromorphic computing is based on a different architecture than the commonly used von Neumann architecture. The event-based nature of neuromorphic computing systems will have new requirements on the computing

[32] https://cordis.europa.eu/project/id/101046878, https://hybrain.eu/.

continuum, since the communication and computing patterns will differ from those of traditional systems.

Potential impact

The integration of neuromorphic hardware into data centres and at the edge could lead to more sustainable and energy-efficient AI processing. This could revolutionize the processing of sensor data at the edge, opening new possibilities for applications in wearables, smart home, and IoT devices. The development and improvement of neuromorphic software can lead to more efficient and powerful AI systems. It can enhance the overall learning performance for specific tasks, moving away from hardware benefits to understanding the potential application benefits of neuromorphic computing. It can also lead to improvements in neuroscience as researchers start to recreate our grey matter in electronics. With its ultra-efficient, event-driven processing capabilities, neuromorphic hardware enables unprecedented power-performance gains for always-on sensing use cases. This could lead to significant advancements in various fields, including chemistry, pharmaceuticals, production lines, traffic and freight transport, and more. Other communities, such as ESA, are also exploring neuromorphic-oriented calls. Event-driven processing, edge computing, and GDPR compliance are essential considerations for successful integration.

6.2 Space edge

At the moment, the situation on both satellite integration and space edge seems to be that they are firmly in scope for the SNS JU, but the Cloud and IoT units might ponder how to find synergy and anchorage collaboration between the SNS projects and the CL4/DATA- or CL4/DIGITAL-EMERGING- ones. Beyond those two points, there are several other points that suggest a convergence and synergy:

- Mention of Edge Computing, with or without "Space" in front, is ubiquitous.
- Flexible service orchestration is also quite prominent.
- Virtualisation and Cloud Computing techniques (e.g., FaaS) overlap with Virtual Network Functions (VNFs).
- Data-driven AI is being considered, if not already adopted as a key enabler for 6G networks.

Ongoing projects and initiatives

SNS-JU RIA: 6G-NTN: (NTN project, using AI, Stream B, Communication Infrastructure Technologies and Devices). This project is about the complete integration of Terrestrial Network (TN) and Non-Terrestrial Network (NTN) in 6G (5G considers TN and NTN as two separate segments and does not try to cross-optimise anything between them). The NTN does not consider only satellites, but anything flying above the surface, so airplanes, drones, and similar flying **vehicles**. There is also mention of space edge in some use cases:

- Public deliverable D2.1: Use Cases Definition. Mention of space edge (page 13, page 15, page 16, page 61).
- Public deliverable D3.1: Report on 3D multi layered NTN architecture (1st version). Description and assessment of the various components of the multi-layered NTN.

The 6G-NTN project has already space edge as one of its potential innovations, particularly relevant to *UC3: Urban Air Mobility* and *UC4: Adaptation to PPDR or Temporary Events.*

There is an SNS call, open now, 2024-STREAM-B-01-03, which should further the work of projects like 6G-NTN. An interesting quote from the call: "*Disaggregation, and virtualisation considering the ground and non-terrestrial segments are in scope. It should enable integrated space and ground edge computing and in-space traffic decision procedures allowing a 'router in the space'.*". There are also mentions of IoT-to-satellite connectivity.

6.3 Integration of quantum computing infrastructure

Context and challenges

The current infrastructure for quantum computing (QC) is limited, making it difficult to fully leverage the capabilities of QC and integrate it with HPC and cloud services to enable efficient AI workloads.

Quantum computing is a promising technology that could revolutionize various sectors. However, the lack of a robust infrastructure, including hardware and software, hinders its full potential. The challenges include developing scalable quantum hardware, creating efficient quantum algorithms, and integrating QC with existing technologies like HPC and cloud services. Some relevant current initiatives include:

- HPCQC EU Project: High Performance Computer – Quantum Simulator hybrid.
- OpenSuperQ: A Quantum Computer for Europe.
- Quantum Flagship (and QUCATS): One of the most ambitious long-term research and innovation initiatives of the European Commission.
- Six EuroHPC sites selected to host first European quantum computers.[33]
- QuIC – Quantum Industry Consortium: QuIC is a pan-European non-profit association dedicated to the advancement of commercial quantum solutions and Europe's competitiveness in quantum technology on the global stage.
- QIA – Quantum Internet Alliance: QIA is a consortium of ~40 world-leading institutions working together to build a global Quantum Internet made in Europe.
- The European Quantum Communication Infrastructure (EuroQCI) Initiative.

R&D priorities

The European Commission could fund projects focusing on addressing research challenges in QC, the development of large-scale quantum computing testbeds, the creation of efficient Open Science (optimised for RISC-V) quantum algorithms, and the integration of QC with HPC and cloud services. This could involve the use of the Modular Supercomputing architecture in

[33] https://eurohpc-ju.europa.eu/selection-six-sites-host-first-european-quantum-computers-2022-10-04_en.

the first EU exascale supercomputer JUPITER, which includes quantum and neuromorphic modules.

Potential impact

By funding these areas, the EC could significantly enhance Europe's computing capabilities, including QC, leading to advancements in various sectors. This could foster innovation, create jobs, and strengthen Europe's position as a global leader in technology. It could also help address societal challenges, such as climate change and healthcare, by enabling more sophisticated data analysis and decision-making. It could also foster innovation, create jobs, and strengthen Europe's position as a global leader in technology.

6.4 Quantum and classical computing fusion

Context and challenges

The primary challenge lies in the inherent limitations of classical computing, which uses binary bits and struggles with problems involving exponential variables. Quantum computing, on the other hand, uses qubits and can process a richer set of possibilities, making it more efficient for certain types of problems. However, quantum computers require specialized environments and operate at cryogenic temperatures, which makes them impractical for personal devices like laptops.

R&D priorities

The solution to making quantum computing widely accessible is through cloud computing. Quantum calculations will be performed in data centres, and the results will be delivered through the cloud. This approach will allow the masses to benefit from quantum computing without the need for personal devices to handle the actual quantum processes. The fusion of Quantum Computing and Classical Computing in the cloud represents an exciting frontier in computing, promising to address problems currently beyond our reach and driving innovation and problem-solving in the future.

Potential impact

The convergence of quantum and classical computing in the cloud is expected to exponentially increase computational capabilities. This integration will enable software to determine the most efficient computing paradigm—classical, AI, or quantum—for different parts of a computation. Users will benefit from this combined power without being directly involved in the complex processes.

7 Conclusions

This first public draft version of the Cognitive Computing Continuum Research and Innovation Roadmap for Europe contains an initial analysis of relevant projects and initiatives, and identifies an initial set of R&D priorities for the cloud-edge continuum. It complements the industrial roadmaps, innovations, and vision from the EU Cloud Alliance and the new IPCEI-CIS, with a view towards R&D needs in the medium to long term. In addition to the four main topics, it identifies several areas that might be further developed in future versions of the roadmap:

- Neuromorphic computing.
- Quantum computing.
- Space edge.
- Hyper-decentralized edge.

Future versions of the roadmap will cover four key transversal topics (aligned with focus areas under the Cloud-Edge WG of the EU Cloud Alliance) in dedicated sections:

- Sovereignty & Open Source.
- Sustainability.
- Interoperability.
- Cybersecurity.

Dedicated Working Groups will be set up as part of NexusForum.EU, corresponding to the main topics and the transversal topics, to further develop and validate this roadmap and engage the broader community in its development. More details about this can be found in the deliverable D4.1.

NEXUS FORUM EU

Consolidating Research and Policy along the Cognitive Computing Continuum

eucloudedgeiot.eu

Funded by the European Union

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI