Cognitive Computing Continuum

Executive Summary

This NESSI[1] position paper analyses the research and innovation landscape of the cognitive computing continuum and makes suggestions for future research directions in this field.

The Horizon Europe Work Programme 2021-2022[2] called for research in the computing continuum to address two main goals: a higher-level operating system for the smart Internet of Things, embedded in a compute continuum from IoT-to-edge-to-cloud, and an AI-enabled cloud-edge framework which automatically adapts to changes in application behaviour and data processing needs. The EC workshop "Digital autonomy in the computing continuum"[3,4] identified further research challenges regarding the computing continuum and stressed the importance of AI-enabled management and orchestration approaches across the entire continuum. NESSI subscribes to these proposals and suggests additional research targets with the aim of developing and exploring the full potential of a cognitive computing continuum.

What is a cognitive computing continuum and why is it important?

A cognitive computing system is often characterised by four main properties: the ability to adapt its behaviour during runtime in response to changing requirements; the capability to interact with its environment, including devices, processes, and people; identifying context and extracting situational awareness; and advanced decision making based on context information and insights learned from the past. Cognitive systems can take decisions and adapt behaviour even in face of ambiguous and incomplete information. The extensive use of AI enables these cognitive properties.

The computing continuum aims to provide compute, storage and networking resources across an infrastructure that comprises centralized public and private clouds as well as distributed edge, fog, hybrid and multi- clouds. The compute, storage, networking, and services resources of such a computing continuum are heterogeneous, dynamically changing, and often spread over multiple administrative domains. Accessing these distributed resources need to be possible on-demand and in a unified, programmatic, and secure way. This unified access will be an important prerequisite to fully automate the software lifecycle across the computing continuum.

When deploying and migrating workloads across the computing continuum, SLAs and workload specific requirements need to be met. These requirements are not limited to the performance or availability of a service, but can also be related for example to privacy protection, low energy consumption, and cloud cost optimization. Cost optimization and the efficient use of cloud resources are becoming more pressing as some software companies spend up to 80% of their revenues[5] on cloud services, and an estimated one third of this

[1] NESSI (Networked European Software and Services Initiative), the European association promoting research, development and innovation in the field of software, data and digital services; https://www.nessi.eu/
[2] https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/wp-call/2021-2022/wp-7-digital-industry-and-space_horizon-2021-2022_en.pdf
[3] https://nessi.eu/digital-autonomy-in-the-computing-continuum/
[4] https://www.h-cloud.eu/news/highlights-of-the-ec-workshop-digital-autonomy-in-the-computing-continuum/
[5] https://a16z.com/2021/05/27/cost-of-cloud-paradox-market-cap-cloud-lifecycle-scale-growth-repatriation-optimization/

cloud spend goes to waste[6] because cloud resources are overprovisioned and not used efficiently. An effective and flexible resource management and orchestration system is essential to address these issues and to cope with the complexity of the computing continuum.

Enriching the computing continuum with cognitive capabilities promises to enable significant enhancements to these highly demanding management tasks. A cognitive computing continuum will be self-learning, able to contextualize information and to use the learnings and context information to make in-the-moment decisions and adaptations. Accessing and using the resources of a cognitive computing continuum will not require specific expertise in cloud, edge or IoT computing. Thus the cognitive computing continuum will help to counter the lack of specialists in these technologies, which is another top challenge that companies face.

Recent developments in AI technologies offer various approaches to providing these cognitive capabilities, e.g. predicting the load of infrastructure nodes through machine learning algorithms trained on historical data, identifying optimum infrastructure configurations by using reinforcement learning and relying on real-time feedback and context information, and generating and adapting management workflows by applying nature-inspired algorithms. Advances in explainable AI will add transparency to the AI-based decisions and will increase the level of trust in these decisions.

Research challenges of the cognitive computing continuum

The EC workshop "Digital autonomy in the computing continuum"[3,4] provided an overview of ongoing and planned research into the development of the computing continuum and the application of AI in this context. The Horizon Europe Work Programme 2021-2022[2] called for research into a management framework for the cognitive cloud, into programming tools and environments for the compute continuum, and into the development of a meta operating system for the edge. All this research is complemented by activities developing open architectures and supporting open-source software projects for cloud-based services and cloud/edge processing. The next Horizon Europe Work Programme will call for piloting and validating the edge programming environments and platforms, and further research into the management of the computing continuum and efficient, secure and 'green' data processing across the continuum.

NESSI considers the cognitive computing continuum as an extension of today's cloud computing and thus as a key enabling technology for future digitalisation. In addition to the challenges already covered or planned to be covered in Horizon Europe, NESSI sees the need for research to address the following challenges.

AI-driven active exploration of potential solutions

A virtual agent of an automated management system in the cognitive computing continuum will face situations that are unexpected and unknown from the past. AI algorithms, such as reinforcement learning and evolutionary computing, are supposed to resolve those situations by exploring possible solutions and identifying an optimal or near-optimal solution based on the feedback information obtained during the exploration. A critical issue for this approach is to ensure that the exploration will not interfere with the normal operation of the computing continuum. Conducting the exploratory AI work based on simulation and digital twins of the computing continuum resources may be one way to resolve this issue. Another possibility may be to leverage safe reinforcement learning, which constrains exploration to known-to-be-safe actions.

Multi-agent, distributed and decentralized intelligence

Management decisions will often made by virtual agents running in different parts of the continuum and using different datasets and knowledge. The challenge is how to coordinate the distributed agents and their

[6] State of the cloud report, Flexera 2022; https://info.flexera.com/CM-REPORT-State-of-the-Cloud

decision-making so that an effective and optimal management of the whole continuum is achieved. Coordinating and controlling the distributed decision making will be even more challenging when the management objectives go beyond quality of service and include objectives related to the resilience, robustness, security, safety, or privacy protection of and within the computing continuum. As with all AI-based complex systems, the overall system behaviour of the cognitive computing continuum emerging from the decisions taken by the distributed autonomous agents might be unexpected and not tolerable. Mechanisms need to be in place to cope with these situations.

Human in the loop and intent-driven management

A cognitive computing continuum manages computing, storage and networking resources for human stakeholders, including application developers and end users, whose requirements and expectations are varying and might be conflicting. The management system of the cognitive computing continuum should provide the best possible service that meets the needs of these human stakeholders. A fundamental research question in this context will be how to support human stakeholders to express their intents and to interact with the management system correspondingly. For example, today's approach using domain-specific languages for describing an intent need to evolve into more advanced, flexible, and business-oriented approaches that allow for description of conditions and preferences for deployment, possible data dependencies, etc.

Trustworthiness of the cognitive computing continuum

Closely related to the above issue of the human in the loop is the trustworthiness of the cognitive system. Transparency of AI-driven processes, being in control of these processes, and clarity regarding who is accountable and liable in case of system failures, are all essential for building trust in a cognitive system. Explainable AI (XAI) will play a key role in this context. However, XAI is still very much a research topic, and additional effort is needed to investigate how emerging XAI approaches can be best embedded in a cognitive computing continuum and how these approaches can be made user-friendly, to effectively support human stakeholders in understanding AI decisions.

Continuum-native software

Today's cloud applications software is designed, developed and implemented to maximise the advantages offered by a cloud. This relies on technologies such as containers, service meshes, microservices, immutable infrastructures and declarative APIs. These enable scalable applications optimized for running in the cloud[7].

With the evolution of the cloud to a cognitive computing continuum, cloud-native software engineering will evolve into cognitive-continuum-native software engineering. There will be a unified management layer across all the heterogeneous and distributed resources of the computing continuum. This management layer will be enriched with cognitive capabilities which are exposed to applications deployed on the continuum. There will be abstraction mechanisms which allow for loose coupling between applications and infrastructure, and support dynamic placement optimization. The application software needs to be engineered to make best use of these capabilities and mechanisms.

Preparing and sharing data

Getting AI to work depends heavily on data and the availability of data. Data and datasets need to be accurate, complete, properly formatted, and accessible. It is also challenging to balance data quality and the protection of data privacy and sovereignty, ensuring that the rights and wishes of the data owners are respected.

[7] https://github.com/cncf/toc/blob/main/DEFINITION.md

In cloud computing, the monitoring of workloads provides cloud providers with a huge amount of raw data that can serve as a basis to train and run AI models for managing cloud resources. However, before being used for AI, datasets need to be pre-processed, including the cleaning and labelling of data. Selecting the relevant data and performing the pre-processing requires not only significant effort but also special expertise combining know-how about AI, cloud management, and the application domains of the workloads. Furthermore, sharing raw and/or pre-processed data as well as sharing of AI models among all stakeholders of a cognitive computing continuum would be needed for its efficient operation. Innovative ecosystems and business and collaboration models going beyond competitors' boundaries are needed to incentivise collaboration and knowledge exchange across the computing continuum.

Recommendations

To meet the challenges identified above, NESSI recommends adding software-related research topics in Horizon Europe to support the following.

- Specific research focusing on approaches for solving multi-dimensional optimization problems in multi-agent, distributed cloud management systems will be essential. The application of digital twin and simulation technologies should be investigated in this context.
- Research across the disciplines of AI and cloud computing is required to develop the cognitive capabilities of the computing continuum. The latest advances in AI technologies and their applicability in the context of the computing continuum need to be explored. Simulations and digital twins will support this exploration. Explainable AI and its integration into distributed cloud systems will play a key role for the trustworthiness of a cognitive computing continuum. Joint work by AI researchers and cloud/edge/networking researchers is needed to address the co-evolution of AI and the management of the computing continuum.
- Developing solutions for the intent-driven management of the cognitive computing continuum calls for interdisciplinary research across application domains, including languages and flexible approaches that allow expressions of intent in diverse application domains and their translation into actionable steps in the continuum management system.
- Developing and exploring continuum-native software engineering principles will be key for building applications that make best use of the capabilities offered by the cognitive computing continuum.
- Dedicated efforts are needed to facilitate the sharing of data and AI models. These efforts should include not only related research activities but also actions aiming at platforms and frameworks which set clear rules for the sharing and trading of data and AI models among the stakeholders involved in the cognitive computing continuum.
- Trials and pilots will be needed to verify the concepts developed and to demonstrate the advantages of the cognitive computing continuum.
- Metrics and benchmarks will be needed to measure the effectiveness of the cognitive computing continuum. The metrics should cover a broad range of diverse indicators including traditional quality of service parameters and also indicators that can be used for cost optimization and the measurement of the sustainability performance of the system.