# Introducing HiPEAC's vision for the future: The next computing paradigm

The HiPEAC Vision seeks to set a long-term vision for the future of computing systems. Hence the main directions are quite similar from one edition to the other, although each edition has inflexions deriving from what is currently going on in computing systems.

As such, the 'races' introduced in the HiPEAC Vision 2023 are still valid (indeed, increasingly so), as is the proposed direction towards the 'next computing paradigm' outlined in the HiPEAC Vision 2024, which even more achievable due to the current advances in science and technology.

It is obvious that the most influential element between the HiPEAC Vision 2024 and this Vision 2025 is the exponential progress of artificial intelligence (AI). This is reflected in this edition, where the two main highlights are how to realize the NCP and the impact of AI. They are even merged into short-term recommendations: using the emergence of distributed agentic AI to set the basis of the NCP technology, i.e. to be the blueprints of what could be a more generic and omnipresent NCP, but one which is adapted to the particular case of distributed agentic AI. We will see in the part of this vision related to artificial intelligence that it is logical because there are close similarities of requirements and technologies between both.

The main focus of the HiPEAC Vision 2025 is therefore the NCP, and its implications in different domains: artificial intelligence, new innovative hardware, tools to develop more efficient hardware and software, cyber-physical systems, cybersecurity, and sustainability. This is complemented by an analysis of the position of Europe, with suggestions of how to improve Europe's position in relation to the global races.

But let's start with an explanation of what the NCP is.

## What will be the future of computing systems (hardware, software and infrastructure)?

The world of computing is evolving at a dramatically fast pace because of the impact of artificial intelligence, cyberattacks and systems that are increasingly integrated with the physical world.

This HiPEAC Vision 2025 describes how these trends could converge into the 'next computing paradigm' based on the federation of distributed elements working and orchestrated together in order to form a complete computing continuum. The NCP aims to play to the strengths of Europe, such as its capacity to develop edge and on-premise devices, and relying on an ecosystem of small and medium enterprises.

From a technical point of view, the NCP emanates from the convergence of multiple foundational technologies, including the web, cyber-physical systems (CPS), cloud computing, the internet of things (IoT), digital twins, artificial intelligence (AI), and more, into a coherent, federated ecosystem. This paradigm is characterized by a deeper integration between machines and humans, creating a 'web of machines' that must interoperate seamlessly with the 'web of humans'. The NCP will not only process data in cyberspace, but will also operate within real-world constraints such as safety, time sensitivity, and location, using technologies like digital twins to optimize efficiency across spatial and temporal dimensions.

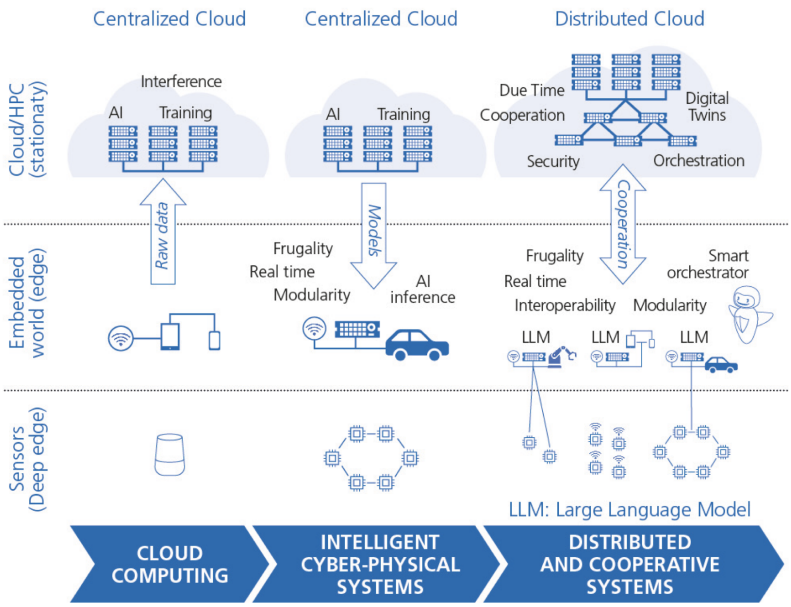


Figure 1: Evolution of computing infrastructures towards the NCP, where services are distributedand cooperate together. Credit: Denis Dutoit, CEA

A key aspect of the NCP is the concept of 'anything as a service' (XaaS), where applications are dynamically composed from various services, often orchestrated by AI-powered systems that ensure efficiency, security, and user trust. These services will be distributed across cloud, edge, and other decentralized environments, depending on the user's needs and global efficiency. The orchestration of these services will be critical, requiring smart systems to manage complex interactions while safeguarding user privacy and data security. This shift also emphasizes interoperability, allowing applications and services to function across different hardware and software platforms, with an increasing focus on modularity, frugality, and real-time processing.

The web has shown that digital resources can be given uniform representations and identities, and can be operated upon by CRUD (create-read-update-delete) service primitives exposed by HTTP verbs. In the next-generation web, which brings together the web of humans with the digital web into a programmable and interoperable hyperspace, the XaaS paradigm becomes a major vector of innovation, which shifts the centre of gravity away from the cloud towards the edge, enabled by 'digital envelopes'.

'Digital enveloping' is the technology-enabled concept by which any item of reality, human, material and immaterial, may be associated with a computable digital representation capable of delegated autonomous action. That capability is provided to individual digital envelopes by the combined operation of three key components: an intelligent digital agent able to pursue goals legitimately assigned to it; sensors, to pull inputs from designated

sources (in the physical world or other digital envelopes); and actuators, to push outputs into designated targets (physical things or digital envelopes).

Digital envelopes have owners, who are the sole entity authorized to communicate goals to them. The digital agent of the digital envelope should receive those goals and translate them into a permissioned orchestration of request-response interactions with other digital envelopes (and thus of the digital agents within them). Thanks to actuators, those interactions may take effect on the physical world or on the digital sphere or both. Those effects might be 'sensed' by other digital envelopes and possibly further 'acted' upon to adjust to emerging needs arising as a function of local and global constraints.

Digital envelopes evolve the concept of 'digital twin' in scope and capability. In scope, no longer confined to an encapsulated digital space, but capable of actuation into the physical world. In capability, via the capability of autonomous planning and execution in pursuit and accomplishment of assigned goals.

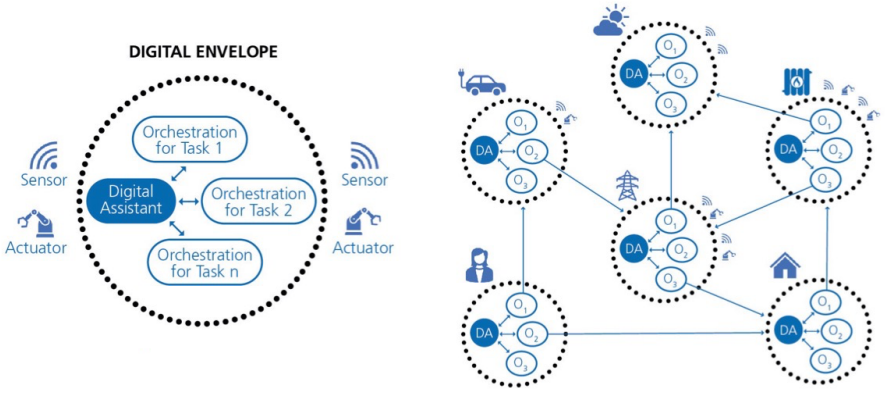


Figure 2: The digital envelopes interacting together

The enactment of the NCP will necessitate advancements in protocols and architectures that support 4D computing—spatial and time-aware operations. This includes enhancing current web protocols to meet the demands of real-time, location-dependent computing environments. AI will play a pivotal role in orchestrating these systems, enabling more natural interactions with humans and ensuring that services are securely and efficiently delivered. Ultimately, HiPEAC envisions the NCP as an infrastructure of highly distributed, cooperative, and intelligent computing ecosystem of federated technologies, which spans diverse sectors and breaks down traditional research and engineering silos, fostering innovation through shared and coordinated resources.

Envisioning the NCP starts from anticipating the evolution towards a 4D computing paradigm that elevates the computing space from the two dimensions of document-based resources into a full-fledged 3D spatial representation, plus time. That will be further enhanced with a coherent continuum of computing that intertwines the real world and its constraints with the cyberworld, incorporating generative AI, enabling dynamic orchestrations of resources in order to achieve what is requested by users. This evolution will create a seamless, multi-level networked cooperative structure where resources are accessed and manipulated as needed with streamlined web-type protocols, and where programs (or 'services') and data flow smoothly onto computing resources that cooperate with each other, enhancing context awareness and efficiency in digital interactions.

## A seamless flow of compute and data across the continuum

Cloud computing has become the dominant model for most end users. Through the offers of ‘software-as-a-service’, ‘platform- as-a-service’ and ‘infrastructure-as-a-service’, it facilitates access to rich applications without the need for significant capital investment and has allowed digital businesses to thrive.

Encompassing the bulk of computing resources, the cloud has therefore become the centre of gravity for computing, with users and data being drawn into its pull. However, vast amounts of computing resources are also available, cumulatively, at the edge of the network and in intermediate layers between datacentres and the edge, where users, usage and data are located. If those resources were pooled together seamlessly, à la cloud, innumerable value-added computations could take place in this continuum of computing rather than in the cloud. This would offer latency and energy reductions, decentralization, personalization, privacy and context awareness in a way the cloud could not possibly match.

Pooling edge resources and joining these with cloud resources gives rise to the edge-cloud continuum, a compute infrastructure where computation may be deployed opportunistically and dynamically, wherever it is most convenient for the user.

Extending the cloud service model to ‘anything-as-a-service’ is another important vector of innovation that shifts the centre of gravity towards the edge. Enabling the ‘anything-as-a-service’ model requires the ability to orchestrate services that execute at various places along the computing continuum from edge to cloud, both in the physical world via IoT sensing and actuating, and in the digital-twin sphere. Services are not only software, but also hardware resources such as compute power, storage, etc. The NCP proposes a dynamic mapping of software services to hardware services, allowing not only the movement of data (like today), but also of code, allowing a real opportunistic edge-to-cloud execution of services. This migration of ‘code’ implies security concerns, hardware enforced silo (trust zones), and compatibility of code to be executed potentially on systems with various instruction sets (ISA).

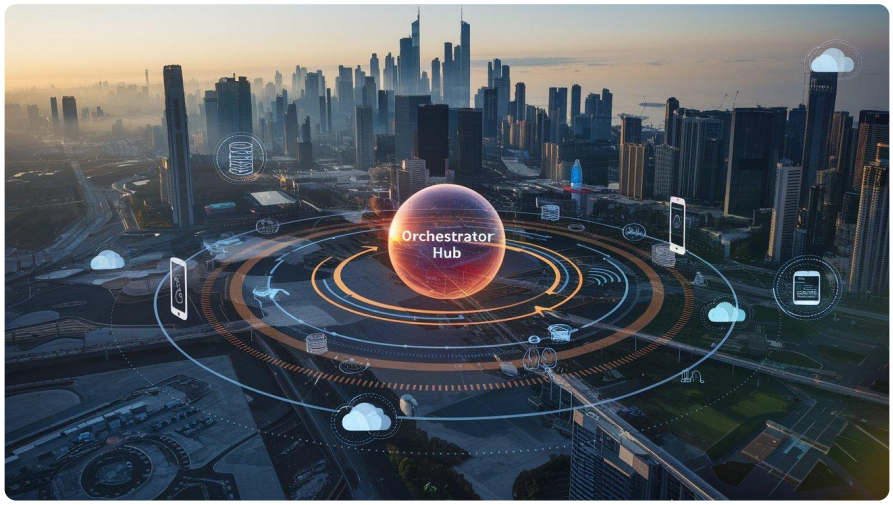


Figure 3: orchestration is at the core of the NCP

The orchestration is in charge to maintain a balance between resource availability (associated with the centre of the cloud) and cybersecurity, privacy, performance, latency, energy, decentralization, personalization and context awareness (all of which are more favourable at the edge). This will need to be more dynamic and adaptive than traditional orchestration at centralized resources in the cloud, and should mean that associated

computations are able to move opportunistically across the continuum in search of the optimal temporary residence.

The envisioned orchestration would require embedding (artificial) intelligence, including generative AI, to do the bidding of individual users at the edge, promoted by user requirements and returning ad hoc programmatic orchestration engines. The underlying infrastructures would also need intelligence to federate opportunely and adaptively available resources within the right timeframe and cybersecurity context.