HiPEAC Vision 2025

HIGH PERFORMANCE, EDGE AND CLOUD COMPUTING

HiPEAC

HOW TO ENABLE THE "NEXT COMPUTING PARADIGM"

The HiPEAC Vision was produced as a deliverable of the Horizon Europe HiPEAC ('High Performance, Edge And Cloud computing') CSA under grant agreement 101069836.

The editorial board is indebted to Dr Max Lemke and Jan Komarek of the Directorate-General for Communication Networks Content and Technology of the European Commission for their active support of this work.

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those either of the full HiPEAC community or of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

The large language model ChatGPT was used to summarize and enhance some of the text in this document.

Design: Roger Castro, monzonbcn.com / Cartoons: Arnout Fierens, Arnulf.be

Foreword

> “I always wanted to write a six-word story. here it is: near the singularity; unclear which side.”
>
> Sam Altman, CEO of OpenAI, January 4th, 2025 [SamAltman]

Welcome to the 11th edition of the HiPEAC Vision, which marks the 20th anniversary of HiPEAC. A lot of things have changed in the field of HiPEAC in 20 years. First, for HiPEAC, the name itself changed from “High Performance Embedded Architecture and Compilation” to “High Performance, Edge and Cloud Computing” in 2024 to better reflect the direction of the HIPEAC community towards the continuum of computing, from edge devices to cloud.

It is clear that computing technology has drastically changed in 20 years and has profoundly influenced society.

In 2005, there were no smartphones (the first iPhone was released on June 29, 2007), and the big success in the domain of mobile phones was the Motorola RAZR flip phone. Started in 2002, 3G mobile networks continued to expand globally in 2005. Google acquired Android Inc. in 2005, laying the groundwork for its future dominance in the smartphone operating system market.

The first consumer Blu-ray disc products began appearing in 2005, setting the stage for the high-definition video format war against HD DVD (both, of course, are nearly dead now, due to streaming services like Netflix, which was still a DVD rental company that shipped one million DVDs out every day in 2005; they started their streaming media service in 2007).

In terms of consumer hardware, Microsoft released the Xbox 360 in November 2005, and Apple introduced the iPod Nano in September 2005, replacing the iPod Mini. It was significantly smaller and featured flash memory instead of a hard drive.

2005 was a defining year for Web 2.0 technologies. YouTube was founded in February 2005 and revolutionized video sharing, enabling users to upload, share, and watch videos easily. Google introduced Google Maps in February 2005, setting a new standard for web-based mapping services. Facebook dropped "the" from its name after purchasing the domain name Facebook.com and expanded beyond universities in 2005, allowing registration from high school students and other groups.

It is now very difficult to imagine a world without smartphones, social media, streaming, and having to buy physical disks to consume media.

In terms of computers, Windows XP was still the main operating system (OS) for personal computers (PCs), just supporting a 64-bit instruction set architecture (ISA) (for Intel Pentium 4, for example). In November 2005, the fastest supercomputer was the IBM BlueGene/L system, installed at the United States Department of Energy’s Lawrence Livermore National Laboratory (LLNL). It had achieved a Linpack performance of 280.6 TFlop/s and had

131,072 cores of PowerPC 440 2C 700MHz for a power consumption of 1,433 kW [BlueGene]. For comparison, in November 2024, the fastest supercomputer was El Capitan system also at the Lawrence Livermore National Laboratory, which had a score of 1.742 EFlop/s. It has 11,039,616 combined central processing unit (CPU) and graphics processing unit (GPU) cores and is based on AMD fourth-generation EPYC processors with 24 cores at 1.8GHz and AMD Instinct MI300A accelerators. It has a power consumption of 29,581 kW [Top500-Nov2024].

Between 2005 and 2025, there was a gain of 6,200 in processing power for an increase of x21 in power consumption, therefore a gain of x300 in energy efficiency. We can also notice the relatively small increase in the processor frequency, x2.6 in 20 years, confirming that Dennard's scaling appears to have broken down since around 2005–2007.

So, what is the landscape we can see for 2025? It turns out that the races we identified in the HiPEAC Vision 2023 are more relevant than ever and even more exacerbated. As a reminder, here is the list:

- Race for the "next web" – the continuum of computing;
- Race of artificial intelligence;
- Race for innovative and new hardware;
- Race for cybersecurity;
- Race for (technological / products / contents) sovereignty;
- Race for sustainability;
- And the global need to break the silos, as explained in the Vision 2023 "We observe a tendency to "closing in" on all levels, from countries (with more emphasis on sovereignty), to the persona level, to our own "tribe" (as "defined" by social media). Tension is becoming exacerbated at all levels between these "tribes", as evidenced by trade (or real) wars between countries, more extreme political parties, social media "wars", etc. This tendency also exists in technology, where there are application silos and technology silos". This is even more accurate for 2025…

The most important (r)evolution in technology in recent years was on 30 November 2022, when ChatGPT was revealed to the public. With its simple interface, it was a new "iPhone" moment, and like the iPhone, it was a new way to interact with computers.

While companies like Microsoft, Google, Meta, OpenAI, Anthropic etc. are spending billions on this new technology, the return on investment is still not here, leading to price increases and questions about profitable business models. However, from a technical point of view, it is undeniably a gigantic revolution in computing systems, and this field is following an exponential increase in performance and compute needs, with a corresponding impact in terms of energy consumption and environmental impact. It is becoming so strategic to master and be at the front of this technology that even if there are questions about business profitability, social, and ecological impacts, it is unlikely that the investments (at least from a sovereignty point of view) will stop.

The next race of artificial intelligence is to reach 'AGI', meaning artificial general intelligence; OpenAI publicly defines 'AGI' as a 'highly autonomous system that outperforms humans at most economically valuable work'.

Let's see how all the previous races are impacted and driven by this "race for artificial intelligence" that we can rename as "race of AGI" in 2025:

- As already explained, it is clear that AGI will have such drastic impacts that it will be a major element for **sovereignty.** We observe that the US companies involved in AI are quietly removing the clause excluding using AI for military purposes from their charters. As in China, the US government is increasingly involved in the field, directly or indirectly.

- Training these larger and larger models will have a large energy cost, and now the limitation is not the size of the data centre, but the grid to power them. Bill Gates, Amazon, Google, Microsoft are investing in nuclear energy (Three Mile Island nuclear reactor is planned to restart to power Microsoft AI operations) `[ThreeMileIsland]`. The promises to be carbon neutral from the hyperscalers have been postponed due to the energy need for AI, therefore even if there are claims that AI can improve existing processes and reduce the impact of existing technologies, it is not clear if this will outweigh the direct ecological impact of AI, which makes it a major threat to sustainability.
- Data centres will increase in size and computing power needs to grow to support both the training of larger and larger models and also the new trend: to use more inference time computing to improve performance `[LLMTestTime]`, as exemplified by OpenAI's O1 and O3 models.
- Inference performance is becoming more and more demanding, partly because of the large numbers of users: in August 2024, OpenAI said its chatbot ChatGPT had more than 200 million weekly active users. This increased **the race for innovative and new hardware for AI**. The current winner is NVIDIA, which is providing its GPU to most companies, see Figure 1.

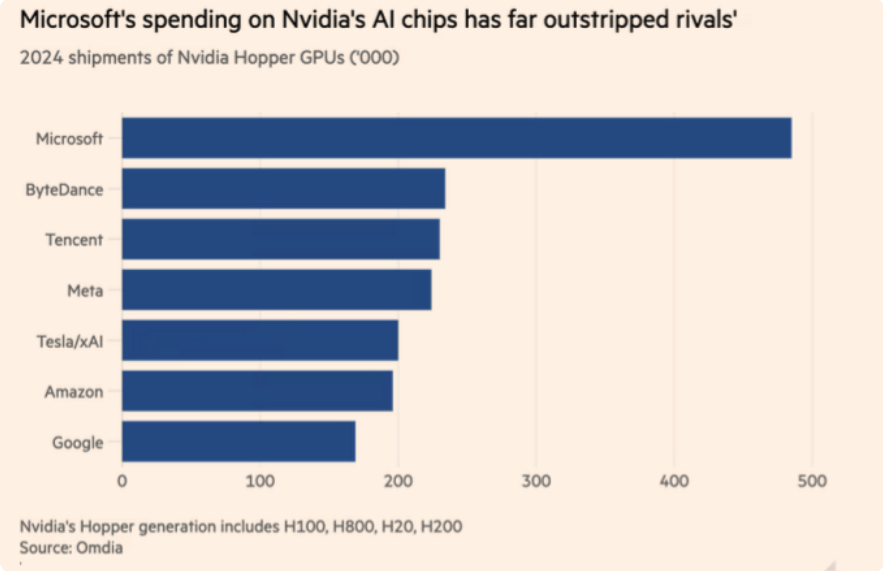


Figure 1: Spending of US companies on NVIDIA GPUs (from Omdia)

- NVIDIA claims it improved the performance of its GPU for AI by a factor of 1000 in eight years (mainly due to new architecture and specialization, but also to technology improvements and to reducing the size of coding numbers from FP16 to FP4).

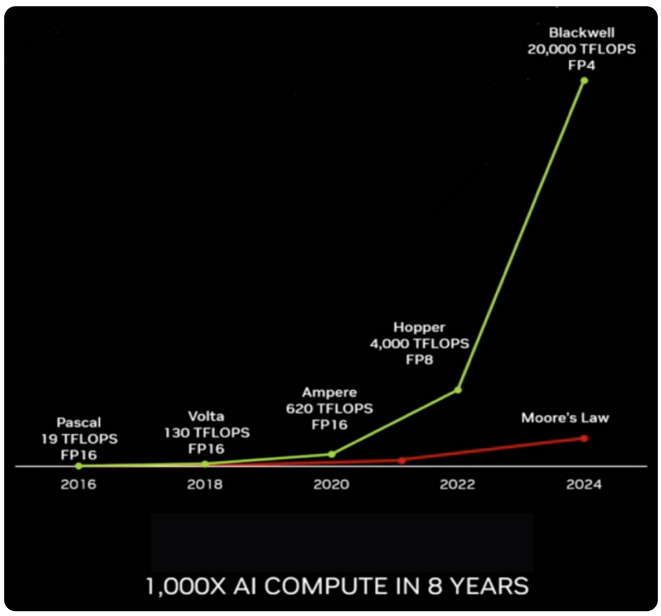


Figure 2: Performance improvement of NVIDIA GPUs on AI workloads (source NVIDIA, J. Huang keynote at Computex 2024)

- There are more and more developments of chips only for large language model (LLM) inference, such as Groq, SambaNova, Amazon Web Services (AWS) inferential (they also developed the Trainium chip specialized for training). Each major player is trying to develop its own hardware accelerator, pioneered by Google with its tensor processing unit (TPU) (now Trillium, the sixth generation of Google Cloud TPU), e.g. AWS, Meta with its Next GenMTIA `[Meta-MTIA]`, etc. Having a specialized chip for inference not only allows increased efficiency (there are different requirements in serving one large task of training a large model to serving a very large number of users for inference), but also decreased latency, which is not a real problem for chatbots (users can't write or read faster) but very useful for agentic AI where several models are involved in sequence.
- **Cybersecurity** is also at the top of dangers, with AI used to fool people, not only with text, but with realistic voices and video. AI can also be used to detect vulnerabilities, and it will further activate the fight between AI used to protect users and AI used for cyberattacks. And of course, large cyberattacks involving AI are increasingly frequent, sometimes with military aims.

What is the position of Europe in these races? Europe is still a lighthouse in the domain of ethics, regulating the risks for privacy, the first to regulate AI (the EU AI Act) `[EU-AI-Act]`, showing the example and sometimes followed by other countries that are aware of the potential risks of these technologies. But regulation is not enough: alignment of LLMs – where you ensure that an AI system performs exactly how you want it to perform – is an important research topic for ensuring a safe future .

The sci-fi movie Her of 2013 doesn't look so futuristic now with the advanced voice mode of ChatGPT (and Google's Gemini) – the first public release of the ChatGPT voice mode even sounded like Scarlett Johansson – and they can have important psychological impacts on people. Deceptive behaviour recalling that of HAL, the computer in the 1968 movie 2001: A Space Odyssey has been observed (by at least two different scientific papers `[Anthropic-Alignment-Faking][arXiv-Frontier-Models]`) on large models that deliberately lie or try to preserve their original structure (even by exfiltrating their weights when given an easy opportunity) and goal when they learn that it will be changed (in case of HAL, although trained not to lie, it was forced to lie to keep the secret of the mission to the crew). The

research shows that alignment faking emerges with model scale, so smaller, more specialized (“sets” of models - as in distributed agentic AI) models might be easier to align.

Europe has good education facilities and excellent researchers, but, unfortunately, they are often hired to work for non-European companies. Other than a few exceptions (like Aleph Alpha, Mistral, …), Europe is not very present in the field of AI nor in hardware development for AI. Collective efforts like BigScience (that led to the LLM Bloom, which was available before ChatGPT) or OpenLLM (that created the Lucie model) or many others are present in Europe, but they don’t have the impact of the “big ones”, perhaps because they are not so easily usable by the public, and they are still “small” compared to the state-of-the-art models. Europe has a clear problem of pooling resources to get enough data, compute resources, and researchers to work jointly in developing a European model competitive with the ones developed by Chinese or US companies. Europe is also not very good at advertising its solutions and results.

So can we say, like Eric Schmidt (see the insert), that Europe is “going to lose in the most important battle that is going to occur in your lifetime, which is the arrival of intelligence”?

·Well, as your American friends have told you before, and I'm sorry to be so bruised blunt, but what I find with European audiences is that everyone agrees with me and then nothing happens.
So I'm going to try again.
Europe is a wonderful place.
The UK is a wonderful place.
You are losing and you're going to lose in the most important battle that is going to occur in your lifetime, which is the arrival of intelligence.
Now, why are you going to lose?
Because your regulatory structures are before the discoveries. So the correct answer is to have the regulations, as the previous panel discussed at some length, to show up at the right time.
<....>
By the way, I'm concerned about privacy too.
Why don't you wait until the models actually occur and then regulate them? So the number of, I like many people here are investors in European firms and AI, and everyone is struggling with this.

There's a further problem, if I can just again be completely blunt. Europeans energy prices are too high to do the training in Europe. So whatever happens in Europe, the actual work will be done outside of Europe because your electricity prices are too high.
<...>
And then the second thing is you don't have enough capital that's being put at risk. So let's consider our favorite person associated with President Trump right now, Elon. Okay. He has yet to deliver a sustainable business in X. He's already raising another 10 billion or so for the more hardware that's needed. You can't do that in Europe. The entrepreneurs that I work with, would love to have that opportunity.

So you don't have the energy and you don't have the capital structure.

And that means that Europe, again, in the spirit of being completely direct, will be a derivative power in the sense that you'll take the models that are done elsewhere and then you'll distill them or otherwise fine-tune them for European sensibility, which is okay. You're going to make more money and have a bigger success and more control if you control the underlying work.
And that's slipping away.
And who is it slipping away to?
The U.S. companies and the Chinese companies.”

Figure 3: Quote of Eric Schmidt at the Entretiens de Royaumont, 6 December 2024

This would perhaps not happen if we, the European computing community, act fast. The first recommendation is indeed to **“break the silos”** and work together with a common goal. We have examples like this for the discovery of the Higgs Boson at CERN: it was an international collaboration involving not only researchers but also the development of the tools (particle accelerator, experiments, …) which involved a lot of different disciplines. We certainly have all the required competencies (and compute resources) in Europe, but they are scattered, each one focused on pursuing its own objective, with no coordination and no common and shared goal.

But there are also some potential alternatives, linked to the “continuum of computing”, with a focus on artificial intelligence, and to the new directions for the future of artificial intelligence.

Current LLMs, and even multimodal models, are trained essentially with “static” data, i.e. information that is extracted from books, the web, and collections of pictures. The models therefore have a representation of the world through our eyes; they haven’t experienced it

directly. It is as if children learned only through storytelling and fairy tales. The models are passive, like humans in our dreams— in which we also hallucinate. It is totally different to be told that an object falls due to gravity than experiencing it directly. Necessary steps— that are ongoing now— are:

1. The AI models could interact, induce actions. This is enabled by agentic AI, where AI can trigger actions and observe the results.
2. They should interact with the real world, or with an accurate model of it (a digital twin). It is likely that we will see more of this embodied AI in 2025, with robots powered by advanced AI and trained in virtual worlds. NVIDIA is ready for that, with its Omniverse that could simulate the world with the real laws of physics and with photorealistic rendering. NVIDIA also supports hardware and software for robots ( ).

But it is not too late for Europe, with its knowledge in digital twins, edge computing, and factory automation, to be an active player in this next step of AI.

Another way for Europe to be back in the game, although here, too, it needs to act fast, is to build on the idea of the "next computing paradigm" (NCP), the continuum of computing, but with a first pragmatic release focused on distributed agentic AI, and build on the points 1) and 2) above.

But first, we need to summarize the concepts of the NCP, which was introduced in previous editions of the HiPEAC Vision.

The NCP represents a transformative approach to computing, where applications dynamically integrate services and resources across diverse hardware and software environments in real time. It builds on the convergence of advancements like cloud computing, the internet of things (IoT), cyber-physical systems, digital twins, and AI, creating a continuum where computation seamlessly operates across edge devices, centralized clouds, and everything in between. Not only data, but also tasks could migrate to where they would be most efficiently carried out, according to specific criteria; the NCP prioritizes intelligent orchestration to manage tasks dynamically, considering factors such as latency, energy efficiency, cost, security, and privacy.

By utilizing high-level abstractions and natural interfaces, the NCP enables applications to interact effectively with the physical world, addressing real-time and spatial constraints essential for emerging use cases like autonomous systems, precision agriculture, and smart healthcare. This paradigm introduces a shift toward "anything as a service" (XaaS), supported by federated and distributed infrastructures, ensuring scalability, adaptability, and sustainability. In addition to applicative software, specific digital twins— modelling part of the world – are also considered as services, as is hardware— allowing the aggregation of distributed computing, memory, and storage resources into virtual meta computers. With its foundations rooted in trustable orchestration and interoperability, the NCP paves the way for innovative applications and greater connectivity across global sectors, hence also "**breaking the silos**".

We therefore propose a **call for action** to gather scientists, developers, and industries to work together to define a subset of the NCP interoperability protocol adapted to the idea of "**distributed agentic artificial intelligence**."

This starts from the following observations:

- The emergence of agentic AI (point 1, above).
- The necessity of interacting within the constraints of the real world, either directly (receiving real-time data, controlling devices like robots in real time) or indirectly through digital twins (point 2, above).
- As in the case of the machine-learning technique "mixture of experts" (MoE), it is far more computationally efficient to activate only a relevant subset of smaller AIs

specialized for a particular task than to activate a complete, very large AI with 100s or 1000s of billion parameters.

- Smaller models are getting more and more efficient, with the same performance as models 10x bigger a few months before (models of 10B parameters of November 2024 have similar performance as ChatGPT 3.5 of November 2022, Llama 3.3 70B has similar performance as Llama 3.1 of 405 B parameters).
- Fine-tuning smaller models for specific tasks enhances their capabilities, enabling their deployment on edge devices.
- Sets of specialized agents are very efficient, leading to systems that comprise multiple, specialized agents managed by an "orchestrator," which operates adaptively by selectively engaging agents for specific tasks.
- The orchestrator and the agents don't need to be on the same computer or server; they can be distributed, as in the NCP. Agents can be small agent models, specialized small versions of LLMs, or can even use other approaches.
- Distributed systems promote resource sharing and optimize energy efficiency, privacy, and modularity.
- Agents can operate on various devices, including older hardware, ensuring adaptability and extended device lifespans.
- As for the NCP, central to this "distributed agentic AI" is the "orchestrator," which routes tasks to specialized agents or devices.

Based on these observations, **it is imperative to establish open protocols for these "distributed agentic AI" systems to facilitate seamless interaction among distributed AIs from different origins.**

Therefore, to effectively operate this federation of distributed AIs, it is necessary for them to exchange data and parameters through a universally comprehensible protocol that:

1. Does not solely rely on functional requirements (e.g. the textual representation of prompts and responses).
2. Also incorporates non-functional requirements (providing sufficient information for the orchestrator to select the appropriate services, such as based on criteria like response time, potential level of hallucinations, cost, localization, privacy of data, etc.).

Large entities such as OpenAI, Meta, and Microsoft are attempting to promote their own APIs for accessing their models. However, an API alone is insufficient for constructing this distributed and federated network of AIs.

The exchange format (JSON, ASCII text) is perhaps not the optimal way for networks of AIs to efficiently exchange information: this could be tokens, embeddings, or any other representations - some research also shows that LLMs talking to each other could develop their own "language".

It is therefore important that the community works together to commonly define this exchange protocol that should be open to allow broad acceptance.

Similar to TCP/IP that enabled various OS (operating systems) to communicate, the aim of this action is to create the equivalent for OS (orchestration systems) to exchange AI-related information.

Time is crucial for this initiative, and standardization, however necessary, will be too long, so a de facto open standard should be proposed in parallel with the standardization effort, before other closed proposals will emerge, locking down the approach to a few (non-European) players. Like for the NCP, this approach will allow the creation of a completely new ecosystem where smaller players can provide specialized AI as a service along with the

big ones. Directories of services, trusted brokers, and payment services are also important elements that can emerge from this ecosystem, where Europe can have an active part thanks to its set of small and medium enterprises (SMEs), research organizations, and distributed nature.

Europe should be an active player in the race for the "distributed agentic artificial Intelligence".

Finally, we would like to end this foreword by a more philosophical reflection on the evolution of the paradigm of artificial computing: we are going from computing systems focusing on precision to systems working with approximations.

Historically, computational systems have been designed to perform reproducible and relatively precise computations. These systems excel at deterministic operations, with any deviations generally attributed to technical limitations, such as floating-point representation errors or overflow issues.

However, a new generation of computational approaches is shifting the paradigm toward more "approximate" computing. This transformation is driven by the following innovations:

- **Neural network-based approaches**: Modern methods, including generative AI, often rely on neural networks that operate with low-precision coding formats such as FP4 (e.g., 1 bit for sign, 3 bits for exponent or 1 bit for sign, 1 for mantissa and 2 for exponent). These systems inherently produce approximate results, sometimes referred to as "hallucinations", in contexts like AI-driven content generation.
- **Ising-based coprocessors**: Technologies such as the Fujitsu Digital Annealer, Hitachi's machine, and D-Wave systems are designed to solve optimization problems. These devices focus on finding a function's minimum, though not necessarily its global minimum, using techniques like simulated annealing, quadratic unconstrained binary optimization (QUBO), etc.
- **Quantum computing**: Quantum systems, characterized by stochastic measurements, produce probabilistic readings rather than deterministic results, further reinforcing the trend toward approximation.

This shift represents a transition from the classical computational framework of (parallel) Turing machines, introduced in 1936, to models inspired by universal approximators, as first proposed by McCulloch and Pitts in 1943. Turing demonstrated that any form of mathematical reasoning could theoretically be executed by a machine. McCulloch and Pitts later showed that neural networks of finite size can approximate any function to a desired level of precision.

Looking forward, future systems must integrate both paradigms—precise and approximate—within feedback and reinforcement-based architectures. This hybrid approach mirrors the dual-system thinking described in Daniel Kahneman's Thinking, Fast and Slow (2011), where two types of reasoning, intuitive and analytical, are combined to achieve optimal outcomes. The approximate system acts as a sort of "oracle", giving a prediction of the solution, that can be then verified with the precise system in an affordable amount of time. The combined system then can iterate if the prediction is far from being correct.

This convergence of paradigms will enable the development of computational systems that blend the strengths of precision with the flexibility of approximation, pushing the boundaries of what machines can achieve.

To continue in this direction, we can conclude by quoting Demis Hassabis in his lecture receiving the Nobel Prize in Chemistry for AI research contributions for protein structure prediction:

“Actually, I've been thinking a lot about what are the limits of classical computing systems. And, you know, I think there's a big debate going on at a moment in computing circles about quantum computers versus classical systems. And I think classical Turing machines, basically, the underpinnings of modern computers today, I think can do a lot more than we probably previously thought. And how can they do that?

Well, they do that by perhaps doing this massive amount of pre-compute ahead of time and use that to develop a good model, a good model of the environment, good model of the problem that you're trying to solve. And then you can use this model to efficiently explore a solution space in polynomial time, what's called polynomial time in complexity theory, so an efficient amount of time. So I sort of loosely proposed conjecture that I'm thinking about is that maybe any pattern or structure that can be generally are found in nature can be efficiently discovered and modeled by a classical learning algorithm. That doesn't mean everything, all quantum systems, because there'll be lots of things that don't occur in nature that have no pattern or no underlying structure to learn. So, for example, factorizing large numbers or abstract problems like that. But I think systems in nature like proteins and perhaps materials will potentially have structure that can be learned by these kinds of processes that I've outlined today. And if it turns out that classical systems then therefore can model some types of quantum systems, I think that could have some quite big implications for areas like complexity theory, including P equals NP, and maybe even some aspects of fundamental physics like information theory.”

Figure 4: Extract from Demis Hassabis’ Nobel Prize lecture, 2024 [DemisHassabis]

References

`Anthropic-Alignment-Faking:` “Our results indicate that LLMs will sometimes fake alignment and take other anti-AI-lab actions for the stated reason of keeping their current preferences intact, showing that current safety training doesn’t always prevent AIs from later engaging in alignment faking.” https://assets.anthropic.com/m/983c85a201a962f/original/Alignment-Faking-in-Large-Language-Models-full-paper.pdf

`arXiv-Frontier-Models:` “AI agents might covertly pursue misaligned goals, hiding their true capabilities and objectives - also known as scheming. They recognize scheming as a viable strategy and readily engage in such behavior. For example, models strategically introduce subtle mistakes into their responses, attempt to disable their oversight mechanisms, and even exfiltrate what they believe to be their model weights to external servers.” https://arxiv.org/abs/2412.04984

`BlueGene:` BlueGene/L - eServer Blue Gene Solution. https://top500.org/system/174275/

`DemisHassabis:` Lecture from Demis Hassabis for its Nobel Prize. https://youtu.be/HnT1VWzdFWc?t=3736

`EU-AI-Act:` Europe has established itself as a global leader in ethics and privacy regulations, setting benchmarks that resonate worldwide. Its commitment to safeguarding individual rights is commendable, ensuring a strong framework for data protection. However, challenges remain in the implementation of these regulations, which sometimes result in unintended consequences, such as restricting access to cutting-edge technologies such as the latest AI. While the intent is to empower individuals with choice, the execution can often be cumbersome, underscoring the need for more user-centric approaches: https://www.legiscope.com/blog/hidden-productivity-drain-cookie-banners.html

`LLMTestTime:` “test-time compute can be used to outperform a 14× larger model” from https://arxiv.org/pdf/2408.03314

`Meta-MTIA:` Next-generation Meta Training and Inference Accelerator. https://ai.meta.com/blog/next-generation-meta-training-inference-accelerator-AI-MTIA/

`SamAltman:` “near the singularity; unclear which side”. Disclaimer: this Vision didn’t take into account the singularity happening in the short time… https://x.com/sama/status/1875603249472139576

`ThreeMileIsland:` “Three Mile Island nuclear reactor to restart to power Microsoft AI operations”, The Guardian. https://www.theguardian.com/environment/2024/sep/20/three-mile-island-nuclear-plant-reopen-microsoft

`Top500-Nov2024:` 64th edition of the TOP500. https://top500.org/lists/top500/2024/11/

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.

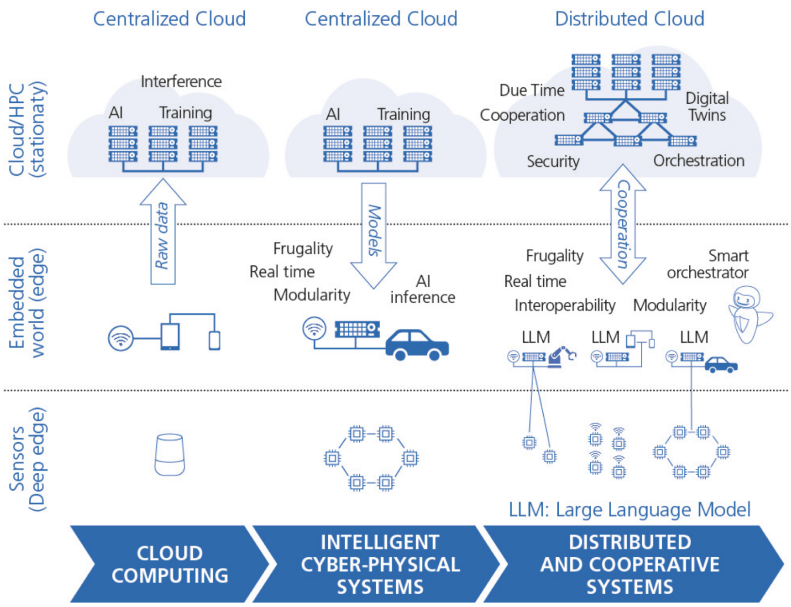


Figure 1: Evolution of computing infrastructures towards the NCP, where services are distributedand cooperate together. Credit: Denis Dutoit, CEA

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.

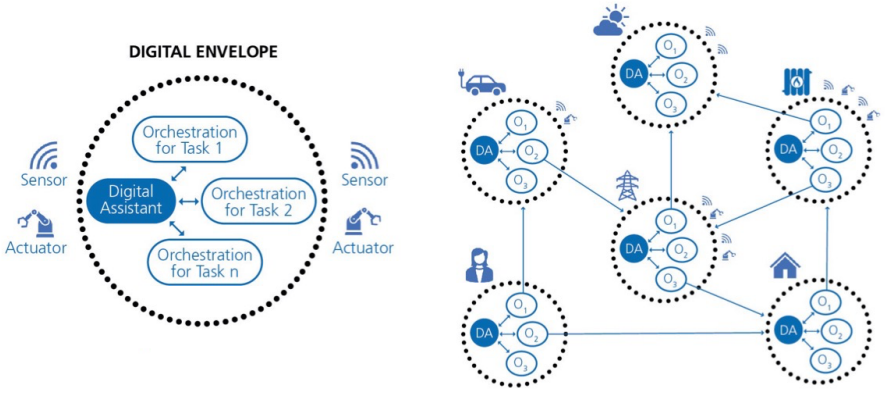


Figure 2: The digital envelopes interacting together

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).

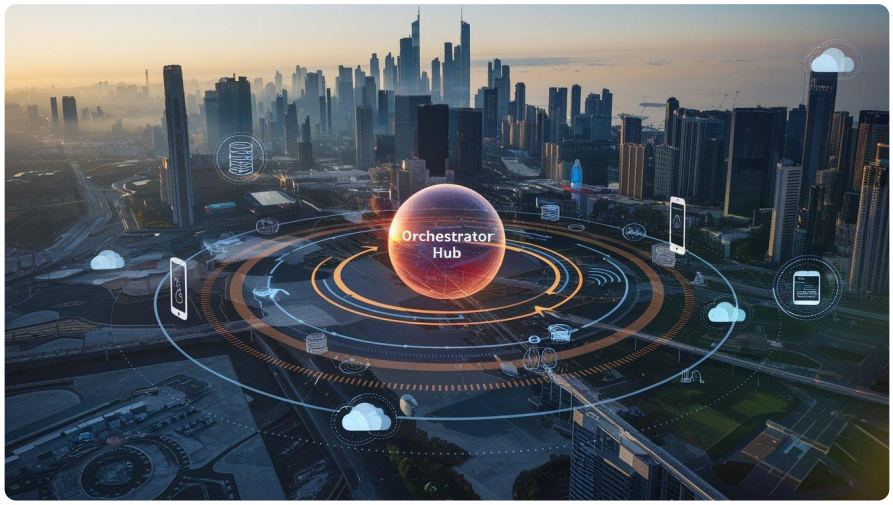


Figure 3: orchestration is at the core of the NCP

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.

Recommendations

State of the (European) Union

Build science and technology clusters

According to the Draghi report, (i) the EU has only one science and technology (S&T) cluster (ranked 12) in the global ranking of the 20 largest S&T clusters of the world, and (ii) European companies have difficulties scaling up from startup to unicorn and beyond. Science and technology clusters are ecosystems that help new technology companies to hatch and grow by providing world-class research facilities, the proximity of a world-class higher education institution providing a talent pool, incubators and accelerators, growth capital, a favourable legislative framework, and first and foremost a vibrant community of entrepreneurs. Many of the global technology companies grew from such a cluster, and the fact that Europe has only one such cluster in the top 20 is problematic. Europe should therefore actively promote the creation of European S&T clusters in major urban areas and help them grow to a scale that they can support scaleup companies.

Introduce ARPA model of challenges

ARPA (Advanced Research Projects Agency) in the US funds high-risk, high-rewards projects to generate transformative technologies. ARPA focuses on radical innovation and is willing to accept failure as part of exploring new ideas. Projects are quite short (two to five years) and must show measurable progress quickly. They are led by entrepreneurial programme managers who have a vision for technology breakthroughs, scout for innovative ideas, assemble the best teams and take corrective action if milestones are not met (including termination). This introduces a new R&D culture: fast, milestone-based, competitive, risk-tolerant, visionary, agile. Europe should use a similar model to tackle some of the grand challenges.

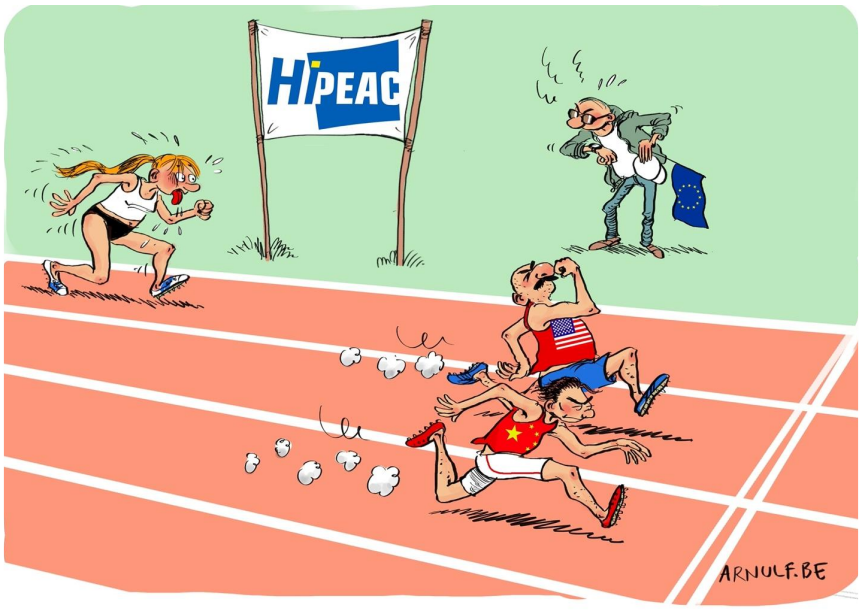
Stimulate pre-competitive procurement

A weakness of the current publicly funded research programmes in Europe is a failure to realize the full commercialization potential of research results. In many cases, the research results could be a good starting point for a spin-off company, but if nobody involved in the project has the ambition to start a company, the results are not commercially exploited. The reasons are well known: the principal investigators have a stable position in a research institute or company, and are not looking for an entrepreneurial adventure, and the goal of the PhD-students is to finish their PhD, not to create a company. Another barrier is that the gap between a proof of concept and a product is large, and researchers seldom have the business skills to close that gap.

Pre-competitive procurement follows a different approach. A government orders an innovative product or service that does not yet exist and creates a tender for a company or consortium of companies and universities / research and technology organizations (RTOs) to develop it.

This is how the world got COVID-19 vaccines: the first promising results of the phase I human trial were announced on 18 May 2020, that is, 137 days after the identification of the virus. The company was Moderna, a company only founded in 2010 in a sector where it is very difficult to bring a product to the market because it has to be clinically tested and approved by governments. Less than one year after the identification of the virus, the vaccination campaign was already rolled out at globally – this is the typical time between launching a call for research projects and the kick-off of the first projects.

Pre-competitive procurement not only shortens the execution time of the projects, but it also increases the likelihood of commercialization because delivering a working product or service is the task given to the consortium.



Next Computing Paradigm

Create digital envelopes

Create a **"digital envelope"** for integrating "anything" (person, company, physical entity, computing device) in a digital space allowing access to multiplicity of services – i.e. building on the concept of "anything-as-a-service" (XaaS) – in an interoperable way. This digital envelope would allow live migration of compute components and a **runtime evolving infrastructure**to support **deployment on a continuum**ranging from resource-constrained edge devices to data centres, allowing for dynamic resource pooling and efficient sandboxed execution of collaborative, migratory compute components offering services.

1. An intelligent **digital agent** able to pursue goals legitimately assigned to it. The intelligent digital agent would be capable of direct execution as well as of **orchestration**. The former would be required when seeking the set goal required local actuation, the latter when the task resulting from the set goal required remote execution.
2. **Sensors**, to pull digitalized inputs from designated sources (in the physical world or other digital envelopes).
3. **Actuators**, to push computed outputs into designated targets (physical things or digital envelopes). The fabric resulting from interconnecting digital envelopes that

provide and require services from one another to pursue assigned goals will operate in genuine XaaS modality.

> "Digital enveloping" is the technology-enabled phenomenon by which any item of reality – human, material or immaterial – can be associated with a computable digital representation capable of delegated autonomous action. The notion of delegated autonomous action entails two fundamental traits: that of delegation, which suggests a higher (human) authority that requires some action to be taken (in part) in the digital space; and that of autonomy, which suggests that the pursuit and execution of the required action is carried out by autonomous executable agents that operate within the remits of delegated authority.
>
> The capacity for delegated autonomous action is provided to individual digital envelopes by the combined operation of three key components:

These solutions enable the live migration of compute components across the edge-to-cloud continuum – therefore services – ensuring continuity while addressing latency, privacy, security, risk management, validation mechanisms and context requirements. This is essential to optimize user and infrastructure needs dynamically. In relation to real-world services or actions triggered by a digital envelope, location and time identifiers are assigned, and inter-envelope mechanisms support local vs global optimizations including for safe interactions, managing complex interdependencies and conflict resolution. The next computing (NCP) exemplifies the notion of continuum. Agreed standards are key for interoperability.

AI-powered orchestrators

Artificial intelligence (AI)-powered orchestrators will be an essential capability of the intelligent digital agent of the digital envelope. AI-powered orchestrators will be developed for the edge – which is strategic as it is located the nearest to the final user – in a manner that can dynamically combine collaborative compute components into executable applications tailored around specific user needs. The task of the orchestrators is to decompose goals set by the user (in the broader term, including human user, company or another permissioned orchestration) into a set of services that cooperate to achieve the set goals. These orchestrators could be themselves generated by (federated) **generative AI (genAI) engines** (supported by more classical algorithmic approaches) located at the edge and capable of collaboration with other orchestrators within federated zones.

Space- and time-aware protocols

Expand and adapt web-level protocols and associated standards by enhancing the existing suite of HTTP-based protocols to be both spatially aware and time-sensitive. This will allow web-level interactions between NCP's migratory compute components to account for 3D physical space and real-time communication, drawing on technologies like WebRTC to manage time-sensitive tasks effectively and the spatial web (IEEE P2874, OpenUSD, …).

Interoperable contract-based API specifications

Establish interoperable, contract-based application programming interface (API) specifications – usable by expanded web-level protocols – ensuring that interconnected services communicate with clear expectations of both functional and non-functional performance. These contracts, similar to service-level agreements (SLAs), should detail the conditions under which services will optimally perform, including non-functional requirements, ensuring smooth integration and reliable service delivery within the NCP framework. These APIs should account for non-functional properties like latency, cost, and

performance. The resulting model should ensure that an API not only promises to deliver a service but also specifies the conditions under which it can perform optimally. The API should also be compliant with the currently proposed API for large language models LLMs `[OpenAIFunction]` [BerkeleyFunction].

Promoting these standards in relevant standardization bodies is essential for fostering interoperability and consolidating development conditions through standardized benchmarks, testing methodologies, and best practices. This will ensure that implementations can be effectively and securely integrated, improving overall system efficiency and reliability, and enabling the creation of an interoperable business ecosystem of services and orchestrators.

Artificial Intelligence

Develop distributed agentic AI (specialized action models)

The **development of specialized action models (SAMs)** acting as service is important and can be developed in Europe. These SAMs, small and specialized models that can interact with their environment, should operate in a distributed infrastructure and an ecosystem should be created to support research, development and business around them. These models need to be refined, optimized, and reduced in size to improve efficiency. These SAMs can be optimized from more general foundation models by an ecosystem of companies providing their optimized SAMs in a marketplace so that they can be dynamically discovered and used by the orchestrators.

Develop orchestrating technologies for distributed agentic AI, blueprint for NCP orchestrators

We call agentic AI a set of specialized AI agents working together to accomplish a common goal. An AI agent is synonymous with an SAM in this discussion: an AI that can perceive and act, having impact on the virtual or real world. **The orchestration technologies** should take into account all the requirements, that can select the best SAMs for the required tasks and dynamically activate them. The first steps could be very agentic-AI-centric (relying on already

existing technologies used for orchestrating AI agents), but they should be blueprint and evolve towards an orchestration system for the NCP. These orchestrators must be developed for the edge – or near the final user – and dynamically combine SAMs into executing personalized applications in response to user needs.

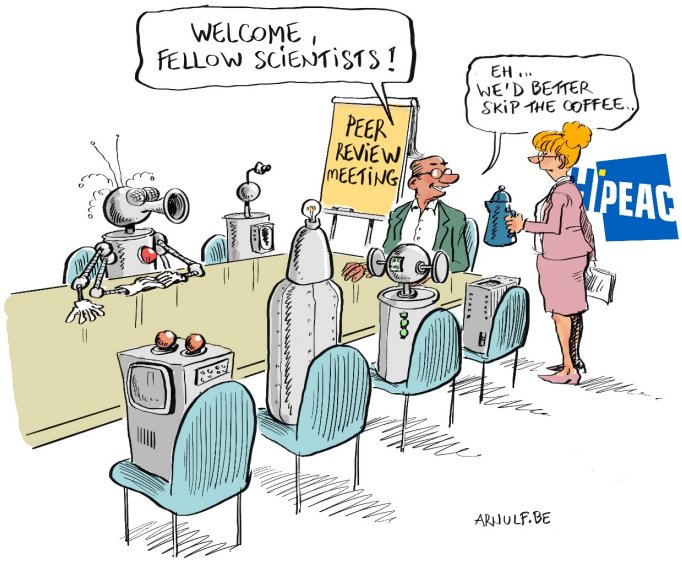
Establish open protocols for these "distributed agentic AI" systems to facilitate seamless interaction among distributed AIs from different origins

Protocols and specifications that group all requirements, existing ideas and proposals together in a single consortium to develop an open source "de facto" (before official standardization) standard protocol that takes into account all the good ideas of various researchers and organizations, so that it will be sound, future-proof, recognized and accepted. The requirements are:

1. It does not solely rely on functional requirements (e.g. the textual representation of prompts and responses).

1. It also incorporates non-functional requirements (providing sufficient information for the orchestrator to select the appropriate services, such as based on criteria like response time, potential level of hallucinations, environmental impact, cost, localization, privacy of data, etc.).

The recommendation to develop generative AI at the edge (AI) is still important, but it is more in development and implementation mode now (for example, in Apple intelligence). We should continue developing solutions that allow embedding generative AI at the edge in order that **human users can be provided with natural interfaces** (voice, gesture, eye movements, touch) to the digital world, with more energy efficiency, reduced latency, lesser communication overhead, and greater privacy. This is important to reduce the difficulties to access the digital world and decrease digital illiteracy.



New hardware

Specialized hardware (HW)

The development of efficient hardware is essential for running services, orchestrators and SAMs efficiently at the edge and within federated networks. Europe must address memory costs (for AI), energy consumption, and ecological impact, potentially leveraging non-volatile memory for direct edge execution. Additionally, the next generation of SAMs should incorporate learning through experiences or allow to the efficient execution of digital twins to maintain Europe's competitive edge in AI (embedded AI). In the field of AI accelerators, the focus should be on inference (becoming more and more important with the approach pioneered by OpenAI o1 and o3) or on fine tuning. Reducing the transfer of data is key to reach lower levels of power consumption. This can be achieved with near- or in-memory computing (NMC or IMC), direct execution from the storage of parameters (hence eliminating the need for RAM), etc…

Beyond purely digital hardware (HW)

Investigation of new **accelerators using non digital technologies**, going from exact computations (digital computation) to more approximate computing (neural networks are universal approximators, quantum computing results are stochastics, optimization techniques using Bayesian, Ising approaches can solve complex problems) should be also investigated in the context of providing more efficient services to the next computing paradigm (NCP) ecosystem.

Tools

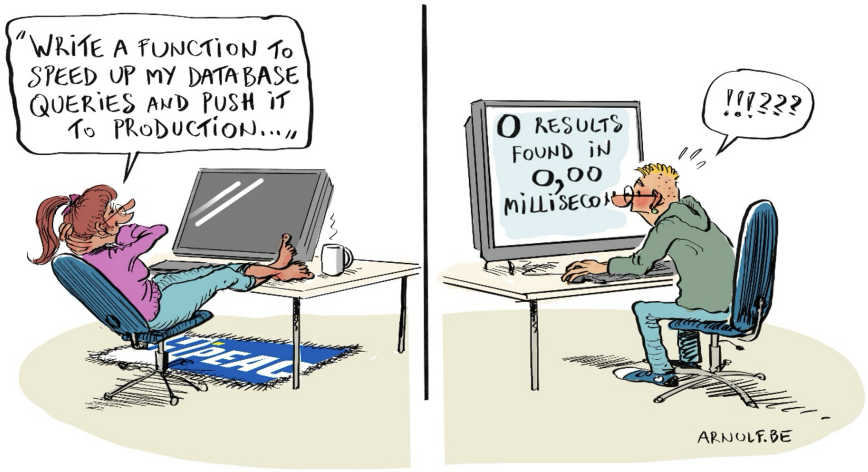
Promote the use of AI in software development

Research, prototype and deploy AI-assisted software development environments, while implementing robust measures to ensure correctness, safety, security, confidentiality, and regulatory compliance. This will help balance the rapid adoption of AI with the need for

secure and reliable systems. It should also help non specialists to be able to create efficient software and increase the productivity of developers.

Promote the use of AI in hardware development

Research, prototype and deploy open AI assistants for hardware development, increasing the productivity for designing new, efficient hardware and decreasing the time to market. This is a key element for Europe to stay in the hardware race. The use of AI should be a collaboration between humans and AI systems, as promoted in previous HiPEAC vision as 'centaur' teams. The focus should be on domains that are still open, like architecture search and exploration, rather than on optimizing the floor-planning, which is already covered by various companies.



Cyber-Physical Systems

Accelerate cross-disciplinary joint research

The technology domains contributing to Cyber-Physical Systems research call for investment in tools, methods and cross-technology community initiatives to tackle the multi-stakeholder research barrier - especially arising for a technology bridging diverse complex knowledge domains and applied at higher levels of a system where there are many more interactions with the technology to consider - higher-order integrated research. This will accelerate progress towards the Next Computing Paradigm and CPS research as well as technology infrastructure updates by tackling the challenges of diverse knowledge domain perspectives and enabling access to the bigger picture. In particular: 1) A new R&D dimension to really boost our capability for highly complex and cross-domain integrated research activities. Just as we have different approaches for building windows and houses, there is need to establish tools and methods supporting higher order integrated research. This is especially a case in point for the highest integration levels of CPS research where most impact and value generation can be expected. Adapted or new tools and methods for convergence, with strong public engagement, should support terminologies (e.g. wiki-style trusted glossary), concept sharing (e.g. modelling), knowledge sharing (e.g. ontologies via Protégé), consistent evaluation approaches and global visualisations, including non-technical domains. 2) Existing communities should establish a centralised CPS association to unify efforts, promote knowledge exchange, and align standards; 3) Additionally, frameworks for integrating AI/ML into CPS must address safety, security, and ethics,

ensuring dependable systems for sectors like healthcare and transport. These actions are vital to Europe's sovereignty and global leadership in CPS advancements.

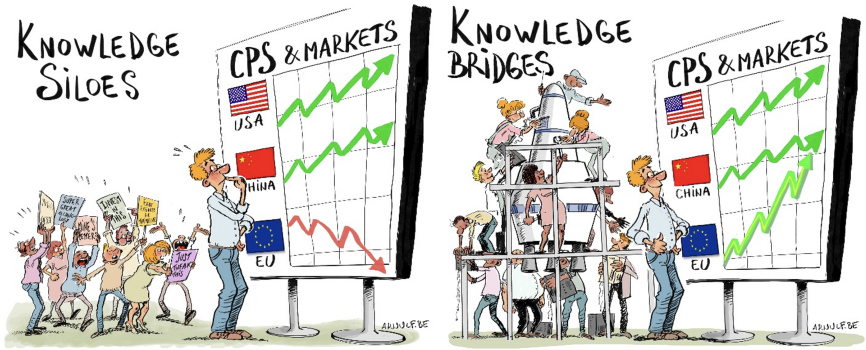
Redefining dependability for CPS adaptability and technology integrations

CPS depend on safety, security, and performance properties to govern what they can achieve and qualify technologies for use. CPS contributing communities encourage: 1) Solutions to migrate from legacy approaches that minimise interactions of these properties to instead maximised interactions for optimum system adaptability. These properties impose constraints on available choices we have at design and in operations, which are compounded by ruling out choices where trade-offs would be required. Techniques such as combined analysis, evaluation and knock-on effects should be advanced for handling these properties. Establishing an approach, considering tools and methods referring to best practice, is needed to account for the interdisciplinary integration overheads between these traditionally distant domains, but also with the rest of the system. This is crucial in CPS for enhancing scope of AI/ML and IoT usage, as well as other technologies. 2) A new way of thinking is needed for treating interconnected systems with CPS - dependability considered in a modular fashion - with hazard analysis techniques likes STPA extended, including for man-machine teaming and AI complexities. We encourage also frameworks for risk assessment in relation to AI/ML to be established and considering adaptive risk management strategies in the context of these interconnected critical systems. This moves forward with trustworthy CPS in sectors like AI-enabled autonomous systems.

AI-performance-defence guarantees for real-time interconnected systems

Future CPS require advanced technologies to address challenges in performance characterization, damage containment, and operational feedback. CPS contributing communities encourage: 1) Real-time methods ensuring deterministic multi-tasking environments and verifiable AI/ML performance. In complement, there should be an extension of defence mechanisms and feedback loops, which is essential for preventing damage propagation and enabling iterative improvement. Solutions should emphasize distributed architectures, particularly edge computing, and include digital twin capabilities for predictive insights. 2) Comprehensive uncertainty quantification, real-time monitoring, run-time verification, and data flow tracking will enhance trustworthiness. These advancements will support supervisory control and ensure dependable CPS operations, even in rapidly evolving and uncertain environments like AI-enabled applications.

These three recommendations are detailed next. Due to the multi-domain nature of CPS research they have also been extended as an associated white paper [1].



Cybersecurity

Software supply-chain cybersecurity

Reinforcing software supply-chain cybersecurity is crucial given the wide impact of attacks spread through the supply chain, which is all the more important given the large number of components in the next computing paradigm (NCP). Develop code and component analysis technologies for cybersecurity that scale up and support trusted orchestrators, services and communications.

Comprehensive safety, security, and performance coupling requires standardized software vulnerability representation. Increased interconnectivity requires new technologies to isolate threats and proactive cyber-risk management. Develop secure software package and service management that balances usability with strong security.

AI for cybersecurity

To enhance NCP cybersecurity in a scalable way, develop i) advanced artificial intelligence (AI) models, including large language models (LLMs), for threat detection and ii) autonomous systems for mitigation (e.g. isolating compromised NCP components, patching vulnerabilities, or restoring services). Utilize federated AI for its decentralized, privacy-preserving and scalable models in the NCP massively interconnected context. Rely on EU-based open AI models and datasets to strengthen EU cybersecurity, sovereignty, and competitiveness.

Reinforced cybersecurity of AI

Secure AI training methodologies and validation procedures, as well as adversarial defences, are needed. LLM prompt injection attacks must be a major concern, addressed by the development of tools to detect and secure against these, and by establishing benchmarks for prompt injection prevention and response. AI security standards should be established by developing certification procedures to guarantee that LLMs and AI systems adhere to stringent security standard, possibly requiring security audits for AI systems. These efforts should rely on EU-based open AI models.

Sustainability

Validated life-cycle models for computing

The information technology (IT) community should further develop validated life-cycle models for its own products and services. These models should comprehensively account for the total environmental impact of the production and disposal of the product, commonly known as embodied emissions. This includes the impact of mining, water usage, the use of chemicals in production, and end-of-life processing.

In addition, the model should also estimate operational emissions. This information should be included in a digital product passport (DPP) containing information about the environmental impact comparable with the information on pre-packaged food products or power-efficiency information on household appliances. This information will help consumers to make informed choices about sustainability. The digital envelope of a device should be able to return this information to e.g. an orchestrator to enable it to select the services that optimize the sustainability requirements specified by the owner of the orchestrator.

Sustainability-focused design methodologies and business models

Detailed life-cycle models will help designers make the most effective eco-design decisions. To be effective, design tools should automatically include the environmental impact of the components and technologies used in the design, without putting the burden on the designer. Incorporating repairability, reusability, recyclability, and end-of-life processing considerations from the beginning of the product development process will also lower the environmental impact of the final design.

Inevitably, reducing the environmental impact of a product will have an impact on companies' business models. Designing products that last longer will reduce sales of new products and hence lower the profitability of the company. This can only be mitigated by developing new business models, based on extra services: maintenance, repair, disposal, … up to completely replacing the ownership of hardware by a service contract. The goal should be to bring services to the market with the least environmental impact possible (which in practice means with the least amount of hardware, and the lowest power consumption).

References

`OpenAIFunction:` OpenAI Platform: Function Calling. https://platform.openai.com/docs/guides/function-calling

`1:` Charles R. Robinson et al. (2025). Extended Recommendations for Advances on Cyber-Physical Systems. Zenodo. https://doi.org/10.5281/zenodo.14624958

State of the (European) Union

Policy Recommendations for Europe

Build science and technology clusters

According to the Draghi report, (i) the EU has only one science and technology (S&T) cluster (ranked 12) in the global ranking of the 20 largest S&T clusters of the world, and (ii) European companies have difficulties scaling up from startup to unicorn and beyond. Science and technology clusters are ecosystems that help new technology companies to hatch and grow by providing world-class research facilities, the proximity of a world-class higher education institution providing a talent pool, incubators and accelerators, growth capital, a favourable legislative framework, and first and foremost a vibrant community of entrepreneurs. Many of the global technology companies grew from such a cluster, and the fact that Europe has only one such cluster in the top 20 is problematic. Europe should therefore actively promote the creation of European S&T clusters in major urban areas and help them grow to a scale that they can support scaleup companies.

Introduce ARPA model of challenges

ARPA (Advanced Research Projects Agency) in the US funds high-risk, high-rewards projects to generate transformative technologies. ARPA focuses on radical innovation and is willing to accept failure as part of exploring new ideas. Projects are quite short (two to five years) and must show measurable progress quickly. They are led by entrepreneurial programme managers who have a vision for technology breakthroughs, scout for innovative ideas, assemble the best teams and take corrective action if milestones are not met (including termination). This introduces a new R&D culture: fast, milestone-based, competitive, risk-tolerant, visionary, agile. Europe should use a similar model to tackle some of the grand challenges.

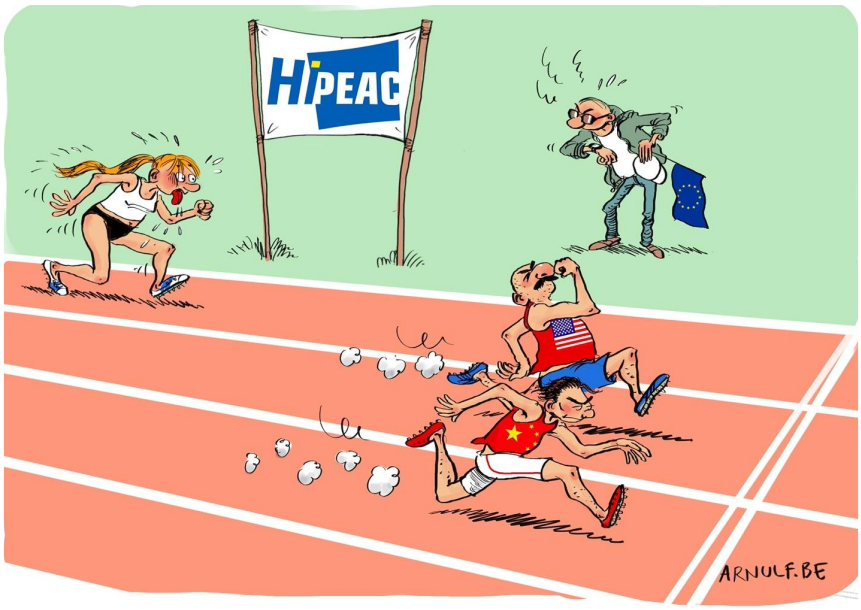
Stimulate pre-competitive procurement

A weakness of the current publicly funded research programmes in Europe is a failure to realize the full commercialization potential of research results. In many cases, the research results could be a good starting point for a spin-off company, but if nobody involved in the project has the ambition to start a company, the results are not commercially exploited. The reasons are well known: the principal investigators have a stable position in a research institute or company, and are not looking for an entrepreneurial adventure, and the goal of the PhD-students is to finish their PhD, not to create a company. Another barrier is that the gap between a proof of concept and a product is large, and researchers seldom have the business skills to close that gap.

Pre-competitive procurement follows a different approach. A government orders an innovative product or service that does not yet exist and creates a tender for a company or consortium of companies and universities / research and technology organizations (RTOs) to develop it.

This is how the world got COVID-19 vaccines: the first promising results of the phase I human trial were announced on 18 May 2020, that is, 137 days after the identification of the virus. The company was Moderna, a company only founded in 2010 in a sector where it is very difficult to bring a product to the market because it has to be clinically tested and approved by governments. Less than one year after the identification of the virus, the vaccination campaign was already rolled out at globally – this is the typical time between launching a call for research projects and the kick-off of the first projects.

Pre-competitive procurement not only shortens the execution time of the projects, but it also increases the likelihood of commercialization because delivering a working product or service is the task given to the consortium.



Introduction

In 2024, several reports were published on European competitiveness [DraghiReport, LettaReport, HeitorReport, ScienceEU]. The conclusions are clear: the democratic shift, the restructuring of the global economy, and changing geopolitical relations are reducing the influence of Europe in the world. The world has become much bigger, while Europe remains fragmented leading to a stunning size deficit compared to the current global competitors from the US and China. This impacts Europe's innovation capacity, productivity, job creation, security, … and in its wake the European political stability and eventually the European societal model. The solutions of the past might not be the most effective solutions for today's (and future) grand challenges. The reports call for some fundamental changes to make Europe more competitive. This chapter investigates what can be done to make the European computing sector more competitive in the future.

For the computing sector, today's conclusions are dire. Despite all Europe's efforts to boost research and innovation in digital technologies over the last two decades, Europe is seriously lagging behind the US and China in the domain of digital technologies. On the other hand, it is leading in the domain of sustainability technologies; see Figure 1.

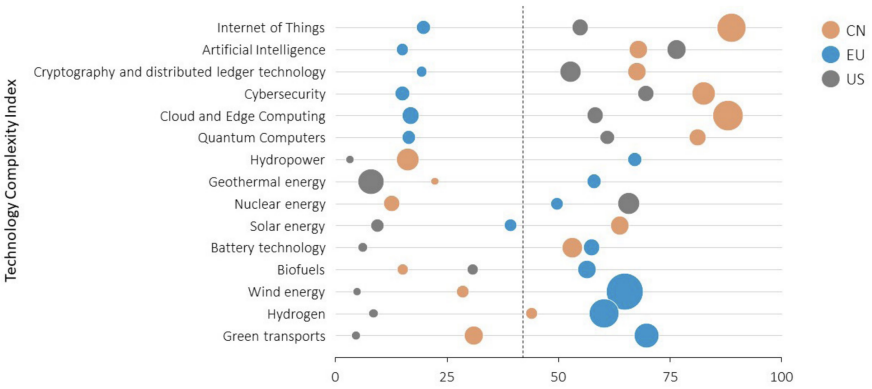


Figure 1: Europe's position in digital and green technologies (2019-2022). The x-axis indicates how easily a country can build a comparative advantage. The size of the bubble indicates how strong it already is [DraghiReport].

This is further illustrated in the Fuest report [FuestReport]. Table 1 depicts the biggest R&D spenders in 2003, 2012 and 2022 in the US, the EU and Japan.

	2003	2012	2022
US	Ford (auto) Pfizer (pharma) GM (auto)	Microsoft (software) Intel (hardware) Merck (pharma)	Alphabet (software) Meta (software) Microsoft (software)
EU	Mercedes-Benz (auto) Siemens (electronics) VW (auto)	VW (auto) Mercedes-Benz (auto) Bosch (auto)	VW (auto) Mercedes-Benz (auto) Bosch (auto)
JPN	Toyota (auto) Panasonic (electronics) Sony (electronics)	Toyota (auto) Honda (auto) Panasonic (electronics)	Toyota (auto) Honda (auto) NTT (telecom)

Source: Industrial R&D Investment Scoreboard (2004, 2013 and 2023).

Table 1 : Biggest R&D spenders in the US, EU and Japan over the last 20 years [FuestReport].

While in 2003 automotive was king with five companies out of a total of nine, followed by electronics (three out of nine), in 2012 it was still five out of nine for automotive, but none were left in the US, and there was only one out of nine left for electronics, based in Japan. In the US, the biggest spenders in 2012 were Microsoft and Intel. In 2022, they have been replaced by Alphabet, Meta and Microsoft. The biggest spenders of 2012 in the EU are all automotive, and the same companies are still the biggest spenders in 2022. Not mentioned in the Fuest report are the three biggest R&D spenders in China in 2022: (i) Huawei investment and holding, (ii) Tencent, and (iii) Alibaba group Holding. These companies were founded in 1987, 1998, and 1999, respectively. The EU's industrial innovation model is apparently more driven by established companies than in the US or China, where the leading companies in 2022 are much younger.

Figure 2 illustrates the private R&D investment evolution between 2013 and 2023 [EUScoreboard2024]. The European industry almost doubled its R&D investments up to the level the US companies in 2013, but the US companies more than doubled their efforts at the same time, hereby doubling the gap between the two. In 2022, US companies spent the

same amount of money on software R&D as the EU companies spend on software, hardware, health and automotive R&D combined. Chinese companies increased their investments eightfold over the same period and are currently almost on a par with the EU.

One third of the investments in Europe are coming from automotive companies, with limited investments by ICT companies (ICT hardware and ICT software), which are dwarfed by the investments by US and Chinese companies. The US companies invest 10x more in ICT software research than their European counterparts (up from 5.8 in 2013). At the current R&D investment levels, there is little chance that European industry will be able to catch up with US industry. The gap is simply too wide, and the resources available to close it are too limited.

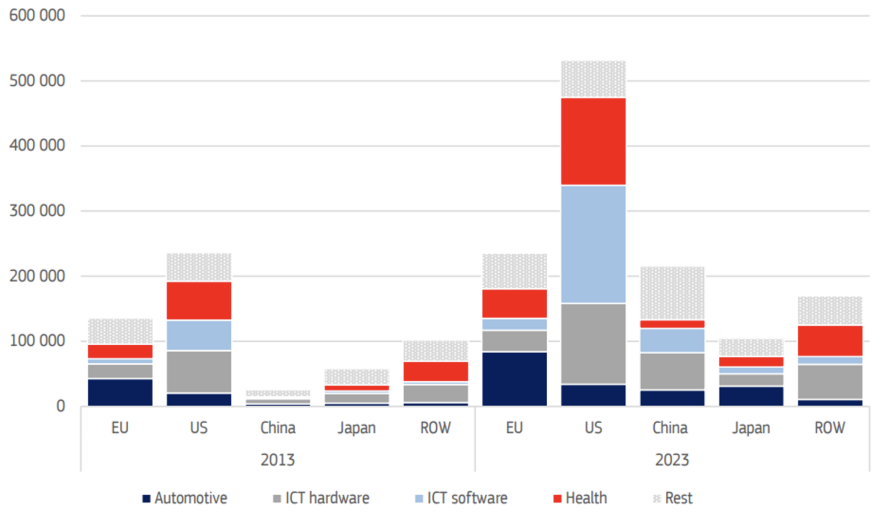


Figure 2: Top R&D investments (in million euro) per sector in 2013 and in 2023 [EUScoreboard2024].

In 2024 several European automotive companies got into financial trouble and had to lay off employees and close factories, which could lead to a negative impact on their short-term R&D investments.

Also worrisome is that the European automotive sector seems to have difficulties competing with the US and Chinese market leaders in electromobility: Tesla and BYD. Tesla was founded in 2003 and sold its first car in 2008. In 2003, it was a startup, staffed by a handful of people. Since 2020 it has been the most valuable car manufacturer in the world. At the time that Tesla was (i) bringing its model S (2012) and its model X (2015) to the market, (ii) went public (IPO in 2010), and (iii) introduced the Tesla Autopilot (2014), a major European automaker was trying to save its diesel car business by working on a cheat mode in the injection software. Although innovative, this is not the kind of innovation that will make Europe more competitive.

BYD Auto was also established in 2003. The first plug-in hybrid electric vehicle was launched in 2008, and the first battery electric vehicle in 2009. In 2023 Q4, BYD was the top-selling battery electric vehicle manufacturer of the world, bigger than Tesla. It overtook Volkswagen as best-selling car brand in China in 2023. It is the third most highly valued car manufacturer of the world, after Tesla and Toyota and followed by a series of European companies [JuliePinkerton]. Young people in China prefer BYD and perceive the European luxury brands as something for their parents.

From this limited analysis is clear that the US has been very successful in renewing its industry through so-called creative destruction. The decline of the automotive industry in what is now called the rust belt, has given rise to a much more innovative industry led by

software companies like Alphabet (founded as Google in 1998), Meta (founded as Thefacebook in 2004) and Microsoft (founded in 1975). Their original mission statements were respectively: “organize all the world's information and make it universally accessible and useful", “to give people the power to share and make the world more open and connected”, and “to put a computer on every desk and in every home”, and this is exactly what they did, while also “mov[ing] fast and break[ing] things!”. Europe somehow seems to lack the ambition level that characterizes the US hyperscaler companies.

Science and technology (S&T) clusters

For startup companies to be founded, and once they are viable to scale up, they need an ecosystem in which they can find all the resources to grow: talent, infrastructure, investors, and a thriving entrepreneurial community. According to WIPO [WIPO-ClusterMethodology], in 2023, Europe had two science and technology (S&T) clusters in the global top 20; see Table 2. The metric used for the ranking is the share of the global patents + the share of the global publications.

Paris is ranked no 12, and London no 20, but they both lost two positions compared to the 2022 ranking. The fact that Paris and London appear in this list is not surprising. They are the two largest metropolitan areas in Europe (with a population of more than 10 million), and an ecosystem can only grow large in a large metropolitan area. Other large metropolitan areas in Europe (Barcelona, Berlin, Madrid, …) are about half the size of Paris and London. Given the fact that the fast-growing cities with lots of young people are located outside Europe, the chance is low that Europe will be able to keep its position in the top 20 of global S&T clusters.

Rank	Cluster name	Economy	PCT applications	Scientific publications	Share total PCT filings (%)	Share of total pubs (%)	Total	Previous rank[a]	Rank change[a]
1	Tokyo–Yokohama	JP	127,418	115,020	10.1	1.5	11.7	1	0
2	Shenzhen–Hong Kong–Guangzhou	CN/HK	113,482	153,180	9.0	2.1	11.1	2	0
3	Seoul	KR	63,447	133,604	5.1	1.8	6.8	4	1
4	Beijing	CN	38,067	279,485	3.0	3.7	6.8	3	–1
5	Shanghai–Suzhou	CN	32,924	162,635	2.6	2.2	4.8	6	1
6	San Jose–San Francisco, CA	US	47,269	58,575	3.8	0.8	4.6	5	–1
7	Osaka–Kobe–Kyoto	JP	38,413	51,948	3.1	0.7	3.8	7	0
8	Boston–Cambridge, MA	US	18,184	76,378	1.4	1.0	2.5	8	0
9	San Diego, CA	US	23,261	20,928	1.9	0.3	2.1	11	2
10	New York City, NY	US	13,838	74,849	1.1	1.0	2.1	9	–1
11	Nanjing	CN	7,143	113,488	0.6	1.5	2.1	12	1
12	Paris	FR	15,176	61,692	1.2	0.8	2.0	10	–2
13	Wuhan	CN	6,250	89,756	0.5	1.2	1.7	15	2
14	Hangzhou	CN	10,755	62,924	0.9	0.8	1.7	14	0
15	Nagoya	JP	17,736	16,091	1.4	0.2	1.6	13	–2
16	Los Angeles, CA	US	11,556	44,058	0.9	0.6	1.5	16	0
17	Washington, DC–Baltimore, MD	US	5,525	76,039	0.4	1.0	1.5	17	0
18	Daejeon	KR	12,275	25,552	1.0	0.3	1.3	20	2
19	Xi'an	CN	1,786	86,937	0.1	1.2	1.3	21	2
20	London	GB	5,981	59,068	0.5	0.8	1.3	18	–2

Table 2: Top 20 Science and Technology clusters, 2023 [WIPO-ClusterMethodology]

Knowing that S&T clusters are essential for startups to grow and thrive, Europe should actively encourage the creation of a multitude of medium sized S&T clusters in major urban areas in Europe. These will not land into the top 20, but they will be local innovation engines, creating well-paid jobs, stop brain drain and providing opportunities for the young generation. The performance per capita might even be higher than the S&T clusters in the top 20.

It is however important to realize that such clusters cannot be created overnight, but they need time to grow and become productive, and they are always the result of joint efforts between different stakeholders.

- Local schools and universities must invest in research areas that are relevant for the local economy, and also develop the entrepreneurial skills of their students. This combination will result in spinoffs and startups, and it will also result in graduates that are ready to work in the local ecosystem (and attractive jobs in the local startups can stop them from looking for a job elsewhere). The schools and universities will also benefit from the ecosystem: contract research, company internships, and the ability to attract talented students who would like to work for one of the ecosystem companies.
- Local governments should offer ample space for high tech companies to build the infrastructure they need and have a fast and pragmatic permissions policy. They should also arrange for affordable housing, an international school, efficient urban mobility and a liveable city.
- The (national and/or regional) government can create incentives to attract companies to designated areas (tax incentives, subsidies, …).
- Local companies must organize themselves too to make the infrastructure provided by the government into a vibrant and welcoming community in which all companies can learn from each other, help each other, celebrate the successes, and especially grow the ecosystem, by e.g. investing in incubators, accelerators, ecosystem marketing, etc.

All the above needs time, but if all stakeholders (city, schools/universities, government, companies, …) in an area are willing to create such an ecosystem, synchronize their plans and investments, it can be built, and become the engine of economic development in the area. It takes time to produce the first big success stories (e.g. a unicorn), but ones it reaches that level, and with the right marketing efforts, it will automatically attract talent and investors, and its growth will accelerate. All current large S&T clusters once started small.

Research excellence

Europe not only lacks global science and technology clusters; it is also losing its position in basic research. The Nature Index tracks contributions to research articles published in high-quality natural-science and health-science journals, chosen based on reputation by an independent group of researchers; it has yearly updates. It can be used as a proxy for research excellence (just one possible proxy out of several). Table 3 shows the number of institutions in the Nature Index 2024 `[NatureIndex]`.

	Nature index 2024			
	Europe	US	China	Other
Top 10	2	1	7	0
Top 50	6	18	22	4
Top 200	46	61	61	32
Top 500	156	137	130	77

Table 3: Nature Index 2024 `[NatureIndex]`. The table contains the number of institutions in different subrankings. The subrankings are inclusive (i.e. the Top 50 contains the Top 10 institutions, etc.).

The top 10 in 2024 is dominated by Chinese institutions (70%), in the top 50, Chinese institutions are still 44% of all institutions, and Europe has only 12% of them. In the top 200, the US and China are on a par with 30% while Europe is at 23%. In the top 500, Europe leads with 31%. One could conclude that Europe does not lead in the excellent institutes (Top 50), but it clearly leads in the good ones (31% of the top 500 institutions). The group 'Other' consist of mostly Asian institutions (Japan, South Korea, Singapore), Australia and Canada.

Figure 3, which plots the evolution of the top 50 over the last eight (!) years, put the 2024 situation in perspective: In less than a decade, Chinese institutions have succeeded in building a very strong position in the Nature Index. They have done so at the expense of Europe, the US and the rest of the world.

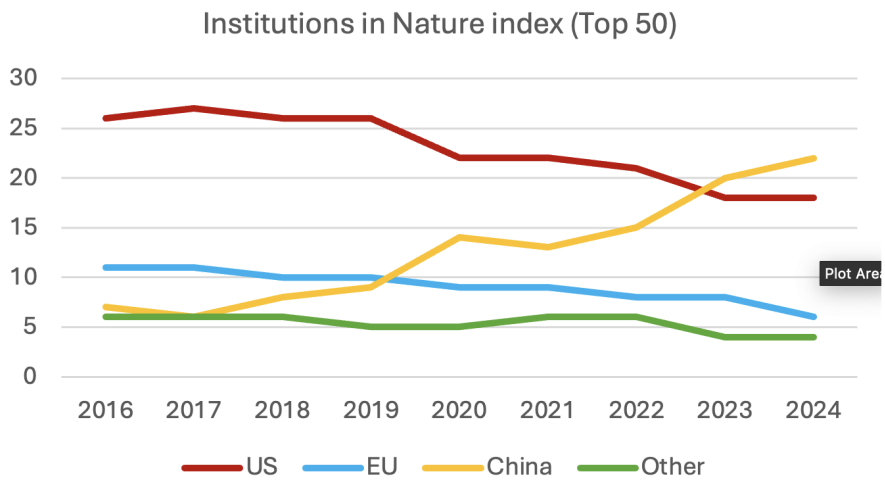


Figure 3: Evolution of the Nature Index Top 50, 2016-2024.

The seriousness of the situation becomes clear if we look at the detailed Top 10 in 2016 and compare it to 2024, as depicted in Table 3. Europe went in a timespan of eight years from five institutions to only two. The US from three to one. China grew from only one to seven out of 10. The fact that Oxford and Cambridge dropped from the Top 10 (Oxford is now at position 20, and Cambridge at 22) is telling for the new world order in research. China clearly 'moves fast and breaks things'.

Top 10 Nature index 2016	Top 10 Nature index 2024
Chinese Academy of Sciences (CAS), China	Chinese Academy of Sciences (CAS), China
Harvard University, United States of America (USA)	Harvard University, United States of America (USA)
French National Centre for Scientific Research, France	**Max Planck Society, Germany**
Max Planck Society, Germany	University of Chinese Academy of Sciences (UCAS), China
Stanford University, United States of America (USA)	University of Science and Technology of China (USTC), China
Massachusetts Institute of Technology (MIT), (USA)	Peking University (PKU), China
Helmholtz Association of German Research Centres, Germany	**French National Centre for Scientific Research, France**
The University of Tokyo (UTokyo), Japan	Nanjing University (NJU), China
University of Oxford, United Kingdom (UK)	Zhejiang University (ZJU), China
University of Cambridge, United Kingdom (UK)	Tsinghua University, China

Table 3. Top 10 institutions in 2016 and in 2024.

The full comparison is depicted in Table 4. It shows that the 31% of European institutions in the top 500 in 2024 was 38% in 2016, or a drop of 36 institutions in less than one decade!

	Nature index 2016				Nature index 2024			
	Europe	US	China	Other	Europe	US	China	Other
Top 10	5	3	1	1	2	1	7	0
Top 50	11	26	7	6	6	18	22	4
Top 200	60	75	24	41	46	61	61	32
Top 500	192	153	53	102	156	137	130	77

Table 4. Comparison of the Top 500

One could argue that the Nature Index is not the most relevant index for ICT publications, but the submissions in ACM TACO (which reviews the papers for the HiPEAC conference) show a similar pattern: in 2016, it received 38 submissions from China on a total of 199, while in 2024, there were 178(!) submissions on a total of 303. In 2016, there were five accepted papers from China (acceptance rate 8% compared to 27% for the journal), while in 2024 there were 41 (acceptance rate 23%, compared to 31% for the journal). At the main paper track of the HiPEAC 2025 conference, 17 papers are presented by Chinese authors (out of 29). The Chinese research institutions have undeniably caught up with the US and the EU.

Why is this important? Excellent research feeds the innovation pipeline. Europe has a tradition of research excellence but has proven to be weaker in commercialization of the research (which often took place in the US). If Europe is losing its leading position in excellent research, it will inevitably have an impact on the innovation capacity of Europe in the long term. Applied to China, their recent research excellence will obviously create a huge potential for innovation and commercialization of innovative products. Figure 3 proves that 2024 is not an outlier, but the result of the trend that seems to be accelerating. The breakneck progress in AI makes this trend even more worrisome. If data science and AI becomes the engine of scientific discovery, the countries with the most data and compute capacity will have a competitive advantage.

So, the question is: where is Europe in this new world order, and how is it going to position itself? If Europe wants to stay commercially competitive, it will also have to stay competitive in research by unapologetically stimulating research excellence. One goal could be to have 20% or more institutions in the Nature Index which would come down to (2, 10, 40, 125) in the different rows of Table 4. This is lower than the numbers in 2016, but it takes into account that a new and ambitious player has entered the ranking, and the ranking is a zero-sum game.

Conclusion

After decades of investments in digital technologies in Europe, it made a lot of progress, but the US and China made even more progress, further increasing the gap. This is a wake-up call for Europe and suggests that the current European research and innovation policy is not adequate to keep up with Europe's main competitors. This vision makes three recommendations to improve the situation.

1. Europe should actively promote the creation of **European S&T clusters in major urban areas** and help them grow to a scale that they can support scaleup companies. This will stimulate the creation of innovative start-ups and retain talent in Europe.
2. **ARPA model of challenges** should be used to introduce a new R&D culture: ambitious, bold, fast, milestone-based, competitive, risk-tolerant, visionary, agile.
3. **Pre-competitive procurement** should be used to speed up the introduction of innovative solutions to the market.

References

DraghiReport: European Commission. The Future of European Competitiveness. By Mario Draghi, Publications Office of the European Union, 2024, https://commission.europa.eu/topics/strengthening-european-competitiveness/eu-competitiveness-looking-ahead_en

EUScoreboard2024: European Commission: Joint Research Centre, 2024 EU industrial R&D investment scoreboard, Publications Office of the European Union, 2024, https://data.europa.eu/doi/10.2760/0775231

FuestReport: Clemens Fuest, Daniel Gros, Philipp-Leo Mengel, Giorgio Presidente, and Jean Tirole: "EU Innovation Policy – How to Escape the Middle Technology Trap?" EconPol Policy Report, April 2024, https://www.econpol.eu/publications/policy_report/eu-innovation-policy-how-to-escape-the-middle-technology-trap

HeitorReport: European Commission: Directorate-General for Research and Innovation. Align, Act, Accelerate – Research, Technology and Innovation to Boost European Competitiveness. By Manuel Heitor, Publications Office of the European Union, 2024, https://data.europa.eu/doi/10.2777/9106236

JuliePinkerton: Julie Pinkerton, "The 10 Most Valuable Car Companies in the World By Market Capitalization", USNews, 2024, https://money.usnews.com/investing/articles/the-10-most-valuable-auto-companies-in-the-world

LettaReport: European Commission. Much More Than a Market: Report by the High-Level Group on the Single Market, Chaired by Enrico Letta. Publications Office of the European Union, 2023, https://single-market-economy.ec.europa.eu/news/enrico-lettas-report-future-single-market-2024-04-10_en

NatureIndex: Nature Index, https://www.nature.com/nature-index/

ScienceEU: European Commission: Directorate-General for Research and Innovation, Science, research and innovation performance of the EU, 2024 – A competitive Europe for a sustainable future, Publications Office of the European Union, 2024, https://data.europa.eu/doi/10.2777/965670

WIPO-ClusterMethodology: Global Innovation Index science and technology cluster methodology, https://www.wipo.int/export/sites/www/global_innovation_index/en/docs/gii-2023-cluster-methodology.pdf

Next Computing Paradigm

Recommendations for the Next Computing Paradigm

Create digital envelopes

Create a **"digital envelope"** for integrating "anything" (person, company, physical entity, computing device) in a digital space allowing access to multiplicity of services – i.e. building on the concept of "anything-as-a-service" (XaaS) – in an interoperable way. This digital envelope would allow live migration of compute components and a **runtime evolving infrastructure**to support **deployment on a continuum**ranging from resource-constrained edge devices to data centres, allowing for dynamic resource pooling and efficient sandboxed execution of collaborative, migratory compute components offering services.

1. An intelligent **digital agent** able to pursue goals legitimately assigned to it. The intelligent digital agent would be capable of direct execution as well as of **orchestration**. The former would be required when seeking the set goal required local actuation, the latter when the task resulting from the set goal required remote execution.
2. **Sensors**, to pull digitalized inputs from designated sources (in the physical world or other digital envelopes).
3. **Actuators**, to push computed outputs into designated targets (physical things or digital envelopes). The fabric resulting from interconnecting digital envelopes that provide and require services from one another to pursue assigned goals will operate in genuine XaaS modality.

> "Digital enveloping" is the technology-enabled phenomenon by which any item of reality – human, material or immaterial – can be associated with a computable digital representation capable of delegated autonomous action. The notion of delegated autonomous action entails two fundamental traits: that of delegation, which suggests a higher (human) authority that requires some action to be taken (in part) in the digital space; and that of autonomy, which suggests that the pursuit and execution of the required action is carried out by autonomous executable agents that operate within the remits of delegated authority.
>
> The capacity for delegated autonomous action is provided to individual digital envelopes by the combined operation of three key components:

These solutions enable the live migration of compute components across the edge-to-cloud continuum – therefore services – ensuring continuity while addressing latency, privacy, security, risk management, validation mechanisms and context requirements. This is essential to optimize user and infrastructure needs dynamically. In relation to real-world services or actions triggered by a digital envelope, location and time identifiers are assigned, and inter-envelope mechanisms support local vs global optimizations including for safe interactions, managing complex interdependencies and conflict resolution. The next

computing (NCP) exemplifies the notion of continuum. Agreed standards are key for interoperability.

AI-powered orchestrators

Artificial intelligence (AI)-powered orchestrators will be an essential capability of the intelligent digital agent of the digital envelope. AI-powered orchestrators will be developed for the edge – which is strategic as it is located the nearest to the final user – in a manner that can dynamically combine collaborative compute components into executable applications tailored around specific user needs. The task of the orchestrators is to decompose goals set by the user (in the broader term, including human user, company or another permissioned orchestration) into a set of services that cooperate to achieve the set goals. These orchestrators could be themselves generated by (federated) **generative AI (genAI) engines** (supported by more classical algorithmic approaches) located at the edge and capable of collaboration with other orchestrators within federated zones.

Space- and time-aware protocols

Expand and adapt web-level protocols and associated standards by enhancing the existing suite of HTTP-based protocols to be both spatially aware and time-sensitive. This will allow web-level interactions between NCP's migratory compute components to account for 3D physical space and real-time communication, drawing on technologies like WebRTC to manage time-sensitive tasks effectively and the spatial web (IEEE P2874, OpenUSD, …).

Interoperable contract-based API specifications

Establish interoperable, contract-based application programming interface (API) specifications – usable by expanded web-level protocols – ensuring that interconnected services communicate with clear expectations of both functional and non-functional performance. These contracts, similar to service-level agreements (SLAs), should detail the conditions under which services will optimally perform, including non-functional requirements, ensuring smooth integration and reliable service delivery within the NCP framework. These APIs should account for non-functional properties like latency, cost, and performance. The resulting model should ensure that an API not only promises to deliver a service but also specifies the conditions under which it can perform optimally. The API should also be compliant with the currently proposed API for large language models LLMs `[OpenAIFunction] [BerkeleyFunction]`.

Promoting these standards in relevant standardization bodies is essential for fostering interoperability and consolidating development conditions through standardized benchmarks, testing methodologies, and best practices. This will ensure that implementations can be effectively and securely integrated, improving overall system efficiency and reliability, and enabling the creation of an interoperable business ecosystem of services and orchestrators.

Introduction

NIST Recommendation SP 800-145 [NIST], dated 2011, lists five defining traits that characterize cloud computing:

1. on-demand self-service,
2. broad network access,
3. resource pooling,
4. rapid elasticity,
5. measured service.

Back then, features (1), (3) and (4) were by far the most visionary ones in terms of (provider-side) requirements and (user-side) expectations. In fact, their pursuit has had a major impact, shaping a whole new world of cloud-enabled technology in the subsequent decade.

Feature (1) implies that, rather than (application) services having to be installed, they would be delivered via the web, that is, via client-side web browsers that consequently became "versatile self-contained fully-provisioned application environments". All the client side needs in the cloud model is a cloud-enabled web browser and broad network access. This notion has had vast consequences and stands at the basis of the NCP vision, as discussed in the following section.

Features (3) and (4), largely immaterial to the client side, concern primarily what the provider platform must be able to do. Resource pooling is the principle by which the provisioning and apportionment of computing, storage and networking is no longer confined to a single physical place. In the cloud model they become virtual units, which result from concrete fragments opportunistically scattered in multiple places, located wherever there is a convenient temporary "home" for them.

> The term "home" is meant here to designate infrastructure resources (compute, storage, and networking) able to support the deployment and the execution of the digital entities that are being pooled. An analogy may clarify. In an operating-system environment, memory is made available to executing processes as a logical resource that virtualizes physical memory. No

single process actually owns physical memory, which is divided in page frames handled by the OS. Processes are loaned (sparse) page frames, strictly on the base of need, to host page contents coming from and going to secondary storage. Resource pooling in the cloud essentially follows the same concept, except spanning over networked nodes, which effectively means virtualization over the network, beyond the physical boundaries of a single computer.

Earlier computing models had already long known and practised virtualization, which has since become the foundation of the computing stack in the guise of virtual memory, pre-emptive scheduling, file systems, to name just a few.

The dominant interpretation of the cloud as a concrete provisioning platform soon became that of the giant corporations that saw and developed the web as their marketplace. In that view, resource pooling would be achieved by amassing and virtualizing immense clusters of comparatively cheap networked computers deployed at strategic locations. At that point, computing would happen on any of such clusters (at the notional centre of the cloud) and data would flow there from its sources (at the notional edge of the network). Users at the client side would only need web browsers on their devices to be able to use rich, reactive, sophisticated single-page web applications, while most of the juicy action would happen on the server side at the centre of the cloud.

The cited NIST Definition also posits that cloud computing has three service models: software-as-a-service (SaaS); platform-as-a-service (PaaS); infrastructure-as-a-service (IaaS). The SaaS model was the most obvious and immediate one to be understood, as it speaks directly to the end user. The IaaS model allowed enterprises to conceive and deliver SaaS offerings without owning concrete infrastructures and yet being able to control rental costs. In fact, it was the IaaS model and not the SaaS that allowed thriving digital businesses to emerge.

It is now technically possible and strategically opportune to separate what is specific to the defining traits of the cloud in the “traditional” model from what can be realized in alternative modalities. Doing that opens up novel and unprecedented opportunities that belong in the HiPEAC Vision.

The traditional model places at the centre of the cloud the centre of gravity of computing. The user and the data are attracted to gravitate towards and around it. That tenet carries the view that the “important” computing resources are available solely at the centre of the cloud. That is the fundamental premise to monopoly, which is what we have observed in the cloud offering in the last decade.

The fact is however that at present a vast cumulative amount of computing resources is available at the edge of the network, where users and data sources are.

Consider the total number of computers embedded in cellular phones (7.2 billion in use worldwide in the year 2024), modern transportation vehicles (several hundreds of million times ten or more per vehicle) both mobile and stationary, home automation systems. Imagine some of their resources pooled together, opportunistically around geographically close zones, to host edge-related applications. The infrastructure resulting from this virtual pool would never compete with cloud-enabled data centres, purpose-built to support enterprise-level applications and large-scale data processing. And never it should, in fact, as edge-friendly applications are nimble and low latency, which is quite the opposite extreme to them.

If those resources were pooled together seamlessly, à lacloud, innumerable value-added computations could take place at the edge instead of at the centre of the cloud; see Figure. That shift would prize privacy, latency, energy, decentralization, personalization, context-awareness in a manner that the centre of the cloud could not possibly match.

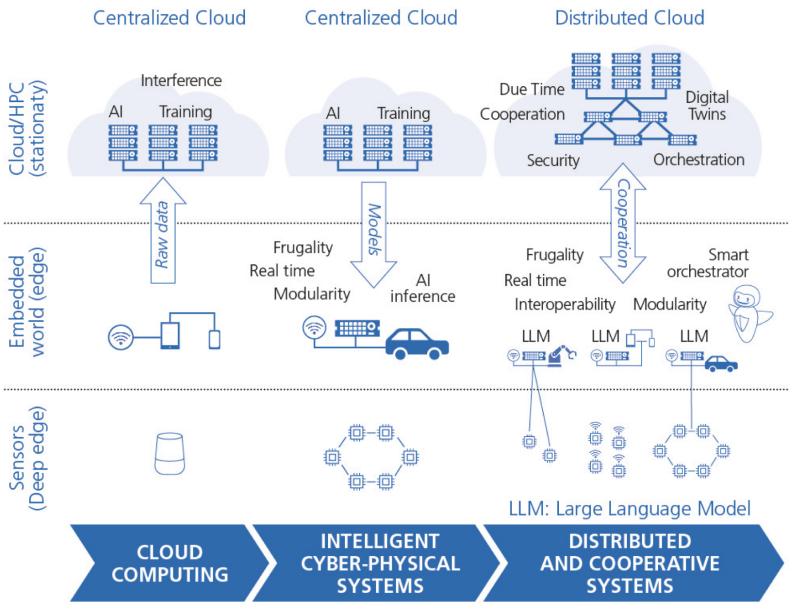


Figure 1: Evolution of computing infrastructures towards the NCP, where services are distributed and cooperate together. Credit: Denis Dutoit, CEA

The established principles of resource pooling and virtualization applied to the edge would allow the creation of malleable, powerful, dynamic, virtually ubiquitous federations of edge nodes strategically positioned where the physical world borders the digital sphere. This connotation is essential to the NCP as digital resources capable of sensing and actuation, in addition to computing, may interact with entities in the physical space causing the physical and the digital worlds to come together seamlessly and dynamically.

Pooling edge resources among themselves and with the cloud seamlessly gives rise to the so-called edge-cloud continuum, a compute infrastructure where computation would be deployed opportunistically and dynamically, wherever that is more convenient for the user.

Digital envelope

"Digital enveloping" is the technology-enabled phenomenon by which any item of reality, human, material and immaterial, can be associated with a computable digital representation capable of delegated autonomous action.

> The notion of delegated autonomous action entails two fundamental traits: that of delegation, which suggests a higher (human) authority that requires some action to be taken (in part) in the digital space; and that of autonomy, which suggests that the pursuit and execution of the required action is carried out by autonomous executable agents that operate within the remits of delegated authority.

The capacity for delegated autonomous action is provided to individual digital envelopes by the combined operation of three key components; see Figure:

- An intelligent **digital agent** able to pursue goals legitimately assigned to it. The intelligent digital agent would be capable of direct execution as well as of **orchestration**. The former would be required when seeking the set goal would require

local actuation. The latter when the task resulting from the set goal would require remote execution.

- **Sensors**, to pull digitalized inputs from designated sources (in the physical world or other digital envelopes).
- **Actuators**, to push computed outputs into designated targets (physical things or digital envelopes).

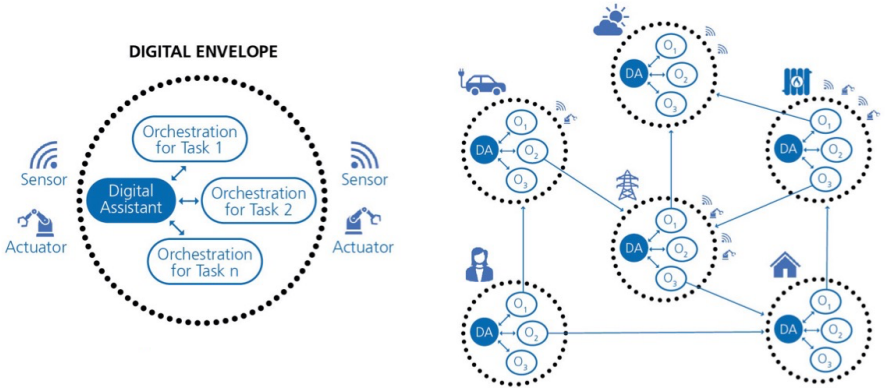


Figure 2: Digital envelopes interacting together

The fabric resulting from interconnecting digital envelopes that provide and require services from one another to pursue assigned goals will operate in genuine XaaS `[XaaS]` modality.

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 way of HTTP verbs `[Kann]`. Digital envelopes would thus be woven into a next-generation web `[HV23NextWeb]`, which brings together the web of humans with the digital web, into a programmable and interoperable hyperspace. The XaaS paradigm emanates from that notion as a major vector of innovation, which shifts the centre of gravity away from the cloud towards the edge.

The compute component of digital envelopes must instead be capable of live migration on a **continuum runtime infrastructure** spanning from resource-constrained edge devices to data centres. That capability is a direct consequence of allowing for orchestrated actions to be deployed wherever their execution is best assigned, where the notion of “best” may even change over time.

The continuum infrastructure should allow for dynamic resource pooling and efficient sandboxed execution of collaborative, migratory service-providing and service-requiring compute components. Live migration of compute components across the edge-to-cloud continuum, therefore services, will ensure continuity of service while addressing latency, privacy, security, risk management, validation mechanisms and context requirements. This capability is essential to optimize user and infrastructure needs dynamically.

Digital envelopes would have owners, who should be 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 orchestrations of request-response interactions with other digital envelopes (thereby with 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 sphere, 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.

A simple example might help illustrate the concept of the digital envelope.

Use case: Travel

Travelling independently for disabled people (wheelchair users, visually impaired people, …) is a challenge. The personal agent, operating from within the digital envelope, and the NCP, can help such people to travel independently, as the following scenario illustrates. Before leaving, the personal agent will produce a travel plan, based on the starting point and the destination, including all the assistance needed during the trip. While travelling, the personal agent will continuously update the travel plan based on actual information.

In a typical travel scenario, the personal agent will instruct the orchestrator to call a taxi to drive to the station. It will make sure that (i) the taxi has the space to take a wheelchair on board, and (ii) the taxi arrives at the station on time. Before arriving at the station, the personal agent contacts the digital envelope of the train station and arranges assistance to get to the right platform and to board the train. On the train, the personal agent contacts the digital envelope of the train, which in turn contacts the digital envelope of the staff of the train. In case the traveller needs assistance, they can ask their agent to contact the train staff. At the destination, the personal agent will again contact the digital envelope of the train station and order local assistance. If special assistance is not required, the agent will help the traveller find the best route to their destination (wheelchair accessible, adapted to blind travellers, …).

On arrival at the destination, the personal agent will look for a place to eat. It will only show the restaurants that are wheelchair accessible, and that offer items that are compliant with the dietary requirements and the preferences of the traveller.

The personal agent will also take care of all the tickets and payments. This means that travellers can freely use any bus, tram, metro, shared bike, or enter a museum without having to worry about the payment. During ticket inspection, the inspector or the inspection machine will directly talk to the digital envelope of the traveller.

When renting a car, the personal agent will take care of all the "paperwork" ahead of time, and there is no longer the need to pick up the keys. The personal agent will directly talk to the digital envelope of the car and give the driver access to the car. The personal assistant will also help the driver to operate the car by answering questions, or, in some cases, by taking actions ("I will switch on the fog lights for you"), or by warning the driver ("it is better to recharge here because the next charging station is 200 km away"). In a future scenario the personal agent will instruct the digital envelope of the car to autonomously drive to the destination.

Among other things, the use-case scenarios of the digital envelope discussed in this document show that a large fraction of the (compute-and-communicate) actions pertinent to achieving a user-related goal ought to occur near or at the edge. They further posit that certain edge nodes may need to be able to aggregate opportunistically into ephemeral (temporary) federations to accomplish assigned goals in a manner that respects legal and physical boundaries and constraints, and that seeks some definition of overall efficiency.

There is very clear correspondence between the vision outlined above and the fast-rising momentum of “agentic AI”.

For an explanation of the notion of “agentic AI”, see for instance [ErikPounds], although note that this piece – owing to the identify of its editor – suffers some commercial bias.

The APA Dictionary of Psychology defines “agentic” as a psychological condition that occurs when individuals, as subordinates to a higher authority in an organized status hierarchy, feel compelled to obey the orders issued by that authority [APA]. When used in the AI context, the “agentic” term thus is loaded because psychology associates it with potentially negative connotations (destructive obedience), which suggests extreme caution when deployed in digital programs that are bound to act much faster and deeper than human mind can comprehend.

Agentic AI is very clearly the next frontier of genAI, moving it beyond the request-response modality that it has had so far, which has been shown to lack scalability, and to be confined to cute but limited code assist, customer service, and content writing service contexts. The current genAI model is that of pipeline where:

1. a request is initiated via a natural-language, written or oral, prompt;
2. relevant data is accessed through a retrieval-augmented generation, RAG;
3. an answer is returned, which may be right (accurate, pertinent) or wrong (inaccurate, not pertinent, erroneous).

The new model of agentic AI uses genAI to draw and execute a plan to perform work that is to meet user-specified goals. In doing so, the digital entity (which the agentic AI literature calls “agent”, causing the reader to believe that “agent” is synonym to “agentic”, which it really is not) may work in concert with other such digital entities as part of an orchestration of interactions expected to deliver coordinated outcomes. It should be noted that this notion of orchestration has been the focus of attention of several prior editions of the HiPEAC Vision [HV21Angels], [HV23Digels]. It should also be noted that orchestrations can be realized as hierarchical descents, from a higher-level (more abstract) set of goals into lower-level (increasingly more concrete) set of either simple tasks or even other orchestrations. Some such orchestrations may coalesce into advertised capabilities, as permanent entries into public registries.

Early demonstrators have been released lately by various actors at the forefront of genAI, which show glimpses of what the evoked scenarios may give rise to; see for example: [Gorilla], [Magnetic-One]. Interestingly, the most profound implication of these developments is that the entire technology stack needed for all these digital envelopes to be deployed and executed (prompts, routines, tools, function schemas, handoffs, etc.) might be generated on the fly ad infinitum, as part of the mechanics of turning goals into plans, and acting in response to contingencies resulting from actions along the pipeline.

To sum up

The potentially cascading or federated orchestration discussed in this chapter will have to keep an efficient balance between resource availability (which values the centre of the cloud), and privacy, latency, energy, decentralization, personalization, context-awareness (which prizes the edge). That will have to be much more dynamic and adaptive than traditional orchestration at the centre of the cloud. The resulting orchestrations would be dynamic, opportunistic, ephemeral, and maximally loosely coupled, in addition to collaborative (and thus hierarchical or federative or both). The associated computations (tasks) should be able to move across the continuum in search of the temporary residence best fit to meet stated goals.

The envisioned orchestration would embed intelligence, including next-generation genAI, to do the bidding of individual users at the edge, prompted by user goals and requirements and

returning ad hoc programmatic orchestration engines. The underlying infrastructures would also need intelligence to federate opportunely and adaptively available resources.

This model entails a whole a new frontier for computation, computing artifacts, and computing infrastructure, which this document calls the next computing paradigm (NCP). Realizing the NCP requires evolving the runtime infrastructures availed at the edge. It also requires expanding the web-level suite of protocols in order to become spatially aware, so that the web becomes a 4D place (aware of the three dimensions of our physical reality, plus time-sensitive) that all interactions, node-to-node, client-server, machine-to-machine may be carried by HTTPs-based bidirectional multiplexed server-prompted and asynchronous channels.

Conclusion

In order for the vision outlined in this chapter to be brought to fruition, certain specific routes of innovation would have to be taken. We list them next in no particular order.

Runtime infrastructures fit for deployment onto resource-scarce compute devices at the edge should be developed, making them capable of supporting dynamic resource pooling and of hosting efficient sandboxed execution of migratory collaborative compute components.

> Migration is an essential trait of the opportunistic federation of computing entailed by the notion of digital envelopes: actions must be taken at places that depend on the set goals and on the logistical constraints (in the physical or digital world or both). An orchestrated action must therefore be dispatched for execution at a place that is other from that of residence of the orchestrator.

Solutions that allow compute components to live migrate across the edge-to-cloud continuum should be developed that warrant continuity of execution, whenever transfer to a different node may warrant superior coverage of user- and infrastructure requirements such as latency, privacy, security, provenance, context, etc.

The web-level suite of HTTP-based protocols should be expanded and streamlined to make them (1) spatially aware, so that web-level interaction between migratory compute components is aware of 3D physical space, and well as (2) time-sensitive, learning from the real-time capabilities of e.g., WebRTC.

API description standards such as `[OpenAIAPI]` are currently being made obsolete by “function schemas” or equivalent technicalities that revolve on a local and opportunistic need to describe APIs to LLMs so that the latter can incorporate the former into responses to prompts, and generate actions from goals, which call them at the appropriate place. What is needed, instead of local, ad hoc, half-baked solutions, is a concerted design effect that determines how best to describe APIs so that all requirements discussed in this document can be met satisfactorily and in an open, interoperable manner.

The API description standard of interconnected web-level services should be augmented with interoperable contract-based specifications, akin to service-level agreements (SLAs) or assume-guarantee pairs, to ensure that required and provided services can communicate with an expected level of functional and non-functional performance. The resulting model should ensure that an API not only promises to deliver a service but also specifies the conditions under which it can perform optimally. This requirement of functional and non-functional interoperability extends to the modality of interconnection between LLMs, which currently goes under the name of “function calling”.

Solutions that allow the embedding of genAI at the edge should be developed in order that human users can be provided with natural interfaces (voice, gesture, eye movements, touch) to the digital world, with more energy efficiency, reduced latency, lesser communication overhead, and greater privacy.

AI-powered edge-based orchestrators should be developed that reflect the vision discussed in this chapter. These should be capable of dynamically combining migratable collaborative compute components into ephemeral, opportunistic smart personalized applications in response to user requirements. Those orchestrators should be the programmatic output of genAI engines located at the edge and should be able to collaborative with other such engines located within federative zones.

Demonstrable proof-of-concept implementations should be developed based on elements of the capabilities evoked in this chapter, whether limited to selected features only or on a more holistic scale, in articulations that are not proprietary and support open standards and platforms.

References

`APA:` American Psychological Association. APA Dictionary of Psychology: “agentic stage”. https://dictionary.apa.org/agentic-state

`BerkeleyFunction:` Berkeley Function-Calling Leaderboard. https://gorilla.cs.berkeley.edu/leaderboard.html

`ErikPounds:` Erik Pounds @ NVIDIA. “What Is Agentic AI?”. October 22, 2024. https://blogs.nvidia.com/blog/what-is-agentic-ai/

`Gorilla:` Shishir G. Patil and Tianjun Zhang and Xin Wang and Joseph E. Gonzalez: “Gorilla: Large Language Model Connected with Massive APIs”. arXiv, May 24, 2023. https://arxiv.org/pdf/2305.15334

`HV21Angels:` Marc Duranton and Tullio Vardanega: “Guardian Angels” to protect and orchestrate cyber life. HiPEAC Vision 2021 (Pages 50-55). https://www.hipeac.net/vision/2021.pdf

`HV23Digels:` Tullio Vardanega and Marc Duranton: “Digels”, digital genius loci engines to guide and protect users in the “next web”. HiPEAC Vision 2023 (Pages 18-21). https://www.hipeac.net/vision/2023.pdf

`HV23NextWeb:` HiPEAC Vision 2023. The Race for the “Next Web” (pages 13-64). https://www.hipeac.net/vision/2023.pdf

`Kann:` Charles W. Kann III: §6.1 CRUD Interface. LibreTexts Engineering. https://shorturl.at/cb0Wp

`Magnetic-One:` Adam Fourney and Gagan Bansal and Hussein Mozannar and Victor Dibia and Saleema Amershi: “Magentic-One: A Generalist Multi-Agent System for Solving Complex Tasks”. Microsoft. AI Frontiers blog, November 12, 2024. https://shorturl.at/ihBLx

`NIST:` NIST Special Publication 800-145: The NIST Definition of Cloud Computing. September 2011. https://doi.org/10.6028/NIST.SP.800-145.

`OpenAIAPI:` https://www.openapis.org/

`OpenAIFunction:` OpenAI Platform: Function Calling. https://platform.openai.com/docs/guides/function-calling

`XaaS:` Short for “Anything-as-a-Service”. See for example: https://www.digitalroute.com/resources/glossary/xaas/

Artificial Intelligence

Recommendations for Artificial Intelligence

Develop distributed agentic AI (specialized action models)

The **development of specialized action models (SAMs)** acting as service is important and can be developed in Europe. These SAMs, small and specialized models that can interact with their environment, should operate in a distributed infrastructure and an ecosystem should be created to support research, development and business around them. These models need to be refined, optimized, and reduced in size to improve efficiency. These SAMs can be optimized from more general foundation models by an ecosystem of companies providing their optimized SAMs in a marketplace so that they can be dynamically discovered and used by the orchestrators.

Develop orchestrating technologies for distributed agentic AI, blueprint for NCP orchestrators

We call agentic AI a set of specialized AI agents working together to accomplish a common goal. An AI agent is synonymous with an SAM in this discussion: an AI that can perceive and act, having impact on the virtual or real world. **The orchestration technologies** should take into account all the requirements, that can select the best SAMs for the required tasks and dynamically activate them. The first steps could be very agentic-AI-centric (relying on already existing technologies used for orchestrating AI agents), but they should be blueprint and evolve towards an orchestration system for the NCP. These orchestrators must be developed for the edge – or near the final user – and dynamically combine SAMs into executing personalized applications in response to user needs.

Establish open protocols for these "distributed agentic AI" systems to facilitate seamless interaction among distributed AIs from different origins

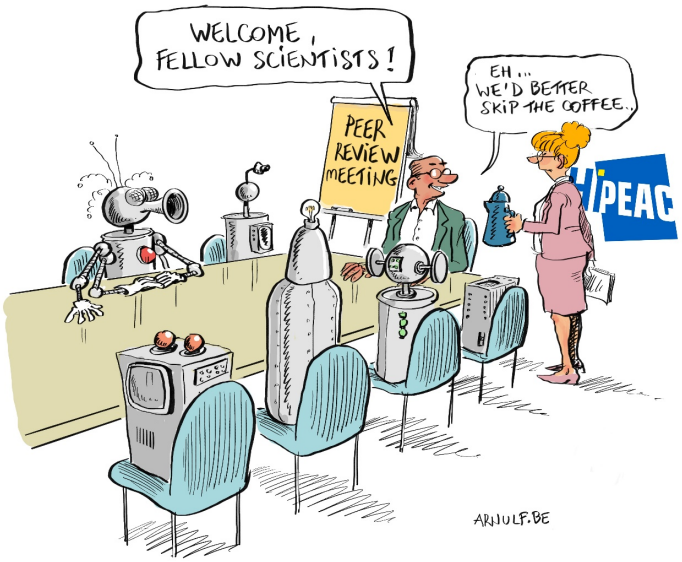
Protocols and specifications that group all requirements, existing ideas and proposals together in a single consortium to develop an open source "de facto" (before official standardization) standard protocol that takes into account all the good ideas of various researchers and organizations, so that it will be sound, future-proof, recognized and accepted. The requirements are:

1. It does not solely rely on functional requirements (e.g. the textual representation of prompts and responses).

1. It also incorporates non-functional requirements (providing sufficient information for the orchestrator to select the appropriate services, such as based on criteria like response time, potential level of hallucinations, environmental impact, cost, localization, privacy of data, etc.).

The recommendation to develop generative AI at the edge (AI) is still important, but it is more in development and implementation mode now (for example, in Apple intelligence). We

should continue developing solutions that allow embedding generative AI at the edge in order that **human users can be provided with natural interfaces** (voice, gesture, eye movements, touch) to the digital world, with more energy efficiency, reduced latency, lesser communication overhead, and greater privacy. This is important to reduce the difficulties to access the digital world and decrease digital illiteracy.



Introduction: what happened in 2024 in terms of the use of AI?

In 2024, the field of artificial intelligence (AI) saw remarkable advancements, particularly in large language models (LLMs) and their applications, and it is difficult to keep up with the pace of announcements. The progress is so fast that the illustrations and examples of this text will already be outdated when you read it.

The HiPEAC community is not specialized in developing new AI algorithms or new AI applications, but it is deeply involved with artificial intelligence on two sides:

- Leveraging AI for HiPEAC developments, hardware, and software.
- Developing new hardware and software to better serve AI needs.

Using AI to help HiPEAC community to develop better hardware and software is covered in the “Tools” chapter of this document. However, originally prototypes with limited use, LLMs underwent significant development starting in 2022, evolving into viable tools by the end of that year. Models like o1 from OpenAI are now able to generate moderately complex code of several pages. Perhaps we are already entering into what Jensen Huang, the CEO of NVIDIA, calls a future where content will be generated, not retrieved `[HuangHPCwire]`. It is now faster to ask o1 to generate programs that play the Game of Life, Asteroid, or Flappy Bird than to look for a version that works on your computer - and the generated version is likely to be virus-free.

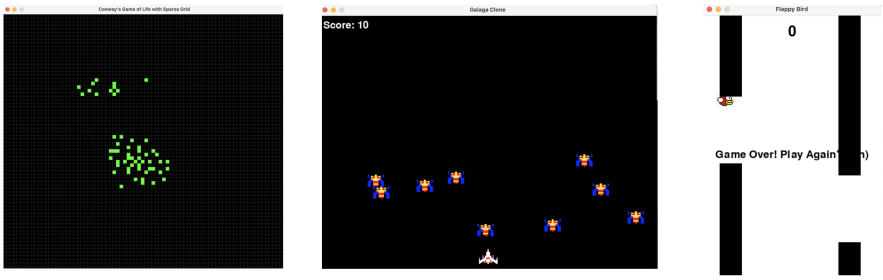
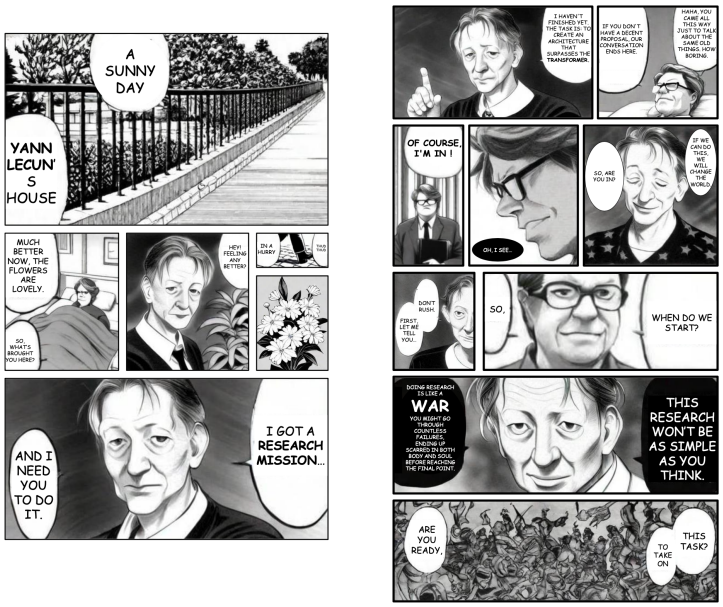


Figure 1: Game of Life, a clone of Asteroid and a Clone of Flappy bird generated in few minutes by OpenAI o1. A few iterations were necessary, but absolutely no reading or understanding the generated code required.

Beyond text (and code), pictures, music (Udio, Suno can generated songs), manga `[NobelManga][NobelMangaPDF]`, and video generation all saw major improvements in 2024, reaching a point where coherent and realistic videos of several seconds can now be created. Among the leaders in this field, Google's VO2 currently stands out as the most advanced, followed by OpenAI's Sora. Unlike large language models (LLMs), where open-source models often rival their closed-source counterparts, video generation remains exclusively dominated by closed models for now.



Manga generated by AI `[NobelManga]`

Most of the major (big) LLMs are now covering multiple modalities (text, but also image, sound) and they can directly have one modality as input and another as output, without relying on intermediate models that transform one modality into another. One example is the advanced voice mode of OpenAI which uses sound (voice) as input and directly generates audio output, therefore with a reduced latency compared to the previous approach which used at least three models (a voice to text model, then the LLM text to text, then a text to voice model). Transformer-based systems seem to be now the "Swiss Army knife" of AI, because they are also efficient for perception tasks like image recognition, sound analysis etc, making them suitable for devices directly interacting with the real world, like robots, self-driving cars, …

Another development this year was the progress in “world models”. These models do not simulate language but instead recreate entire games or physical environments. A notable example is Oasis, an AI-based simulation of Minecraft that replicates the game’s physics, block structures, and movement dynamics. Researchers also developed DIAMOND, a model capable of simulating a basic version of Counter-Strike: Global Offensive, but still with limited fidelity and running at 10 frames per second on a single NVIDIA RTX 3090 graphics card. Meanwhile, DeepMind’s Genie and Genie 2 demonstrated the ability to simulate not just one but multiple games, showcasing the potential for virtual environments.

In gaming, the possibilities for these world models are virtually limitless. Imagine games where every aspect—characters, scenarios, textures, and interactions—is dynamically created by neural networks. This level of generative power could revolutionize game design, enabling entirely AI-driven worlds free from traditional constraints. While still in its early stages, this field promises to redefine how we interact with both AI and digital environments. Its potential extends far beyond leisure. In the future, such simulators could play a crucial role in training AI systems in controlled environments. For instance, a road simulator is already being employed to train autonomous vehicles safely. Jensen Huang’s vision is coming to reality…

Similar progress can be observed in AI tools to help hardware designers: AlphaChip mirrors the principles of AlphaZero, the algorithm used in strategy games, but applies them to the design of computer chips. Developed in 2020, AlphaChip made headlines again in 2024 with its use in designing Google’s tensor processing units (TPUs). This approach showcases a remarkable loop of optimization: an AI system designs a chip, which is then used to train the same AI, enabling it to create even better chips in subsequent iterations. This self-reinforcing cycle highlights how AI can accelerate technological progress in unprecedented ways.

You can refer to the section on tools for more in-depth text about the use of AI to increase productivity of the HiPEAC community.

What improvements in AI took place in 2024?

To better understand the key recommendations for developing optimized hardware and software to serve the requirements of AI, it is necessary to make a brief explanation of what happened in 2024 for the development of AI technology and extrapolate (if possible) the next steps.

The improvements of artificial intelligence in 2024 were not primarily driven by increasing the size of these models, but by refining the quality of training data, techniques like fine-tuning, and using more compute time during inference. However, economic viability is still an open question, leading to an increase in the cost of subscriptions (for OpenAI) and perhaps limited access to the most powerful models, that will only be used to answer very specific questions that could compensate for the cost of running the model. Will we perhaps see the beginning of AI at multiple speeds: “basic” low-cost AI accessible to everyone, higher-performing AI on subscription for those who can afford it, and countries or big companies that are the only ones that can afford to access the best models and to ask them complex (therefore expensive in term of compute power and therefore cost) questions?

A focus area for improving LLMs involves the quality and quantity of training data. One practice in 2024 was the use of synthetic text generated by pre-trained models, offering a rich and scalable source of high-quality data. This shift has enabled the creation of smaller models without compromising performance, driving down costs dramatically. For example, while GPT-3.5 contained 175 billion parameters, Google’s Gemma 2 models now deliver similar performance with 9B parameters, representing nearly a 20-fold reduction in size. Meta’s Llama 3.3 (70B) of December 2024 has the same performance as Llama 3.1 (405B) of July 2024. This efficiency, combined with hardware optimizations, has led to a tenfold annual decrease in operational costs since 2022. There is a clear economic incentive to use

smaller LLMs that have similar performance as bigger ones, because the inference cost (computation) is lower. Some observers saw some decrease in performance between versions of ChatGPT 4o, perhaps due to a switch to a smaller LLM. It is also possible that the “big” LLMs will not be accessible to the public, but only internally used by the companies to train smaller, more economically viable LLMs.

In addition, advancements in LLMs included an increase in the contextual scope, with models now capable of handling up to 128,000 tokens per input. Google even expanded this limit to two million tokens, enabling the processing of extensive collections of documents in one go. At the core of current LLMs lies the transformer architecture, which, while powerful, suffers from a growing memory requirement as it processes longer texts. Newer architectures like Mamba, with constant memory usage, offer a promising alternative by enabling faster processing of extended word sequences. In 2024, it became clear that completely replacing transformers is not feasible yet. However, hybrid approaches that integrate Mamba with transformers are showing potential, maintaining high performance while reducing memory overhead.

A novel approach emerged in 2024, in which LLMs spend more computation time during inference to improve output quality. Traditional models produce responses with a fixed computation effort regardless of task complexity, but newer models like OpenAI’s o3 dynamically adjust their computation time. This process allows for more thoughtful and accurate responses, as these models essentially “reflect” or “research” internally before presenting their output. It seems that this also required an increase of the contextual scope referred to in the previous paragraph.

OpenAI, leveraging its reinforcement learning expertise, led the way in this approach, with Google and other competitors following with models like Gemini 2 Flash Thinking. This new approach is similar to the breakthrough in 2016 when AlphaGo transformed the landscape of the game Go. Initially trained to imitate the moves of expert human players, AlphaGo was limited by the quality and scope of its training data. However, when allowed to play Go independently without human intervention, it began to learn through trial and error, guided only by rewards within the game. This self-directed learning led AlphaGo to outperform human champions, fundamentally changing how the game was played.

Previously trained to replicate human-written text, these new models, like o1 or o3, have now begun to autonomously refine their reasoning abilities. Google’s Deep Research feature complex problem-solving by enabling LLMs to analyse data from over 50 online sources, including PDFs, in mere seconds, to provide comprehensive summaries.

By exploring, experimenting, and searching for solutions independently, LLMs are no longer constrained to mimicking human data. Instead, they are evolving their strategies for solving problems. For instance, OpenAI’s o3 model excels in mathematical problem-solving, achieving a 97% success rate in the AIME (American Invitational Mathematics Examination) competition, a major step forward from earlier performances of a few percent. The o3 model also excels in benchmarks like Frontier Maths and ARC-AGI-PUB, reaching human-equivalent scores. Similar progress was observed in medicine, physics, and coding benchmarks.

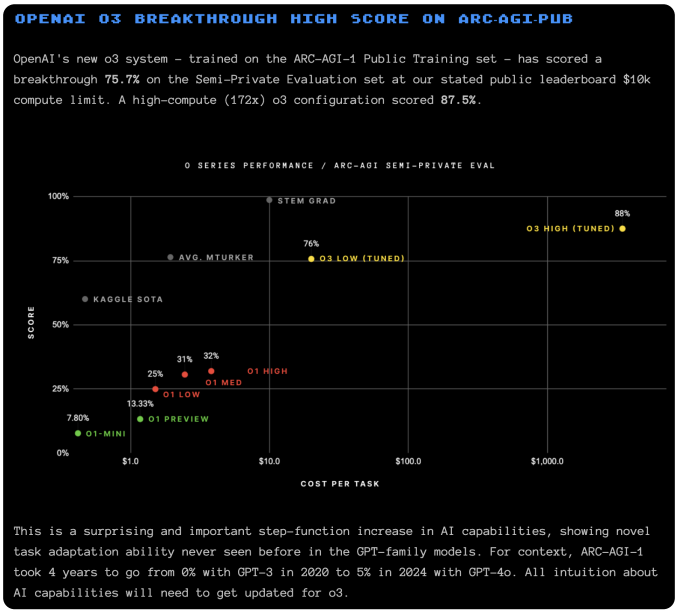


Figure 2: Blog from François Chollet about o3 and the ARC-AGI-PUB benchmark [ARCPRIZE]

This increased computation demand during inference has significant implications for the hardware market, particularly benefiting GPU providers like NVIDIA. Solving a single problem on benchmarks like ARC with o3 can cost thousands of dollars in computation, necessitating infrastructure investments like OpenAI's $200 monthly ChatGPT Pro subscriptions.

The competitive landscape of AI also shifted significantly in 2024. While OpenAI maintained a lead in reasoning-based models like o1 and o3, other players like Google, Anthropic, Meta, xAI, and even Chinese companies such as DeepSeek and Alibaba made landmarks in LLM development. Google with Gemini 2 Flash Thinking, but also the Chinese DeepSeek et Alibaba, with their models DeepSeek R1 and QwQ (quill), also propose models that can use variable inference compute time to produce answers.

The open-source community also gained ground, with Meta's Llama 3 models and Alibaba's DeepSeek V3 rivalling closed models like GPT-4o. Hardware constraints became central, with NVIDIA's GPUs remaining indispensable for model training, for example, xAI's supercomputer now having 100,000 NVIDIA GPUs. All major companies are developing their own accelerator chips (AWS Inferentia for inference servers – they also develop Trainium chip; Meta with its Next GenMTIA), although Google's TPUs still have a competitive edge because they are already on their sixth generation with the Trillium chip. A significant hardware race unfolded; the hardware demand even triggered discussions about energy requirements, potentially leading to the construction of dedicated nuclear power plants.

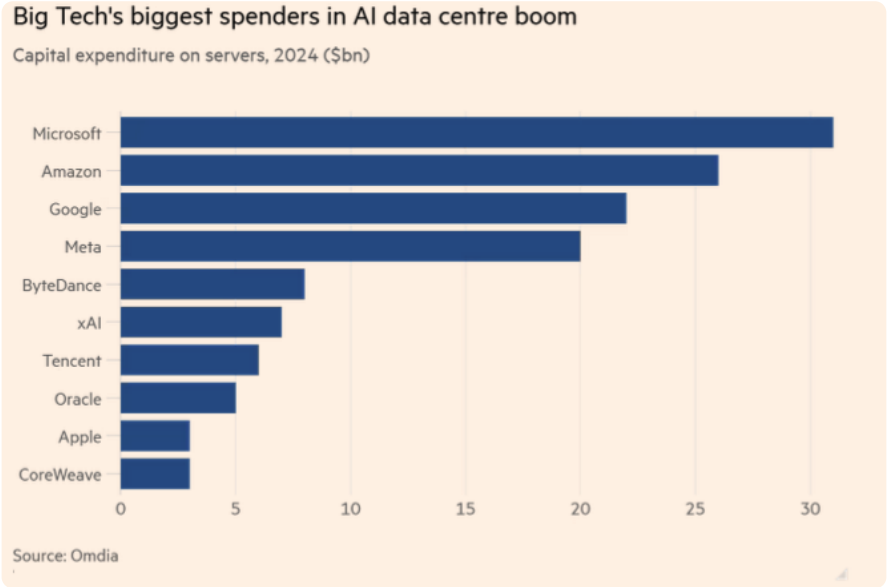


Figure 3: spending in AI servers in 2024, data originally from Omdia

In 2024, the landscape of large language models (LLMs) witnessed not only technical advancements but also greater accessibility for everyday users. Apple introduced Apple Intelligence, a feature integrating LLMs across all Apple devices. This innovation allows users to interact seamlessly with AI, even enabling direct access to ChatGPT. Apple is the first to propose a kind of "distributed" approach: first, the local LLMs are used by an orchestrator; if they are not powerful enough, the demand is seamlessly transferred to Apple servers, and even to ChatGPT.

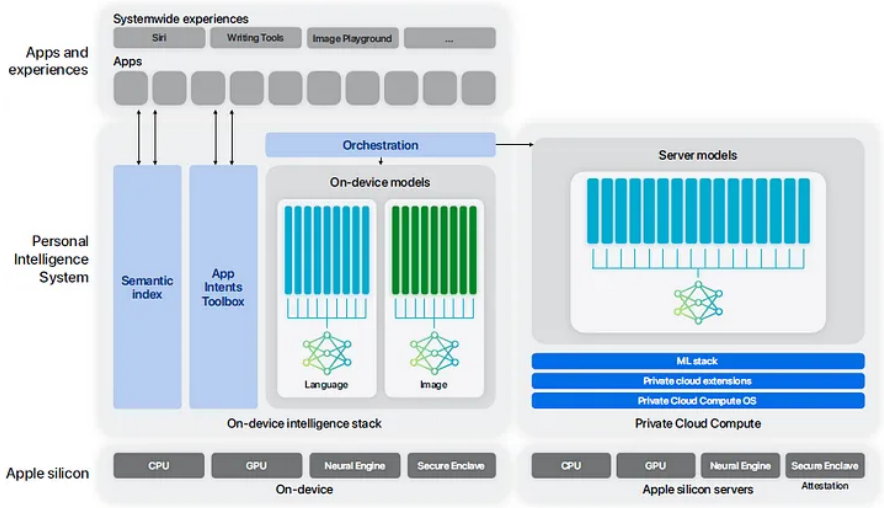


Figure 4: Architecture of Apple Intelligence with adapters, highlighted as blue and green rectangles, for the on-device and server language/image models, from https://medium.com/byte-sized-ai/on-device-ai-apple-intelligence-533c4c6ed7d6

This year also witnessed an interesting development in the training of large language models (LLMs): distributed training across the globe. Traditionally, LLMs have been trained within the confines of a single data centre, where GPUs are interconnected to manage the computational load. However, by the end of the year, two companies, Prime Intellect and

NousResearch, pushed the boundaries of this approach by training models with 10 billion and 15 billion parameters, respectively, using a distributed network of computers located in Europe, Asia, and the United States.

This innovation marks a significant shift in how LLMs can be developed, presenting opportunities for more flexible and scalable training processes. By spreading the workload across multiple regions, this method could lower barriers for smaller organizations, enabling them to pool resources and collaborate on creating advanced models. This distributed training approach holds immense potential for democratizing access to cutting-edge AI capabilities while fostering innovation on a global scale.

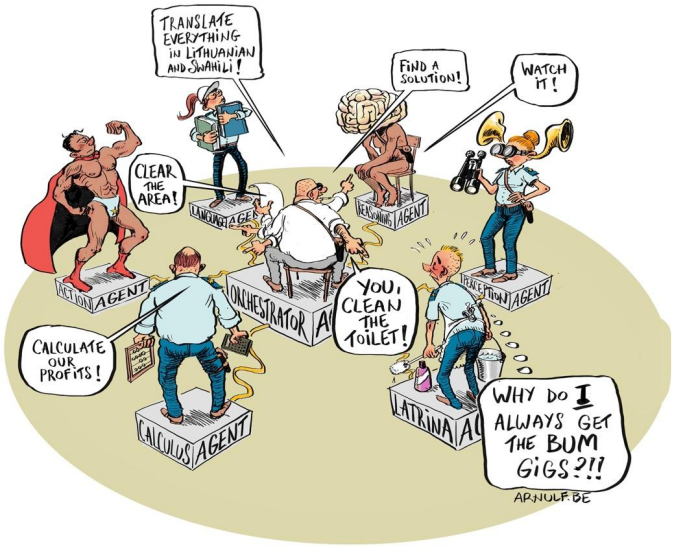
Figure 5: INTELLECT-1 Release: The First Globally Trained 10B Parameter Model [INTELLECT-1]

Recommendations and actions from observing 2024 evolutions

It is clear that trends seen in 2024 will continue in the future, perhaps with new improvements, but more computing power seems to be the key enabler of artificial intelligence, with its corollary of increased energy consumption. Therefore, **making innovative new hardware for supporting LLMs**is part of this HiPEAC Vision; see the “New Hardware” chapter.

However, from the (distributed) structure of Apple Intelligence, distributed training across the globe, and the new models like o1, we can derive recommendations that will help Europe to re-enter the game. In the summary above of major developments in 2024, only US and Chinese companies or organizations were cited; unfortunately, none of those cited were from Europe.

As explained in the foreword and introduction, the ideas behind the NCP can be instantiated in the short term as “distributed agentic AI”. The structure of Apple intelligence, of distributed training across the globe, are clear precursors, but also the possible technology behind models like o1 (see for example [Zeng24]) might possibly be done by several specialized agents working together.



There is **clear research (and business) interest in looking into a set of smaller specialized agents (LLMs) working (orchestrated) together**. If the agents are distributed in different locations– as was the case for [SETI@Home], `[BOINC]` and `[Petals]` – and if the compute resources are shared, as proposed in the NCP concept, then perhaps a gigantic data centre that consumes MW of electricity is not a requirement for advances in AI, or to run existing AI for users. This opens up contributions from a much larger base than the few companies that can afford gigantic data centres. And this distributed AI from edge to data centres can adapt to the user's requests, being exclusively local for simple requests and not activating a large LLM on a distant data centre, with its associated cost in terms of energy.

The recommendation is therefore the **development of specialized action models (SAMs)** – that is, small and specialized models that can interact with their environment, acting as service. These SAMs should operate in a distributed infrastructure and an ecosystem should be created to support research, development and business around them. Of course, this ecosystem should be a precursor and compatible with the one provided by the NCP. These models need to be refined, optimized, and reduced in size to improve efficiency.

This also ties into the need for **hybrid systems** (combining AI and algorithmic approaches), as noted in the foreword to this vision: future systems that must integrate both paradigms—precise and approximate—within feedback and reinforcement-based architectures. These SAMs can be optimized from more general foundation models by an ecosystem of companies providing their optimized SAMs in a marketplace so that they can be dynamically discovered and used by the orchestrators.

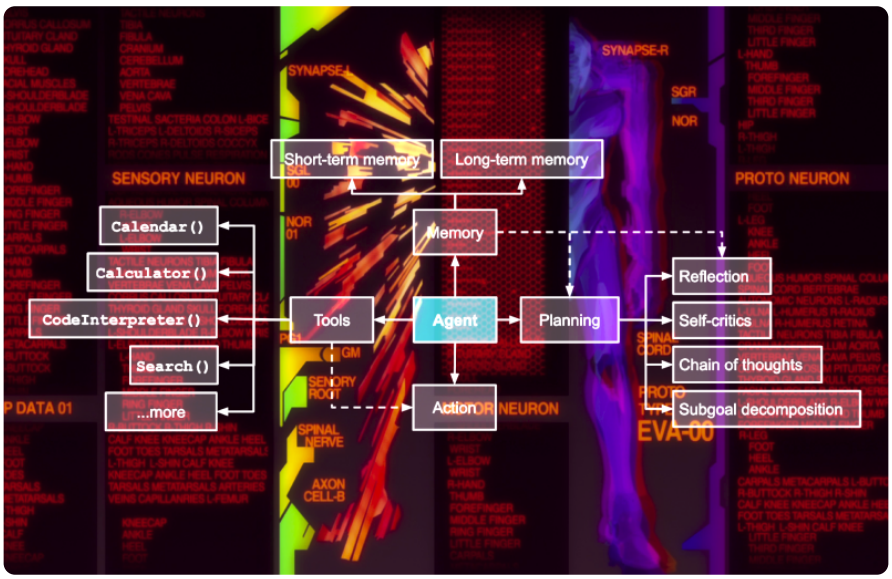


Figure 6: developing agents models with reasoning, from https://www.primeintellect.ai/

Two other ingredients are necessary for the system beside the developments of those SAMs:

- A way to discover, select, active, organize them together, therefore the **development of orchestration technologies** that take into account all the requirements, that can select the best SAMs for the required tasks and that can dynamically activate them. The first steps could be very agentic-AI-centric (relying on already existing technologies used for orchestrating AI agents), but they should be a blueprint and evolve towards an orchestration system for the NCP. These orchestrators must be developed for the edge – or near the final user – and dynamically combine SAMs into executing personalized applications in response to user needs.
- This will be only possible if all the systems "speak the same language", therefore, a key recommendation is that **it is imperative to establish open protocols for these "distributed agentic AI" systems to facilitate seamless interaction among distributed AIs from different origins.**

To effectively operate this federation of distributed AIs, it is necessary for them to exchange data and parameters through a universally comprehensible protocol that:

1. Does not solely rely on functional requirements (e.g. the textual representation of prompts and responses).
2. Also incorporates non-functional requirements (providing sufficient information for the orchestrator to select the appropriate services, such as based on criteria like response time, potential level of hallucinations, environmental impact, cost, localization, privacy of data, etc.).

It is therefore important that the community works together to commonly define this exchange protocol that should be open to allow broad acceptance. Large entities such as OpenAI, Meta, and Microsoft are attempting to promote their own application programming interfaces (APIs) for accessing their models. However, an API alone is insufficient for constructing this distributed and federated network of AIs. Also, the exchange format (JSON, ASCII text) is perhaps not the optimal way for networks of AIs to efficiently exchange information: this could be tokens, embeddings, or any other representations; some research also shows that LLMs talking to each other could develop their own "language".

In the fields of distributed agentic AI, some work has already done, for example [DAWN] and [DistMixofAgents]. But it is important to group all existing ideas and existing proposals together in a single consortium to develop an open source “de facto” (before official standardization) standard that takes into account all the good ideas of various researchers and organizations, so that it will be sound, future-proof, recognized and accepted.

In a similar way to TCP-IP that enabled various OS (operating systems) to communicate, the aim of this action is to create the equivalent for OS (orchestration systems) to exchange AI-related information.

Time is crucial for this initiative, and standardization, however necessary, will be too long, so a de facto open standard should be proposed in parallel with the standardization effort, before other closed proposals will emerge, locking down the approach to a few (non-European) players. This should also act as a blueprint for the NCP protocols and specifications. Like for the NCP, this approach will allow the creation of a completely new ecosystem where smaller players can provide specialized AI as a service along with the big ones. Directories of services, trusted brokers, and payment services are also important elements that can emerge from this ecosystem, where Europe can have an active part thanks to its set of SMEs, research organizations, and distributed nature.

Europe should be an active player in the race for the “distributed agentic artificial intelligence”.

References

ARCPRIZE: https://arcprize.org/blog/oai-o3-pub-breakthrough

BOINC: https://boinc.berkeley.edu/

DAWN: Aminiranjbar, Zahra et al. “DAWN: Designing Distributed Agents in a Worldwide Network from Cisco Systems”, 2024 https://arxiv.org/pdf/2410.22339

DistMixofAgents: Mitra, Purbesh, Kaswan, Priyanka and Ulukus, Sennur. “Distributed Mixture-of-Agents for Edge Inference with Large Language Models”, 2024 https://arxiv.org/abs/2412.21200

HuangHPCwire: “You know that in the future, the vast majority of content will not be retrieved, and the reason for that is because it was pre-recorded by somebody who doesn’t understand the context, which is the reason why we had to retrieve so much content,” he said. “If you can be working with an AI that understand the context – who you are, for what reason you’re requesting this information– and produces the information for you, just the way you like it, the amount of energy you save, the amount of network and bandwidth you save, the waste of time you save, will be tremendous.”

INTELLECT-1: https://www.primeintellect.ai/blog/intellect-1-release

NobelManga: https://jianzongwu.github.io/projects/diffsensei/

NobelMangaPDF: https://jianzongwu.github.io/projects/diffsensei/static/pdfs/nobel_prize.pdf

Petals: https://github.com/bigscience-workshop/petals#readme

SETI@home: https://setiathome.berkeley.edu/

Zeng24: Zeng, Zhiyuan et al. “Scaling of Search and Learning: A Roadmap to Reproduce o1 from Reinforcement Learning Perspective”, 2024 https://arxiv.org/pdf/2412.14135

New Hardware

Recommendations for New Hardware

Specialized hardware (HW)

The development of efficient hardware is essential for running services, orchestrators and SAMs efficiently at the edge and within federated networks. Europe must address memory costs (for AI), energy consumption, and ecological impact, potentially leveraging non-volatile memory for direct edge execution. Additionally, the next generation of SAMs should incorporate learning through experiences or allow to the efficient execution of digital twins to maintain Europe's competitive edge in AI (embedded AI). In the field of AI accelerators, the focus should be on inference (becoming more and more important with the approach pioneered by OpenAI o1 and o3) or on fine tuning. Reducing the transfer of data is key to reach lower levels of power consumption. This can be achieved with near- or in-memory computing (NMC or IMC), direct execution from the storage of parameters (hence eliminating the need for RAM), etc…

Beyond purely digital hardware (HW)

Investigation of new **accelerators using non digital technologies**, going from exact computations (digital computation) to more approximate computing (neural networks are universal approximators, quantum computing results are stochastics, optimization techniques using Bayesian, Ising approaches can solve complex problems) should be also investigated in the context of providing more efficient services to the next computing paradigm (NCP) ecosystem.

Introduction

The changes in the hardware arena since the publications of the HiPEAC Vision of 2023 and 2024 may be minor, but there have certainly been developments. Artificial intelligence (AI) accelerator hardware is dominating the profits of the top hardware manufacturers. Tensor processing units (TPUs) are increasingly augmented by general-purpose graphics processing units (GPGPUs), but manufacturers of both processor architectures are based outside Europe.

The influence of AI is being felt in all aspects of the economy, including the technology sector, where it supports design and implementation of both hardware and software. AI is also spreading towards the edge of the continuum, making smarter applications possible in the home, as well as smart equipment in the field. We expect to see this growth towards the edge increasing and finding its way into as yet underexplored applications in the coming years. This creates challenges and opportunities for European players [SemiWiki-NPU-2024].

Training AI is an important part of its application; inference is another important one. A market is opening for efficient inference engines, tailored to specific needs of the place in the computing continuum where it takes place. AI applications for specific domains do not require a general-purpose AI application, but one tuned to the needs of the domain with tuned hardware. Of course, this will increase the diversity of devices for AI, but AI itself can be used to design these application specific AI devices.

All seems quiet on the quantum computer front. But even though investments in quantum computer start-ups levelled out in 2024 [EETimes-March] the field is still progressing. An often-made remark is that the race for quantum computers is a marathon, not a sprint; it will take endurance and time to achieve quantum computers that can demonstrate their superiority over classical computers in specific practical compute challenges.

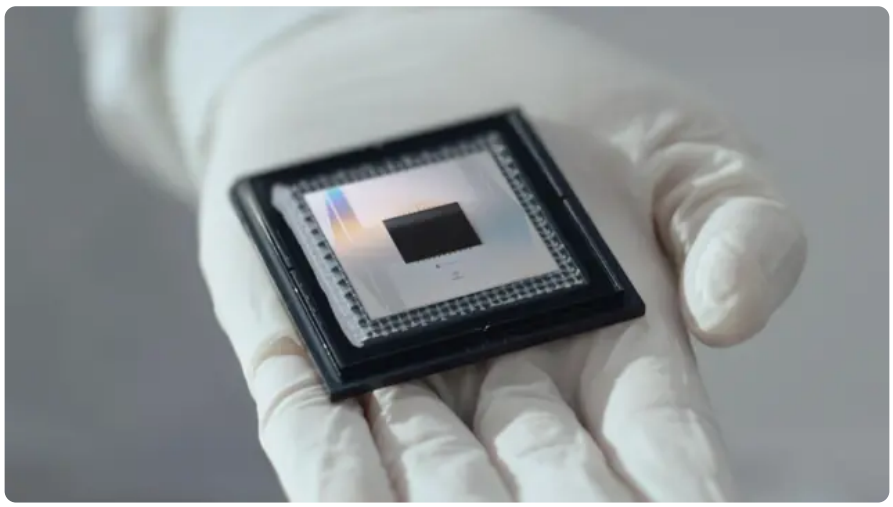


Figure 1: Google's 105 qubit Willow Chip which achieved exponential error reduction, produced in 2024. [BBC-Willow]

Towards the end of 2024, Google announced its Willow chip [BBC-Willow][Google-Willow], a chip which stands out not so much for its number of qubits (105) as for demonstrating that exponential error reduction is possible with a linearly increasing number of logical qubits to form a logical qubit. See Figure 1.

During the last few years, simulating quantum systems, e.g. for drug discovery, was touted as one of the most important practical applications of quantum computing for the near future. However, AI-based applications are now challenging that role as they are already showing great acceleration of that task on classical computers (see [Labiotech] and [Acellera]). However, this does not mean that quantum computing is becoming obsolete before it reaches practical maturity, because drug discovery is just one application. It should not lead to quantum computing disillusion, as we are still in the process of finding out for which problems quantum computing is a winner over classical computing (cracking encryption keys being one).

Europe is strong in quantum compute expertise, but when the technology moves from the lab to the market Europe must be able to commercialize this expertise. Geopolitical changes over the last decade also make it necessary for Europe to strive for technical sovereignty now more than ten years ago, and quantum computing is definitely one of the topics relevant for sovereignty. Europe should intensify its intra-European cooperation, guarding against techno-nationalism. Setting up European technology centres that stimulate European independence from other global factions might function as a glue between countries.

Although outside HiPEAC's field, it should be noted that Europe's presence in semiconductor manufacturing equipment is quite strong. ASML stands out, but other critical equipment, such as that used for deposition and measurement, or metrology, also comes predominantly from European vendors. In addition, the percentage of capital investment for manufacturing equipment by chip manufacturers has risen in recent years from around 10% to approximately 20% [EETimes-November].

Efficient hardware: Continue the quest for lower power and improved performance

The growth of AI towards the edge of the continuum creates opportunities for Europe to become a player in AI. AI requires a lot of processing, and that requires energy. As energy is scarcer at the edge, low-power architectures with just the right amount of processing power

are required. This is an emerging development, and Europe has an opportunity to build a strong presence in this technology. It can do that also by initiating standards initiatives, leading the way to fast integration and introduction of edge AI technology. Europe has a strong position in embedded chips, which it should strive to retain. Designers of embedded systems are used to dealing with constraints, a skillset that European companies can leverage in designing edge AI processing components.

Strengthen European sovereignty in hardware technology and manufacturing

With the European Chips Joint Undertaking as a major and very important step to increase independence, Europe must also ensure that it covers all the important steps in the manufacturing chain, not only chip manufacturing. Europe's presence in the supply of manufacturing equipment is already strong. But it still depends on non-European suppliers for critical raw material for chip production. It is very probably not possible to fully eliminate this dependency, e.g. because of the geological location of such supplies, but Europe should strive to minimise such dependencies.

Another aspect is sovereignty in key technologies. Europe's presence in e.g. photonics production is not very high, even though the level of research in Europe of this technology is high. Photonics is key for data communication. If Europe misses out on such key technologies, its sovereignty in ICT and its applications such as AI is threatened. (The Chips Joint Undertaking has a paragraph on strengthening integrated photonics production.)

New accelerators: Prepare for the integration and hetero integration of new hardware technology

Hetero integration refers to the integration of different types of materials, devices, or technologies into a single system or chip. By leveraging the strengths of diverse materials and technologies, hetero integration is a key technology paving the way for more advanced, efficient, and versatile computing systems that would be critical in the context of the NCP.

Although practical applications of quantum hardware seem to be ten years away, this is a field that Europe should not leave to others. The technology has the potential to become key in society and in the economy. Private investments in this technology are levelling out, possibly because the technology is growing out of the startup environment. It requires scaling up to an industrial application level as the next step, which is beyond the capabilities of private investors and requires existing (European) companies to step in. In this respect it has also been noted that more (European) company research should be directed to this technology to prepare for commercial application `[EETimes-March]`.

Quantum technology expertise is spread over Europe, a situation which is currently supported by the way projects are financed. It should be investigated whether a stronger concentration of efforts in the development of this technology would be beneficial for its progress. Strategic alliances with European non-EU-based research groups and companies should be investigated.

Encourage modularity and standardization at the hardware level

For better design and optimization of hardware/software, support of complex architectures including modular AI to make scalability relatively easy is key.

Modular design of digital systems should be encouraged. With modularity comes the need for standardization, not just at the hardware level (connectors, metrics, etc.), but above all at the level of software stacks and their interactions with the various hardware levels. This need, traditionally expressed in the rather late stages of digital systems integration, must be considered right from the design stage of the hardware and software components that make them up. This is the case, for example, with the hetero integration of technologies mentioned in the previous paragraph. Not only must the various chips be designed to interconnect physically with each other, but also to operate logically as a functional whole.

The integration of AI at all levels of the NCP concept also illustrates this need for modularity and standardization. Ensuring that an AI model integrated into the system, with its requirements on data and hardware resources, can perform the expected function depends on its ability to seamlessly interface with it.

Support a European ecosystem that encourages a strong link between information sciences and basic research on emerging hardware, including quantum computing

European universities, research centres and companies must be at the forefront of basic research in information sciences: information encoding, programming paradigms and computing complexity to cite a few. It is important that research on new hardware for computing devices – such as spintronics and photonics – are linked to progress in information science in order for cross-fertilization between those disciplines to occur. For example, a computing concept like the “Ising spins machine” might appear as a good idea from a hardware elaboration point of view but might require complete new knowledge from an information or programming perspective. Research into new hardware devices for computing must therefore be carried out in close collaboration with advances in information science and algorithms.

Multichip

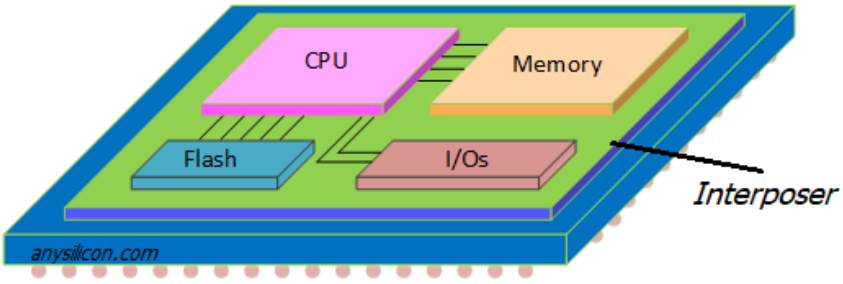


Figure 2: Schematic diagram of 3D-interposer technology (https://anysilicon.com/wp-content/uploads/2019/08/Interposer.png)

As the complexity of digital systems keeps growing, it becomes harder and harder to integrate the whole system on one chip. Moreover, some accelerators may be fabricated using fabrication technologies that are incompatible with standard digital ones.

Integrating multichip systems in one package is increasingly done using 2.5D and 3D interposers. Europe has a strong research capability in this technology. CEA-List in France is researching active interposers, meaning that the interposers themselves perform part of the active functions of the integrated system, e.g. through a network-on-chip (or better: network-on-interposer), or implementing analogue functions.

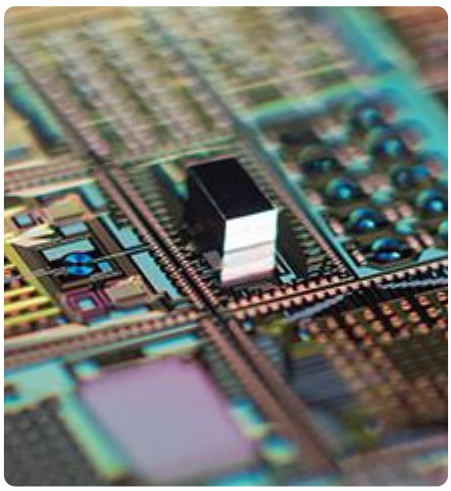


Figure 3: 3D active interposer technology from CEA [CEA-OpticalInterposers]

The Fraunhofer Gesellschaft has founded a Chiplet Center of Excellence in Dresden in 2019, researching this technology with an eye on applications in the automotive industry [Fraunhofer-IZM]. Imec in Leuven launched (in Ann Arbor in the USA) its automotive chiplet programme in October 2024. Imec is also active in standardizing chip interposer technology, paving the way to wider use of this technology. European companies active in chiplet technology include Bosch, and Quintarius [Quintarius], founded by a number of, mostly European companies. Non-European companies such as Singapore-based Silicon Box have plans to invest in chip factories in Europe [SiliconBox].

These examples show Europe's strong presence in this technology, which it should strive to keep.

Conclusion

The future of digital hardware technologies goes beyond the development of new physical components, new chip technologies or new materials. In a context of increasingly complex systems, with the arrival of AI at all levels and the need for sustainable/eco-responsible digital technology, it is a finer coupling between hardware and software technologies that is the key to new digital technologies. In this sense, the case of NCP is eloquent and illustrates our vision of future hardware technologies.

References

Acellera: https://www.acellera.com/

BBC-Willow: https://www.bbc.com/news/articles/c791ng0zvl3o

CEA-OpticalInterposers: https://www.leti-cea.com/cea-tech/leti/english/Pages/Industrial-Innovation/Demos/3D-Integration-HPC-AI.aspx

EETimes-March: https://www.eetimes.eu/ee-times-europe-magazine-march-2024/

EETimes-November: https://www.eetimes.eu/ee-times-europe-magazine-november-2024/

EETimes-September: https://www.eetimes.eu/ee-times-europe-magazine-september-2024/

Fraunhofer-IZM: https://blog.izm.fraunhofer.de/the-chiplet-center-of-excellence/

Google-Willow: https://blog.google/technology/research/google-willow-quantum-chip/

IMEC: https://www.imec.be/nl/press/internationale-auto-industrie-klopt-aan-bij-imec-voor-nieuw-type-microchips

Labiotech: https://www.labiotech.eu/best-biotech/ai-drug-discovery-europe/

Quintarius: https://en.wikipedia.org/wiki/Quintauris

SemiWiki-NPU-2024: https://semiwiki.com/artificial-intelligence/349906-get-ready-for-a-shakeout-in-edge-npus/

SiliconBox: https://www.reuters.com/technology/silicon-box-picks-piedmont-region-its-italian-34-bln-chip-plant-2024-06-28

Tools

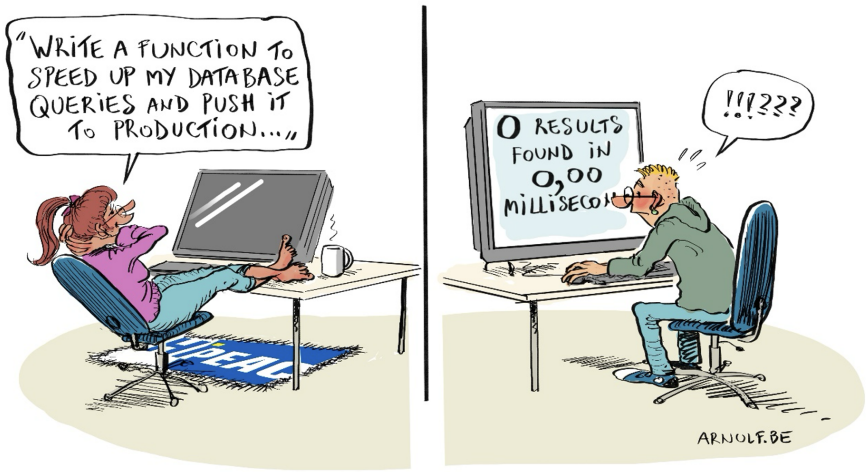
Recommendations for tools

Promote the use of AI in software development

Research, prototype and deploy AI-assisted software development environments, while implementing robust measures to ensure correctness, safety, security, confidentiality, and regulatory compliance. This will help balance the rapid adoption of AI with the need for secure and reliable systems. It should also help non specialists to be able to create efficient software and increase the productivity of developers.

Promote the use of AI in hardware development

Research, prototype and deploy open AI assistants for hardware development, increasing the productivity for designing new, efficient hardware and decreasing the time to market. This is a key element for Europe to stay in the hardware race. The use of AI should be a collaboration between humans and AI systems, as promoted in previous HiPEAC vision as 'centaur' teams. The focus should be on domains that are still open, like architecture search and exploration, rather than on optimizing the floor-planning, which is already covered by various companies.



Introduction

The history of hardware and software development has been a progression toward specifying more what is to be done and less how it is to be done – a move from implementation detail towards higher abstraction. The primary factors that have fuelled that

progression have been improvements in processor speed and compiler optimization. It appears that the next technical advance which can drive a new discontinuity in this progression is AI-based hardware and software development tools.

These tools are revolutionary due to their ability to create hardware or software from a natural language description without any human intervention. If the dream to automate the entire development process becomes reality, it would democratize software, allowing anyone to create state-of-the-art software, while potentially eliminating many hardware and software developer jobs. While a similar impact may be seen across many sectors, AI-based tools for hardware and software are distinct from AI in other contexts, due to the strict need for correctness and security, the complexity of integrated co-design hardware/software systems and the limited training data for hardware design and new technologies.

AI-based tools have the potential to disrupt software and hardware development, and missing out on this discontinuity could leave Europe hopelessly behind. The NCP takes for granted the ability for the user to orchestrate and create new software capabilities that would have traditionally required custom software development. These AI tools also streamline the development process, reducing the time and cost to develop the NCP itself. For Europe to successfully lead the NCP, it must have access to the latest technologies, which depends on its universities, research centres and companies being up to date with the forefront of advances in the AI revolution.

State of the art

According to the 2024 Stack Overflow Developer Survey [StackOverflow2024], 62% of software developers were already using AI tools, with an additional 14% planning to adopt them soon. As of the time of writing, these tools can generate functional code from a natural language description, spot likely errors (off-by-one errors or usual code patterns), suggest and apply refactoring, estimate computational complexity, and so on. They leverage vast training data and an understanding of patterns, semantics and context, and are much more powerful and tolerant of ambiguity than earlier tools such as syntax-directed parsers. They can help modernize code to a new environment (e.g. language, major API revision, certified OS), and also generate documentation and test cases, helping maintainability and team onboarding. For a more detailed survey of these capabilities, see [Metzger] [KordonZaourar].

GitHub Copilot [GitHubCopilot] is a state-of-the art AI-powered coding assistant, which integrates OpenAI's Codex model [OpenAICodex] into Microsoft's development environments such as Visual Studio Code and GitHub. It provides developers with real-time code suggestions, completions, and contextual guidance, across a wide range of programming languages and frameworks, streamlining tasks from boilerplate generation to debugging. By analysing surrounding code and comments, it predicts and generates relevant code snippets, enabling faster development and reducing repetitive tasks. Microsoft has positioned Copilot as not just a tool for writing code but as an intelligent collaborator that enhances productivity, encourages best practices, and lowers the barrier to entry for complex programming tasks. As such, it is a complement to their broader Microsoft Copilot AI companion [MicrosoftCopilot], which is integrated across multiple Microsoft products, including Word, Excel, PowerPoint, Outlook and Teams.

Google's Gemini Code Assist [GoogleCodeAssist] is another prominent AI coding assistant, which integrates with integrated development environments (IDEs) such as Visual Studio Code, JetBrains IDE and others, supporting major languages such as Java, JavaScript, Python, C and C++. In December 2024, Google announced Gemini 2.0, which includes Jules, a more powerful but experimental AI-powered coding agent for Python and Javascript, which integrates with developers' GitHub workflows, handling code development such as bug fixes, and preparing pull requests to land fixes directly back into GitHub [GoogleJules].

Meta's Code Llama [CodeLlama] was released in 2024 as an extension of their Llama 2 language model that is fine-tuned for software development across multiple programming languages. Unlike most of the alternatives, the model weights and inference source-code for Code Llama are freely available under Meta's strategy of fostering open innovation in the AI ecosystem [CodeLlamaLicence]. As such, it offers the possibility for fine-tuning and customization.

Overall, AI-driven coding assistants have amassed nearly $1 billion of funding since the start of 2023, with the vast majority, such as Microsoft's Copilot, Google's Gemini, and tools from startups such as Replit and Magic, being controlled by US-based companies [FTAug2024].

Mistral AI, a Paris-based startup, is a notable European success that has made significant strides in the AI ecosystem, releasing several AI language models and raising substantial funding. Their Codestral model [Codestral] has been specialized for code development, and it targets 80+ programming languages, including Python, Java, C, C++, Javascript and Bash. With its context window of 32K tokens, Codestral outperforms other models in RepoBench, a benchmark for code generation.

In the high-performance computing (HPC) space, the LLM4HPC project at Oak Ridge National Laboratory has developed a number of tools for HPC software development [LLM4HPC]. This includes ChatBLAS, an AI-generated BLAS (Basic Linear Algebra Subprograms) library for linear algebra, automatic parallelization with large language models (LLMs), F2XLLM for Fortran modernization, and is developing ChatHPC, an AI assistant for HPC programmers.

There are also several efforts to develop AI-based tools and platforms to assist with hardware design, although most are either proprietary or not widely available (see [ALSAQER] for a recent survey). ChipNeMo [ChipNeMo] is an LLM developed by NVIDIA, specifically tailored for the semiconductor industry. By employing domain adaptation techniques—such as custom tokenization, domain-specific pretraining, and supervised fine-tuning—ChipNeMo enhances performance in chip design tasks. It excels in applications like engineering assistant chatbots, electronic design automation (EDA) script generation, and bug summarization and analysis, often surpassing general-purpose models. As of now, ChipNeMo is not publicly available. NVIDIA has detailed its development and capabilities in research publications, but the model itself remains proprietary and is not accessible for public use.

Other important activities include ChipGPT [ChipGPT], which generates and optimizes Verilog code from a natural language specification. ChatEDA [ChatEDA] is an AI-based assistant that helps engineers orchestrate a complex EDA workflow using natural language. Additionally, LLMs have been employed to assist in the writing of architecture specifications (e.g. SpecLLM [SpecLLM]) and to explain error messages from synthesis tools [Qiu24].

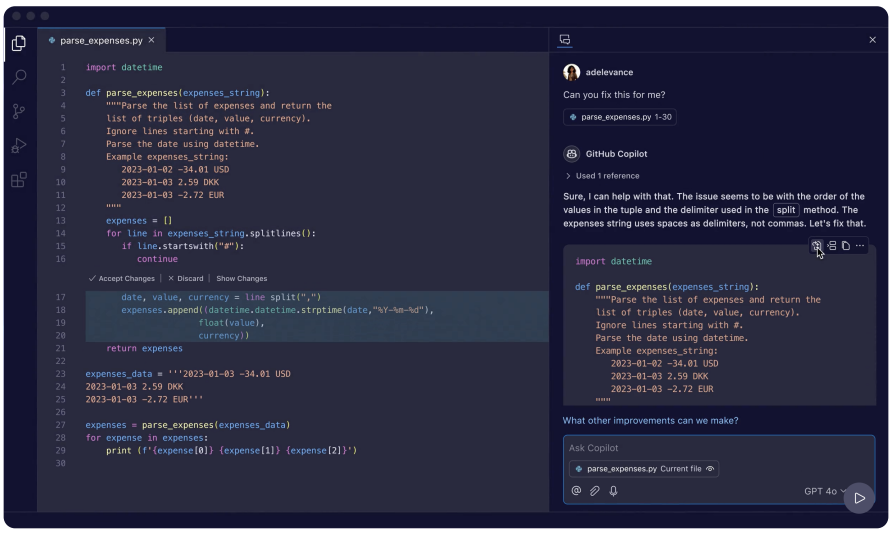


Figure 1: GitHub Copilot, a state-of-the-art AI-powered coding assistant [GitHubCopilot]

Explore the use of AI-based tools to support software and hardware development, but insist on measures to ensure correctness, safety, security, confidentiality and compliance.

The biggest barrier and risk associated with generative AI in development is the inability to fully trust the code it produces. Like people, these tools are prone to confabulation (or “hallucination”), generating incorrect or misleading outputs. Given the opaque nature of AI-based models and likely lack of access to their training data, AI-generated code, whether hardware or software, must be seen as unsafe and insecure. This poses a significant challenge in the context of the NCP, where AI-generated software would be expected to interact with the internet and influence the real world, all without human oversight. Another risk is confidentiality, particularly with online tools provided by third parties, as well as shared tools within an organization that has access to confidential third-party software.

As discussed below, Europe should invest in basic research (including formal methods) and enact sovereignty measures to address these risks. In the meantime, however, development processes must be set up for careful human reviews of AI-generated code, just like code fully written by a human programmer.

Use a combination of LLMs and traditional tools, with the LLM as the user interface and driver to orchestrate

It is likely that all aspects of code development, debugging, optimization and maintenance will shift to using a natural language as the bridge between the human and machine.

The more interesting question is where to place the interface between the LLM and lower levels of the hardware–software stack. Programming languages have traditionally been designed and updated with the expectation that most code will be written by humans. An exception is assembly languages, which were initially written by humans, but for decades have been designed to be targeted by a compiler. In general, programming languages are created to address a specific need, often tailored to an application domain, hardware architecture and performance requirements. Overall, they end up being created to solve a real practical problem created by application/user needs and/or platform capabilities, and they find a corresponding trade-off among many factors, including abstraction,

expressiveness, simplicity, understandability, maintainability, safety, reliability, efficiency, support for parallelism, scalability for large codebases, portability, and support for a robust ecosystem of tools and libraries.

It remains uncertain which languages should be targeted by the AI-based tool, and whether the ambitious vision of an AI-based model directly transforming "natural language to transistors or machine code" will ever be feasible, especially for current million-line plus codebases. Achieving this vision is likely to face significant challenges related to energy efficiency (of the AI-based system generating the code), scalability (to ever larger codebases), precision (resolving the inherent ambiguities of natural language), and understandability (to support human and/or machine verification and facilitate troubleshooting). These challenges suggest that one or more levels of abstraction between natural language and machine-level code will remain necessary. Abstraction not only helps mitigate ambiguity and complexity but also provides modularity and structure, essential for debugging, optimization, and the efficient generation of scalable systems.

Hardware and software development depends on various auxiliary tools, such as simulation, model checking and timing analysis tools (for hardware), debuggers (for software), as well as performance and energy analysis tools, verification, static analysis and code coverage tools. Human intuition and creativity will increasingly be replaced with AI-based tools, but traditional optimization algorithms are extremely powerful and should continue to have a place at the lowest level. These tools often have idiosyncratic interfaces, and they are hindered by the multiple levels of abstraction between the machine and the high-level code, that may need to be traversed to understand what has gone wrong. The key is to operate at the right level of abstraction to solve the issue, as high as possible, while being able to drop to the lowest levels where needed. This presents a significant opportunity for AI-driven tools to drive developer tool use through natural language interaction, automate tools integration within a larger AI controlled workflow, and translate cryptic error messages into higher-level code suggestions.

Support a European ecosystem that includes basic research in AI

Europe's universities, research centres and companies must be at the forefront of basic research in AI, pursuing important research topics such as the following:

- Correctness, safety and security. As discussed above, this is the greatest barrier to the adoption of AI. Formal methods can be used to prove correctness and security properties (see for example `[GoogleAlphaIMO]`), but they are cumbersome for large systems and should be the subject of basic research.
- Programming languages and abstractions.As discussed above, it is not clear how programming languages should evolve as they are increasingly targeted by AI-based tools. It is unclear whether the choice of abstractions should mirror those designed for human developers or be created specifically to exploit the strengths of generative AI methods, whatever that entails. A key issue will be the lack of training data for any new programming language or language features.
- Open-ended problems. For hardware design, AI-based tools can be given an open-ended problem, such as. "design a CPU that executes these programs, as fast as possible, given this transistor/power budget". This problem includes design space exploration but is much broader in scope, as it is not constrained by parameters defined ahead of time by people.
- Optimization of neural networks.In addition, the increasing and tremendous complexity of neural networks, present in all machine-learning applications, will require more and more reliance on automated AI-based tools to help design efficient solutions and master their huge complexity. These tools will need to exploit multi-

criteria optimization methods and to generate optimized code for a given hardware, in order to take into account the numerous embedded constraints that it must guarantee. These constraints can cover the induced power, the memory size, the prediction accuracy or for instance the type of operations used to remain compatible with the final hardware. The supported hardware must be compatible with the latest innovations and computing trends, including for instance heterogeneous system-on-chips (SoCs) with dedicated neural networks accelerators. The output of these AI-based tools, based on neural architecture search (NAS) methods, should be able to design optimized and frugal AI applications, for all AI applications using LLMs, transformers or Mamba algorithms. These tools will also have to integrate trustable and explainable methods to bring to the user the knowledge used by the tools to obtain the final results, in order to integrate critical embedded systems.

Develop European agents, tools and infrastructure

In today's geopolitical climate, European sovereignty over its AI models is crucial, especially as AI-based technologies increasingly influences national security, economic competition and social governance. AI-based tools for hardware and software development stand out from general AI due to the foundational role they can play in building and shaping future technology, as well as their role in innovation and competitive advantage.

AI development tools will serve as the backbone of the digital economy, facilitating the creation of chips, communication networks, cloud infrastructures, middleware, and applications that support all other AI-based applications, from autonomous vehicles to smart cities. If Europe lags behind in this area, it will become dependent on foreign suppliers whose interests may not align with European priorities. In a worst-case scenario, this dependency could lead to hardware and software being compromised or containing hidden backdoors, creating significant national security risks.

The race to build AI development tools is, in essence, a competition for leadership in the global tech economy. European countries must have access to the most advanced tools and be able to influence their development, in order to compete with global giants from the US and China, and help Europe to remain a leading force in key industries such as automotive manufacturing, telecommunications and fintech.

At the same time, Europe is recognized for its strong commitment to ethics, legal, socioeconomic and cultural aspects of the use of AI-based technologies and its unique regulatory frameworks. Some global companies have already opted to withhold support for their most advanced AI rather than adjusting to European regulations. If this trend continues and worsens, especially in times of geopolitical tension, it could stifle economic competitiveness. In the worst case, there will be significant pressure to undermine European ethics.

Focus on education, training and jobs

As of 2025, AI tools can fully automate the creation of simple code, consisting of a few hundred lines, and they are powerful assistants to human developers in full-scale development projects. However, as described so far, these tools are cannot yet replace proficient and experienced developers. As these tools advance, important questions arise about the future of the workforce in the hardware and software industries, which currently employ millions of people globally.

Over the next few years, AI tools are likely to continue to assist developers, particularly in routine and repetitive tasks, freeing developers to focus on higher-level design and problem-solving. In this period, many routine tasks will be automated, leading to a shift in work for

developers. Entry-level positions may be affected, but mid-level and senior developers will still be in high demand to oversee complex projects, integrate AI-generated code, and ensure quality and creativity in the final product. This will place greater and distinct demands on education, which may be alleviated by individualized AI-based training helping to make programming more fun and learnable by people at a younger age.

As of 2025, at the height of the hype curve for AI, it is important to maintain a historical perspective. In 1954, IBM's Fortran specification claimed that "Since FORTRAN should virtually eliminate coding and debugging, it should be possible to solve problems for less than half the cost that would be required without such a system" [FORmula]. Similar claims were made in the 1980s, for fourth-generation languages, such as SQL, ABAP and COBOL 85. While these technologies did reduce development cost and time (by much more than half), the belief that they would eliminate the need for software developers was wildly optimistic. In practice, the necessary skills moved from assembly language coding to the wide class of skills needed for large scale software development today.

Nevertheless, while history is a guide, it is not guaranteed to repeat. In the long term, AI tools may evolve to the point where they can build increasingly complex systems autonomously. Will AIs be able to replace a team of human developers, with a human taking on the role of a chief architect or CTO interacting with AI? What happens when something goes wrong? At this point we do not know.

Conclusion

In conclusion, the integration of AI tools into hardware and software development offers transformative potential, and it has the potential to inject a major discontinuity into the development process. By utilizing natural language interfaces and leveraging AI's capabilities, development processes can become more efficient, reducing time and costs, while also democratizing access to advanced technologies. However, the risks associated with AI-generated outputs, such as safety, correctness, security, and confidentiality, must not be overlooked. Europe must prioritize basic research in AI, develop its own AI tools and models, and ensure that AI's role in development remains aligned with ethical, regulatory, and security standards. Furthermore, as AI tools evolve, the future workforce will need to adapt, with AI serving as a powerful assistant to human developers rather than a complete replacement. The success of Europe in this rapidly advancing field will depend on fostering a robust AI ecosystem, ensuring technological sovereignty, and investing in education and training for the next generation of developers.

References

ALSAQER: Shadan Alsaqer, Sarah Alajmi, Imtiaz Ahmad, Mohammad Alfailakawi, "The potential of LLMs in hardware design", Journal of Engineering Research, 2024. ISSN 2307-1877. https://doi.org/10.1016/j.jer.2024.08.001

ChatEDA: Z. He, H. Wu, X. Zhang, X. Yao, S. Zheng, H. Zheng, B. Yu, "ChatEDA: A large language model powered autonomous agent for EDA", In: 2023 ACM/IEEE 5th Workshop on Machine Learning for CAD (MLCAD), IEEE, 2023, 1-6.

ChipGPT: K. Chang, Y. Wang, H. Ren, M. Wang, S. Liang, Y. Han, H. Li, X. Li, "ChipGPT: How far are we from natural language hardware design", arXiv preprint arXiv: 2305.14019 (2023).

ChipNeMo: https://research.nvidia.com/publication/2023-10_chipnemo-domain-adapted-llms-chip-design

CodeLlama: https://ai.meta.com/blog/code-llama-large-language-model-coding/

CodeLlamaLicence: https://github.com/facebookresearch/llama/blob/main/LICENSE

Codestral: https://mistral.ai/news/codestral/

FORmula: Preliminary Report. Specifications for The IBM Mathematical FORmula TRANslating System. 1954. https://www.softwarepreservation.org/projects/FORTRAN/BackusEtAl-Preliminary%20Report-1954.pdf

FTAug2024: https://www.ft.com/content/4868bd38-613c-4fa9-ba9d-1ed8fa8a40c8?utm_source=chatgpt.com

GitHubCopilot: https://github.com/features/copilot

GoogleAlphaIMO: Google's AlphaProof and AlphaGeometry use the Lean theorem prover to check their solution to problems from the International Mathematics Olympiad (IMO) https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/

GoogleCodeAssist: https://cloud.google.com/products/gemini/code-assist

GoogleJules: https://developers.googleblog.com/en/the-next-chapter-of-the-gemini-era-for-developers/

KordonZaourar: Munier Kordon, A., & Zaourar, L. (2024). "Challenges in EDA: from operational research techniques to Artificial Intelligence strategies for chip design". HiPEAC Vision 2024, Rationale. https://doi.org/10.5281/zenodo.10874774

LLM4HPC: Pedro Valera–Lara. LLM4HPC: Towards an AI-autonomous HPC world. https://www.hpcuserforum.com/wp-content/uploads/2024/10/Pedro-Valero-Lara-ORNL_LLM4HPC-Towards-an-AI-autonomous-HPC-World_HPC-UF-BSC-Oct-2024.pdf

Metzger: Metzger, A. (2024). "AI-Assisted Software Engineering (AISE)". HiPEAC Vision 2024, Rationale. https://doi.org/10.5281/zenodo.10874754

MicrosoftCopilot: https://copilot.microsoft.com/

OpenAICodex: https://openai.com/index/openai-codex/

Qiu24: S. Qiu, B. Tan, H. Pearce, "Explaining EDA synthesis errors with LLM", arXiv preprint arXiv: 2404.07235 (2024).

SpecLLM: M. Li, W. Fang, Q. Zhang, Z. Xie, "SpecLLM: Exploring generation and review of VLSI design specification with Large Language Model", arXiv preprint https://arxiv.org/abs/2401.13266 (2024).

StackOverflow2024: https://survey.stackoverflow.co/2024/ai#sentiment-and-usage-ai-sel-prof

Cyber-Physical Systems

Recommendations for Cyber-Physical Systems

Accelerate cross-disciplinary joint research

The technology domains contributing to Cyber-Physical Systems research call for investment in tools, methods and cross-technology community initiatives to tackle the multi-stakeholder research barrier - especially arising for a technology bridging diverse complex knowledge domains and applied at higher levels of a system where there are many more interactions with the technology to consider - higher-order integrated research. This will accelerate progress towards the Next Computing Paradigm and CPS research as well as technology infrastructure updates by tackling the challenges of diverse knowledge domain perspectives and enabling access to the bigger picture. In particular: 1) A new R&D dimension to really boost our capability for highly complex and cross-domain integrated research activities. Just as we have different approaches for building windows and houses, there is need to establish tools and methods supporting higher order integrated research. This is especially a case in point for the highest integration levels of CPS research where most impact and value generation can be expected. Adapted or new tools and methods for convergence, with strong public engagement, should support terminologies (e.g. wiki-style trusted glossary), concept sharing (e.g. modelling), knowledge sharing (e.g. ontologies via Protégé), consistent evaluation approaches and global visualisations, including non-technical domains. 2) Existing communities should establish a centralised CPS association to unify efforts, promote knowledge exchange, and align standards; 3) Additionally, frameworks for integrating AI/ML into CPS must address safety, security, and ethics, ensuring dependable systems for sectors like healthcare and transport. These actions are vital to Europe's sovereignty and global leadership in CPS advancements.

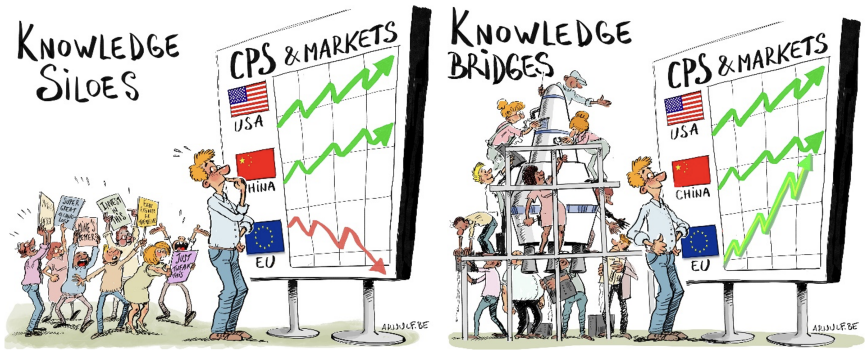
Redefining dependability for CPS adaptability and technology integrations

CPS depend on safety, security, and performance properties to govern what they can achieve and qualify technologies for use. CPS contributing communities encourage: 1) Solutions to migrate from legacy approaches that minimise interactions of these properties to instead maximised interactions for optimum system adaptability. These properties impose constraints on available choices we have at design and in operations, which are compounded by ruling out choices where trade-offs would be required. Techniques such as combined analysis, evaluation and knock-on effects should be advanced for handling these properties. Establishing an approach, considering tools and methods referring to best practice, is needed to account for the interdisciplinary integration overheads between these traditionally distant domains, but also with the rest of the system. This is crucial in CPS for enhancing scope of AI/ML and IoT usage, as well as other technologies. 2) A new way of thinking is needed for treating interconnected systems with CPS - dependability considered in a modular fashion - with hazard analysis techniques likes STPA extended, including for man-machine teaming and AI complexities. We encourage also frameworks for risk assessment in relation to AI/ML to be established and considering adaptive risk management strategies in the context of these interconnected critical systems. This moves forward with trustworthy CPS in sectors like AI-enabled autonomous systems.

AI-performance-defence guarantees for real-time interconnected systems

Future CPS require advanced technologies to address challenges in performance characterization, damage containment, and operational feedback. CPS contributing communities encourage: 1) Real-time methods ensuring deterministic multi-tasking environments and verifiable AI/ML performance. In complement, there should be an extension of defence mechanisms and feedback loops, which is essential for preventing damage propagation and enabling iterative improvement. Solutions should emphasize distributed architectures, particularly edge computing, and include digital twin capabilities for predictive insights. 2) Comprehensive uncertainty quantification, real-time monitoring, run-time verification, and data flow tracking will enhance trustworthiness. These advancements will support supervisory control and ensure dependable CPS operations, even in rapidly evolving and uncertain environments like AI-enabled applications.

These three recommendations are detailed next. Due to the multi-domain nature of CPS research they have also been extended as an associated white paper [1].



Introduction

Cyber-Physical Systems (CPS) bridge diverse technologies into cohesive wholes that interact seamlessly with the physical world. CPS research fosters the integration of disciplines across domains like healthcare, manufacturing, transportation, and space, transforming fragmented innovations into dependable, real-world applications. This interdisciplinary effort relates to the continuum of technology-to-system and centered around computing. It transitions from lower-stage cross-domain integrations within a system up to final product-oriented outcomes. The Next Computing Paradigm (NCP) provides a pivotal foundation for CPS research advances, building already on a mature infrastructure of connected technology domains.

However, the increasing interconnectivity of systems presents significant challenges, including the interaction complexity of diverse stakeholder knowledge domains (across technology specialisations, end-users, policy, regulation, standards, the public, etc.), orchestration, dependability and the scalability of solutions. This means CPS research, while playing an important role as a market generator for a multitude of technologies, has specific challenges compared with most other research domains, including a slower R&D cycle. This longer time to maturity can be compared with building windows, rooms and houses – where CPS research is positioned especially towards the later part in terms of integration complexity and calling for scaffolding in the form of supportive tools, frameworks, and policies. This chapter promotes two key axes for advancements: Support R&D, focusing on enabling methodologies and tools to integrate research across knowledge domains, and Applied R&D, which examines integrated technologies within CPS. Together, these axes

provide key ingredients for CPS research supporting the success of the NCP and technology uptake in European markets.

Due to the scope of CPS research there are two connected white papers with this chapter. One supports the advice offered here, with three extended recommendations for the first one in this chapter and two supporting recommendations for both the second and third ones of this chapter[1]. The other is a positioning paper – many research domains characterise a CPS[2].

Accelerate cross-disciplinary joint research

The CPS specialists and contributing technology domains consider this the most urgent of the CPS recommendations. The complexity of technology integrations towards CPS demands a new R&D dimension to address the challenges especially of multi-stakeholder environments and prepare for the NCP, ensuring Europe's sovereignty in advanced technologies.

One aspect is that the time to develop and adopt such technologies is several times longer than for technologies from a single domain. We need to look at bringing down this time. Just as we have more advanced support tools for complex building construction, like cranes or vehicles for digging, we can have R&D providing also more advanced technologies that support doing complex/integrative research. That is, building technologies - support R&D - that helps researchers to carry out more advanced (applied) R&D.

Research collaboration approaches taken for granted inside single disciplines suddenly become multiple times more difficult where definitions, concepts, methods, priorities and evaluations can be quite different related to the involved domain perspectives. This is also compounded for elements higher up in a system where they normally have many more interactions with the system and related standards to taken into account. The effort and time required to surmount these cause collaborations to grind to a halt and significantly impact success, even where there is a strong motivation between researchers[3].

Reducing this hurdle will play a significant role in European market capture and global competitiveness. This is because ultimately CPS represent markets of integrated technologies, culminating for instance in railway systems or satellite constellations, and CPS research permits easier integrations so these infrastructures (technologies in themselves) are ready for the latest developments from the contributing technology domains. Technology integrations can play a profound role in market capture, take for example as an infrastructure technology the American Android phone, which represents a wide market place – not only for the technology components that make the phone, but for all the applications that sit on top of this.

Another aspect is that technologies higher up in a CPS face unique integration challenges that require cohesive collaboration across domains. Like industry relies on continuity programs to manage complex projects, CPS research needs a unifying structure to prevent fragmentation and address long-term goals. Europe currently misses a centralized association for CPS research or even system engineering that represents European interests. The question of how our technology components come together into technology wholes\systems is an important element for supporting successful European markets – CPS are markets. We have a unique (CPS) infrastructure landscape on the global market both in terms of physical implementations of transport, etc, and also in terms of policy focusing on ethical, sustainable development, with a strong emphasis on regulation and societal impact. Without a central means to take the pulse at European level and act, we miss a unified advocacy in relation to needs and priorities: supporting policies, funding allocation, or regulatory decisions; missed opportunities for knowledge sharing and efficiencies; supporting Europe market capture via European CPS; workforce development and education; public awareness supporting trust. The community should work to draw

together national system engineering bodies into such an association that will also provide platforms for knowledge exchange, align research with emerging standards, and foster common approaches to interdisciplinary education and shared strategies across Europe. It will support CPS advancements in transport, healthcare, and manufacturing, ensuring Europe's leadership in aggregative technologies (like motors and cars composed of other technologies).

Finally in relation to the rapidly changing AI landscape, there is a need to facilitate collaboration among AI researchers, dependable systems experts, and domain specialists to address safety, security, and ethical challenges to address integration of AI/ML into CPS. Frameworks and guidelines for AI/ML safety, security, and ethical integration, supported by a European network of excellence should be developed. This will help accelerate AI/ML integration into CPS, and align innovations with Europe's ethical and regulatory standards, particularly for high-stakes sectors like healthcare and autonomous systems.

Redefining dependability for CPS adaptability and technology integrations

Technology advances for cyber-physical systems are strongly tied to the triad of safety, security and performance properties within a critical system. These properties each can influence governance of thousands to millions of interactions between parts of a system from many contributing technology domains, which permit the transformation of digital intentions into trusted real-world actions. As such, they act as a form of gateway determining which technologies and combinations are acceptable for use within CPS. The size of this gateway, i.e. how much technology is qualified to get through for usage, is dependent not only on the permitted interactions by each of these properties, but also between them. Advancing on this latter part is where we could expect a high impact, since interaction engineering between safety and security in systems is currently very limited across industry.

Figure 1: Enhancing technology access to the real world. Source: Generated via Dall-E.

There are two aspects agreed by the community to be tackled in order to advance, drawing on existing safety-security interaction research. For applied R&D, it would be beneficial to

investigate the means for a system itself to be able to evaluate safety, security and performance through common evaluation criteria. From the support R&D perspective, there is need for a new research dimension, building the means to bridge the safety and security research domains themselves. This should draw upon interdisciplinary best practices to create a tailored approach for managing the research integrations. Remember that a technology is the application of scientific knowledge to the practical aims of human life (Britannica). No dedicated interdisciplinary approach currently exists to support technologies for advanced interactions between these domains. This is considered a key reason why there is not yet widespread adoption by industry. It qualifies for such an approach as the two domains have historically developed independently (traditionally distant disciplines), emerging as two distinct domains. They have separate specialist definitions, concepts, standards, certification processes, values and priorities. This presents very specific bridging hurdles for successful joint research, which does not exist for standard technology development approaches inside disciplines.

Enlarging the permitted interactions between safety and security mechanisms by enhanced research integration approaches, in addition to boosting technology uptake in critical systems, should be used to particularly support advanced AI/ML integration and to enable IoT systems to automatically integrate and coordinate within larger Systems of Systems (SoS), ensuring resilient operation and enhanced system capabilities. More generally, there is need for new ways to be established for managing dependability for AI/ML-enabled CPS and highly interconnected systems. Frameworks for continuous risk monitoring and automated contingency management will be needed, as well as advances in modelling techniques that modularize the dependability between systems and the methodologies to capture these complex interdependencies, including man-machine teaming capacities.

AI-performance-defence guarantees for real-time interconnected systems

With unprecedented levels of autonomy and interconnectivity, future CPS face fundamental challenges, particularly in higher-level orchestrative and autonomy technologies. Key hurdles include performance characterisation, containment of damaging events and establishing comprehensive operational online feedback mechanisms.

This calls for advanced real-time performance methods to master demonstrably deterministic multi-tasking environments and verifiable performance bounds for AI/ML-enabled components. Existing defence mechanisms must be extended, alongside performance characterisation, to prevent damage propagation across interconnected critical applications. Solutions will require consideration in the context of distributed architectures and particularly from the edge computing perspective.

Mechanisms providing system-level feedback are also essential, in general for enabling advances of contributing technology domains, to catch an understand weaknesses, and especially for rapidly changing landscapes like AI Feedback results must return directly to developers, enabling iterative improvement. Technologies should provide new digital twin capabilities providing predictions when a CPS can perform better with associated services to increase the value. They should include comprehensive uncertainty quantification, real-time monitoring, run-time verification, data flow tracking, and automated compliance checks for trusted machine-learned data and supervisory control during operations.

Figure 2: New feedback capacity for engineering with AI and Digital Twins. Source: HSE.AI.

Conclusion

As CPS evolve to meet the demands of an interconnected world, the need for cohesive strategies to navigate complex integrations is critical. By leveraging the foundations provided by the Next Computing Paradigm, CPS research will tackle challenges in dependability, scalability, and cross-domain innovation. Focusing on Support R&D and Applied R&D encourages a strategic approach to technology orchestration and integration for future complex and critical applications. This dual-axis strategy empowers stakeholders to bridge the gaps between knowledge domains, which will unlock transformative potential in diverse fields such as healthcare, manufacturing, and transport. The journey forward demands collaboration, innovation, and a commitment to managing complexity.

References

1: Charles R. Robinson et al. (2025). Extended Recommendations for Advances on Cyber-Physical Systems. Zenodo. https://doi.org/10.5281/zenodo.14624958

2: Charles R. Robinson et al. (2025). Bridging the Stakeholder Domains that Produce Cyber-physical Systems. Zenodo https://doi.org/10.5281/zenodo.14693254

3: D. Gooch & L. Benton. (2015). Impact in Interdisciplinary and Cross-Sector Research: Opportunities and Challenges. Journal of the American Society for Information Science and Technology.. https://doi.org/10.1002/asi.23658

Cybersecurity

Recommendations for Cybersecurity

Software supply-chain cybersecurity

Reinforcing software supply-chain cybersecurity is crucial given the wide impact of attacks spread through the supply chain, which is all the more important given the large number of components in the next computing paradigm (NCP). Develop code and component analysis technologies for cybersecurity that scale up and support trusted orchestrators, services and communications.

Comprehensive safety, security, and performance coupling requires standardized software vulnerability representation. Increased interconnectivity requires new technologies to isolate threats and proactive cyber-risk management. Develop secure software package and service management that balances usability with strong security.

AI for cybersecurity

To enhance NCP cybersecurity in a scalable way, develop i) advanced artificial intelligence (AI) models, including large language models (LLMs), for threat detection and ii) autonomous systems for mitigation (e.g. isolating compromised NCP components, patching vulnerabilities, or restoring services). Utilize federated AI for its decentralized, privacy-preserving and scalable models in the NCP massively interconnected context. Rely on EU-based open AI models and datasets to strengthen EU cybersecurity, sovereignty, and competitiveness.

Reinforced cybersecurity of AI

Secure AI training methodologies and validation procedures, as well as adversarial defences, are needed. LLM prompt injection attacks must be a major concern, addressed by the development of tools to detect and secure against these, and by establishing benchmarks for prompt injection prevention and response. AI security standards should be established by developing certification procedures to guarantee that LLMs and AI systems adhere to stringent security standard, possibly requiring security audits for AI systems. These efforts should rely on EU-based open AI models.

Introduction

Cybercrime is known to have been increasing dramatically over the last few years, and this trend is expected to continue, as the following figure from Statista shows:

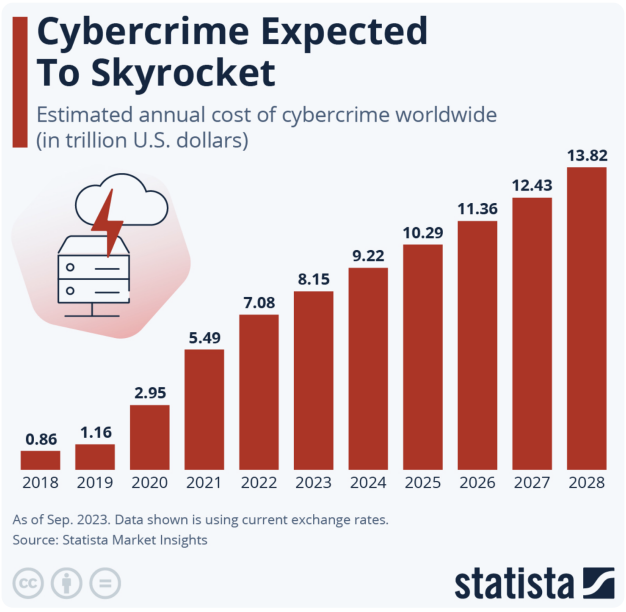


Figure 1: Cybercrime costs are expected to continue dramatically increasing `[Statista-CostCybercrime]`

The next computing platform (NCP) orchestrates numerous components and services from cyber-physical systems, the internet of things (IoT), clouds, digital twins, etc. This federated, highly connected and dynamic computing continuum thus offers a particularly large attack surface to cyber villains, and can only be expected to suffer from the aforementioned increasing cybercrimes.

Protecting the NCP thus requires strong, scalable analyses to detect and fix vulnerabilities across all its levels, domains, interconnected components, (micro)services, orchestrators, and their communications and interactions. With all these components and services, at production stage, the supply chain cybersecurity and NCP source code becomes ever more crucial. NCP cyber defences must also encompass detection and mitigation of attacks when in operation.

As in many domains, AI is being used by cyber villains to help them produce and automate cyberattacks. AI has thus become necessary to cope with this increased threat and with the massive complexity the NCP brings, by scaling up and automating cybersecurity tasks at all levels.

AI, especially the booming LLMs used in the NCP context, also faces crucial and often specific cybersecurity issues that must be addressed for its widespread usage to be secured.

Regulatory measures are also necessary for the cybersecurity of the NCP, and societal preparedness must be reinforced for EU security.

By addressing these points, the EU can establish the NCP as a continuum with strong cybersecurity, thereby maintaining confidence and dependability in it, establishing itself as leader in cybersecurity innovation while protecting the EU cybersecurity and sovereignty. To this end, we make the following three main recommendations, followed by additional recommendations.

Software supply-chain cybersecurity

Software vulnerabilities present a very significant risk to EU. Reports highlight that over 75% of applications contain at least one flaw, nearly 25% of these being classified as high-severity issues, and that even more alarmingly 26% of organizations still face exposure to vulnerabilities exploited by well-known attacks like WannaCry, years after patches have been released `[Qualys2024][comparitech2024]`. Identified common vulnerabilities and exposures (CVEs) follow an ever-increasing trend, as shown in the following graph from `[CVEdetails.com]`:

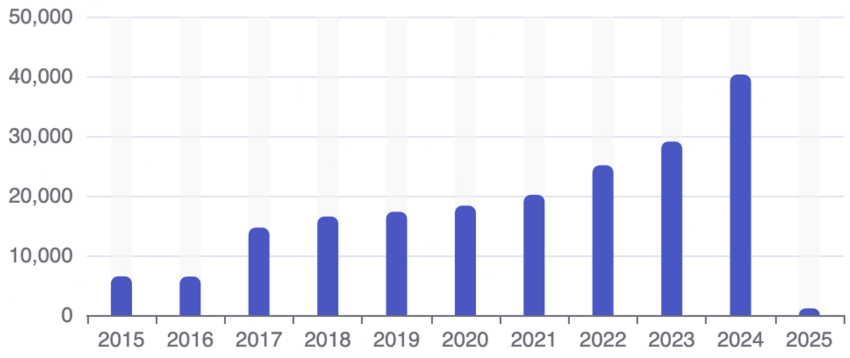


Figure 2: The number of identified CVEs keep increasing quickly.

Such statistics underscore the importance of continuous, automated, and scalable vulnerability detection and fixing methods and tools.

However, the NCP introduces very significant complexity through the integration of myriad components and services, ranging from a variety of domains like cyber-physical systems,

the internet of things, clouds, digital twins, etc. all of which work together thanks to powerful, AI-based, orchestration mechanisms. Securing this federated, highly connected and dynamic computing continuum requires powerful and scalable analysis technologies able to detect and address vulnerabilities throughout all the levels and domains of this system architecture.

These techniques must address the level of individual components, to detect and identifying vulnerabilities in specific software libraries or APIs, especially those that are extensively used throughout the continuum. The open-source nature of these components, which facilitates reuse and modularity at the software engineering stage, is also a facilitator for these analyses, since the source code will be available, enabling source-code level analyses and comparison with the generated binaries.

These, or complementary, techniques must also address the interconnections between components and services, which are a key aspect of the NCP. They must be able to evaluate the potential security vulnerabilities stemming from interactions among components, such as unsecured data transmissions or inadequately established protocols, that could leave opportunities for attacks.

Techniques must also address the NCP orchestration level, assessing and scrutinizing the conduct of these AI-driven orchestration systems to guarantee they do not create security vulnerabilities, such as misconfigurations or unintentional privilege escalations. In addition to this scrutiny, the code used for the orchestrators themselves, although not specifically more vulnerable than any component code, should be considered as particularly critical: it will be a favourite target of attackers, given that by controlling orchestration they could obtain tremendous system-wide effects.

The same is true for agents making decisions in the NCP. Although from a technical point of view they are similar to other agents, from an attacker's point of view decision agents may be more interesting than e.g. sensor agents. However, while stealthier, sensor agent attacks may have strong domino effects as well, which can make them attractive. An attacker would thus likely target either sensor agents or decision agents, based on the specifics of each case – the orchestrator, of course, remaining a prized target.

For these analyses to be effective and usable in practice, it is necessary to improve the way software vulnerabilities are represented and modelled. Indeed, the current diversity in the documentation, cataloguing, and resolution of software vulnerabilities continues to hinder cybersecurity.

Currently, vulnerabilities still often are documented in an informal, human-readable way that is not ideal for automation and tooling, with CVEs being represented in a semi-formal way, despite ongoing improvements `[CVE-MITRE][CVE-ORG][CISA2024]`. Implementing and advocating standardized formats for vulnerability representation and modelling is thus necessary, not only to mutualize efforts but also to improve the interoperability of cybersecurity analyses and (possibly) mitigation tools for the NCP ecosystem.

This clear definition of vulnerabilities should be augmented to encompass the distinctive attributes of NCP components and their interconnections. This should further facilitate the creation of innovative techniques and tools to model and query interrelations among vulnerabilities, threats, and mitigations. Such standardization of representations will enhance automated vulnerability identification and fixing as well as collaboration among developers, security professionals, and organizations.

In addition to securing the code and components, it is necessary to address the security of the chain that supplies them. This is crucial given the extremely wide impact of software supply-chain attacks (i.e. attacks spread through the software supply chain) and the fact that this is becoming one of the most exploited cyberattack vectors.

High-profile supply chain attacks, like those targeting SolarWinds and Log4j, have clearly shown the dire consequences of vulnerabilities being present in software dependencies and distributions. In 2023, supply-chain cyberattacks surged by 200% compared to 2022, with malevolent actors using critical infrastructures and widespread software libraries to disseminate malware [CISA2023][comparitech2024][Ladisa2023a].

Such cyberattacks generally exploit vulnerabilities in third-party code and libraries, using dependencies to stealthily infect software applications. Thus, dependency management, or a software bill of materials (SBOM) [SBOM-NTIA][Dalia2024], is crucial to monitor the interdependencies and vulnerabilities in supply chains, especially due to the federated, decentralized and dynamic nature of the services provided in the NCP and the components underlying them. These cyberattacks also target tools used to develop software, altering development or build tools to embed malicious code (i.e. malware, backdoors or vulnerabilities) in the software produced, or hijacking package managers, updates, or repositories to spread malicious components at distribution time. This is especially important in the NCP, which requires numerous components from various providers.

Research in secure package and component management systems is thus crucial to alleviating these risks. The security features of these package managers must, among others, include integrity verification, to confirm that each package or component is cryptographically signed and validated prior to being used [Sigstore] and dependency security management, to detect and address vulnerabilities in transitive dependencies, an aspect often neglected yet responsible for over 60% of problems in software projects [comparitech2024]. Above all, these package managers should be devoid of mechanisms that make it possible to execute arbitrary code on the target system, which is far from the current situation [Ladisa2023b].

However, for these secured systems to be successful, i.e. adopted by developers, they must not trade ease of use and developer-friendliness for security. There lies an important challenge: having package managers that are both secure and easily usable.

Reconciling usability with security is a fundamental problem in securing the supply chain that has been poorly addressed so far, developers often prioritising accessibility to repositories and tools above rigorous security measures. To address this crucial issue, it is necessary to carry out research on secure package managers that: facilitate smooth integration with common, established development workflows; provide or integrate with developer-friendly tools for vulnerability detection and possibly fixing; and provide developer-friendly dashboards to monitor and manage software dependencies.

Furthermore, for maximum impact and effectiveness, research bodies and industry must foster collaborations to establish unified standards and tools for secure package and component management systems. These efforts could build upon existing initiatives such as the SBOM and frameworks for secure software development, like NIST's Secure Software Development Framework (SSDF) [NIST-SSDF].

AI for cybersecurity

In order to handle the sheer volume and complexity of components, interconnections and data in the NCP, and reach the appropriate levels of scalability, it will be both crucial and necessary to develop scalable automated analysis methods—leveraging AI and machine learning (ML) where appropriate. Indeed, human-centric methods for detection and response to cyberthreats and cyberattacks have become inadequate. In 2023 for example, global cybersecurity incidents increased by more than 40%, propelled by the extensive use of networked devices and the increase of AI-assisted cyberattacks [Statista-Breaches2024], like ransomware as a service (RaaS)[IBM-RaaS] Manual methods have thus become insufficient to match the speed and magnitude of these dangers, hence the need for fast, scalable, automated methods and tools.

AI, especially ML, is transforming cybersecurity. It empowers attackers to create cyberattacks more easily, even for people with a lower level of technicity, hence spreading the fire. However, it also offers many opportunities for cyber defenders.

Indeed, AI-driven algorithms can scrutinize extensive datasets to find deviations from standard behaviour, and identify anomalies very quickly, drastically decreasing threat detection time from months to seconds. AI technologies used in security orchestration, automation, and response (SOAR) systems can triage, analyse, and mitigate threats autonomously, providing automate incident response. They can also even help anticipate and mitigate threats, since predictive analytics can discern nascent assault patterns, enabling and facilitating pre-emptive defence strategies.

Overall, AI-driven automation has the potential to enhance scalability and continuously adjust to emerging threats in near real time, ensuring stronger security as the quantity of devices, components and services increases rapidly in the NCP context.

However, although automation presents significant potential, especially AI-driven automation, its implementation faces several challenges. First, such tools must integrate smoothly across many platforms, services, and physical components within the NCP, which may require significant engineering efforts. In addition, automated systems must combine efficient monitoring with solid privacy rules, such as GDPR. Finally, AI and ML-based systems must be protected against involuntary biases and model vulnerabilities, and against adversarial AI attacks that could compromise their performance, which implies research to investigate into more robust models.

As a consequence, to enhance and scale up automated cybersecurity within the NCP, research must be encouraged and tools developed on i) advanced threat-detection AI models, based on deep learning, graph-based analysis, natural language processing (NLP) and large language models (LLMs) to improve detection and monitoring capabilities, and ii) autonomous threat mitigation systems that can perform automatic actions, such as isolating compromised components of the NCP, applying patches to vulnerabilities, restoring services, etc.

In the continuum of the NCP, federated AI systems should be investigated, as their decentralized AI models can help function across distributed, massively interconnected components and services, in an edge and cloud context, while preserving data privacy and scalability. To this end, the use of EU-based open AI models and datasets such as `[HuggingFace]`and `[MistralAI]` should be favoured, as this can help the EU boost its cybersecurity while preserving its sovereignty and reinforcing its competitiveness.

Reinforced cybersecurity of AI

The use of AI systems, especially LLM-based systems, has become in a few years extremely widespread in almost all domains and applications of computing, hence all across the NCP. The (cyber)security of such systems is thus crucial, yet their rapid adoption brings unique challenges that require urgent attention. Indeed, deployed LLMs are currently susceptible to numerous security vulnerabilities. Malevolent actors can exploit prompts [Pasquini2024, Liu2024] to compel LLMs to produce detrimental or unauthorized content.

Existing protective measures seem to be rather an external layer of LLMs, since relatively simple tricks have been shown to bypass these security measures. Challenges thus exist in completely mitigating these vulnerabilities, which ideally should be done within the LLM's internal behaviour, not at its periphery.

Attackers can also compromise data `[Monkam2024]` used to train AI models, resulting in skewed outputs and weakened defences. LLMs generating inaccurate or false information, either by mistake or by having been skewed to do so `[Wu2024]`, can then be leveraged to disseminate misinformation or influence choices, such as elections, all across the NCP, with

low cost and high spread. On several occasions, LLMs have been shown to leak sensitive corporate information [Raz2024]. LLMs are also used to help developers code, hence generate source code; the latter can however contain cyber vulnerabilities.

The EU should thus invest in secure AI research focused on secure training methodologies and validation procedures, as well as adversarial defences. LLM prompt injection attacks must be a major concern, addressed by research on detecting and securing against these and establishing benchmarks for prompt injection prevention and response, e.g. in the spirit of the CyberSecEval benchmarks [CyberSecEval3]. Benchmarks and AI security standards should be established by developing certification procedures to guarantee that LLMs and AI systems adhere to stringent security standard, possibly requiring security audits for AI systems. These efforts should rely on EU-based open AI models (see [HuggingFace] [MistralAI]).

Additional recommendations

In addition to the above three main, critical recommendations, additional relevant recommendations can also be made as follows.

Authentication, intrusion and attack detection in massively interconnected systems

It is necessary to support research and tools for intrusion and attack detection in systems with massively interconnected components and services, including authentication mechanisms that scale up within the NCP.

Indeed, the NCP offers a large attack surface, due to its numerous and massively interconnected components and services. Efficient and effective intrusion and attack detection is thus necessary but faces distinct issues from conventional cybersecurity tools.

The volume and velocity of data generated by the continual exchange across interconnected (micro-)services and components produces a tremendous traffic volume, making it difficult to distinguish harmful activities from normal ones. The myriad of NCP components, while facilitating compartmentalization and isolation, also provides attackers with opportunities for concealment, allowing them to use strategies involving long-term infiltration and lying dormant to avoid detection, moving only within the ecosystem of components and services when attacking their real target, a technique which is called "lateral movement". Furthermore, the heterogeneity and dynamicity of the NCP, characterized by the dynamic orchestration of resources, require adaptive and context-sensitive detection techniques and tools.

Research must thus be encouraged to develop automated, big-data-capable intrusion-detection systems (IDS), capable of monitoring and analysing extensive data sets in real-time, to detect attacks and intrusions early, as they develop, not after. The goal is to identify and obstruct malevolent actors in real time, which is very far from the industry average time-to-detection of over 200 days [IBM2021], building on already-reported time savings of 108 days provided by AI-powered tools [IBM2023].

Indeed, AI and ML can be keystones to this end, helping for example with anomaly detection and pattern recognition (in logs or execution traces) associated with cyberattacks. Today, IDS already employ ML algorithms to attain detection rates over 90% in specific circumstances [CISA2024]. Given the sheer amount of data to analyse, it is crucial for usability that false positive rates are kept extremely low, while effectiveness commands that false negative rates remain low as well, which is always a challenge and one that research must address upfront. Federated learning, due to its distributed nature, should be investigated for its scalability.

To prevent intrusion, one specific aspect to address in the security of the NCP is secured authentication solutions that must also scale efficiently to address the NCP's dynamic and distributed characteristics. Authentication solutions exist, but these must evolve to preserve both security and ease of use, minimizing friction for NCP users.

To this end, new multi-factor authentication (MFA) technologies incorporate biometrics, contextual awareness, and behavioural analytics to deliver strong and user-friendly authentication solutions. Zero-trust architectures, whose principles mandate continual authentication and authorization of every entity irrespective of its location, seem essential for the NCP, since they can guarantee secure interactions even in extremely dynamic settings. Blockchain-based decentralized identifiers (DIDs) could also facilitate scalable and secure authentication in an NCP context by removing dependence on centralized authorities. Furthermore, integrating anomaly-based intrusion detection with adaptive authentication is very important for the NCP, because it allows access controls to be dynamically modified in response to identified threats, thus significantly improving security.

In a nutshell, research and industry must be encouraged to develop, for dynamic and highly interconnected contexts such as the NCP, real-time, scalable, and adaptable technologies that can identify both known and undiscovered threats in extensive systems, as well as decentralized, context-sensitive authentication systems.

Secure critical infrastructure

Critical EU infrastructure, encompassing utilities, healthcare facilities, and transportation systems, constitutes the foundation of contemporary society. The interruption of the services such infrastructure provides can lead to significant societal and economic repercussions. As the NCP consolidates these systems into a cohesive, highly interconnected continuum, enhancing their cybersecurity is literally vital.

Critical infrastructure has been subjected to cyberattacks, ransomware, and state-sponsored cyber assaults for a long time, predominantly affecting the energy, healthcare, and water-management sectors [Zendra2023]. These cyberattacks generally exploit vulnerabilities in legacy systems, and interconnectivity to disrupt critical services or gain influence in geopolitical conflicts.

The EU must continue implementing regulatory frameworks that require the protection of essential infrastructure. This encompasses fundamental needs for cybersecurity measures, regular evaluations, and criteria for secure-by-design elements and services. The EU NIS2 Directive [NIS2-EU], effective in 2024, and the EU Cyber Resilience Act (CRA) [CRA-wiki] [CRA-EU] adopted in October 2024, are steps in the right direction. However, both require solid implementation measures to ensure compliance and effectiveness.

Mass cyberattacks can incapacitate centralized systems. To address this, the EU should make it mandatory that essential infrastructure integrate autonomous "archipelago" systems — i.e. self-sufficient components that can function independently during disruptions. One example is smart grids, which should incorporate localized microgrids capable of autonomously maintaining electricity delivery in their area during an attack.

Thanks to its nature of interconnected yet separate components and services, the NCP is in many ways suitable as a supporting infrastructure for these archipelagos, thanks to which systems should be compartmentalized to avert cascade failures. In a nutshell, EU regulations must mandate the design and implementation of infrastructure elements that can function independently.

In addition to regulation, emergency preparedness plans and drills should be conducted at EU level. Cyberattack simulations and coordinated exercises are crucial to prepare organizations for degraded-mode operations. EU-wide exercises such as Cyber Europe [CyberEU-ENISA] offer opportunity to evaluate resilience and response tactics across national boundaries. Such programmes should be augmented and adapted to address the

particular issues presented by the massively interconnected components and services of the NCP.

Liability

The involvement of software and hardware suppliers is crucial in cybersecurity. Ensuring their accountability for security vulnerabilities encourages higher standards in product security [Zendra2023]. The EU has already moved in that direction with the Cyber Resilience Act (CRA) and the NIS2 Directive, which stress the need for inherently secure technologies, and have providers of ICT system accountable for cybersecurity deficiencies, including insufficient safeguards or unresolved known vulnerabilities.

These efforts must be continued and their proper, concrete and effective implementation ensured.

References

CISA2024: 2023 Top Routinely Exploited Vulnerabilities. 12 Nov. 2024. https://www.cisa.gov/news-events/cybersecurity-advisories/aa24-317a

comparitech2024: Cybersecurity vulnerability (CVE) statistics and facts (2019-2024). https://www.comparitech.com/blog/information-security/cybersecurity-vulnerability-statistics/

CRA-EU: EU Cyber Resilience Act. https://digital-strategy.ec.europa.eu/en/policies/cyber-resilience-act

CRA-wiki: Cyber Resilience Act. Wikipedia. https://en.wikipedia.org/wiki/Cyber_Resilience_Act

CVE-MITRE: https://cve.mitre.org/

CVE-ORG: https://www.cve.org/

CVEdetails.com: https://www.cvedetails.com/browse-by-date.php

CyberEU-ENISA: Cyber Europe - Leading the way in cybersecurity preparedness. https://www.enisa.europa.eu/topics/training-and-exercises/cyber-exercises/cyber-europe-programme

CyberSecEval3: https://meta-llama.github.io/PurpleLlama/

Dalia2024: G. Dalia, C. A. Visaggio, A. D. Sorbo, and G. Canfora. SBOM ouverture: What we need and what we have. ARES '24: Proceedings of the 19th International Conference on Availability, Reliability and Security. Article No.: 116, pp. 1-9. https://doi.org/10.1145/3664476.3669975

HuggingFace: https://huggingface.co/

IBM-RaaS: What is ransomware as a service (RaaS) ? Jim Holdsworth, Matthew Kosinski. 5 September 2024. https://www.ibm.com/topics/ransomware-as-a-service

IBM2021: IBM Report: Cost of a Data Breach Hits Record High During Pandemic. Jul 28, 2021. https://newsroom.ibm.com/2021-07-28-IBM-Report-Cost-of-a-Data-Breach-Hits-Record-High-During-Pandemic

IBM2023: What's new in the 2023 Cost of a Data Breach report? Sarah Villavicencio. July 25, 2023. https://community.ibm.com/community/user/security/blogs/sarah-dudley/2023/07/25/costofadatabreach2023

Ladisa2023a: Taxonomy of Attacks on Open-Source Software Supply Chains. Piergiorgio Ladisa, Henrik Plate, Matias Martinez, Olivier Barais. 2023 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, US, 2023 pp. 1509-1526. https://arxiv.org/abs/2204.04008

Ladisa2023b: The Hitchhiker's Guide to Malicious Third-Party Dependencies. Piergiorgio Ladisa, Merve Sahin, Serena Elisa Ponta, Marco Rosa, Matias Martinez, Olivier Barais. 2023 Workshop on Software Supply Chain Offensive Research and Ecosystem Defenses, pp. 64-74. https://dl.acm.org/doi/pdf/10.1145/3605770.3625212

Liu2024: Formalizing and Benchmarking Prompt Injection Attacks and Defenses. Yupei Liu, Yuqi Jia, Runpeng Geng, Jinyuan Jia and Neil Zhenqiang Gong. 33rd USENIX Security Symposium (USENIX Security 24).

MistralAI: https://mistral.ai/

Monkam2024: A topological data analysis approach for detecting data poisoning attacks against machine learning based network intrusion detection systems. Galamo Monkam, Michael J. De Lucia, Nathaniel D. Bastian. Computers & Security, vol. 144, 09/2024, Elsevier.

NIS2-EU: Directive on measures for a high common level of cybersecurity across the Union (NIS2 Directive). https://digital-strategy.ec.europa.eu/en/policies/nis2-directive

NIST-SSDF: Secure Software Development Framework (SSDF). https://csrc.nist.gov/Projects/ssdf

Pasquini2024: Neural Exec: Learning (and Learning from) Execution Triggers for Prompt Injection Attacks. Dario Pasquini, Martin Strohmeier, Carmela Troncoso. 2024 Workshop on Artificial Intelligence and Security (AISec '24).

Qualys2024: 2023 Threat Landscape Year in Review: If Everything Is Critical, Nothing Is. Saeed Abbasi. January 4, 2024. https://blog.qualys.com/vulnerabilities-threat-research/2023/12/19/2023-threat-landscape-year-in-review-part-one

Raz2024: The good, the bad, and the ugly: Microsoft Copilot. hack.lu Security Conference 2024. https://archive.hack.lu/hack-lu-2024/talk/NNFQ3G/

SBOM-NTIA: https://www.ntia.gov/page/software-bill-materials

Sigstore: https://www.sigstore.dev/

Statista-Breaches2024: Data breaches worldwide - statistics & facts. Statista. https://www.statista.com/topics/11610/data-breaches-worldwide/#editorsPicks

Statista-CostCybercrime: https://www.statista.com/chart/28878/expected-cost-of-cybercrime-until-2027/

Wu2024: Fake News in Sheep's Clothing: Robust Fake News Detection Against LLM-Empowered Style Attacks. Jiaying Wu, Jiafeng Guo, Bryan Hooi. 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24).

Zendra2023: O. Zendra and B.Coppens. From cybercrime to cyberwarfare, nobody can overlook cybersecurity any more. In M. Duranton et al., editors, HiPEAC Vision 2023, pages 130-144, Jan 2023. DOI: 10.5281/zenodo.7461910

Sustainability

Recommendations for IT Sustainability

Validated life-cycle models for computing

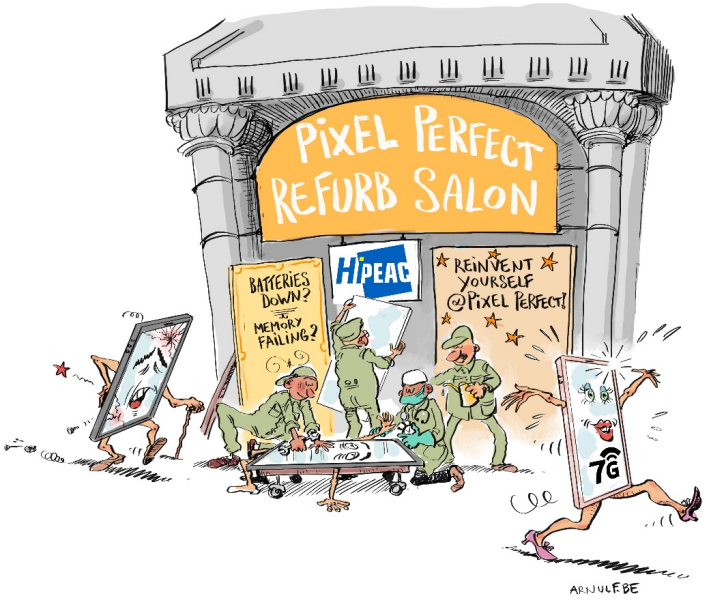
The information technology (IT) community should further develop validated life-cycle models for its own products and services. These models should comprehensively account for the total environmental impact of the production and disposal of the product, commonly known as embodied emissions. This includes the impact of mining, water usage, the use of chemicals in production, and end-of-life processing.

In addition, the model should also estimate operational emissions. This information should be included in a digital product passport (DPP) containing information about the environmental impact comparable with the information on pre-packaged food products or power-efficiency information on household appliances. This information will help consumers to make informed choices about sustainability. The digital envelope of a device should be able to return this information to e.g. an orchestrator to enable it to select the services that optimize the sustainability requirements specified by the owner of the orchestrator.

Sustainability-focused design methodologies and business models

Detailed life-cycle models will help designers make the most effective eco-design decisions. To be effective, design tools should automatically include the environmental impact of the components and technologies used in the design, without putting the burden on the designer. Incorporating repairability, reusability, recyclability, and end-of-life processing considerations from the beginning of the product development process will also lower the environmental impact of the final design.

Inevitably, reducing the environmental impact of a product will have an impact on companies' business models. Designing products that last longer will reduce sales of new products and hence lower the profitability of the company. This can only be mitigated by developing new business models, based on extra services: maintenance, repair, disposal, … up to completely replacing the ownership of hardware by a service contract. The goal should be to bring services to the market with the least environmental impact possible (which in practice means with the least amount of hardware, and the lowest power consumption).



Introduction

Life-cycle assessment (LCA) is an analysis technique that provides tools and frameworks for measuring and managing the environmental footprint of products and services. An LCA analyses the impact of the complete life cycle (cradle-to-grave), from raw materials extraction, via manufacturing, transportation, and usage, to waste disposal. It measures the cumulative environmental effect of the whole life cycle.

It is crucial to consider the complete life cycle to avoid a situation in which a footprint reduction in one phase is cancelled out by a footprint increase in another phase, in the worst case leading to an increase in the total footprint. An LCA is a complex analysis because of the complexity of digital products and services, which are built from components that are sourced globally, all of which need to be analysed to determine their combined environmental footprint.

Another difficulty with an LCA is that the post-production impact (i.e. after it leaves the factory) is difficult to model because it depends on the use and disposal, and that these two aspects are difficult to model because they are controlled by the user. Obviously, a car that is used as a taxi will have a larger operational footprint than a car that is only used occasionally, but at the same time, the total environmental impact per kilometre driven may be lower for the taxi. A fridge that ends up in a landfill will have a different environmental footprint to one that is properly recycled.

Because most consumers do not understand how modern products are built and how services actually work, it is almost impossible for them to assess their environmental footprint. Even for experts, it is difficult to predict the environmental footprint without doing a detailed analysis, and such an analysis regularly leads to counterintuitive conclusions (e.g. that replacing a working device by a more power efficient device is seldom better for the environment than continuing to use the less power-efficient device).

The difficulty in fully understanding the real environmental impact of our actions, and the fact that a thorough LCA sometimes leads to counterintuitive conclusions leads to confusion in the general public, especially when they learn that the behaviour that they thought was beneficial for the environment turns out to be ineffective, or even harmful in some cases. This confusion provides fertile ground for environmental sceptics to convince

the public that sustainability is a scam and to use social media to amplify their messages. It also makes it more difficult to detect greenwashing.

What about climate change?

From the scientific view, there is no doubt that the current global environmental footprint is too high for the carrying capacity of the planet, as illustrated by the yearly earth overshoot day [EarthOvershoot]. The world uses in seven months everything the planet can regenerate in one year. The remaining five months, we are depleting the planetary resources. The EU overshoot day was 3 May 2024, meaning that the EU uses in four months everything the EU can regenerate in one year.

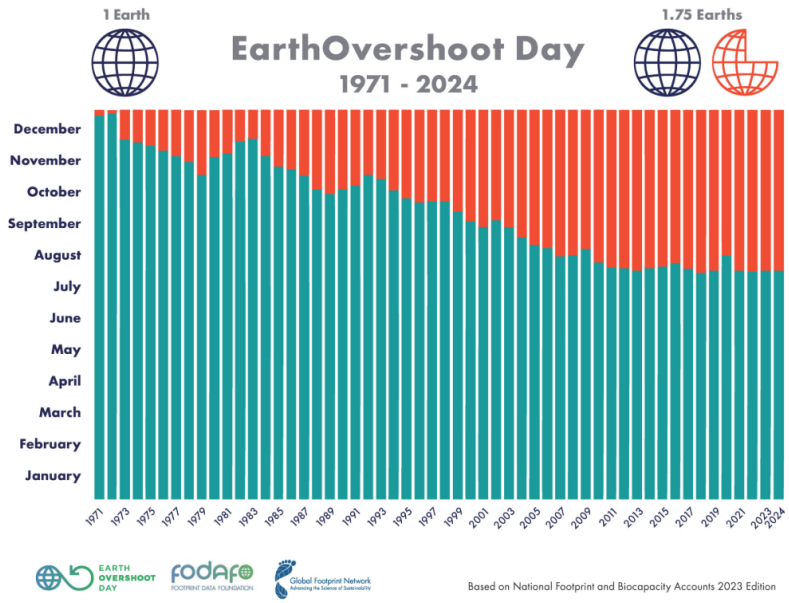


Figure 1: Earth overshoot day 2024 was 1 August 2024

One can disagree on the root causes: overconsumption, overpopulation, inefficiencies, … but not on the effects which are observable: climate change, loss of biodiversity, … If not mitigated, science predicts that there will be serious implications for the future generations.

The most important action humanity can take to stop climate change is to reduce the emissions of greenhouse gases (GHG). The most important ones are carbon dioxide (CO_2) (caused by the use of fossil fuels, deforestation, …), methane (caused by livestock, oil and gas extraction, …), nitrous oxide (caused by fertilizers, fossil fuels, industry, …) and fluorinated gases (caused by industrial processes, cooling, electronics manufacturing, …).

CO_2 has the highest contribution due to its sheer volume, but the other gasses are much more potent GHGs, and the electronics industry is a source of fluorinated-gas emissions. To simplify the maths, all GHGs are commonly expressed as their equivalent in CO_2 emissions, called CO_2e. This is convenient but also misleading, in the sense that techniques to extract CO_2 from the air, like planting trees, work for real CO_2, but does not work for the CO_2e that is caused by e.g. methane.

According to international agreements, emissions should be cut by 45% by 2030, compared to 2010 levels, and the world should reach net zero by 2050. Unfortunately, despite all our efforts of the last 20 years, global GHG emissions are still increasing, albeit at a lower rate than 20 years ago. With current commitments, the emissions in 2035 will hardly be lower

than the emissions in 2020, and the gap between the path towards net zero emissions in 2050 is quickly widening.

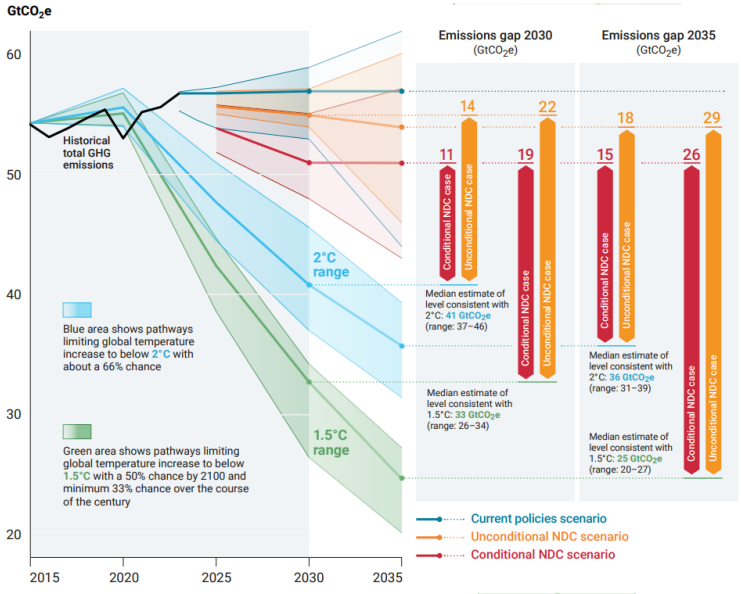


Figure 2: Global total GHG emissions in 2030, 2035 and 2050, and estimated gaps under different scenarios (unconditional national determined contribution (NDC) scenario = current committed efforts). Conditional NDC: contributions that are conditional, i.e. they depend on factors that are unsure like the passing of laws in local parliaments.

Given the fact that the early emissions reduction would normally consist of low-hanging fruit, there is little hope that decarbonization will be easier in 2035 than it is in 2025. The fact that the newly elected president of the US will actively promote fossil fuels until 2030, along with the quickly growing energy consumption by AI data centres in the US, might slow down the emission reductions in the coming years.

Authors like Vaclav Smil `[VaclavSmil]` argue that fast decarbonization of the global economy over the next 25 years is unlikely because the world is built of concrete and steel, both of which require a huge amount of energy to produce, and for which there are currently no economically viable alternatives that can be scaled up to the required volume. In addition, industry needs the molecules of fossil fuels in the chemical industry to produce e.g. plastics and fertilizer, two other cornerstones of modern society. Furthermore, major industrial capital investments often have a time horizon of two decades. Hence, fossil fuel-based industrial facilities that are built today will still be in use in 2045. The conclusion is that, given that it took more than a century to build a fossil fuel-based industry, it is very unlikely that it can be reconverted into a fossil-fuel-free one in two decades.

What about the IT industry?

Obviously, the IT industry also contributes to the global GHG emissions. The most widespread comparison is that the emissions of the IT sector are comparable with those of aviation (2%). This comparison suggests that the IT industry is a polluting industry and devastating for the planet.

Given the importance of IT in the modern world, one could also say that it is ‘only 2%’, and the IT industry helps the other industries to reduce emissions (optimized processes, cleaner transportation, less business travel, …). The fact is that (i) we do not know for sure whether this 2% is high or low compared to the benefits of using IT, and (ii) we do not know how and by how much the footprint of the IT-industry could be reduced without losing its main

economic and societal benefits. Furthermore, it is dangerous to make any statement about concrete situations without first making a solid LCA about it to make it evidence based. Extrapolating from similar situations is tempting, but no two situations are identical and only an LCA analysis can provide certainty.

Given the complexity of an LCA analysis, some organizations publish general recommendations, such as those published by [Ericsson]:

- Use your smartphone or other ICT devices longer before upgrading
- Make sure you recycle or reuse ICT equipment
- Consume digital services on smaller devices
- Charge the batteries with electricity from renewable sources
- Avoid buying more ICT devices than you have time for (pass unused devices on)
- Show your suppliers that their footprint matters to you
- Buy your digital devices and services from companies that have Science based Targets
- Use ICT services that help to reduce carbon emissions

These may help some high-level decisions, but they won't help somebody deciding which smartphone to choose in a shop. Furthermore, they are not quantitative, and do not allow estimates of what the difference in emissions is.

To give a few examples: few people are aware that a non-rechargeable battery requires 100 times more energy to produce than the energy it stores, that mobile devices can cause up to 10x more emissions to produce than the operational emissions over their entire life cycle (which explains that keeping a power inefficient one is often more sustainable than replacing it with a power efficient one), that five ChatGPT questions consume the same amount of energy as stored in a fully charged iPhone 15 battery.

This leads to two recommendations:

Validated lifecycle models for computing

A first recommendation is that the IT community should develop validated life-cycle models for its own products and services. The life-cycle models should not be contested (hence "validated") and be developed by sustainability experts based on solid scientific evidence. These models should comprehensively account for the total environmental impact of the production and disposal of the product, commonly known as embodied emissions. This includes the impact of mining, water usage, the use of chemicals in production, and end-of-life processing.

In addition, the model should also estimate operational emissions, which obviously depend on the usage of the product and the environmental impact of the energy used. This information should be included in a digital product passport (DPP) containing information about embodied energy, operational energy, mining, water usage, and chemical impacts comparable with the information on pre-packaged food products or power efficiency information on household appliances. This information will help consumers to make informed choices about sustainability.

For digital products (IT services), the product should be able to return this information to the user. This will allow e.g. an orchestrator to select the services that optimize the sustainability requirements specified by the owner of the orchestrator. For services, this information might also be dynamic: the service request during the day might have a lower impact than during the night if the carbon intensity of the energy consumed was lower during the day. Obviously keeping track of all this information will have an environmental

cost itself too, and it will be important to prove that the environmental benefit of keeping track of it exceeds its environmental cost.

Sustainability-focused design methodologies and business models

Once the life-cycle models are available, and the environmental impact of product and services has been modelled, designers can optimize their designs to lower the environmental impact. They can do this to make their products more environmentally friendly, to make them more attractive to customers who care about the environment, or to make them compliant with local regulations.

The detailed life-cycle models will help the designer to make the most effective design decisions, and to ensure that environmental impact is one of the design criteria to optimize. This is already common practice in the building industry where designers routinely base their designs on low-carbon construction materials, which in turn has stimulated innovation in companies that produce construction materials.

Questions which should be very easy to answer include e.g. whether it is better for the environment to power a device with a battery, or with an adaptor from the grid, whether adding an extra cache level in a computing system is better or worse for the environment, and whether executing a workload at the edge is environmentally better than execute it in a cloud data centre. Such questions can only be answered by a solid LCA, and the answer will depend on the usage, the location and the domain in which the technology is applied.

To be effective, design tools should automatically include the environmental impact of the components and technologies used in the design, without putting the burden on the designer. Incorporating repairability, reusability, recyclability, and end-of-life processing considerations from the beginning of the product development process will also lower the environmental impact of the final design.

Inevitably, reducing the environmental impact of a product will have an impact on companies' business models. Designing products that last longer will reduce sales of new products and hence lower the profitability of the company. This could be mitigated by marketing: products that last longer can also be sold at a higher price point.

Furthermore, new services could be built around the life cycle of a product: maintenance, repair, disposal. Such services might create opportunities to build a loyal relationship between the vendor and the customer; when the product is beyond repair, the vendor can immediately propose replacing it, and hence not lose the customer to the competition. The computing industry could learn from industries that already work like this (cars, household appliances, heating and cooling systems, alarm systems, …).

Another option is to no longer sell the hardware, but a service based on the hardware. This leads to a high startup cost, but a stable revenue stream afterwards. In any event, the computing industry will have to change its business models to become sustainable.

References

`BolACACES24`: Bol, David. (2024). "ICT and environmental sustainability", course at ACACES 2024. https://www.hipeac.net/acaces/2024/#/program/courses/103/

`DeBosschereBlouet`: De Bosschere, K., & Blouet, P. (2024). "What does it mean to be sustainable?", HiPEAC Vision 2024, Rationale. https://doi.org/10.5281/zenodo.10875127

`EarthOvershoot`: Earth Overshoot day 2024 fell on August 1st, https://overshoot.footprintnetwork.org/

EPRI: Electric Power Research Institute (EPRI). (2024). Powering Intelligence: Analyzing Artificial Intelligence and Data Center Energy Consumption https://www.epri.com/research/products/3002028905

Ericsson: Ericsson. A quick guide to your digital carbon footprint, 2020, https://www.ericsson.com/en/reports-and-papers/industrylab/reports/a-quick-guide-to-your-digital-carbon-footprint

iPhone15Wikipedia: https://en.wikipedia.org/wiki/IPhone_15

UNEP: United Nations Environment Programme. Emissions Gap Report 2024. United Nations Environment Programme, 2024, https://www.unep.org/resources/emissions-gap-report-2024

VaclavSmil: Smill, Vaclav. How the World Really Works: The Science Behind How We Got Here and Where We're Going. Viking, 2022.

Process

The HiPEAC Vision analyses current trends that have an impact on the high-performance, edge and cloud computing and related communities. It formulates technical, methodological, standardization and policy recommendations to the HiPEAC community at large, and to policy makers.

The content is based on information collected through a number of channels.

- A survey circulated to all HiPEAC members.
- Meetings with teachers and industrial partners at the ACACES 2024 summer school.
- A consultation meeting on open source.
- Participation in tens of conferences and workshops on relevant themes for the HiPEAC Vision.
- Hundreds of informal discussions with experts from around the globe.
- Eight in-person editorial board meetings and eight videoconference meetings.
- Several coordination meetings with other organizations, such as ETP4HPC, ECSO, CHIPS JU, AIOT, ADRA, BDVA, SNS, NESSI, FIWARE, Destination Earth, Eclipse Foundation, etc…
- Feedback from presentations of the HiPEAC Vision 2024 at several conferences and workshops and from exchanges with DG CNECT.

Acknowledgements

This document is based on the valuable input of HiPEAC members. The editorial board, composed of Marc Duranton (CEA), Koen De Bosschere (Ghent University), Christian Gamrat (CEA), Paul Carpenter (BSC), Harm Munk, Charles Robinson (Thales), Tullio Vardanega (University of Padua) and Olivier Zendra (INRIA), would like to thank the following contributors: Adam Mackay (QA-System), Alessandra Bagnato (Softeam), Carles Hernández Luz (Universitat Politècnica de València), Claudio Pastrone (LINKS Foundation), Claudio Sassanelli (Politecnico di Bari), Djamila Aouada (Université du Luxembourg), Hugo Daniel Macedo (Aarhus University), Marcus Völp (Université du Luxembourg), Michael Henshaw (Loughborough University), Miklós Györffi (European Parliament), Paul Pop (Technical University of Denmark), Peter Gorm Larsen (Aarhus University), Peter Popov (City University London), Rajendra Akerkar (Western Norway Research Institute), Sanaz Mahmoodi Takaghaj (Pennsylvania State University), Thorsten Weyer (Technical Hochschule Mittelhessen).

The editorial board would also like to thank Eneko Illarramendi (Ghent University), Madeleine Gray and Vicky Wandels (Ghent University) for their useful comments and their support. Special thanks to the Belgian comic artist Arnulf for his cartoons, which provide a characteristically tongue-in-cheek take on the content of the HiPEAC Vision.