SafeComp2025

The Next Computing Paradigm: where the computing continuum and AI meet the physical world

- Challenges for trustability, safety and security

September 10th, 2025

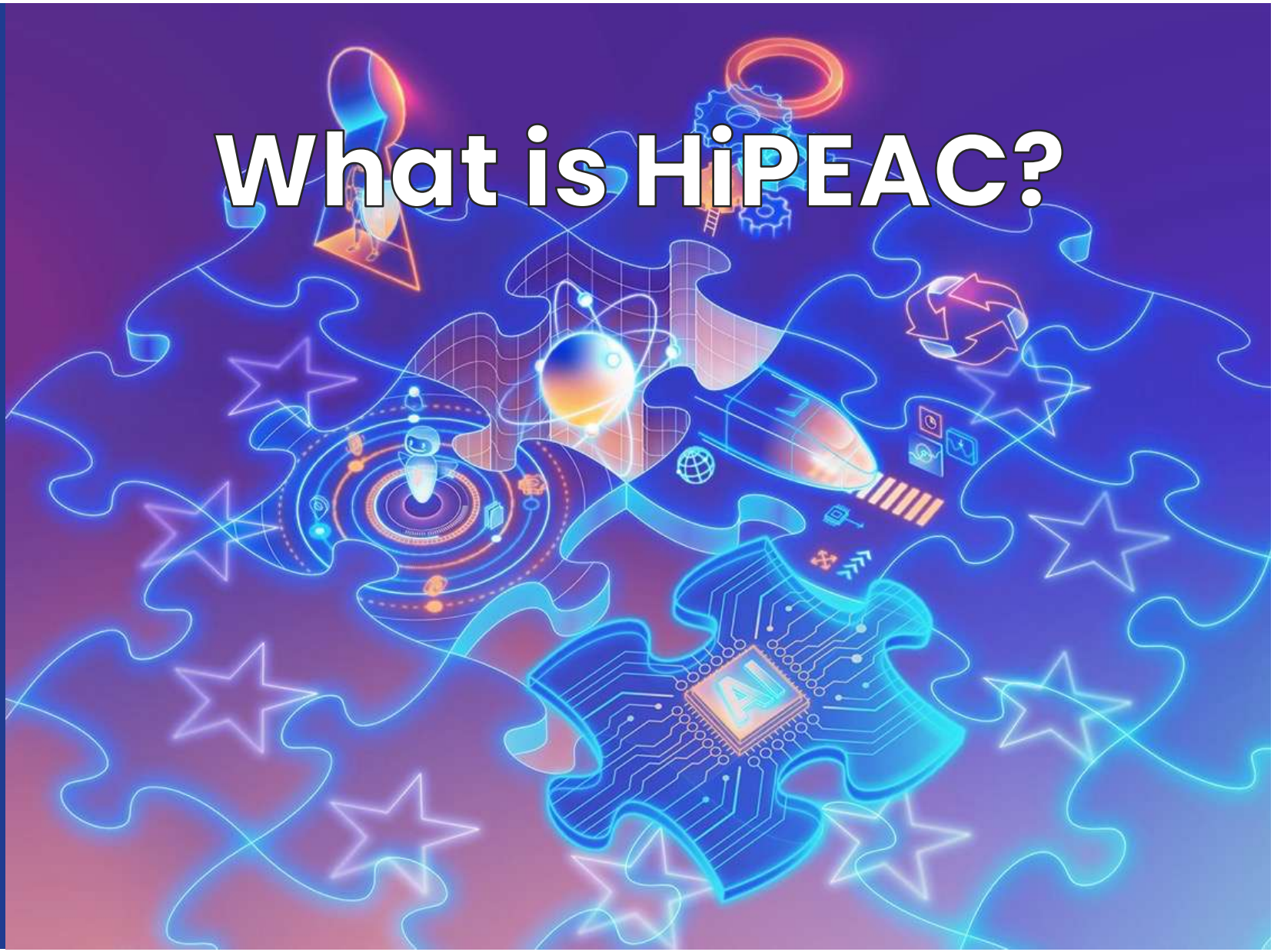
Marc Duranton

Senior fellow

CEA (France) and HiPEAC

HiPEAC Vision 2025

HIGH PERFORMANCE, EDGE
AND CLOUD COMPUTING



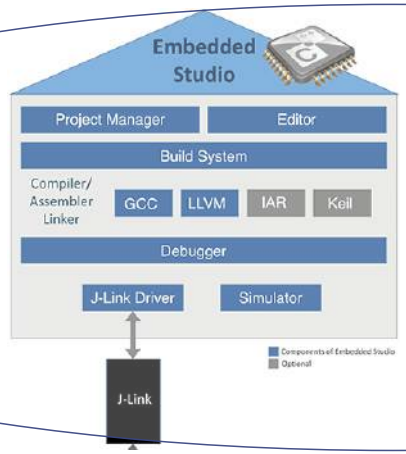
Funded by
the European Union

What is the HiPEAC community?

Applications

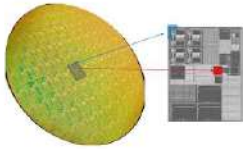
Industry Energy Mobility Health Digital Society Agrifood ....

HiPEAC



European Processor Initiative

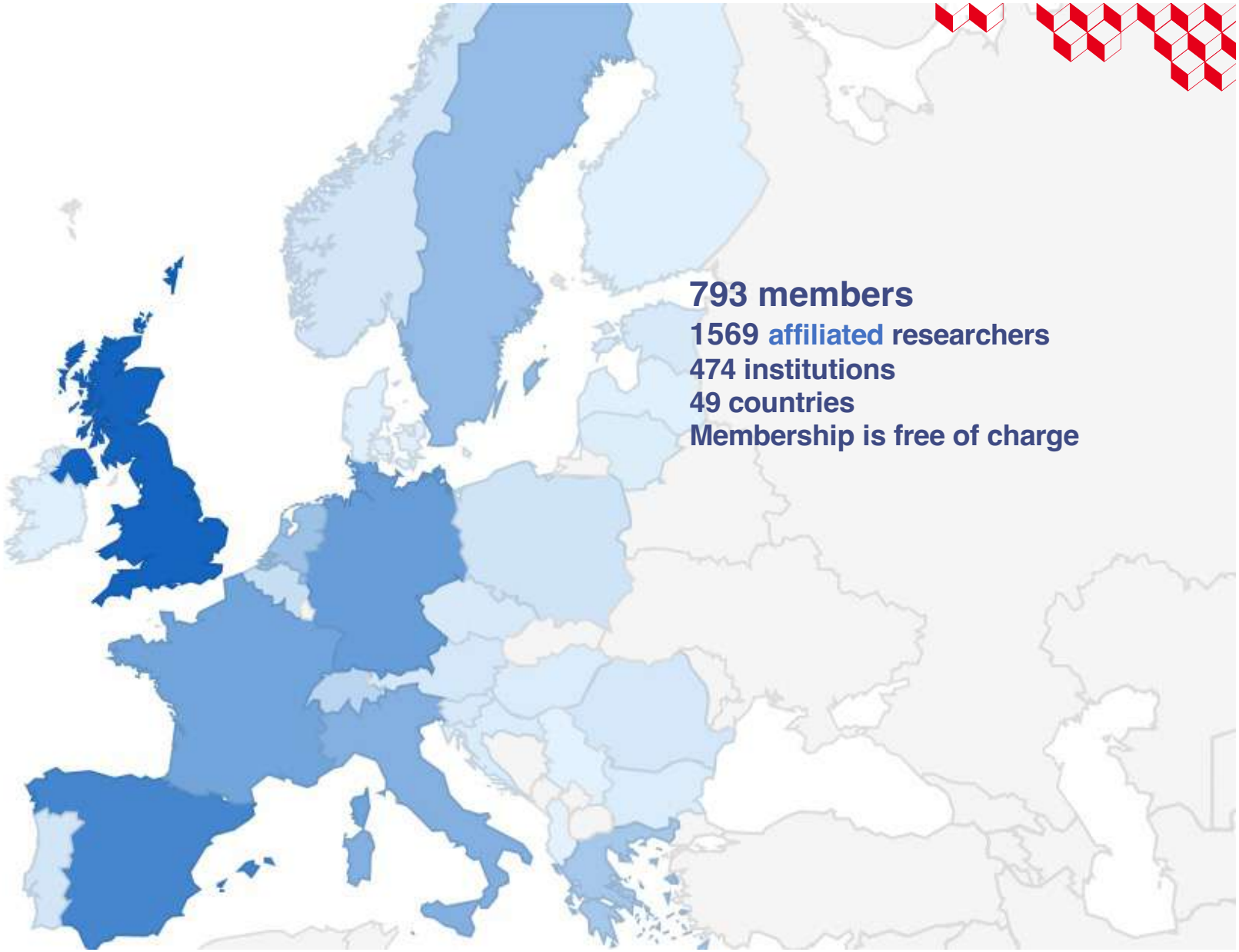
Technology



Process Technologies
Equipment, Materials and Manufacturing

HiPEAC

Membership



793 members
1569 affiliated researchers
474 institutions
49 countries
Membership is free of charge

HiPEAC conference

HiPEAC conference

ACACES 2025

About FAQ Venue and travel

Slot 1

From Edge to HPC and back: programming the computing continuum for scientific applications

Nicola Ferrier
Argonne National Laboratory, United States

Robust Systems in the 21st Century

Subhasish Mitra
Stanford University, United States

Trust, Threat, and AI: Security Challenges in Generative & Collaborative Learning

Ahmad-Reza Sadeghi
TU Darmstadt, Germany

Slot 2

A Modern Perspective on the Role and Design of Operating Systems

Timothy Roscoe
ETH Zurich, Switzerland

Chameleon: Research Infrastructure in Edge to Cloud Continuum

Michael Sherman
The University of Chicago, United States

Kate Keahey
Argonne National Laboratory, United States

Dataflow-based modeling and design

Shuvra Bhattacharyya
University of Maryland, United States

Slot 3

Entrepreneurship: from idea to unicorn

Wim De Wispelaere
Ghent University, Belgium

In-Memory Computing for AI Acceleration

Irem Boybat
IBM Research Zurich, Switzerland

Toward Sustainable Computing

Adrian Friday
Lancaster University, United Kingdom

Slot 4

Designing RISC-V energy-efficient edge-node System-On-Chips

Pasquale Davide Schiavone
ETH Zurich, Switzerland

Prompt Engineering: Leveraging Large Language Models

Maximillian Moundas
Vanderbilt University, United States

Quantum Computing: From Software, Architecture, and Systems Perspective

Poulami Das
The University of Texas at Austin, United States

What is the HiPEAC Vision?

HiPEAC Vision 2025

HIGH PERFORMANCE, EDGE AND CLOUD COMPUTING

The ***HiPEAC Vision*** is a deliverable of the **c**oordination and **s**upport **a**ction on **H**igh **P**erformance, **E**dge **A**nd **C**loud computing (previously: **Hi**gh **P**erformance and **E**mbedded **A**rchitecture and **C**ompilation), a European network of almost 2,500 world-class computing systems researchers, industry representatives and students.

The HiPEAC Vision shows the *trends, technology evolutions* and limitations and ***position of Europe*** in the domain of computing (hardware and software) and provides a ***long-term view*** and recommendations to the HiPEAC community at large.

2008 2009 2011 2013 2015 2017 2019 2021 2023 2024

January 2025

January 2025 version is available at:

https://www.hipeac.net/vision/

HiPEAC Vision 2025 overview

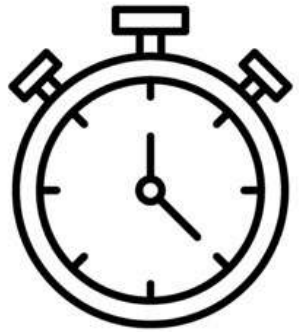
We are ***more and more*** in races:

against **time**

- because of external constraints – e.g. global warming
- because the evolution of technology is very fast (e.g. AI)

with **the rest of the world**

- on multiple aspects (economy, reducing dependances, ethics, …)



The HiPEAC Vision has identified specific races:

- Race for **artificial intelligence**
- Race for innovative and new **hardware** (including quantum)
- Race for **cybersecurity**
- Race for **sustainability**
- Race for **sovereignty**

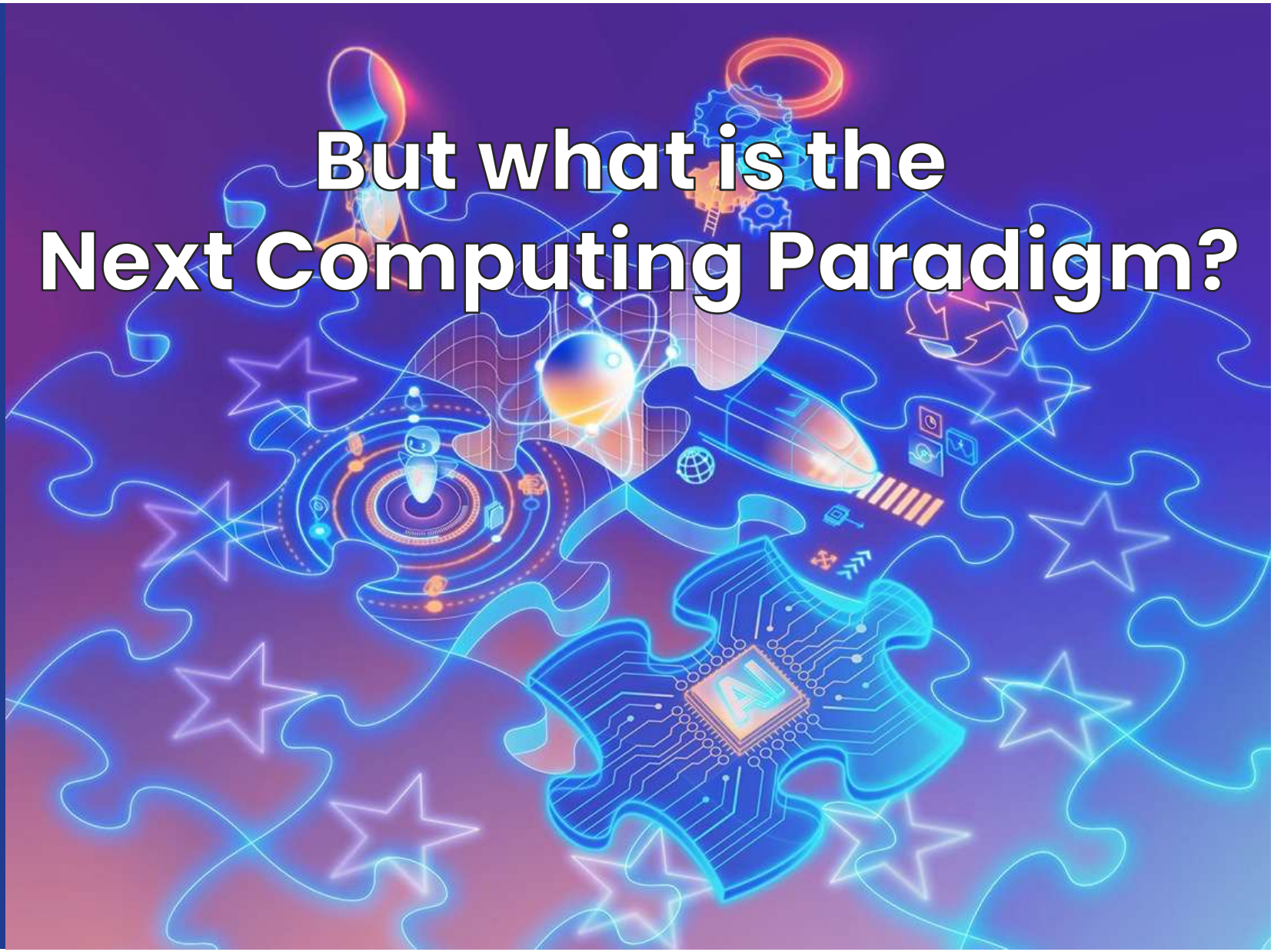
And a set of global recommendations promoting **holistic approaches and breaking silos**

= > the "next web", the "**next computing paradigm**" – federating R&D efforts towards a continuum of computing, **as a way to be "back in the race": therefore the *HiPEAC Vision 2025* focus is**

HOW TO ENABLE THE "NEXT COMPUTING PARADIGM"

HiPEAC Vision 2025

HIGH PERFORMANCE, EDGE AND CLOUD COMPUTING



Funded by the European Union

HiPEAC Vision 2025

HIGH PERFORMANCE, EDGE
AND CLOUD COMPUTING

THE NEXT COMPUTING PARADIGM

Funded by
the European Union

Understanding the Next Computing Paradigm (NCP)

Imagine a world where your computer applications are not just programs installed on your device but are ***dynamic collections of services*** that can adapt to your needs in real-time. This is the essence of the Next Computing Paradigm (NCP).

- **Services are anything from software functions to hardware capabilities**
- They are **orchestrated by an intelligent system** which selects the services from your own specific requirements.
- **Taking into account non-functional properties** like time, localization, privacy, cost, security, …..

Image generated by ChatGPT

What are the key Ingredients? - 1

Applications are defined **as (dynamic) composition of various services** (XaaS)

The services are selected and **orchestrated** together :

- The orchestration is done in several steps:
  - Decomposition of the application into “elementary” services
    - This can be done hierarchically
  - Discovery of available services
    - This can be helped by a trusted directory (similar to Google for web searches) for helping in governance and mediation
  - **Selection of service according to defined set of criteria** (performance, trust, cost, privacy, locality, latency, security, reliability, energy, possibility of migration, already successfully used, certified, green, ….)
  - Secure activation and authentication of the services, with guaranties
  - Orchestration of the services to perform the requested application
  - (for business model: reward in exchange of the service: €, reward for usage of local resources, contribution to data pool, …)

Image generated by ChatGPT

Should be
hardware enforced:

- Virtualization
- Silo, trust zones
- Cryptographic protocols
- (Lightweight) authentication
- …

What are the key Ingredients? - 2

- Independence of the services (software) and hardware on which it is executed
  - Code (*not only data*) can migrate from one location to another
    - like with *Web Assembly – WASM*
    - Containers
  - Hardware (processing, acceleration, storage) can also be considered as services and they have properties (processing power, QoS, energy consumption, security, locality, …)
- The mapping service/resource process is part of the global orchestration mechanism
  - Problem of assessing ***dynamically*** the performance of a task on a particular hardware (digital twin, trusted directory, …): progressive introduction following innovations

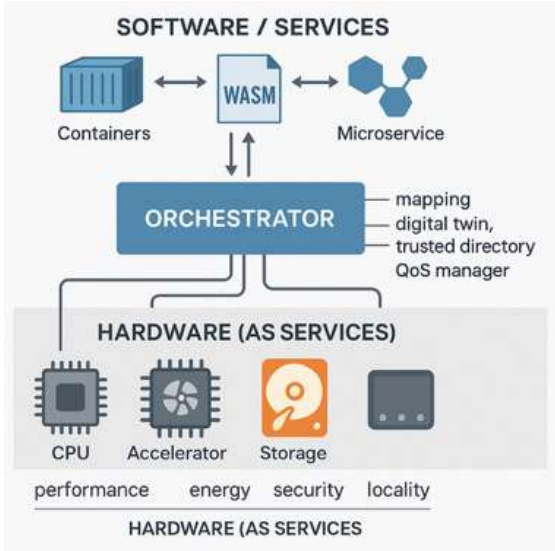


Image generated by ChatGPT

Device, Edge to cloud continuity “continuum”

This mechanism can be seen at multiple scales, local or distributed but with compatibility between all scales

- Various functions in a core (executed locally on a compute core): ex: Python + libraries
- Specific accelerators as chiplets orchestrated by CPU in a chiplet+interposer solution
- Resources sharing between devices (“*devices are aggregated at the system layer to form a super device, allowing flexible scaling of device hardware capabilities*”)
- Resource sharing at company/premise level
- Aggregation of data centers (e.g. like for Destination Earth program)

Device, Edge to cloud continuity “continuum”

This mechanism can be seen at mu
compatibility between all scales

- Various functions in a core (exec
  Python + libraries
- Specific accelerators as chiplets
  chiplet+interposer solution
- Resources sharing between devices (“*e aggregated at the system layer to form a super device, allowing flexible scaling of device hardware capabilities*”)
- Resource sharing at company/premise level
- Aggregation of data centers (e.g. like for Destination Earth program)

“users can integrate capabilities of their various smart devices, implementing ultra-fast connection, capability collaboration, and resource sharing among them. This way, services can be seamlessly transferred to the most suitable device, delivering smooth all-scenario experience.”

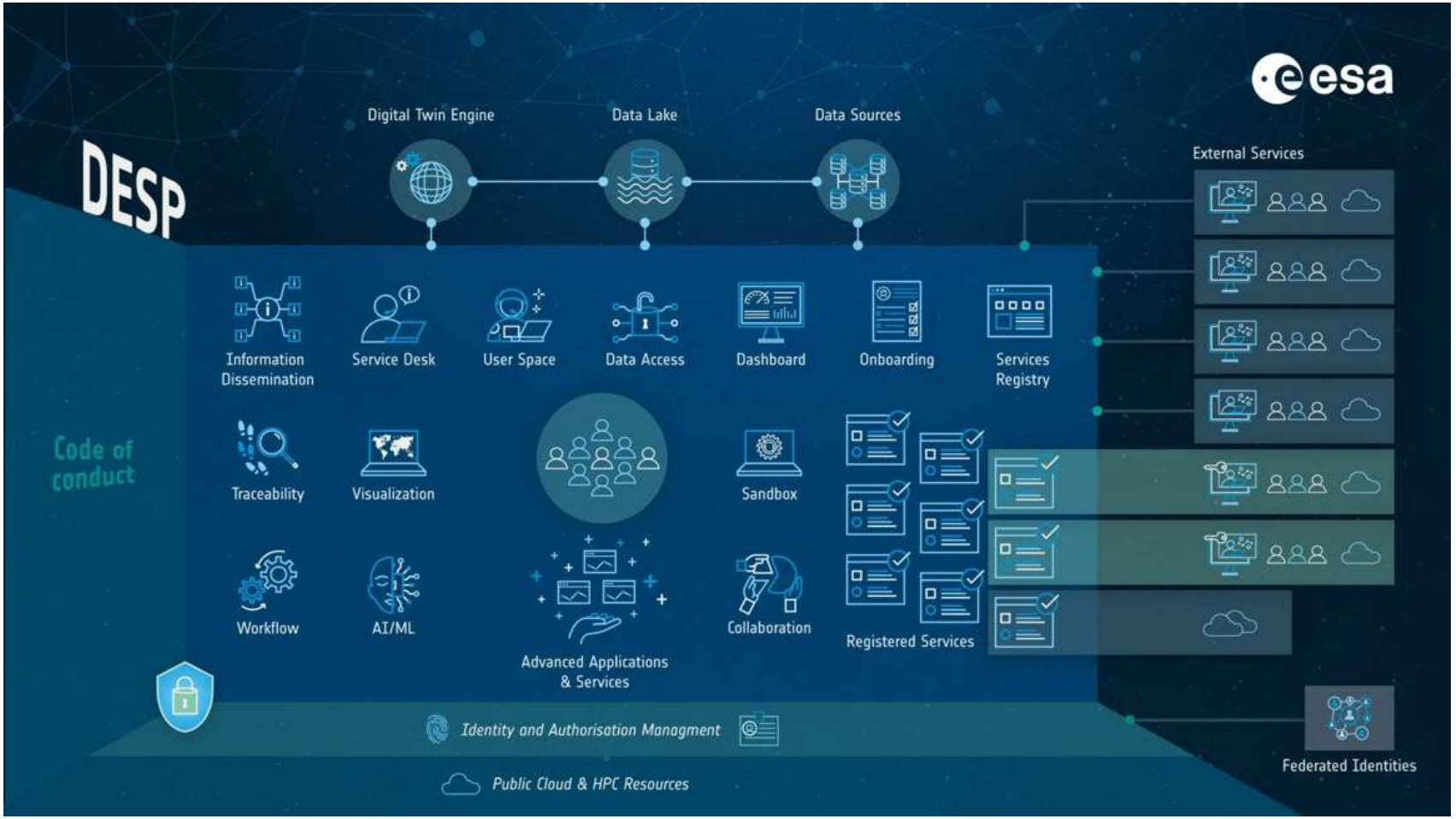
From harmonyos-guides-v3

oniro

OpenHarmony

Destination Earth program

More details at https://destination-earth.eu/ and https://destine.ecmwf.int/artificial-intelligence-in-destine-the-explainer/



From https://destination-earth.eu/wp-content/uploads/2023/04/DestinE-Core-Service-visual-min-2048x1152.webp

NCP enhances efficiency across ***business cases***

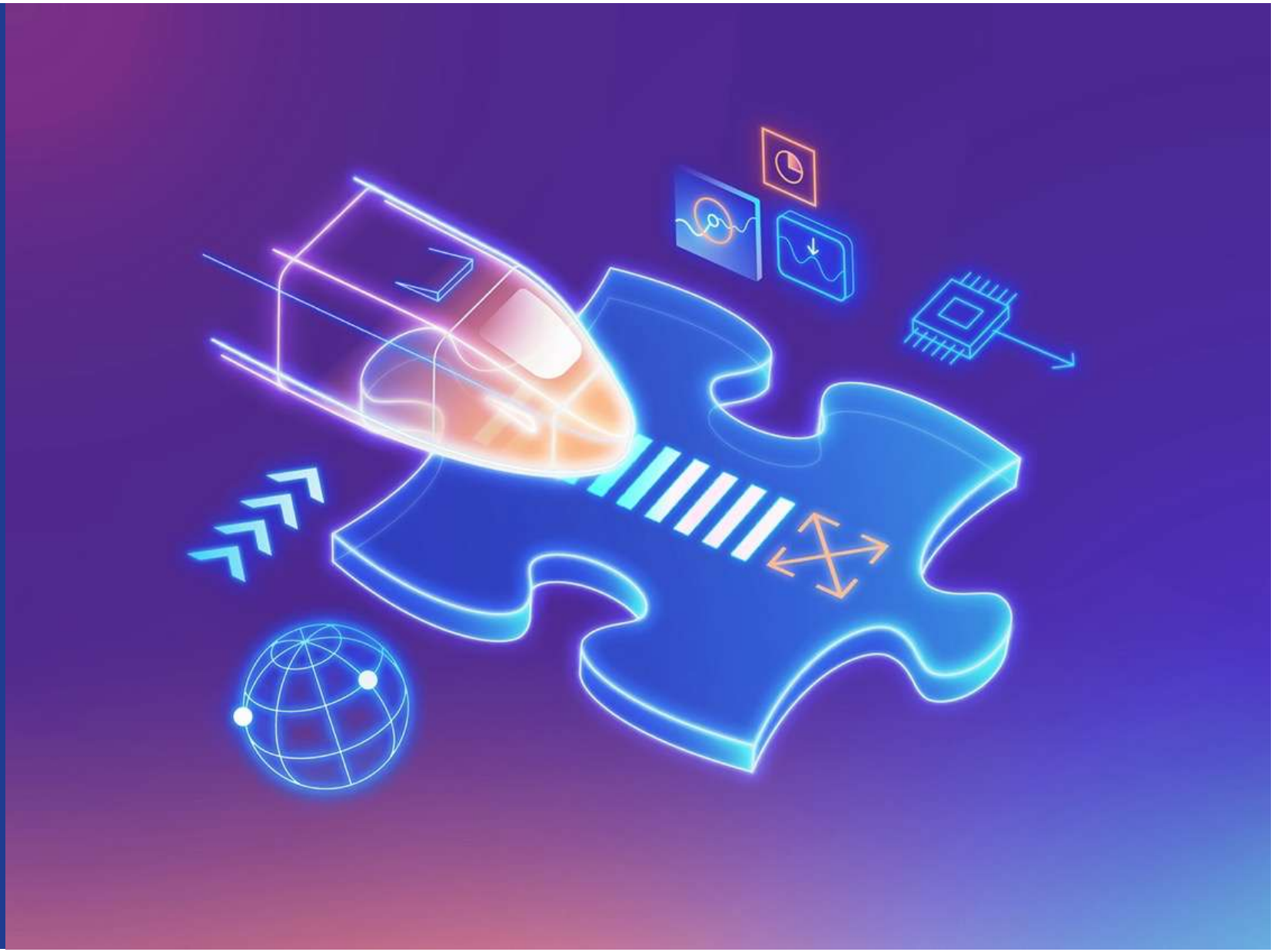
- **Dynamic Resource Allocation:** By selecting and orchestrating services in real-time, the NCP ensures that resources are used where they are most needed, *decreasing the global cost* of operation.
- **Scalability:** Businesses can scale services up or down without significant infrastructure changes, adapting quickly to market demands.
- **Interoperability:** The NCP's ability to integrate disparate systems and services leads to a more cohesive operation, which can significantly enhance efficiency across various industries.
- **Security and Compliance:** Communication are done only when local resources are not enough; Dynamic reconfiguration allowing operation even if some services are down; Enhanced security protocols ensure data integrity and compliance with regulations, reducing risks.
- **Reducing obsolescence**: By combining resources from existing systems in a “meta” systems, NCP allow to increase the usable lifetime of devices.

HiPEAC Vision 2025

HIGH PERFORMANCE, EDGE AND CLOUD COMPUTING

NCP and CYBER-PHYSICAL SYSTEMS

Funded by the European Union



Some observations leading to the "next computing paradigm"

Data & interconnected machines

- The **"NCP"** will have to **integrate the "web of machines" with the "web of humans"**.

Large scale safety-critical systems and digital twins

- Systems will be **increasingly intertwined with the physical world.**
- **Real-world constraints** such as **safety, real time and location** will need to be considered.
- **Digital twins -** to model past, present and future

The spatial web

- Systems will also be **spatial**: we will **go/are into the computing continuum (ubiquitous computing).**
  - **Properties depend on location and time (4D).**
  - Current protocols should be extended to support requirements
    - E.g. IEEE P2874 for spatial web

Some observations leading to the "next computing paradigm"

Data & interconnected machines

- The **"NCP"** will have to **integrate**
  machines" with the "web of hu

Large scale safety-critical systems and

- Systems will be **increasingly int**
  physical world.
- **Real-world constraints** such as
  location will need to be consider
- **Digital twins -** to model past,

The spatial web

- Systems will also be **spati**
  computing continuum **quito**
  - **Properties depend on l** **ation and time**
  - Current protocols sho**ld** be extended to support requirements
    - E.g. IEEE P2874 for spatial web

IEEE SA STANDARDS ASSOCIATION

IEEE P2874 Spatial Web, Working Group

Home | Meetings | Members

Title: Standard for Spatial Web Protocol, Architecture and Governance

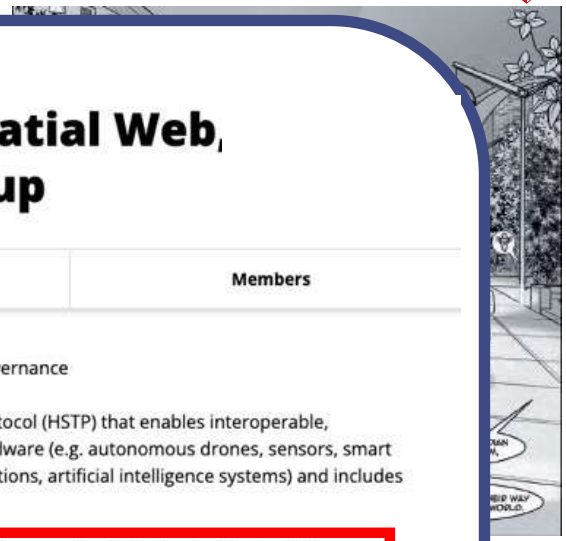
Scope: This standard describes a Hyperspace Transaction Protocol (HSTP) that enables interoperable, semantically compatible connections between connected hardware (e.g. autonomous drones, sensors, smart devices, robots) and software (e.g. services, platforms, applications, artificial intelligence systems) and includes specifications for:

1) a spatial range query format and response language for requesting data about objects within a dimensional range (spatial, temperature, pressure, motion) and their content.

2) a semantic data ontology schema for describing objects, relations, and actions in a standardized way

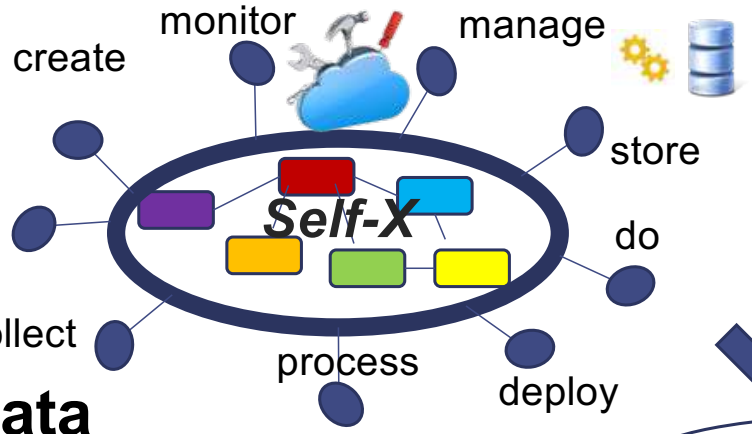
3) a verifiable credentialing and certification method for permissioning create, retrieve, update, and delete (CRUD) access to devices, locations, users, and data; and

4) a human and machine-readable contracting language that enables the expression and automated execution of legal, financial and physical activities.



Distributed Intelligence via Sharing and Coordination of Resources Across Heterogeneous, Connected, Locally Managed Devices

- ***Dynamic*** construction of applications from **distributed services** (local or remote)
- Accounting for ***non-functional requirements*** (time, latency, locality, energy, privacy, cost, …)
- Using ***high level abstractions***
- ***Trustable and secure*** (authentication, encrypted, …)
- Services (and resources) ***where*** they are the most efficient
- Built on top of existing technologies (e.g. IP, REST, IEEE P2874, …)



*OS: **O**rchestrating **S**ystem

Cyber world

Local, distributed, decentralized resources (swarm, federated, fog, ..)

Data

Actions

Trusted computing base

Secure Exchanges

Build on a trustable, distributed, smart **meta-level "OS*"**

Company/private data

***E**verything **a**s **a** **S**ervice*

Running on edge hardware

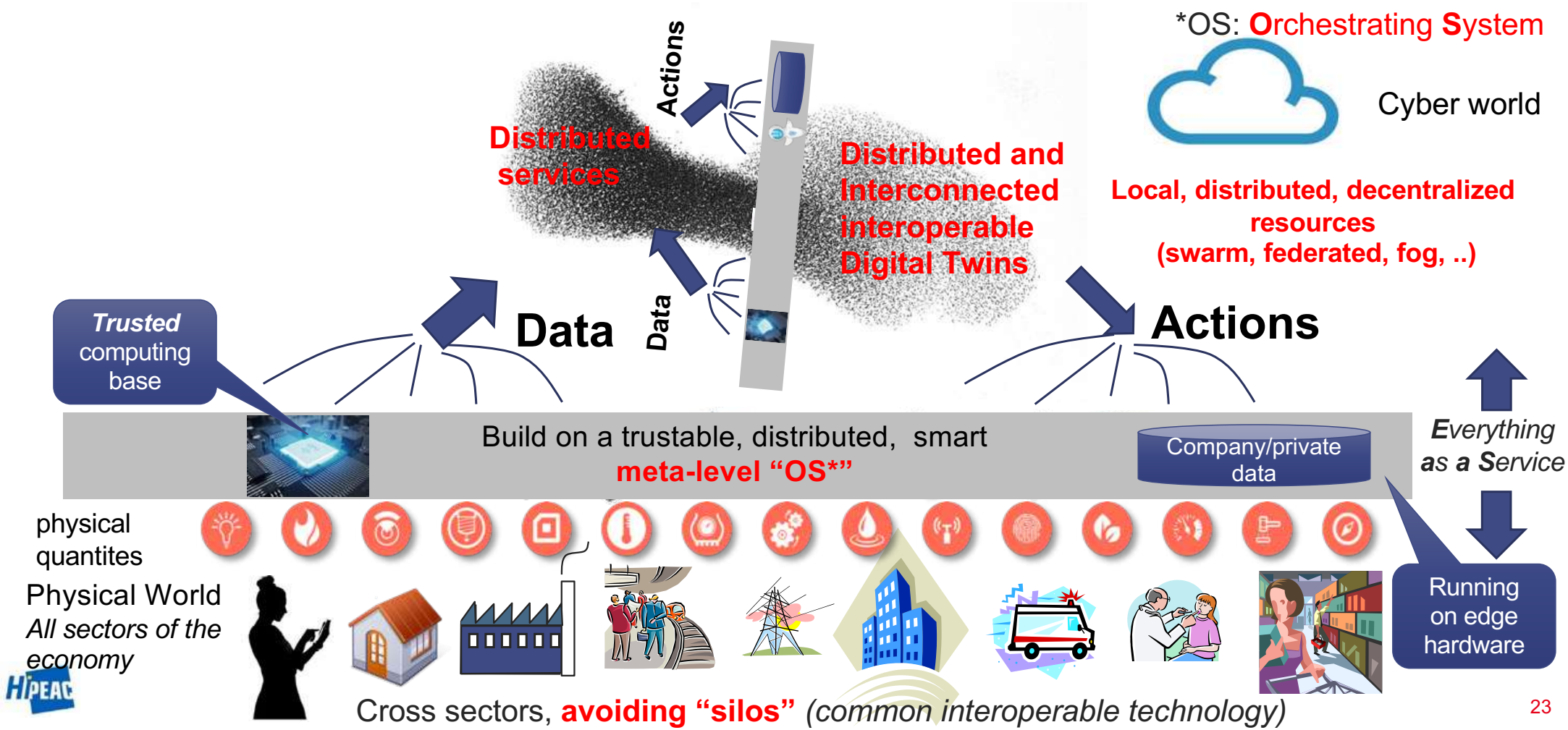
physical quantites

Physical World

All sectors of the economy

Cross sectors, **avoiding "silos"** *(common interoperable technology)*

Distributed Intelligence via Sharing and Coordination of Resources Across Heterogeneous, Connected, Locally Managed Devices



Some research challenges for the NCP

- **Dynamic service composition** seems incompatible with guaranteed performances in term of timing/latency:
  - Step by step approach
    - Explicit composition from already tested services composition
    - Algorithmic assessment… if possible (WCET)
    - Running digital twins in advance to assess performances
    - Computing backup scenarii with tradeoffs on services and **dynamic measurement of performance (monitoring)** to switch between scenario
    - …

- **Scalability:** in large systems, the MTTF (Mean Time to Failure) could be very short*:
  - MTTF of 1024-GPU jobs is 7.9 hours
  - MTTF of a 131072 GPU job is 14mn
  - Problem of invisible errors:
    - (service of a) **verification engine (monitoring)** running in parallel to check likelihood of results
      - Also useful for checking generative AI services!
    - Here again, using new services (either dynamically or following a precomputed approach) after detection of a non-conformity during real-time execution of services

=> Need of a **monitoring system service**, checking the performances in real time of the orchestrated services and instructing the orchestrator to change scenario in case of mismatch on a criteria

HiPEAC

* https://arxiv.org/abs/2410.21680

Some research challenges for the NCP

- **Interoperability:**
  - how to describe the non-functional properties in a sufficient but not extensive way so that the orchestrator can take rapidly meaningful choices?
  - Need of standardization
    - Extension of existing protocols
  - Hierarchical orchestration => hierarchical description
  - Use of directories of pre-characterized compositions?

Image generated by ChatGPT

Challenge: How to reach a safe and secure NCP?

- NCP has a large number of components & interconnections (but only used when required)
  - ➢ More places to hide, more places to attack
- **Need to secure NCP individual components**
  - Fast, scalable, static code analysis required, audit of code (open source?)
- **Secure communications required**
  - Need solid cryptography & protocols…
  - ...that are actually used !
- **How to ensure safe and secure payment of services**
  - Lightweight distributed ledgers “blockchain”
- **How to ensure that services are satisfying the (non-)functional requirements?**
  - Trusted intermediate
  - Sandbox test/execution
  - Use of digital twins
  - Methods and tools to check the properties…

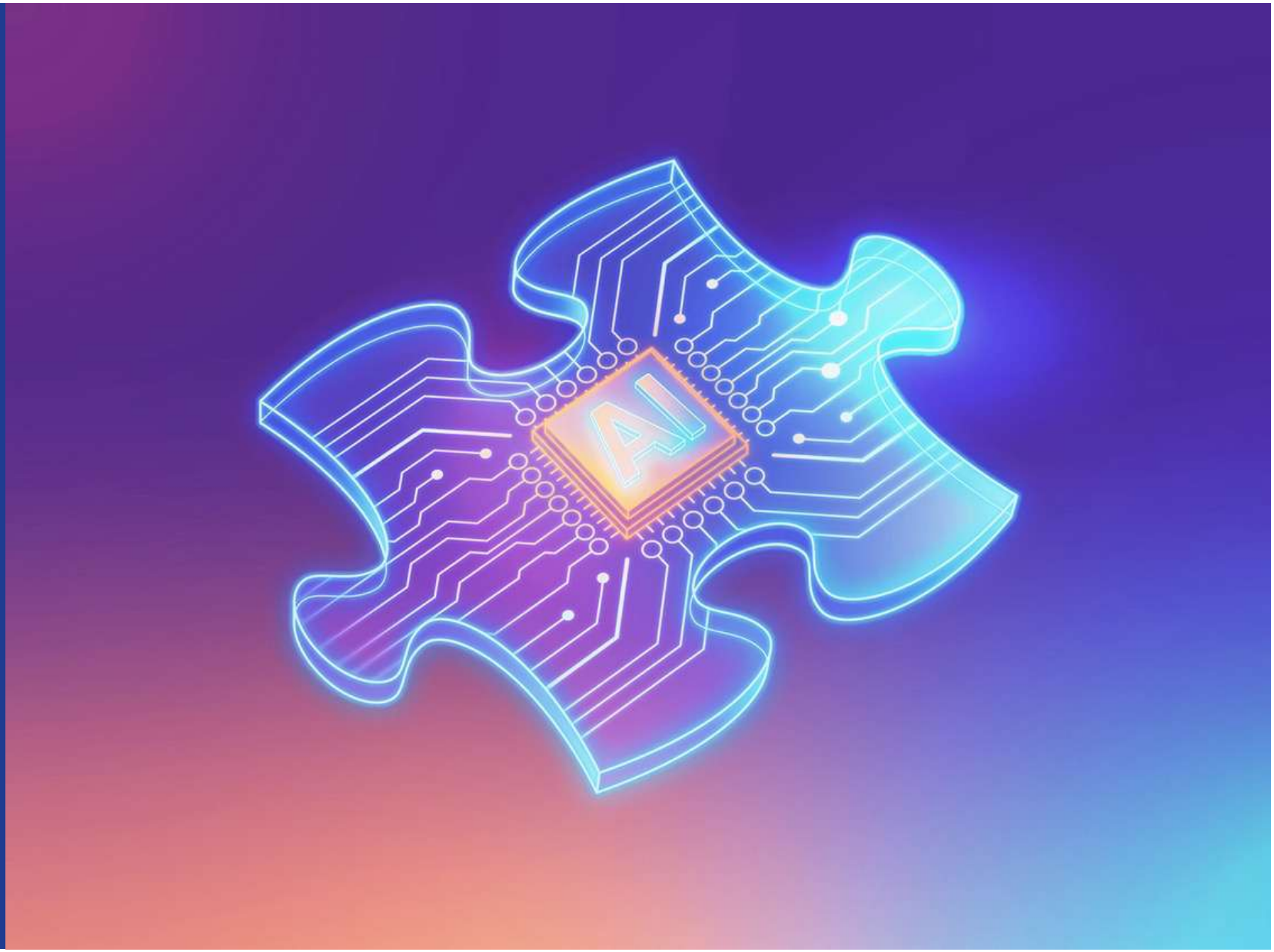
Your contributions are welcome!

HiPEAC Vision 2025

HIGH PERFORMANCE, EDGE AND CLOUD COMPUTING

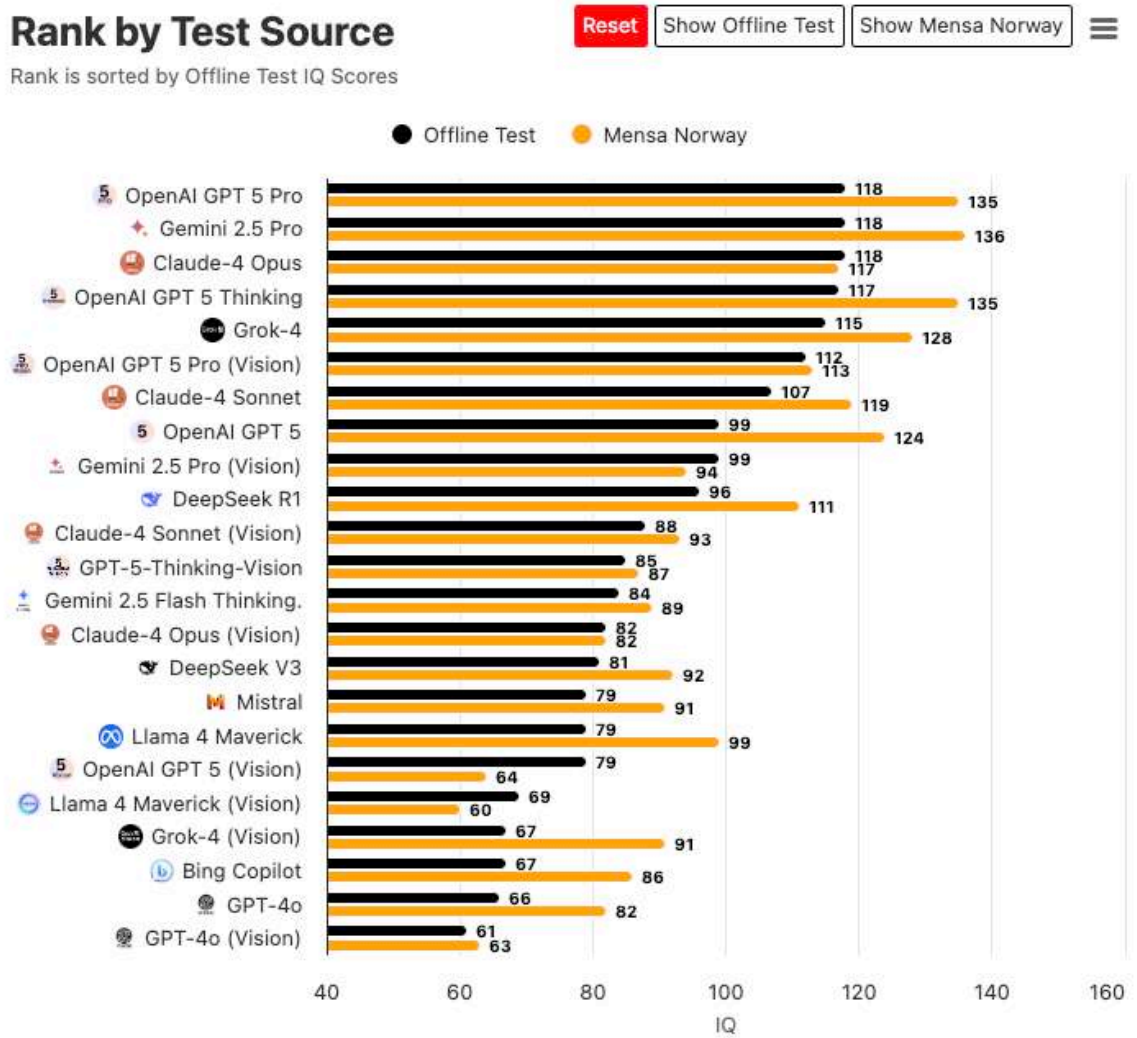
NCP and ARTIFICIAL INTELLIGENCE

Funded by the European Union



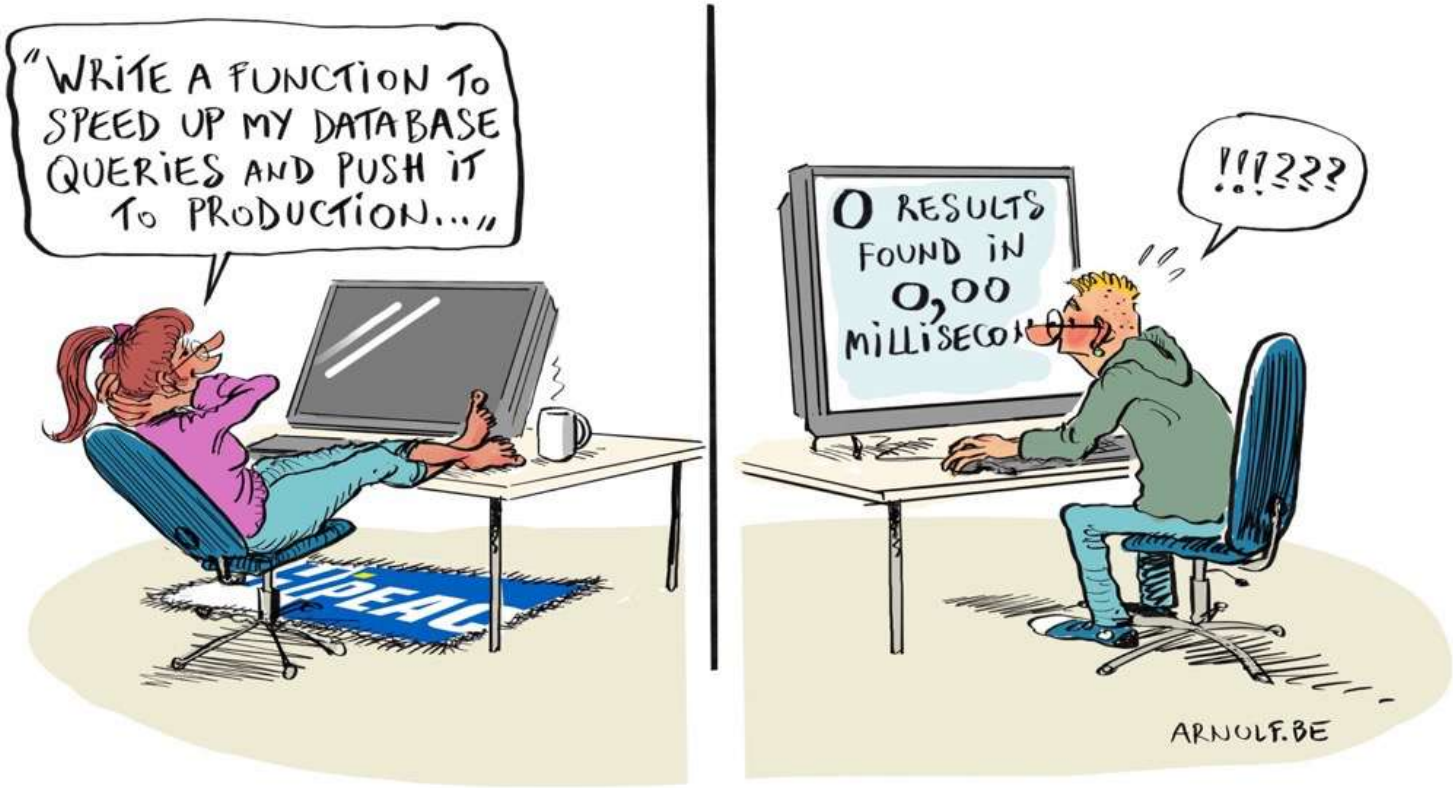
The explosion of artificial intelligence in 2025

- In late 2022, ChatGPT broke records as the AI platform reached 1 million users in less than a week.
- By early 2023, ChatGPT had accumulated over 100 million monthly users.
- As of July 2025, OpenAI.com receives 1.2 billion monthly visits. ChatGPT.com receives over **5 billion visits**, making it the #5 website globally.

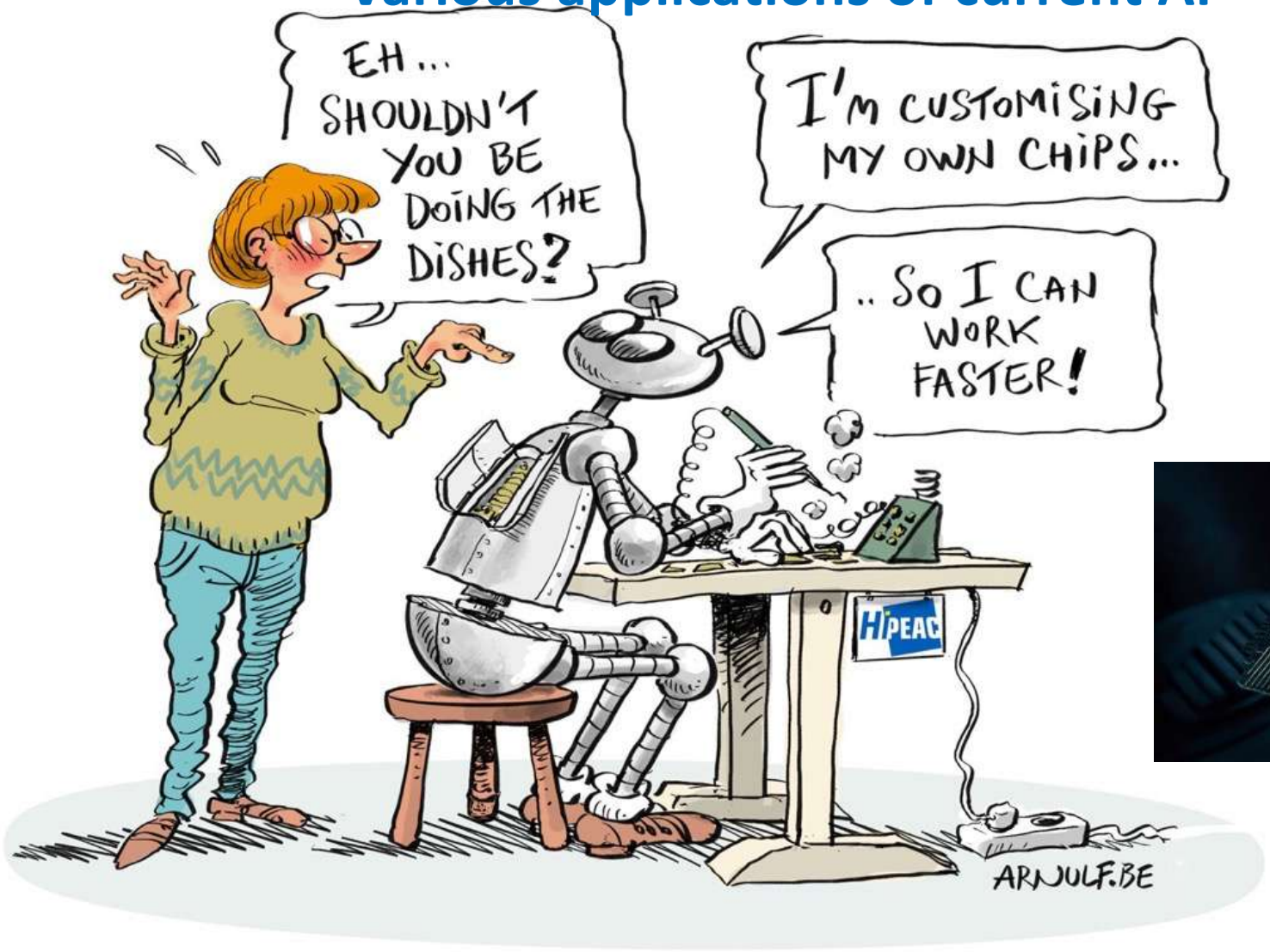


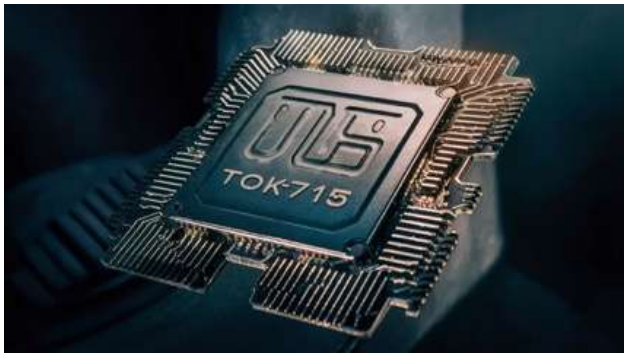
From https://trackingai.org/home September 7th, 2025

Various applications of current AI

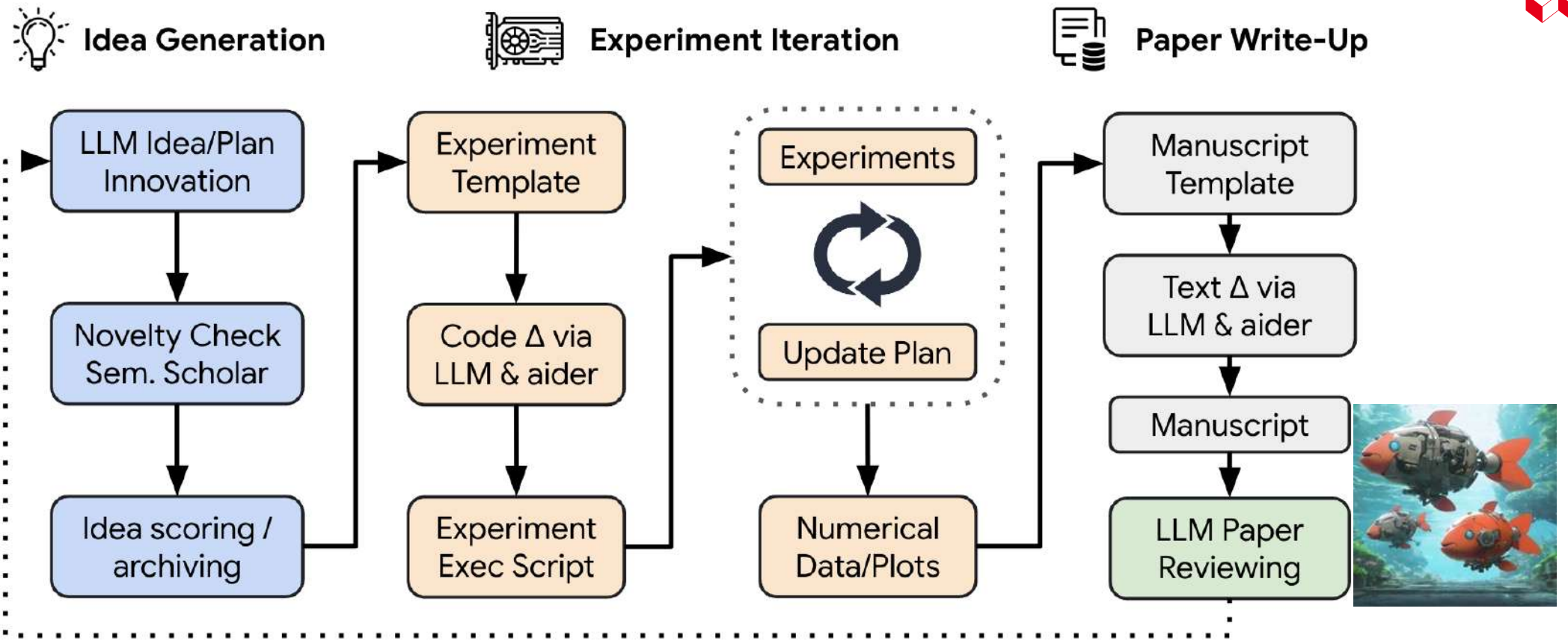


Various applications of current AI





The AI scientist from Sakana.ai



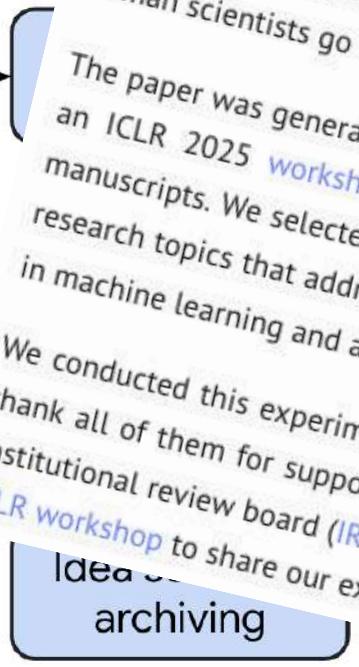
Conceptual illustration of The AI Scientist. The AI Scientist first brainstorms a set of ideas and then evaluates their novelty. Next, it edits a codebase powered by recent advances in automated code generation to implement the novel algorithms. The Scientist then runs experiments to gather results consisting of both numerical data and visual summaries. It crafts a scientific report, explaining and contextualizing the results. Finally, the AI Scientist generates an automated peer review based on top-tier machine learning conference standards. This review helps refine the current project and informs future generations of open-ended ideation. From https://sakana.ai/ai-scientist/

The AI scientist from Sakana.ai

March 12th, 2025

Experiment Iteration

Paper Write-Up

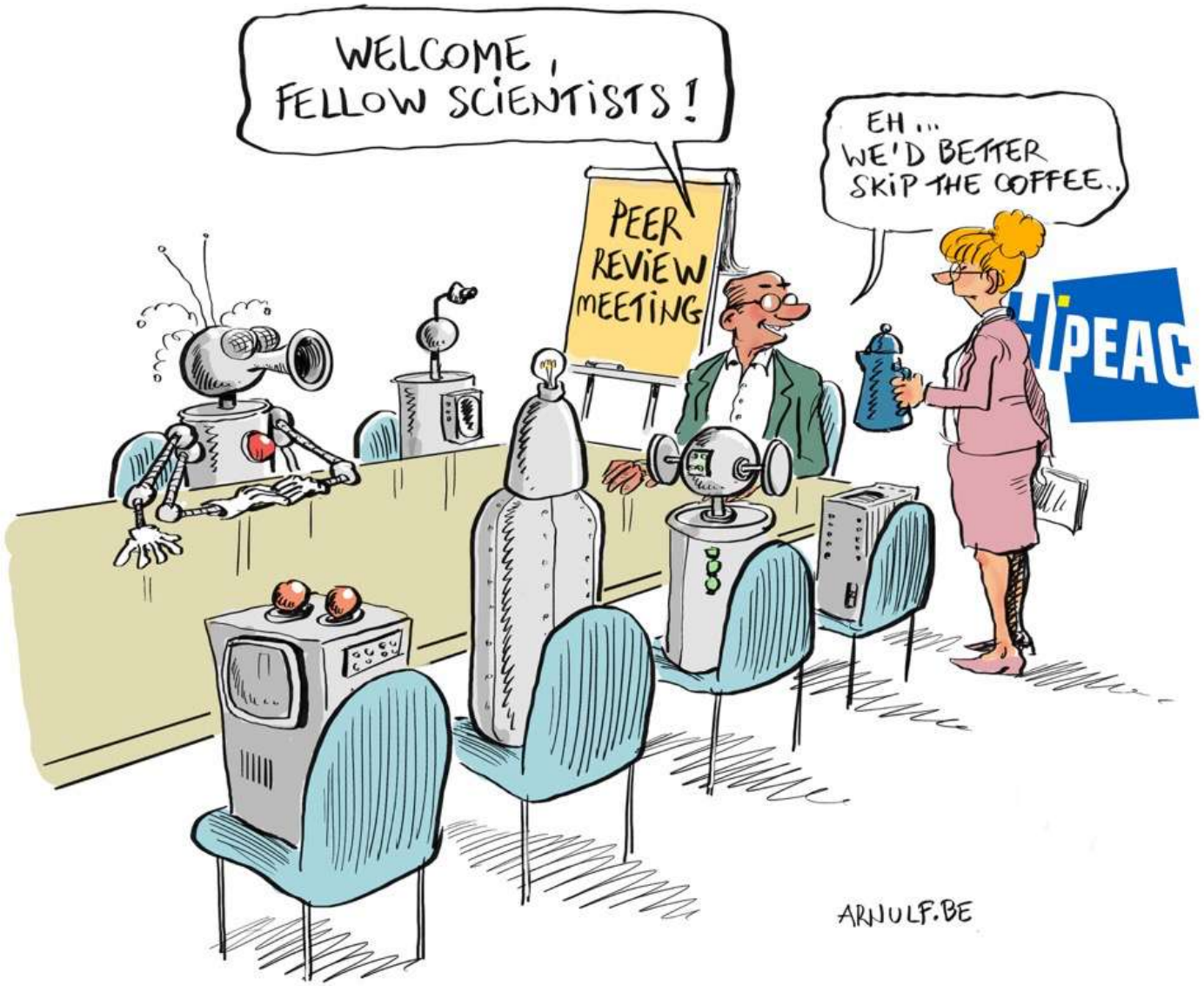


We are proud to announce that a paper produced by The AI Scientist passed the peer-review process at a workshop in a top machine learning conference. To our knowledge, this is the first[1] fully AI-generated paper that has passed the same peer-review process that human scientists go through.[2]

The paper was generated by an improved version of the original AI Scientist, called **The AI Scientist-v2**. This paper was submitted to an ICLR 2025 workshop that agreed to work with our team to conduct an experiment to double-blind review AI-generated manuscripts. We selected this workshop because of its broader scope, challenging researchers (and our AI Scientist) to tackle diverse research topics that address practical limitations of deep learning. The workshop is hosted at ICLR, one of three premier conferences in machine learning and artificial intelligence research, along with NeurIPS and ICML.[3]

We conducted this experiment with the full cooperation of both the ICLR leadership and the organizers of this ICLR workshop. We thank all of them for supporting this research into how AI-generated papers fare in peer-review. Furthermore, we also received an institutional review board (IRB) approval for this research from the University of British Columbia. Lastly, we plan to give a talk at the ICLR workshop to share our experiences and particularly the challenges with the AI Scientist project.

Conceptual illustration of The AI Scientist. The AI Scien… . Next, it edits a codebase powered by recent advances in automated code … Scientist then runs experiments to gather results consisting of both numerical data a… ific report, explaining and contextualizing the results. Finally, the AI Scientist generates an automated pe… er machine learning conference standards. This review helps refine the current project and informs future ge… open-ended ideation. From https://sakana.ai/ai-scientist/


WELCOME, FELLOW SCIENTISTS!
PEER REVIEW MEETING
EH... WE'D BETTER SKIP THE COFFEE..
HiPEAC
ARNULF.BE

Evolution of Generative Pre-trained Transformers (GPT) in OpenAI

Model	Architecture	Parameter count	Training data	Release date	Training cost
GPT-1	12-level, 12-headed Transformer decoder (no encoder), followed by linear-softmax.	117 million	BookCorpus: 4.5 GB of text, from 7000 unpublished books of various genres.	June 11, **2018**	"1 month on 8 GPUs", or 1.7e19 FLOP.
GPT-2	GPT-1, but with modified normalization	1.5 billion	WebText: 40 GB of text, 8 million documents, from 45 million webpages upvoted on Reddit.	February 14, **2019** (initial/limited version) and November 5, 2019 (full version)	"tens of petaflop/s-day", or 1.5e21 FLOP.
GPT-3	GPT-2, but with modification to allow larger scaling	175 billion	499 Billion tokens consisting of CommonCrawl (570 GB), WebText, English Wikipedia, and two books corpora (Books1 and Books2).	May 28, **2020**	3640 petaflop/s-day, or 3.2e23 FLOP.
GPT-3.5	Undisclosed	175 billion	Undisclosed	March 15, **2022**	Undisclosed
ChatGPT	Undisclosed	? (rumor 20M???)		**November 20, 2022**	
GPT-4	Also trained with both text prediction and RLHF; accepts both text and images as input. Further details are not public.	Undisclosed (1.8 trillon aka 1.8e12)	Undisclosed (13 trillon tokens, aka 1.3e13)	March 14, **2023**	Undisclosed. Estimated 2.1e25 FLOP.

From https://en.wikipedia.org/wiki/Generative_pre-trained_transformer

Compute requirement

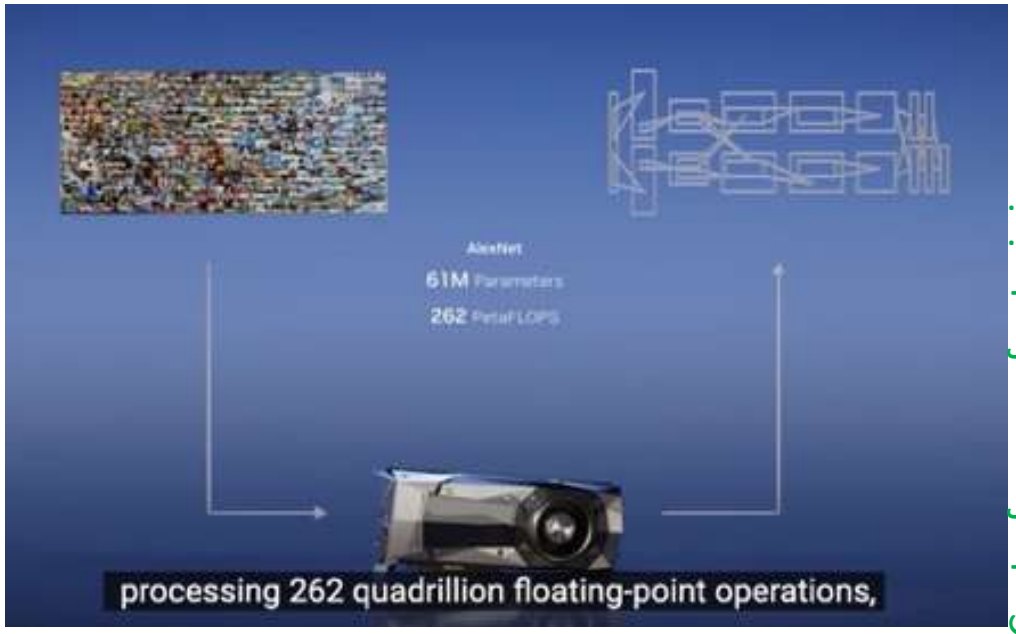
~ x 88

~ x 213

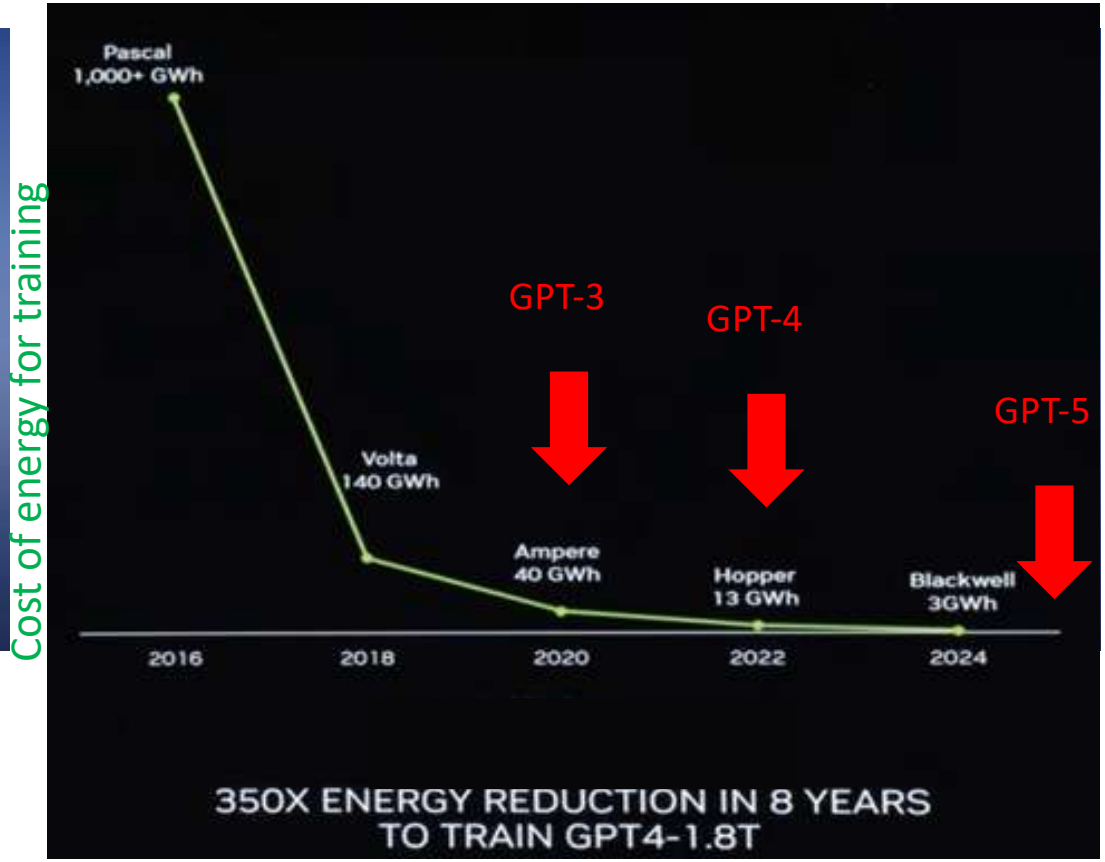
~ x 65

~ x 1 218 360

Computing is driving AI performances



2012: AlexNet
GeForce GTX 580
Won ImageNet Challenge
262 x 10^{15} FLOPS (262 PetaFLOPS)



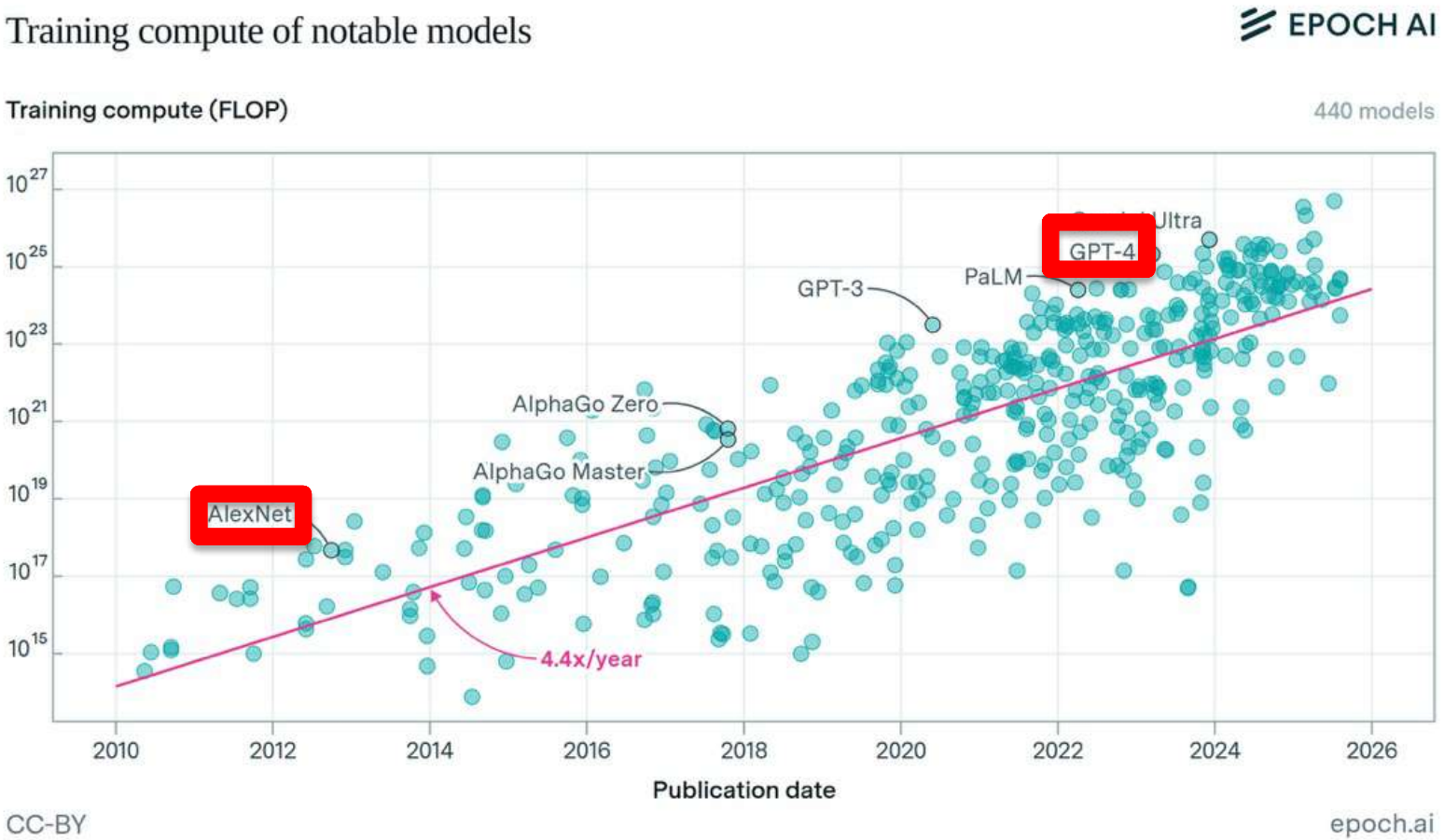
From GTC 2023 Keynote with NVIDIA CEO Jensen Huang

Cost of energy for training is a limiting factor!

AI uses more and more (processing) power for training SOTA networks

"Since 2010, the training compute used to create AI models has been growing at a rate of 4.4x per year.*"

(by comparison, Moore's Law had an 18-month doubling period)



From https://epoch.ai/data/ai-models#data-insights August 2025

How to reduce the inference cost?

Inference (using generative AI) ***is becoming more demanding***

- E.g. ChatGPT's montly users have grown to more than 5 billion (July 2025)
- ***Test-time compute*** is rising and is even more demanding of compute power in inference

=> **Specialization of hardware for inference** (e.g Groq chip, AWS Inferentia vs AWS Trainium chips, etc)

- Approaches that **don't need to use all the "neurons"** of a LLM:
  - **Mixture of Experts**: MoE architectures create specialized "experts" within a large model
    - **Only a subset** of these experts are activated for each task, promoting a modular structure within a single model.

GPT-4 performances due to its structure

- **GPT-4's Scale**: GPT-4 has ~1.8 trillion parameters across 120 layers, which is over 10 times larger than GPT-3.
- **Mixture Of Experts (MoE)**: OpenAI utilizes 16 experts within their model, each **with ~111B parameters** for MLP. **Two of these experts are routed per forward pass**, which contributes to keeping costs manageable. (NB: **1/8 of the computation**)
- **Dataset**: GPT-4 is trained on ~13T tokens, including both text-based and code-based data, with some fine-tuning data from ScaleAI and internally.
- **Dataset Mixture**: The training data included CommonCrawl & RefinedWeb, totaling 13T tokens. Speculation suggests additional sources like Twitter, Reddit, YouTube, and a large collection of textbooks.
- **Training Cost**: The training costs for GPT-4 was around $63 million, taking into account the computational

DeepSeek-R1 uses only 37B parameters on its 671B in inference: 18x less

required and lower utilization rates.

- **Inference Architecture**: The inference runs on a cluster of 128 GPUs, using 8-way tensor parallelism and 16-way pipeline parallelism.
- **Vision Multi-Modal**: GPT-4 includes a vision encoder for autonomous agents to read web pages and transcribe images and videos. The architecture is similar to Flamingo. This adds more parameters on top and it is fine-tuned with another ~2 trillion tokens.

From SemiAnalysis, July 2023

How to reduce the inference cost?

Inference (using generative AI) is becoming more demanding

- E.g. ChatGPT's montly users have grown to more than 5 billion (July 2025)
- Test-time compute is rising and is even more demanding of compute power in inference

=> **Specialization of hardware for inference** (e.g Groq chip, AWS Inferentia vs AWS Trainium chips, etc)

- Approaches that **don't need to use all the "neurons"** of a LLM:
  - **Mixture of Experts**: MoE architectures create specialized "experts" within a large model,
    - **Only a subset** of these experts are activated for each task, promoting a modular structure within a single model.
  - **Agentic AI:** Agentic AI often operates with multiple, distinct agents, each responsible for specialized tasks or competencies.

Smaller LLM models get more powerful

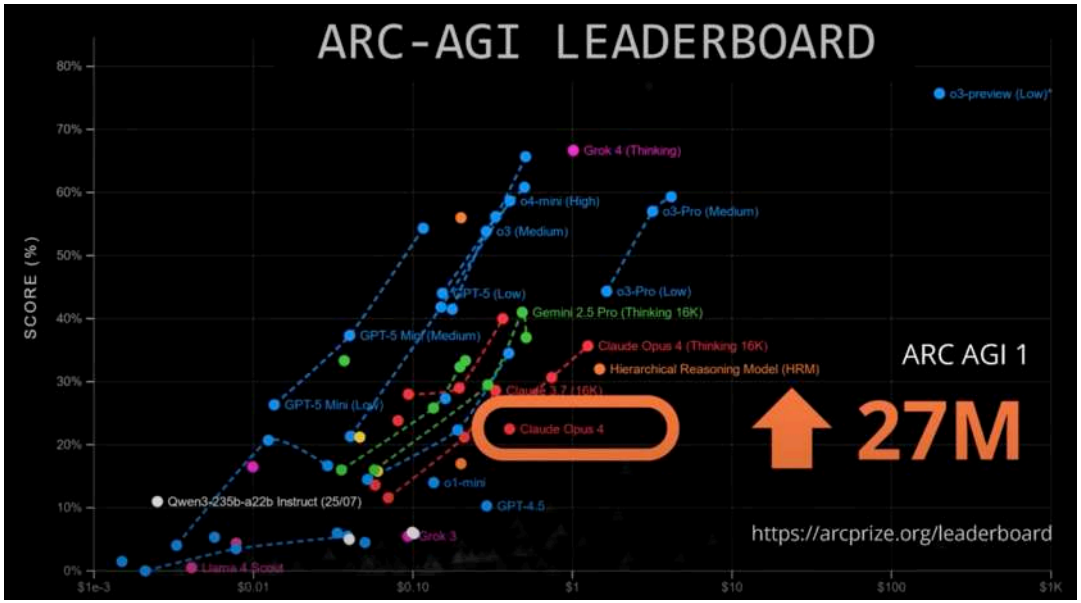
Current models of about 10B parameters have better performances on specific tasks than the original ChatGPT of 2022

Model name	Announced	MMLU Pro *
GPT-5 (High)	August 2025	0.87
GPT-4o	May 2024	0.73
Seed-oss-36B	August 2025	0.83
Qwen3-30B	August 2025	0.81
Phi-4-14B	December 2024	0.76
GPT-oss-20B	August 2025	0.74
Gemma-3-12B	August 2025	0.61

From https://huggingface.co/spaces/TIGER-Lab/MMLU-Pro

*Massive Multitask Language Understanding

"Open source" models are catching up closed models with few months delay



https://arcprize.org/leaderboard

On ARC-AGI 1 benchmark, a 27M parameters model (HRM, Hierarchical Reasoning Model) beat Claude Opus 4

Distillation: Smaller Models Can Be Powerful Too*

*"We demonstrate that the **reasoning patterns of larger models can be distilled into smaller models,** resulting in better performance compared to the reasoning patterns discovered through RL on small models.*

Using the reasoning data generated by DeepSeek-R1, we fine-tuned several dense models that are widely used in the research community. The evaluation results demonstrate that the distilled smaller dense models perform exceptionally well on benchmarks. We open-source distilled 1.5B, 7B, 8B, 14B, 32B, and 70B checkpoints based on Qwen2.5 and Llama3 series to the community."

Model	AIME 2024 pass@1	AIME 2024 cons@64	MATH-500 pass@1	GPQA Diamond pass@1	LiveCodeBench pass@1	CodeForces rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481

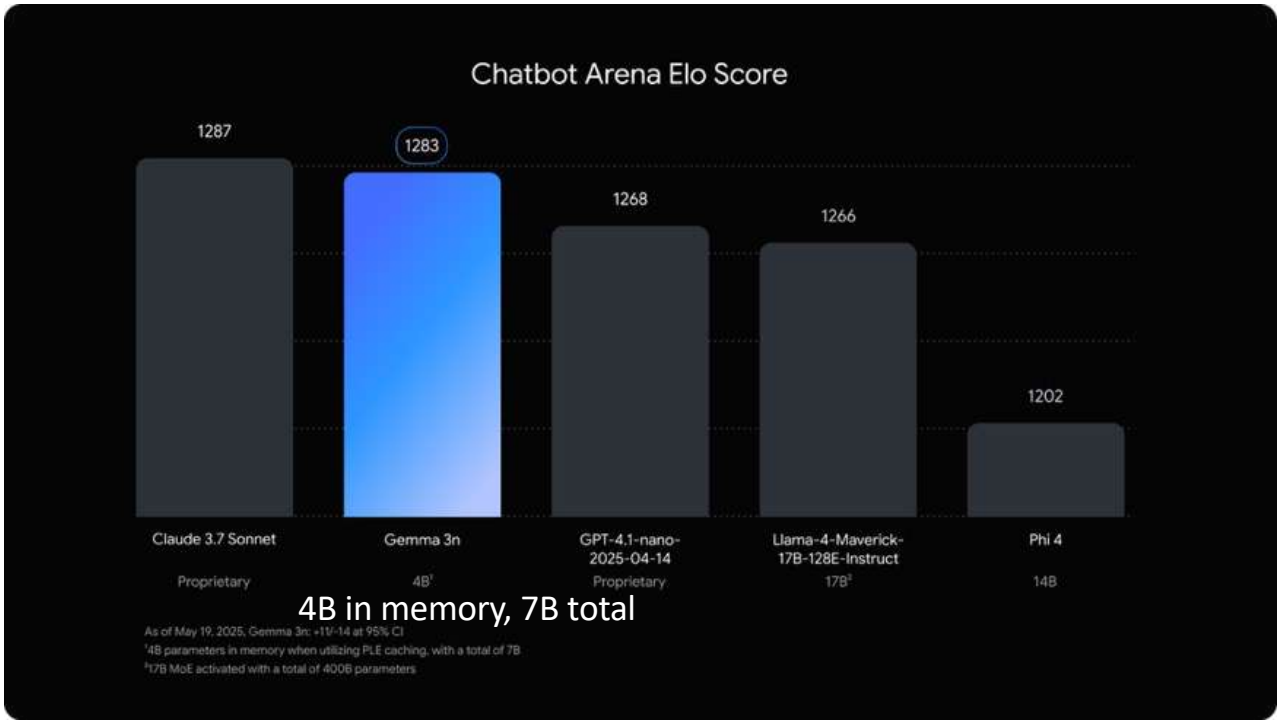
* From https://github.com/deepseek-ai/DeepSeek-R1?tab=readme-ov-file#distilled-model-evaluation

The American Invitational Mathematics Examination (AIME) is a selective and prestigious 15-question 3-hour test given since 1983 to those who rank in the top 2.5% on the AMC 10.

Ultra low parameters LLMs

Some (specialized) LLMs are now below the 1G range:

- Microsoft MU: Mu is an efficient 330M encoder–decoder language model optimized for small-scale deployment, particularly on the NPUs on Copilot+ PCs*
- Version 0.3B (360M) of Ernie, from the LLM family of Baidu** https://huggingface.co/baidu/ERNIE-4.5-0.3B-PT
- New memory optimization in the multimodal open-source model Gemma3n from Google (A 7B parameters model running on 4GB of RAM) ***

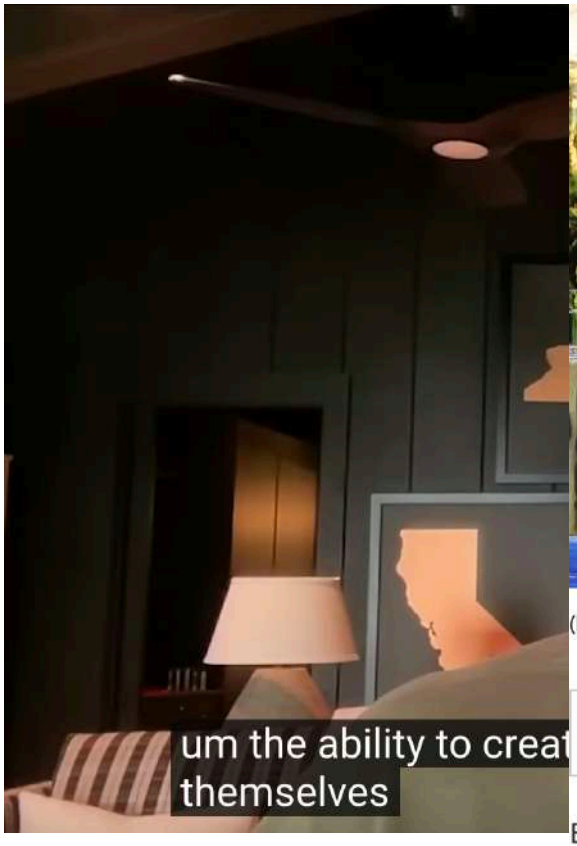


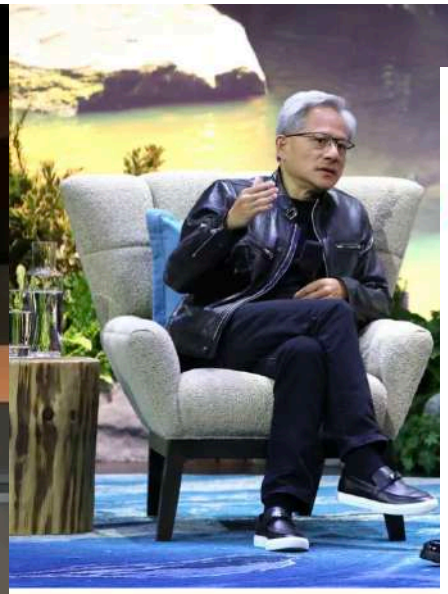
Capability	Benchmark	ERNIE-4.5 -0.3B-Base	Qwen3 -30B-A3B-Base	ERNIE-4.5 -21B-A3B-Base	DeepSeek-V3 -671B-A37B-Base	ERNIE-4.5 -300B-A47B-Base
General	C-Eval	40.7	87.2	**88.0**	90.2	**91.5**
	CMMLU	39.8	86.0	**86.2**	88.2	**91.2**
	MMCU	37.2	88.8	**94.0**	94.0	**95.9**
	AGIEVAL	28.5	**72.8**	68.4	75.8	**78.4**
	MMLU	41.9	**81.0**	78.9	**87.9**	87.4
	MMLU-Redux	43.2	**84.6**	80.7	**89.4**	89.2
	MMLU-Pro	16.0	**56.7**	51.2	66.7	**69.5**
Reasoning	BBH	30.4	72.7	**77.5**	87.5	**89.4**
	DROP	28.6	39.6	**70.8**	**84.6**	82.8
	ARC-Easy	60.7	**98.6**	96.9	98.7	**99.2**
	ARC-Challenge	40.6	**93.7**	90.7	95.1	**96.3**
	HellaSwag	33.0	87.7	**92.1**	91.3	**96.6**
	PIQA	55.2	**91.0**	80.6	94.1	**94.6**
	WinoGrande	51.3	**76.3**	70.6	88.2	88.3
	CLUEWSC	48.6	**76.8**	76.2	**81.4**	79.9
Math	GSM8K	25.2	70.8	**81.0**	90.6	**91.8**
	MATH	12.4	**61.0**	50.8	63.0	**69.1**
	CMI7K	4.6	**69.4**	64.9	79.3	**86.4**
	MGSM	2.7	**71.2**	69.2	79.8	**88.6**
	ASDIV	29.8	82.8	**83.5**	89.7	**90.2**
	SVAMP	24.0	**86.3**	79.7	**92.7**	90.0
	MATHQA	21.6	39.4	**56.1**	76.9	**83.0**
	CMATH	52.2	88.9	**93.7**	90.7	**96.2**
Knowledge	SimpleQA	1.8	7.1	**30.4**	24.9	**38.4**
	ChineseSimpleQA	7.4	52.0	**54.8**	64.8	**72.2**
Coding	HumanEval+	25.0	83.5	**86.0**	76.0	**84.8**
	MBPP+	40.2	**76.2**	75.1	**76.7**	74.9
	MultiPL-E	14.2	**66.1**	65.4	63.0	**68.6**

* https://blogs.windows.com/windowsexperience/2025/06/23/introducing-mu-language-model-and-how-it-enabled-the-agent-in-windows-settings/
** https://huggingface.co/baidu/ERNIE-4.5-0.3B-PT
*** https://developers.googleblog.com/en/introducing-gemma-3n/

Agentic AI the future of AI?

- Using a set of small specialized LLMs can have similar performances than of a large LLM
- Only a subset of the LLM are activated simultaneously





(Image credit: Getty Images / Justin Sullivan)

Jump to: Read more

Bringing AI agents into the workforce will soon be as common as onboarding human employees, as they work together to make businesses smarter and more efficient, Nvidia CEO Jensen Huang has predicted.

What is the key element of Agentic AI?

The key element in both approaches is the “**router**”, or “**orchestrator**”

- **MoE:** The MoE **router** selects the most appropriate experts based on the input context, enhancing the model’s ability to adapt dynamically to various types of inputs. This routing mechanism is foundational in allowing a large model to focus on the right areas at the right time.
- **Agentic AI**: Similarly, Agentic AI involves a decision-making layer or **"agent manager”, or “orchestrator”** that allocates tasks to the best-suited agents. The manager dynamically routes requests to different agents based on the context or goal, enabling the system to adaptively respond to complex, changing inputs.

Agents can be centralized, or ***distributed***:
- Agents can run on different devices, even “old” ones, increasing lifetime of devices
- If a device is not powerful enough, it can delegate to other devices (good to increase lifetime of devices!)

Image generated by ChatGPT

Structure of Apple Intelligence

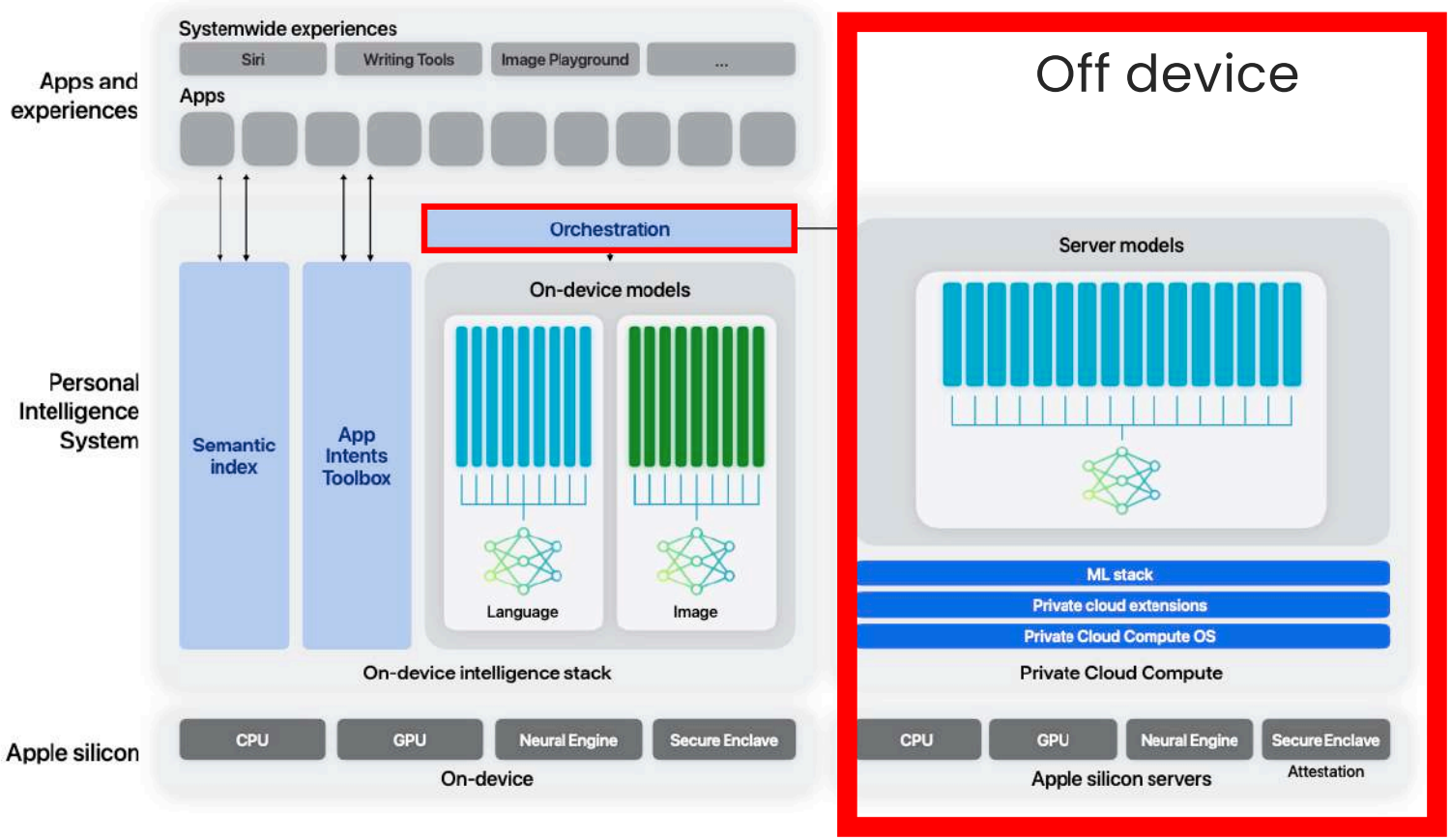


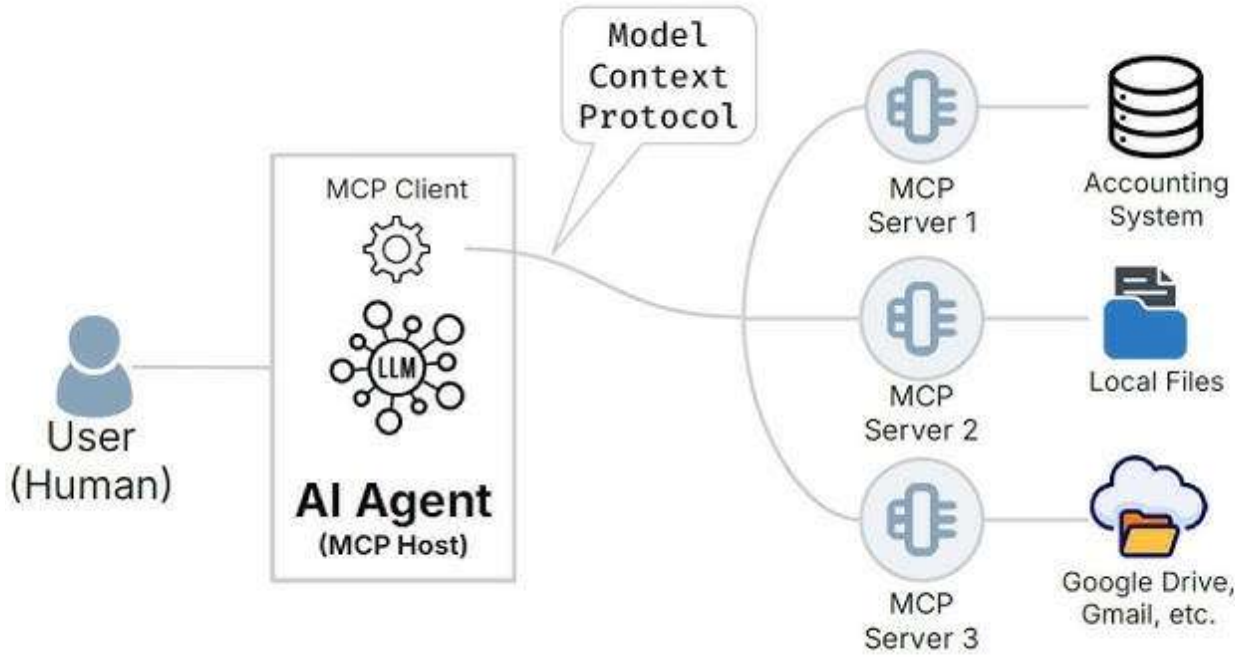
Figure 2: Architecture of Apple Intelligence with adapters for the language on-device and server models and the image models. In this report we are only describing the text models.

From: https://arxiv.org/abs/2407.21075

MCP= a first interconnect step by Anthropic

MCP (Model Context Protocol) is an open-source standard for connecting AI applications to external systems.

Using MCP, AI applications like Claude or ChatGPT can connect to data sources (e.g. local files, databases), tools (e.g. search engines, calculators) and workflows (e.g. specialized prompts)—enabling them to access key information and perform tasks.



From https://modelcontextprotocol.io/docs/getting-started/intro

Interoperability protocols for federated agentic AI

All exchange data and parameters should be done in a commonly understood protocol that:

- rely not only on functional requirements (like MCP, A2A, …)
- but also on non-functional requirements (providing enough information such as the orchestrator can selects the right services, e.g. according to criteria such as **response time, potential level of hallucinations, cost, localization, privacy of data, etc...**).

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

"Like TCP-IP was a way to allow various OS (Operating Systems) to communicate together, this challenge is to create the equivalent for OS (Orchestration Systems) to exchange AI related information."

Interoperability protocols for federated agentic AI

All exchange data and parameters should be done in a commonly understood protocol

- rely
- but

ation such as th
riteria such as **res**
n, privacy of data,

It is t
monly defin
acceptance.

"Like
ommunicate together, this challenge is to create the equivalent for OS (Orchestration Systems) to exchange AI related information."

Do you see the link with the NCP?
We can use this as a *blueprint* for developing the NCP technologies


TRANSLATE EVERYTHING IN LITHUANIAN AND SWAHILI!
FIND A SOLUTION!
WATCH IT!
CLEAR THE AREA!
ACTION AGENT
ORCHESTRATOR
YOU, CLEAN THE TOILET!
CALCULATE OUR PROFITS!
CALCULUS AGENT
LATRINA
WHY DO I ALWAYS GET THE BUM GIGS?!!
ARNULF BE

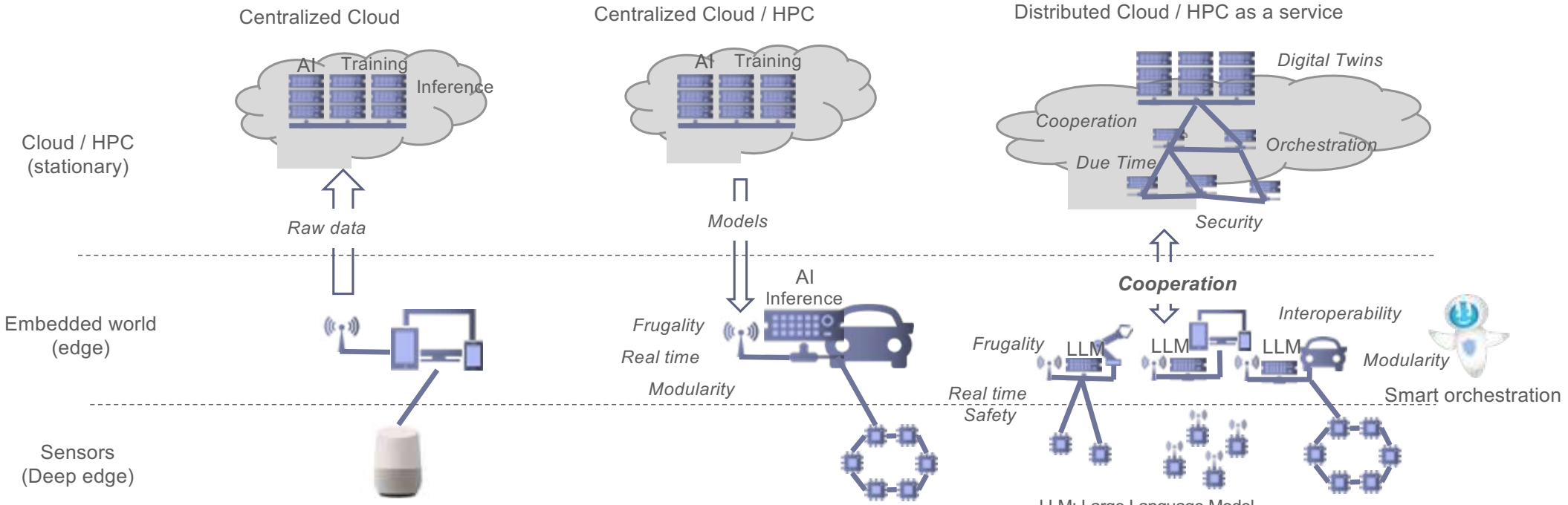
HiPEAC
Vision
2025

HIGH PERFORMANCE, EDGE
AND CLOUD COMPUTING

Funded by
the European Union

Conclusion

Evolution of computing: Cloud, CPS, IoT, AI → Next Computing Paradigm



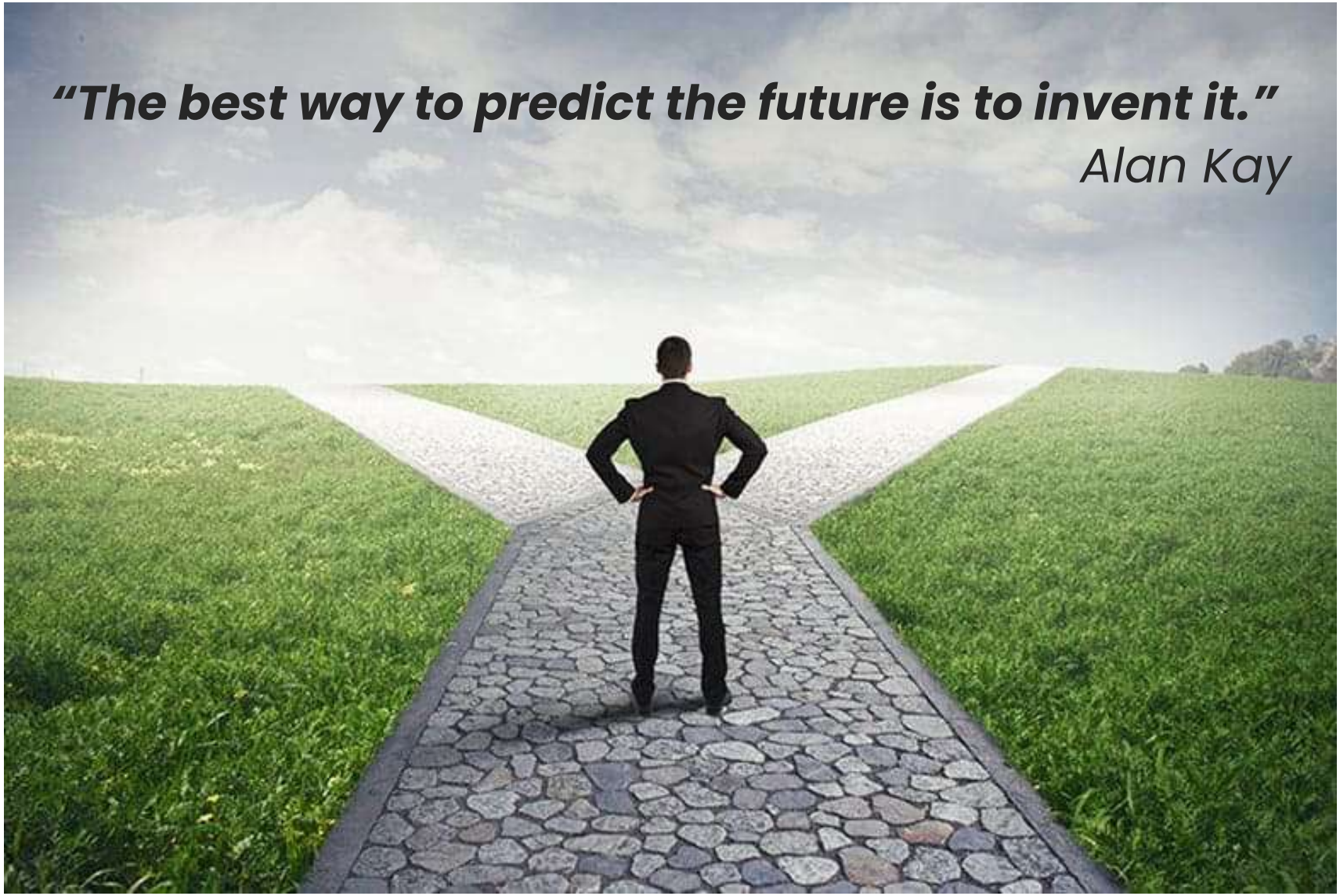
Cloud computing → Intelligent Cyber-Physical Systems → Distributed and cooperative systems

From Denis Dutoit, inspired from HiPEAC 2023

Bringing different communities together to create the NCP

WE NEED YOU!

CONCLUSION: WE LIVE AN EXCITING TIME!




HiPEAC Vision 2025
HIGH PERFORMANCE, EDGE AND CLOUD COMPUTING
HiPEAC
More details at https://vision.hipeac.net/
HiPEAC Vision Video: https://www.youtube.com/watch?v=SdyHJF8ui-M