EPI Forum

Paris, 6-7 October, 2025

European Processor Initiative

Empowering European Digital and AI Sovereignty

October 7th, 2025

Marc Duranton

Senior fellow

CEA (France), HiPEAC and ETP4HPC

DISCLAMER

The opinions and proposals in this presentation are solely those of the author and do not necessarily reflect the views, positions, or strategies of any employer or affiliated organization.

They are synthesized from ongoing exchanges and roadmap discussions across industry, academia, and public initiatives.

MY DEFINITION OF “SOVEREIGNTY”

“Classical” definition: “**Sovereignty** is the supreme authority to make and enforce rules over a community—most often a state—without being subject to any higher earthly power. »

Supremacy* vs Sovereignty: « Supremacy is the ranking of laws or institutions within a given legal or political order. It says which rule prevails in case of conflict (e.g., constitutional norms over statutes; EU law over conflicting national law; federal law over state law). »

No “kill switches” controlled by other parties

In this presentation, I will take a **weaker definition of sovereignty** (related to technology) as

“Having a say at the negotiating table” -> What technologies and approaches should Europe develop to be less dependent from outside?

It only captures an aspect of sovereignty (voice in decisions that affect you), which seems pragmatic in the field of technology

But sovereignty is more than participation—it’s the right and capacity to have the final say over your own affairs, to enforce those decisions, and to choose when to join, shape, or refuse joint decisions (including delegating powers and taking them back).

This is more in the political and legislative hands…

* In this context: controlling a complete technology chain in order to be independent from other countries on this particular topic

WHAT WENT WRONG IN THE 2000'S?

Strong B2C "vertical" companies: Philips, Siemens, Nokia, Ericsson, …

Several European processors architectures pushed to the market: ARM (1990 as "Acorn Risc Machine", founded by Acorn computers, Apple computers and VLSI Technology), ARC (1997), Inmos (1985 for Transputers), Philips (Trimedia – 1987)…

Silicium has a high value but a small volume and weight: creation of an international ecosystem where transportation cost is negligeable vs. specialized and localized factories: wafer production, masks, factories, slicing, packaging, testing, integration on a PCB board, making a system, …

A "chip" could make several turns around Earth before the final product is delivered to its final customer!

- Unlike CRTs for example

European managements chose to focus the companies where ***they perform the best,*** forgetting the rule of "***the sum of local optimizations is lower than a global optimization***"

From totally integrated companies, to fab light, to fab less…

Unlike the current big companies such as Samsung, Apple, …

WHAT IF…

- “In 1986, Philips sign a joint venture contract with Taiwan to put up $58 million, transfer its production technology, and license intellectual property in exchange for a 27.5 percent stake in a new company. In addition to capital, Philips played a crucial role by transferring semiconductor manufacturing technology, intellectual property, and patents, enabling the company to scale more rapidly. Philips also provided the first CEO, James E. Dykes, who had previously worked at Philips North America. This partnership represented an early example of the “fab-light” strategy”*

This company is ***TSMC***…

- ***ASML*** was founded in 1984 as a joint venture between the Dutch technology companies Philips and ASM International. Later, nobody believed the EUV lithography would be possible, but ASML persisted…
- Philips bought VLSI technology in 1999

What missed opportunities to be at the negotiating table…

In other domains, European companies are still leaders (e.g. Airbus) and have created an ecosystem around them

* Mainly from https://en.wikipedia.org/wiki/TSMC

THERE HAS BEEN MORE CREATIVE DESTRUCTION IN THE US THAN IN 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)

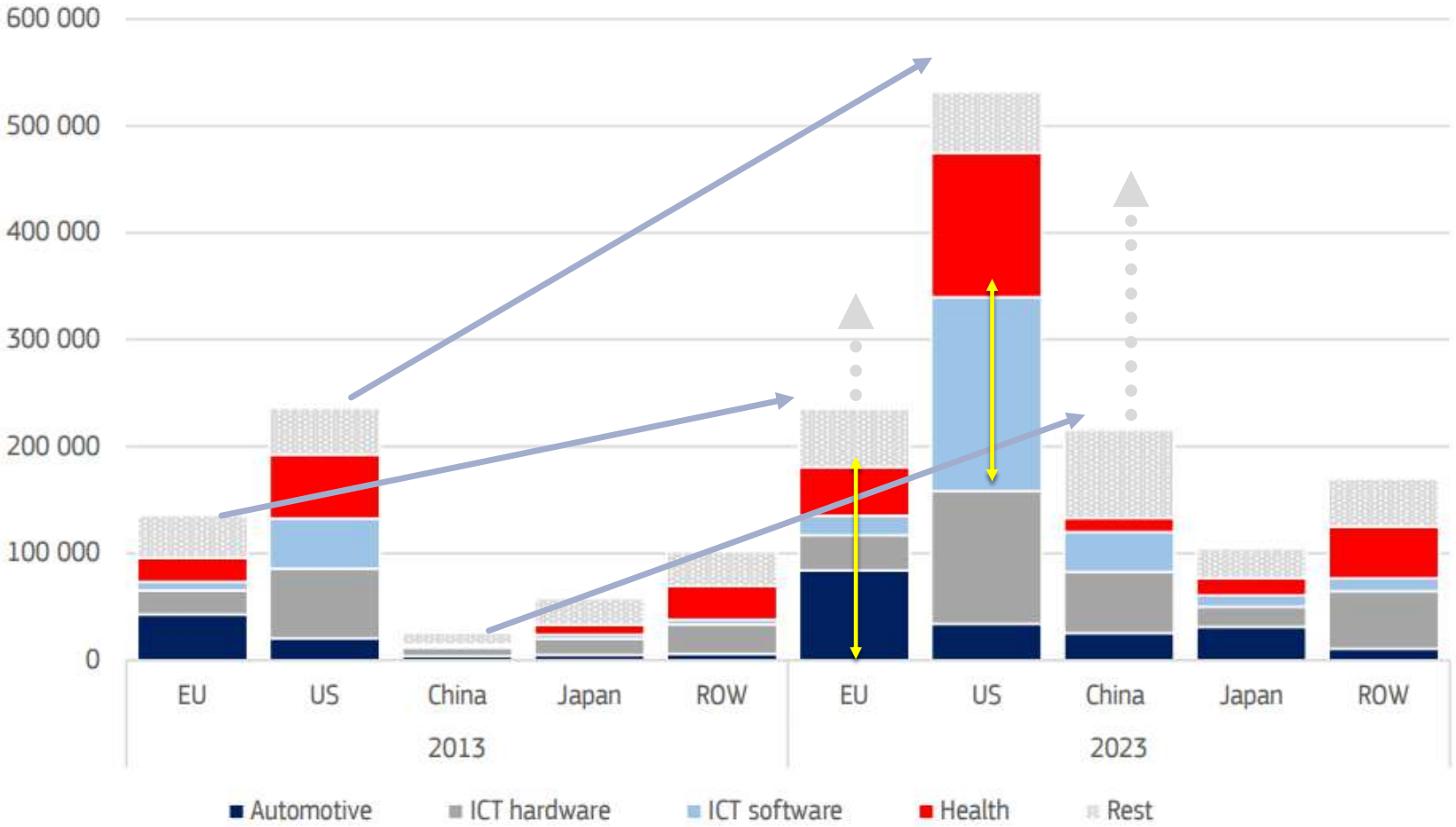
Source: Industrial R&D Investment Scoreboard (2004, 2013 and 2023).

Most of the "big giants" in the digital domain are not making the "final" product for the customer, but ***enabling technologies*** ("tools") allowing others to make final products:

- Microsoft -> OS, (and Office)
- Google -> Web search, App platform (Android)
- Apple -> App platform, computers
- OpenAI -> AI LLM technology (API)
- Nvidia -> hardware enabler of AI
- Intel, AMD: GP processors

They all serve therefore a large "customer" base.

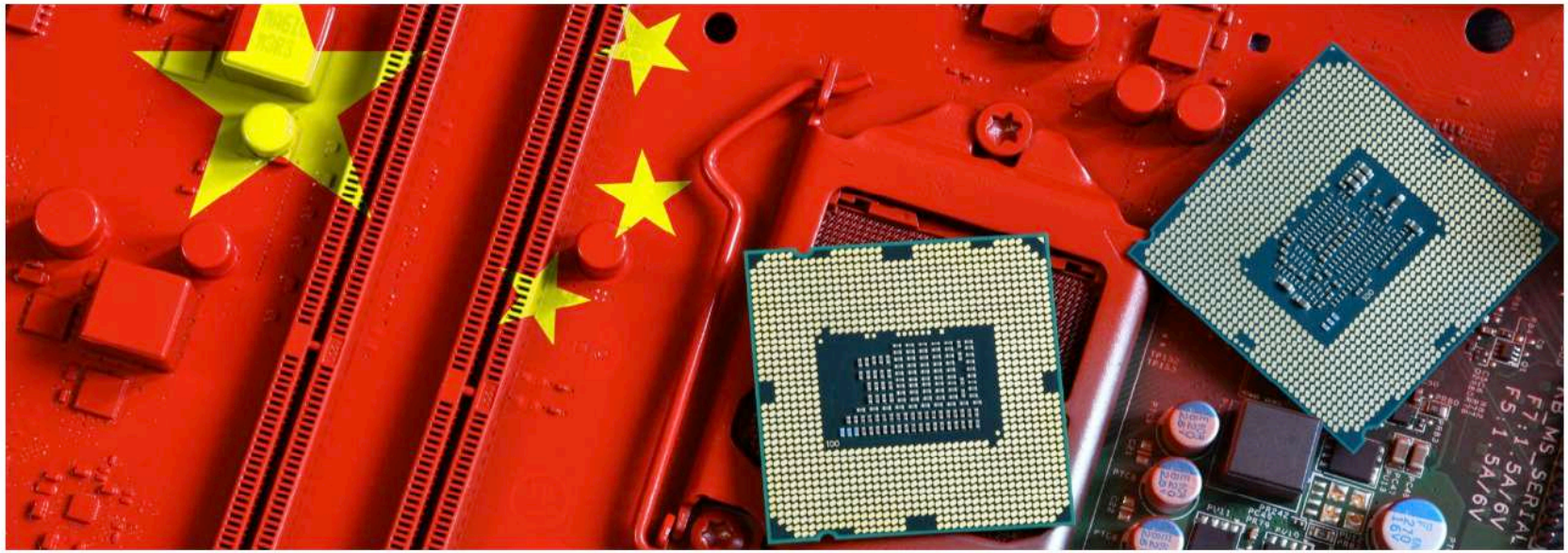
THE EU DOUBLED ITS R&D INVESTMENTS, BUT THE US DOUBLED IT TOO, AND CHINA OCTUPLED…



THE EU DOUBLED ITS R&D INVESTMENTS, BUT THE US DOUBLED IT TOO, AND CHINA OCTUPLED…

China to deploy $98bn in AI investment this year amid US tech rivalry

Dashveenjit Kaur June 26, 2025



Share this story:

Tags:

Categories::
AI & INTELLIGENCE

- **China's AI investment could reach $98 billion in 2025, a 48% growth.**
- **Government funding will dominate at $56 billion, while internet companies contribute $24 billion to AI.**

China's AI investment is poised to reach unprecedented levels in 2025, with forecasts suggesting the mainland could deploy between $84 billion to $98 billion in AI capital expenditure this year, according to a new Bank of America report.

The surge represents a potential 48% growth from 2024 levels, underscoring the escalating technological competition between China and the rest of the world.

PERSISTENCE OF THE VISION

A long-term vision is essential for success!

Disruptive approach can “win” in an “established” market e.g. Apple iPhone, Tesla, …

If you put real innovation on the market, market analysis are not so useful…

“It’s really hard to design products by focus groups. A lot of times, people don’t know what they want until you show it to them.” — BusinessWeek interview of Steve Jobs, May 25, 1998.

“We do no market research… We just want to make great products.” — Fortune interview of Steve jobs, Mar 7, 2008.

Constancy, persistency versus short term business ups and down

My experience: don’t expect a competitive chip at the first tape out…

A foundry will not have high yield just when it opens…

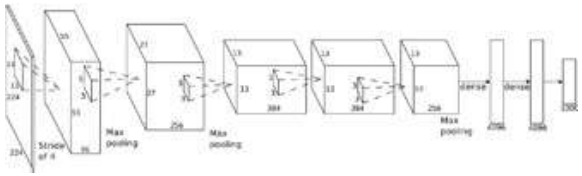
““We build in zero-billion-dollar markets—spaces with no customers and no competitors—until they become billion-dollar industries.” paraphrasing Jensen Huang*

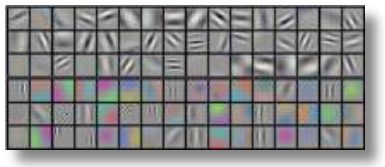
* From https://blogs.nvidia.com/blog/jensen-huang-caltech-commencement-address/?utm_source=chatgpt.com

2012: DEEP NEURAL NETWORKS RISE AGAIN

- "Geoff Hinton, Alex Krizhevsky, and Ilya Sutskever used Nvidia CUDA GPUs to train AlexNet and shocked the computer vision community by winning the 2012 ImageNet challenge," Jensen Huang said, "This was the big moment, the big bang of deep learning. A pivotal moment that marked the beginning of AI revolution."

- Top-5 17% error rate (huge improvement) in 2012 (now ~ 3.5%)





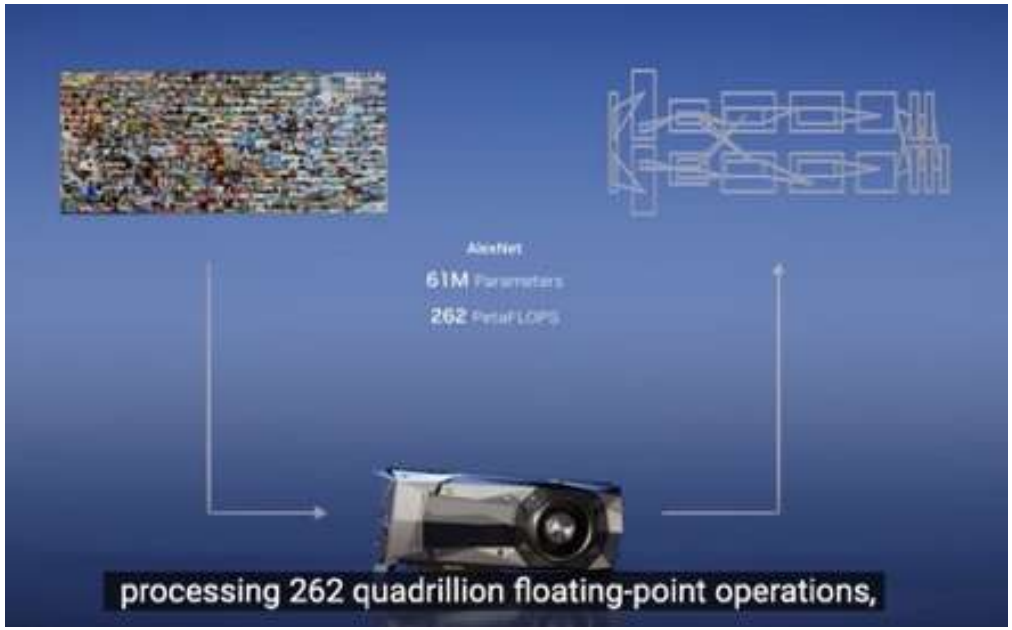
"**Supervision**" network
Year: 2012
650,000 neurons
60,000,000 parameters
630,000,000 synapses

- The 2018 **Turing Award recipients** are Google VP Geoffrey Hinton*, Facebook's Yann LeCun and Yoshua Bengio, Scientific Director of AI research center Mila.

* He was also awarded with John Hopfield the 2024 Nobel Prize in Physics for "foundational discoveries and inventions that enable machine learning with artificial neural networks"

Figure 2. **Outline of the *DeepFace* architecture**. A front-end of a single convolution-pooling-convolution filtering on the rectified input, followed by three locally-connected layers and two fully-connected layers. Colors illustrate feature maps produced at each layer. The net includes more than 120 million parameters, where more than 95% come from the local and fully connected layers.

COMPUTING IS DRIVING AI PERFORMANCES



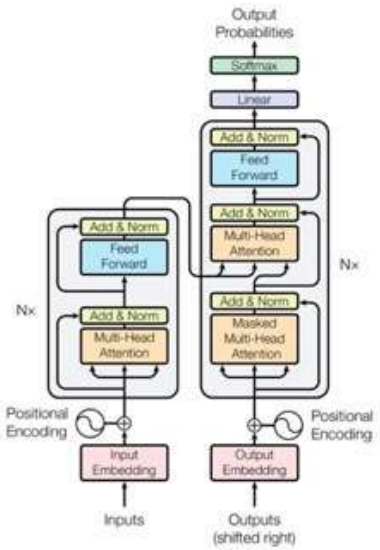
2012: AlexNet
GeForce GTX 580
Won ImageNet Challenge
262×10^{15} FLOPS (262 PetaFLOPS)

2020: GPT-3
323×10^{21} FLOPS (323 ZetaFLOPS)
X 1 000 000 more floating point operations

From GTC 2023 Keynote with NVIDIA CEO Jensen Huang

2017: THE “TRANSFORMER” PAPER FROM GOOGLE

We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.



Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

NVIDIA Delivers DGX-1 Supercomputer in a Box to OpenAI

Aug. 15, 2016

By: Michael Feldman

OpenAI, a non-profit research company devoted to advancing artificial intelligence, has become one of the proud owners of a DGX-1, NVIDIA's so-called "supercomputer in a box," a server specifically designed for machine learning work. The system, which was hand-delivered to the company's headquarters in San Francisco by NVIDIA CEO Jen-Hsun Huang, will be used to run some of OpenAI's most computationally challenging applications.

NVIDIA CEO Jen-Hsun Huang and OpenAI Co-Founder Elon Musk with DGX-1

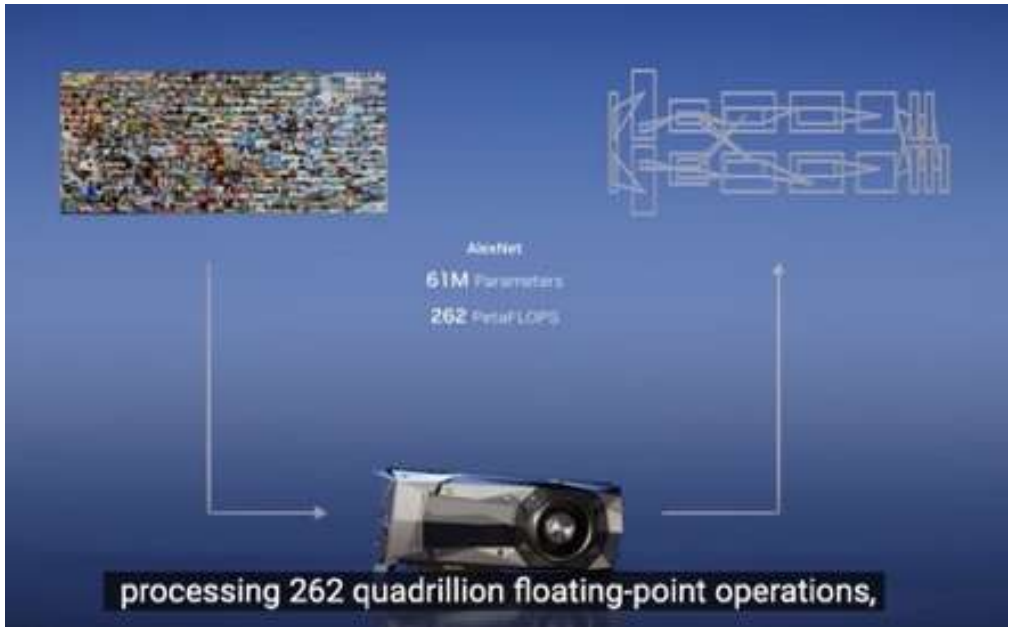
European Processor Initiative

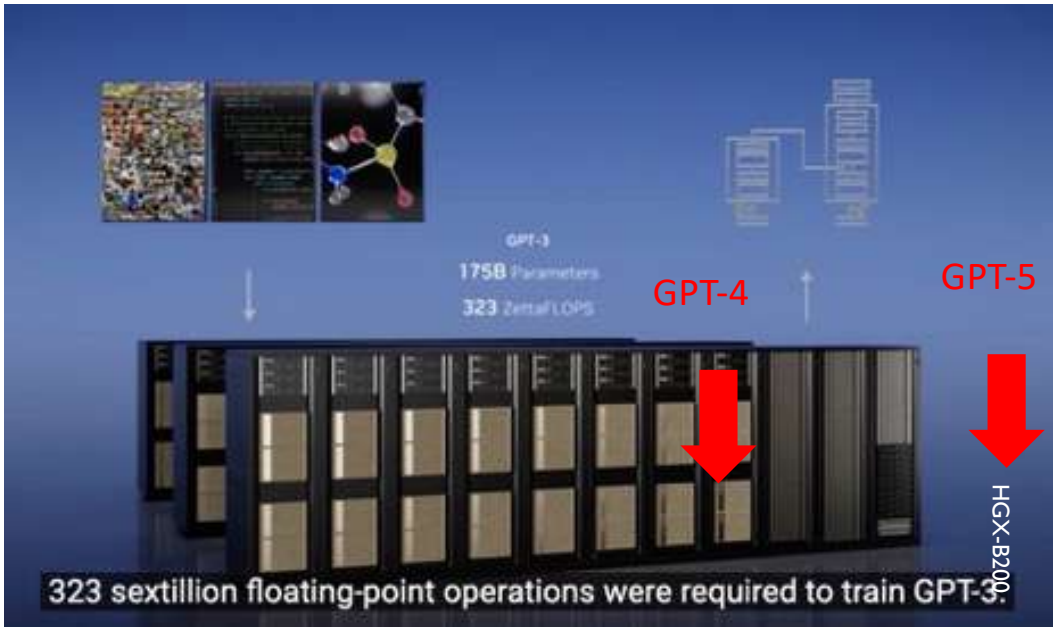
More generally, the DGX-1 will be used to support the company's mission, namely to "advance digital intelligence in the way that is most likely to benefit humanity as a whole, unconstrained by a need to generate financial return." The non-profit is being backed by Silicon Valley icons like Elon Musk and Peter Theil, and managed to attract more than a $1 billion worth of funding at the time it was founded in December 2015. The company only expects to spend a tiny fraction of that amount over the next few years.

One of the premier uses cases is creating a digital entity that can hold a conversation with humans. That requires building a model that encapsulates an understanding of how language works and how it's used to engage people when they talk. Understandably, such a model is hard to train and takes bushels of computing power, which is where the DGX-1 comes in. OpenAI Research Scientist Ilya Sutskever characterized the NVIDIA machine as "a huge advance" and would allow them to explore new sets of problems that could not be attempted before.

From https://www.top500.org/news/nvidia-delivers-dgx-1-supercomputer-in-a-box-to-openai

COMPUTING IS DRIVING AI PERFORMANCES





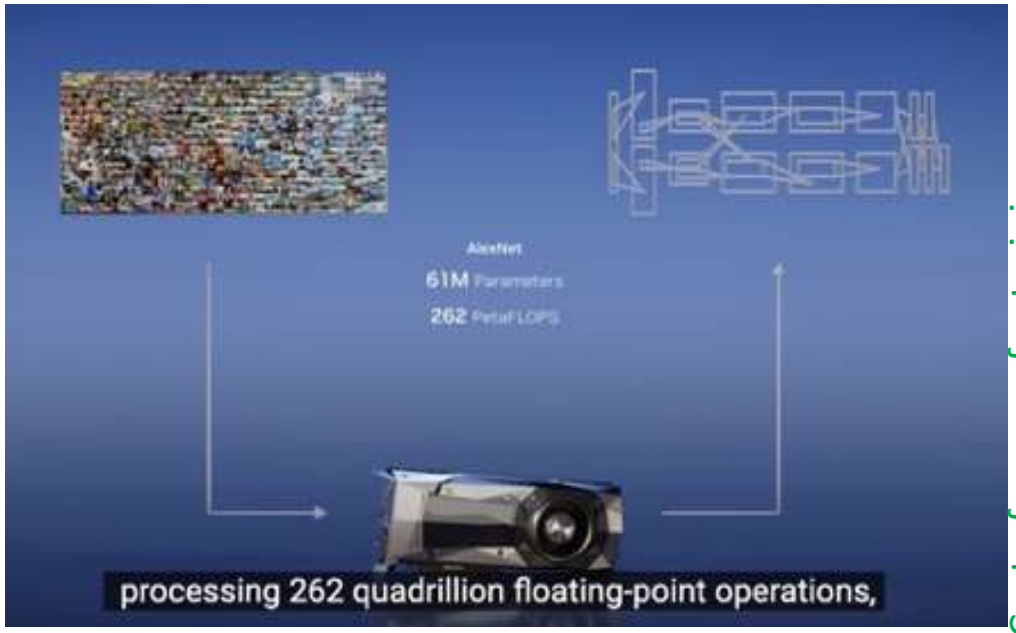
2012: AlexNet
GeForce GTX 580
Won ImageNet Challenge
262 x 10^{15} FLOPS (262 PetaFLOPS)

2020: GPT-3
323 x 10^{21} FLOPS (323 ZetaFLOPS)
X 1 000 000 more floating point operations

From GTC 2023 Keynote with NVIDIA CEO Jensen Huang

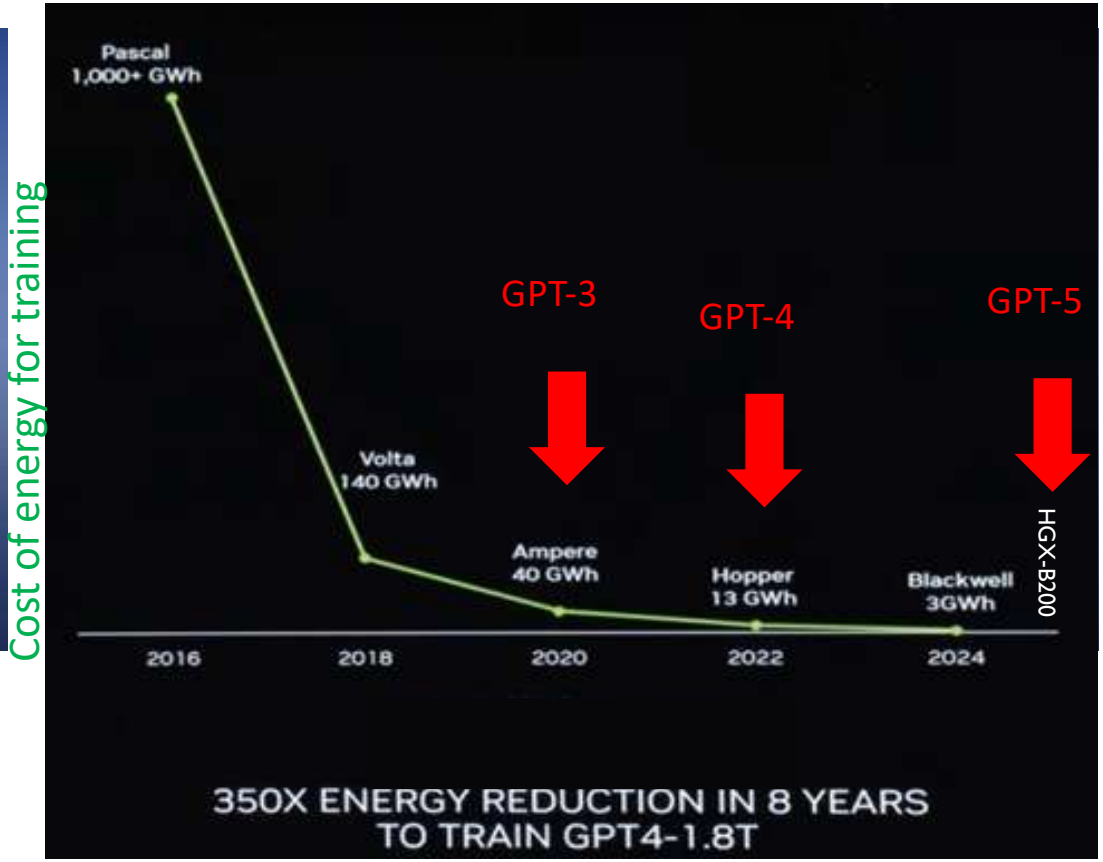
Cost of energy for training is a limiting factor!

COMPUTING IS DRIVING AI PERFORMANCES



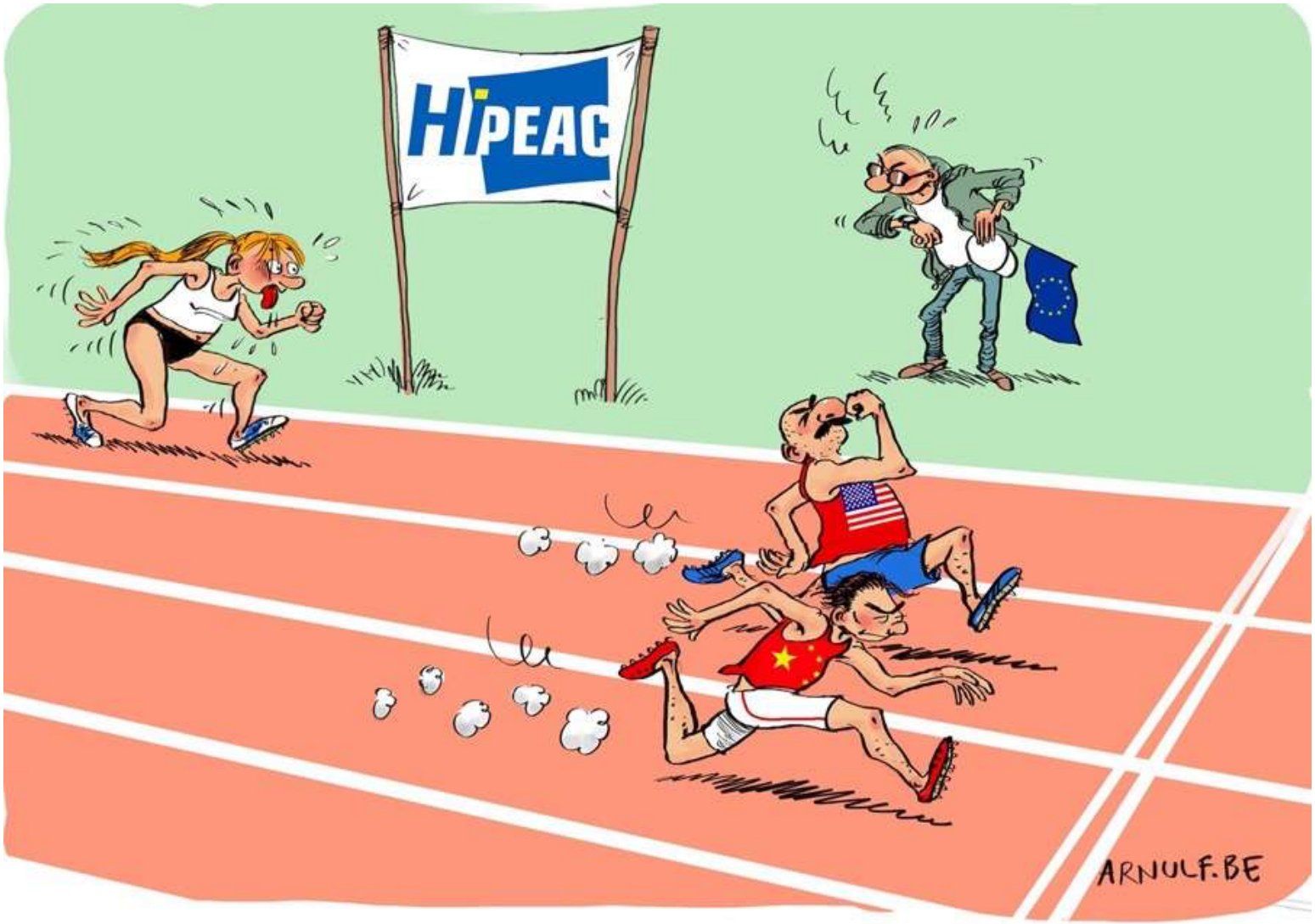
2012: AlexNet
GeForce GTX 580
Won ImageNet Challenge
262 x 10^{15} FLOPS (262 PetaFLOPS)

From GTC 2023 Keynote with NVIDIA CEO Jensen Huang



Cost of energy for training is a limiting factor!

IS THE CONCLUSION…



OR HOW CAN EUROPE...

HOW EUROPE CAN BE BACK IN THE RACE?

- Learn from the past (see previous slides), what Europe did wrong, what other did well…
- The weakness of Europe is its fragmentation

The strength of Europe is its fragmentation!

- diversity
- small flexible structures (SMEs)
- a specific market

We need to ***work together*** with a long-term vision and accept to make some concessions for the "***global optimization***" – no CEO or government to give the direction…

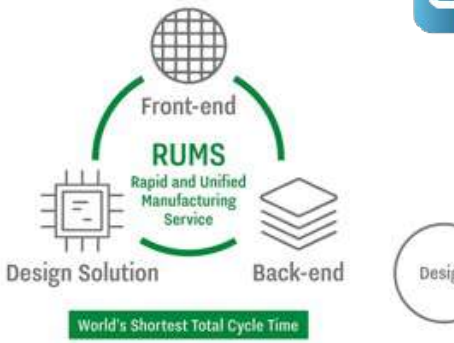
Need of real « Important Projects of Common European Interest » ☺ like for airplanes (Airbus)

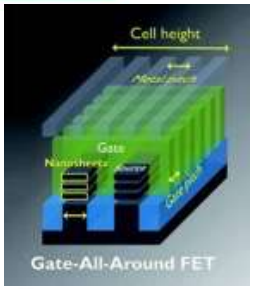
And ***Europe should not necessarily copy what US and China are doing*** but have an adapted strategy

EXAMPLE OF RETHINKING SEMICONDUCTOR MANUFACTURING

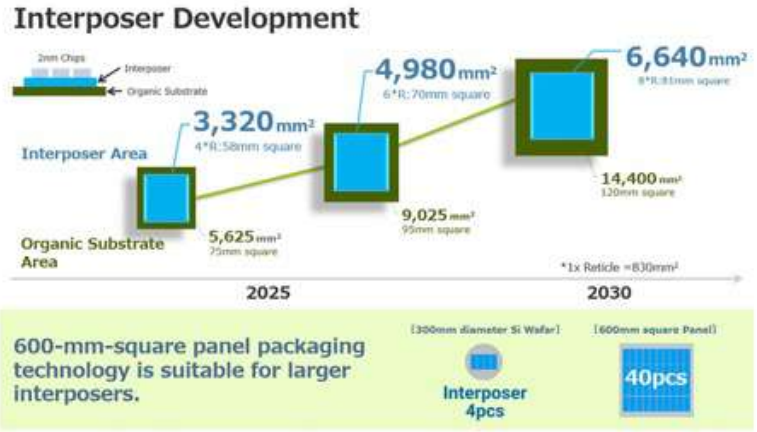
The Japanese Rapidus approach*:

- RUMS (Rapid and Unified Manufacturing Service)
- “Raads” (Rapidus AI-assisted Design Solution) supports design optimization using AI, and “DMCO” aims to mutually optimize design and manufacturing
- “GAA for the front-end process and “chiplets” for the back-end process were developed to deliver the world's fastest cycle time.
  - The ***single-wafer process*** can flexibly respond to the production of a wide variety of specialized products.

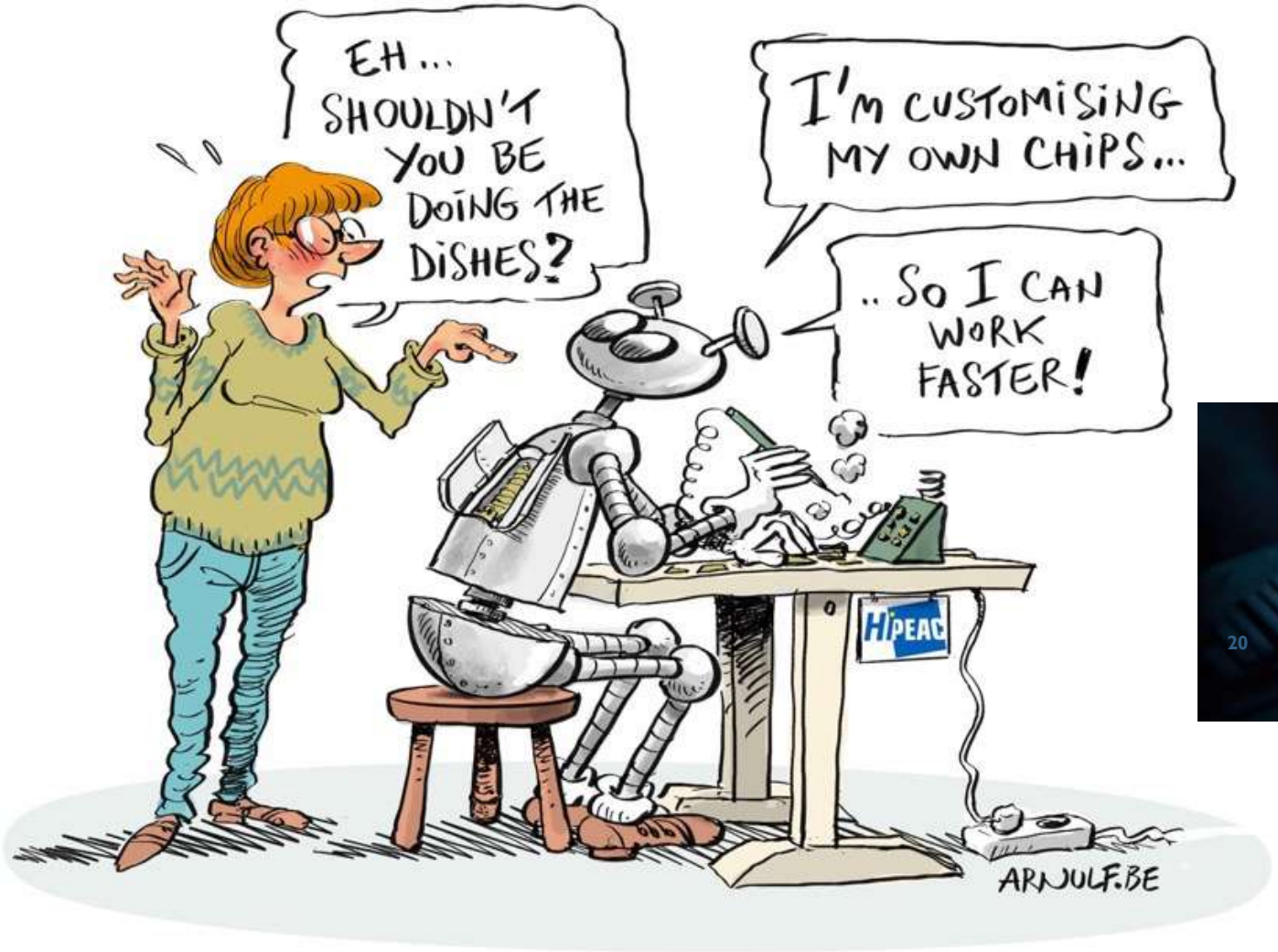


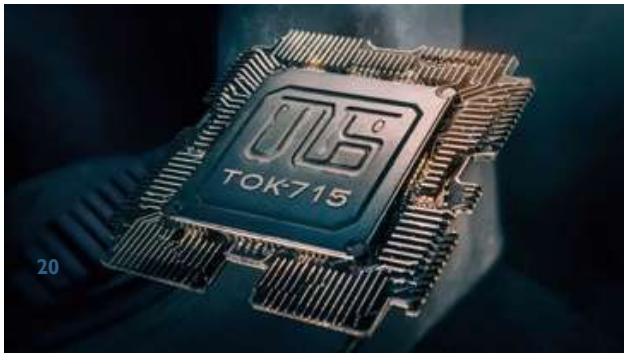


On July 18, 2025, Rapidus unveiled its first wafer featuring a gate-all-around (GAA) transistor fabricated using a 2nanometer (nm) process. This milestone was reached just over three months after transporting an extreme ultraviolet (EUV) lithography system in December 2024 by air from the Netherlands.
By June 2025, the first production lot had been processed, and the resulting GAA transistor wafer was showcased at a customer event in July. Rapidus attributes this exceptionally rapid establishment of production capability to its defining manufacturing strength.



- *From https://www.rapidus.inc/en/business/#technology
- ** From https://www.rapidus.inc/en/interview/it0003/

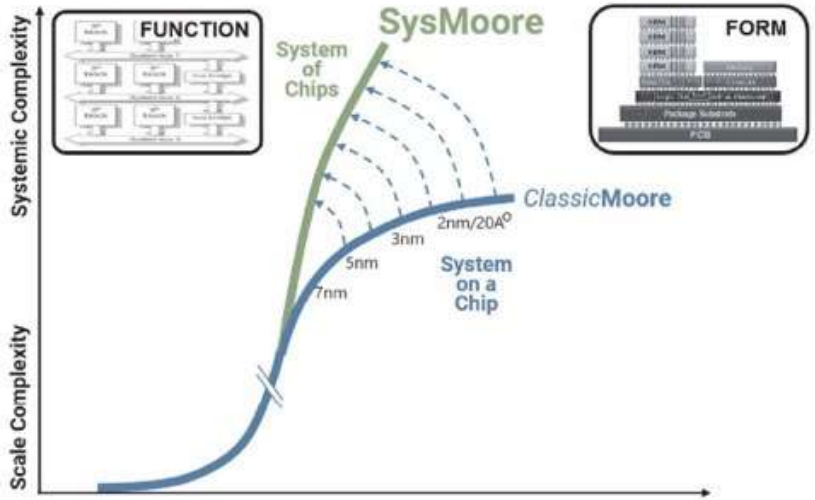

EH... SHOULDN'T YOU BE DOING THE DISHES?
I'M CUSTOMISING MY OWN CHIPS...
..SO I CAN WORK FASTER!
HiPEAC
ARNULF.BE

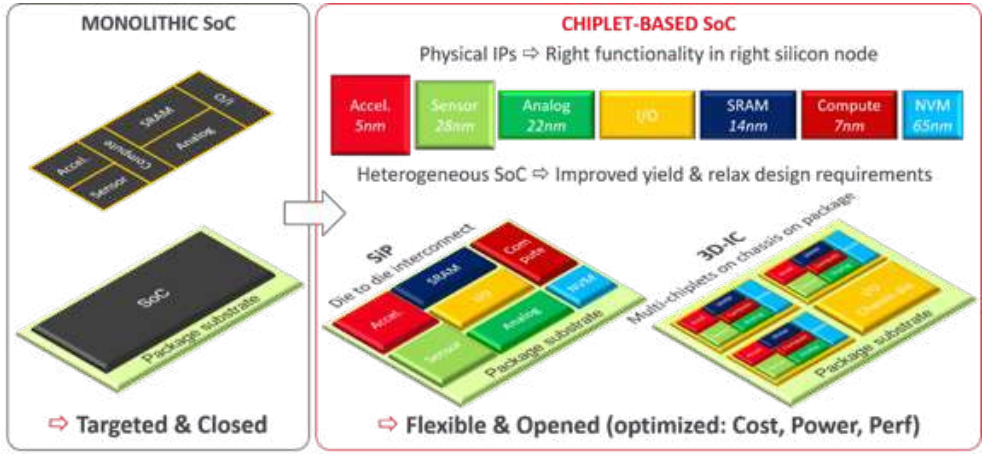

TOK-715

MOVING TOWARDS SYSTEM-OF-CHIPLETS

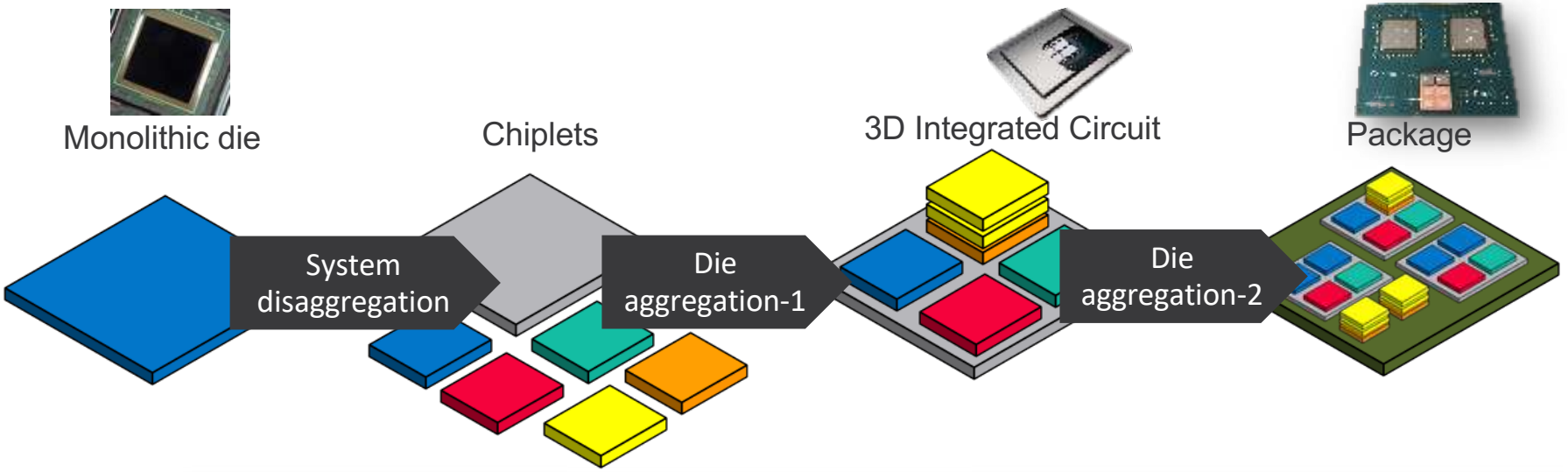
- **Application trends**
  - More processing & data (AI, Security…)
  - Reuse of legacy (ISA, IO interface…)
  - More modularity & scalability (low to high end)

- **System-of-chiplet enables Flexible, Faster and Cheaper designs with sovereignty.**
  - Cost reduction (right silicon node per function)
  - Enhanced customization (modularity)
  - Enhanced integration of additional functionality (design & system scalability)
  - Faster Time-to-Market (chiplet reuse)
  - Reduced sourcing dependency
  - Adding differentiating features





CHIPLETS: THE NEW IC DESIGN PARADIGM



Architecture

Design
Design flow
Wafer-level integration

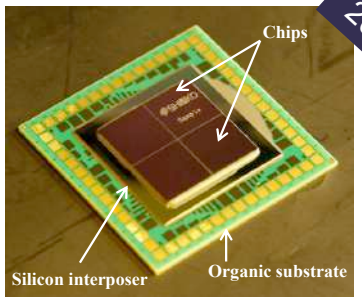
Design
Design flow
Package-level integration

System Technology Co-Optimization

Up to 100x gain on Power Efficiency with 3D

HPC AND AI CONVERGING ROADMAPS AT cea

European Processor Initiative



2012

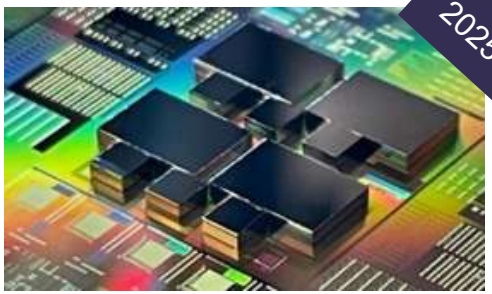
Metallic Passive interposer

- ✓ Chip-to-chip side-by-side communication

2019

Active interposer (ENoC) Intact

- ✓ Extended communication capability (increased distance, routing, power management, …)



2025

Photonic interposer (ONoC) Starac

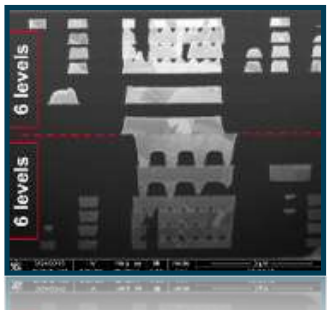
Optical communication:
- ✓ Reduction of on-chip latencies
- ✓ Higher throughput
- ✓ Lower energy consumption
- ✓ Scalability

More than 14 years expertise on Silicon Interposers

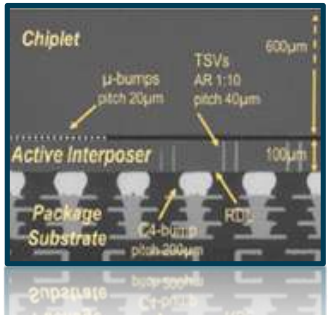
Y. Thonnart et al., "POPSTAR: a Robust Modular Optical NoC Architecture for Chiplet-based 3D Integrated Systems," Proc. DATE, 2020, p. 6.
D. Saint Patrice et al, "Process Integration of Photonic Interposer for Chiplet-Based 3D Systems" Proc. IEEE ECTC, 2023

INTERCONNECT TECHNOLOGIES FOR HETEROGENEOUS INTEGRATION : TECHNOLOGICAL DEVELOPMENTS

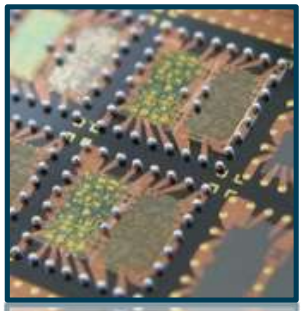
3D sequential Integration
2 tiers including Photodiodes, LED, High & low temp. Transistors



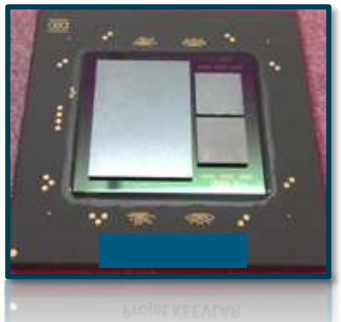
Direct hybrid bonding
Wafer-to-wafer
Min. pitch = 200nm
Die-to-wafer
Min. pitch = 4 µm
Self-assembly



Active interposers
Through-silicon-Vias
TSV 0.3 to 80µm Φ
RDL 20µm pitch
Cu pillars 10µm Φ



Fan-Out-Wafer-Level-Packaging
(FOWLP)
Heterogeneous
System-in-Package
Thermal extraction



Die-level packaging
Defence
Power modules
Cryo-packaging for quantum computing

nm — 1 µm — 10 µm — 100 µm — mm

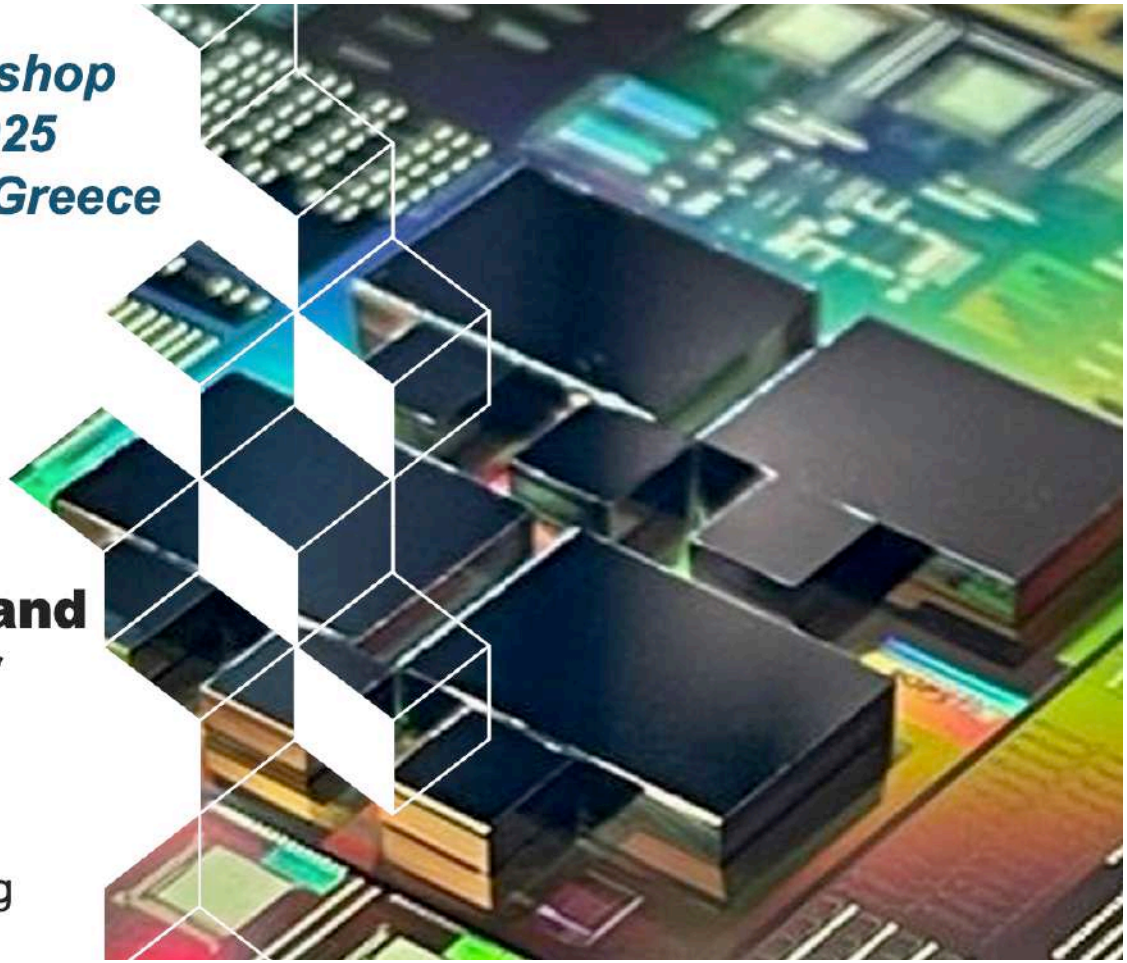
Device ← Large range of interconnect dimensions & materials → System

Source : E. Ollier, Chiplet Summit 2025

MORE TECHNICAL DETAILS IN

European Processor Initiative

EPI Codesign Workshop
10-11 September 2025
FORTH, Heraklion, Greece



Chiplets and Interposers: Architectural Trends, Integration Technologies, and Strategic Opportunities for Europe

Denis Dutoit, CEA-List

Program Manager – Advanced Computing

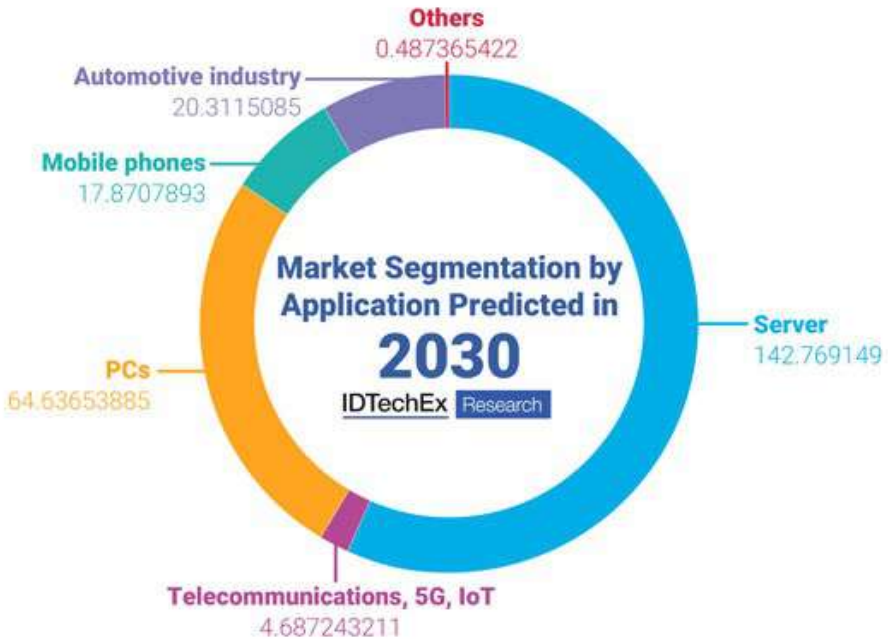
DIE-TO-DIE INTEROPERABILITY

- Ensuring interoperability requires
- **PHY compatibility**: same physical interface (number of communication links, side-band signals), same link training scheme, electrical & impedance compatibility, compatibility of data rates, frequencies, transmission schemes (SDR, DDR) and modulations (NRZ, PAMx), timing settings (ΔUI) and compatibility of pitches and bumps arrangements
- **Data link compatibility**: identical packet organisation (number of flits, field position in header, payload, etc.) and exchange protocol (Flow Control, Credits exchange, etc.).
- **Protocol and transport layer compatibility**: identical Bus (AXI, AHB, CHI, CXL, ...) or Streaming (AXI Stream, ...) communication protocol, identical organization of the transmission of the various communication channels on the link (i.e. AXI = 5 channels to transfer), compatible memory map, etc…
- For instance:
  - Two UCIe supporting the AXI protocol may not interoperate if the communication flits transporting the AXI fields are not defined exactly the same (still not defined by UCIe standard)

FOCUS ON UCIE

- UCIe specification, first issued in February 2022 and promoted by a large consortium of major players in the semiconductor market, has rapidly established itself as the standard in the HPC field, and is gaining increasing traction in the automotive sector.
  - UCIe specification is the only one covering not only the PHY but the full stack up to the protocol layer, with focus on interoperability and quick technical adoption for fast business ramp-up.
  - Since Q2/20022, IP Vendors developments and new SoC design announcements are all targeting UCIe only
  - Version 2.0 of the specification issued in August 2024 introduced new features targeting industrial deployment
    - Link manageability, DFx, Enhancement for the Automotive market, ..
  - Version 2.0 of the specification issued in August 2024 introduced advanced features for 3D integration (vertical links)

THE MARKET WILL REACH US$411 BILLION BY 2035, DRIVEN BY HIGH-PERFORMANCE COMPUTING DEMANDS ACROSS SECTORS SUCH AS DATA CENTERS AND AI.



- **Server**: High performance computing demands drive significant adoption
- **Telecommunications, 5G, IoT**: Chiplets enable efficient network solutions
- **PCs**: Enhanced performance and customization are key drivers
- **Mobile Phones**: Chiplets contribute to advanced functionality and efficiency
- **Automotive Industry**: The integration of diverse functionalities supports automotive innovations
- **Others**: Various applications benefit from the modularity of chiplets.

Source: https://www.microwavejournal.com/articles/42896-chiplet-market-to-reach-us411-b-in-semiconductor-industry-by-2035

EUROPEAN LANDSCAPE

- EuroHPC projects pioneered heterogeneous integration: ExaNoDe, Mont Blanc 2020, EPI, DARE

Common platform

RISC-V chiplet-based
HPC HW and SW

- ChipsJU chiplet-related projects are ramping up:
  - E2PACKMAN (package oriented including SiP for chiplets)
  - Coming soon, chiplet design for automotive

Packaging: E2PACKMAN (European Consortium for Accelerating Innovations in Electronic Package Manufacturing

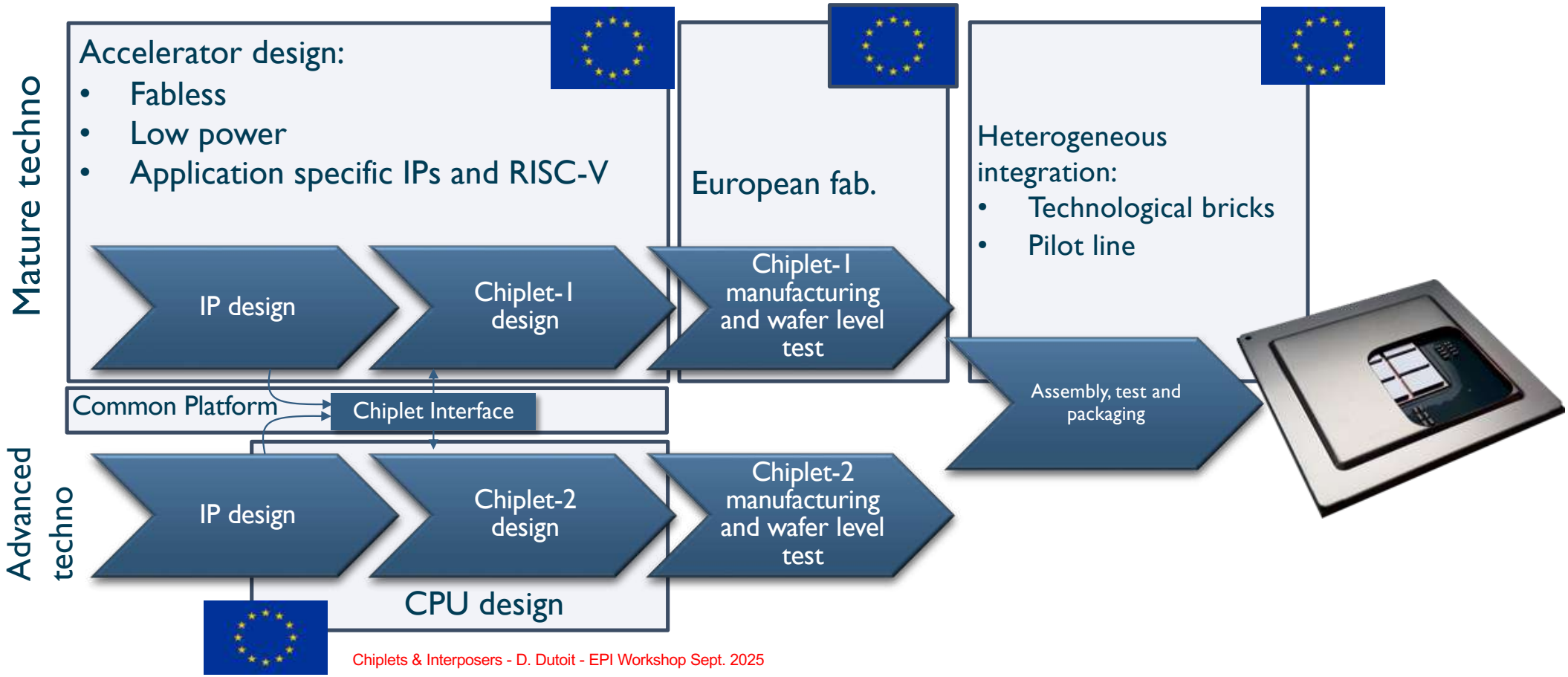
Chiplet design for automotive: under gant negociation

- European Defence Fund: ongoing call for chiplets for defence

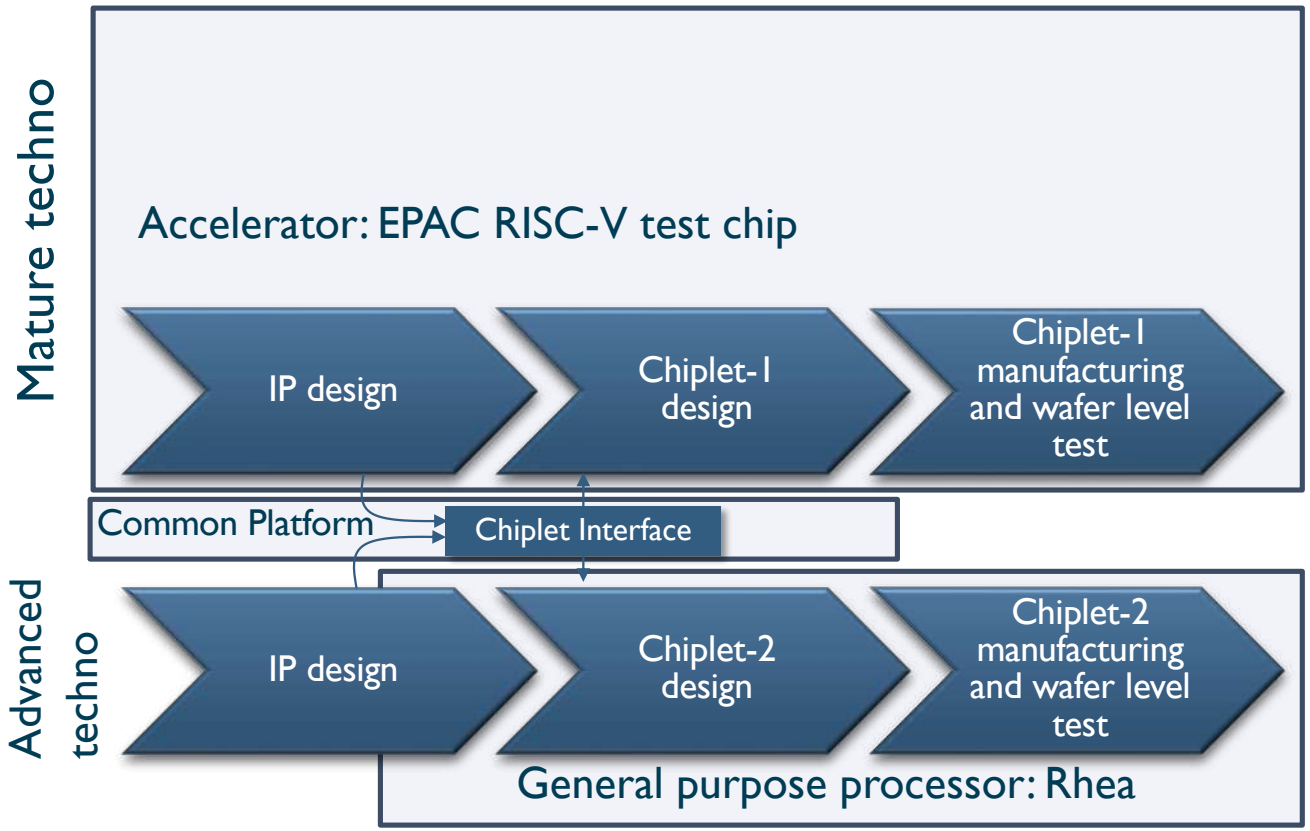
Chiplet for defence: ongoing call

OPPORTUNITIES TO REGAIN SOVEREIGNTY IN EUROPE

Chiplets open the door for multi-techno computing components



EPI CONTRIBUTIONS

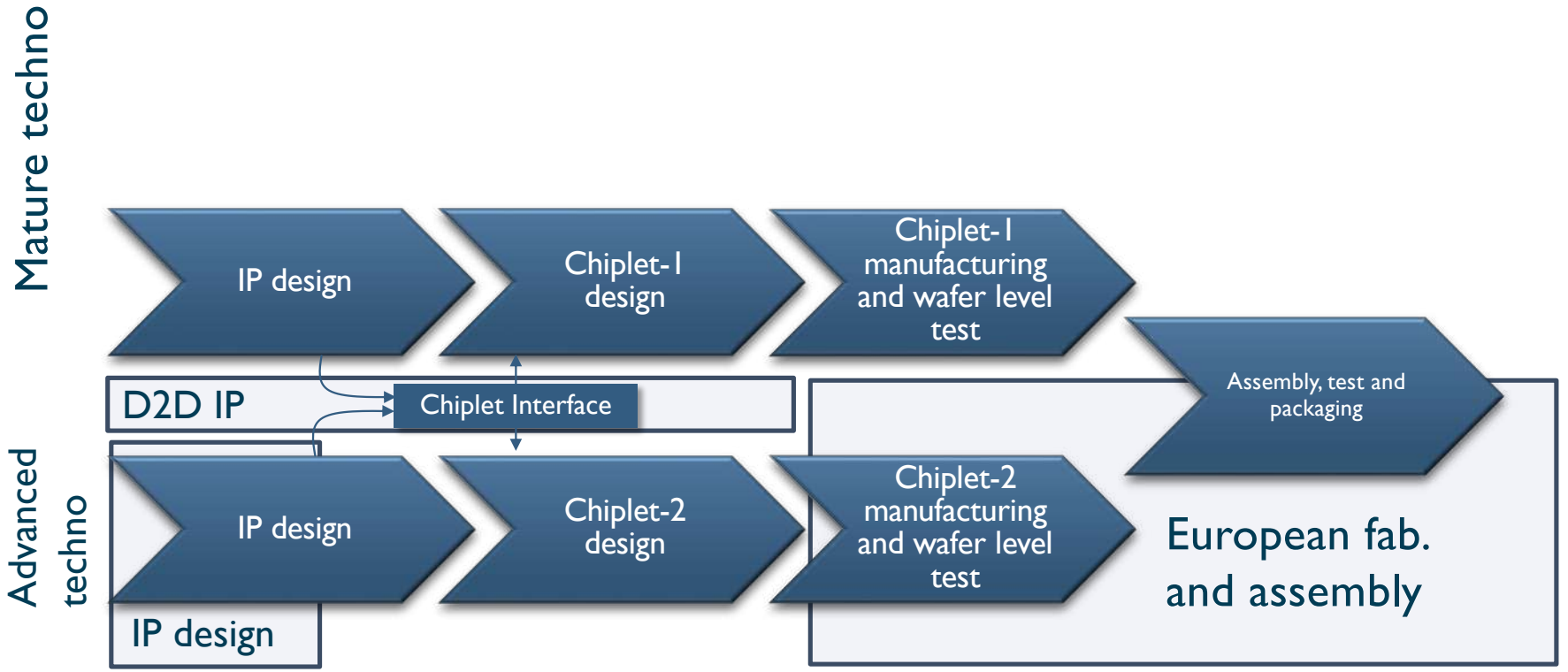


European Processor Initiative

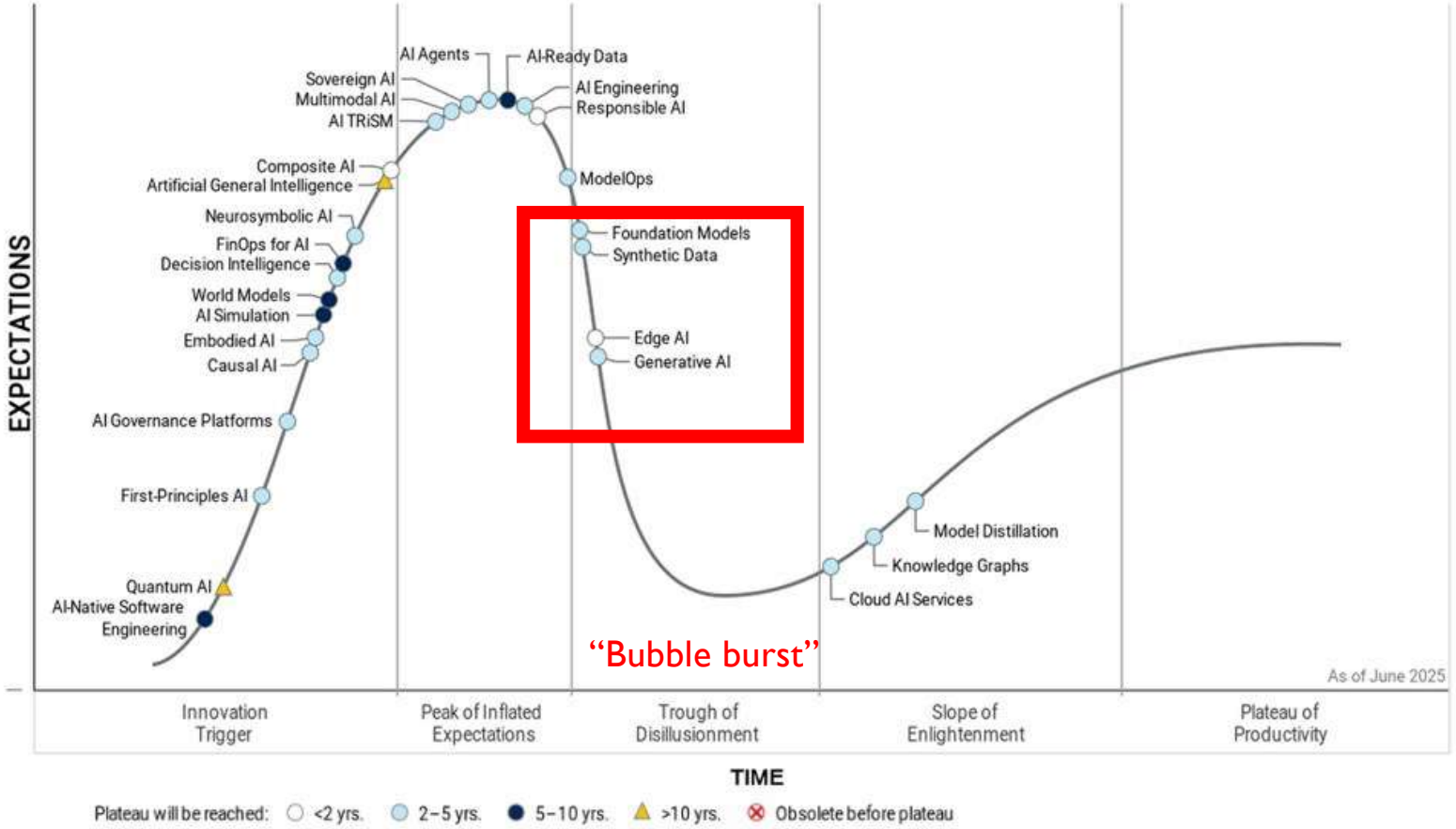
EuroHPC Joint Undertaking

STILL MISSING IN EUROPE

IP portfolio inc. die-2-die, advanced technology node foundries, OSAT



AND WHAT ABOUT ARTIFICIAL INTELLIGENCE?



* From https://www.gartner.com/en/newsroom/press-releases/2025-08-05-gartner-hype-cycle-identifies-top-ai-innovations-in-2025

COMPUTING CLUSTERS

The race for bigger clusters (for AI) in US is mainly driven with a ***supremacy drive***:

ASI – Artificial Super Intelligence- is supposed to need such compute power and is believed to be a tool for ensuring supremacy (“event horizon” or “the singularity”)

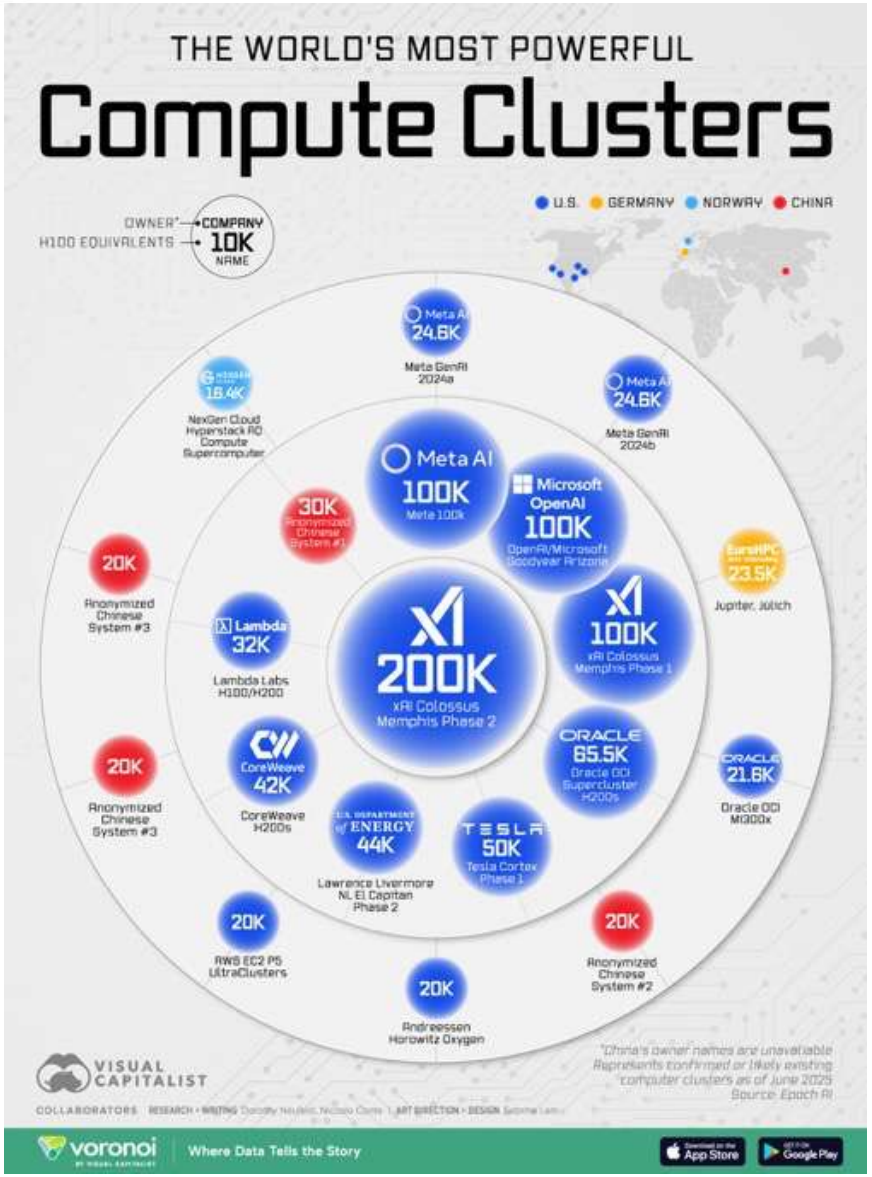
Is it a new Manhattan project?

It is like HPC versus cloud:

- 1 load (training) using a lot of compute resources (HPC) versus several loads (Inference for various users) sharing a node (Cloud)

Current HPC mainly running digital twins (like Destination Earth)

But it is perhaps not what Europe should do for ***offering AI to its citizens*** (as we don’t have big cloud providers in Europe) …



European Processor Initiative

Risk management*: how to ensure the position of Europe if ASI is possible?

* Not in the scope of this presentation

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.

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

- **Inference Cost**: GPT-4 costs 3 times more than the 175B parameter Davinci, due to the larger clusters 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

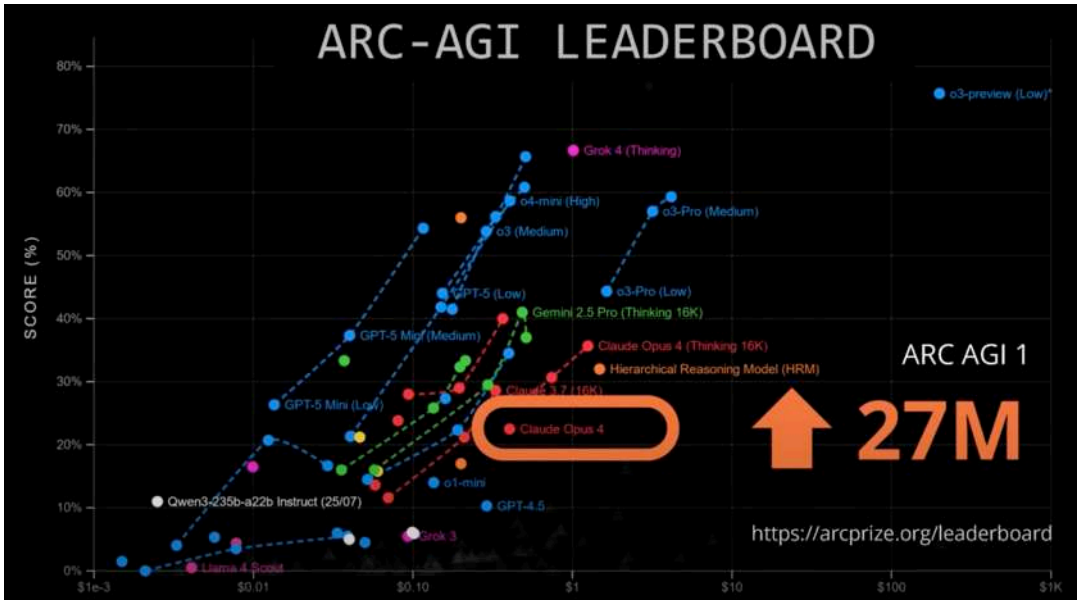
Current models of about 10B parameters have better performances on specific tasks than the original ChatGPT of 2022

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 weight" models are catching up closed models with few months delay.
They become political assets (China vs. US)



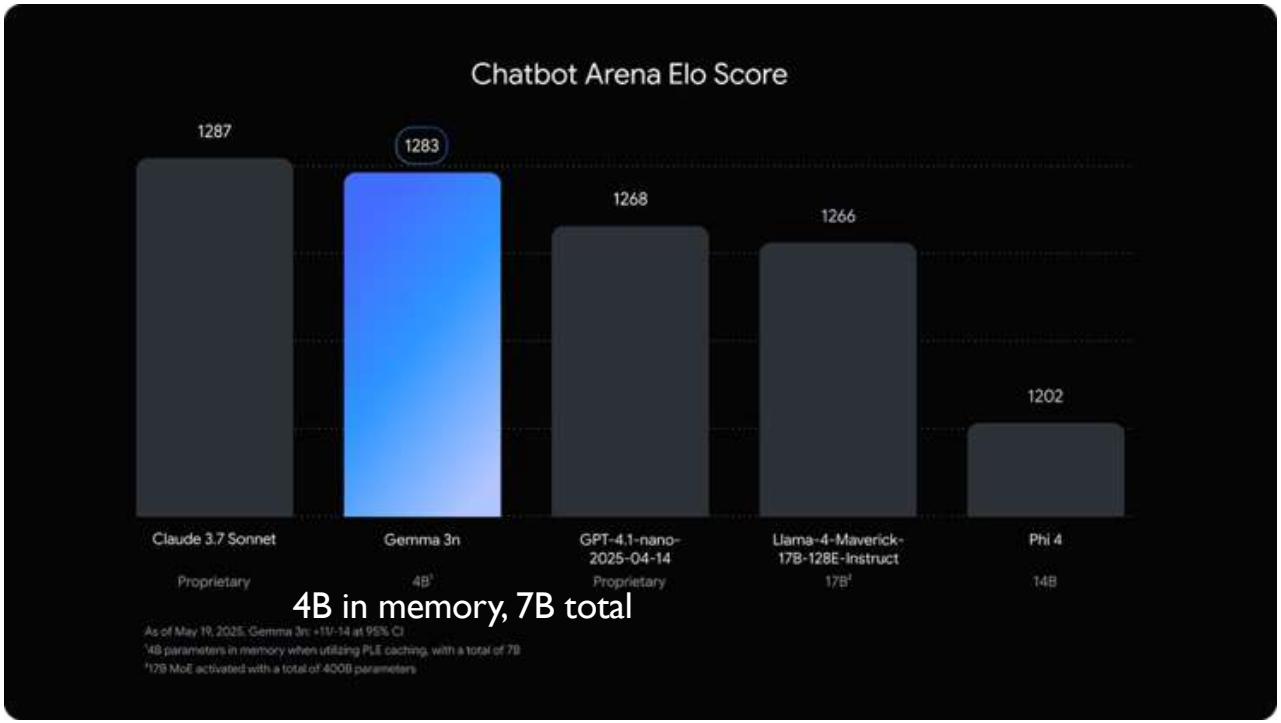
https://arcprize.org/leaderboard

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

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) ***

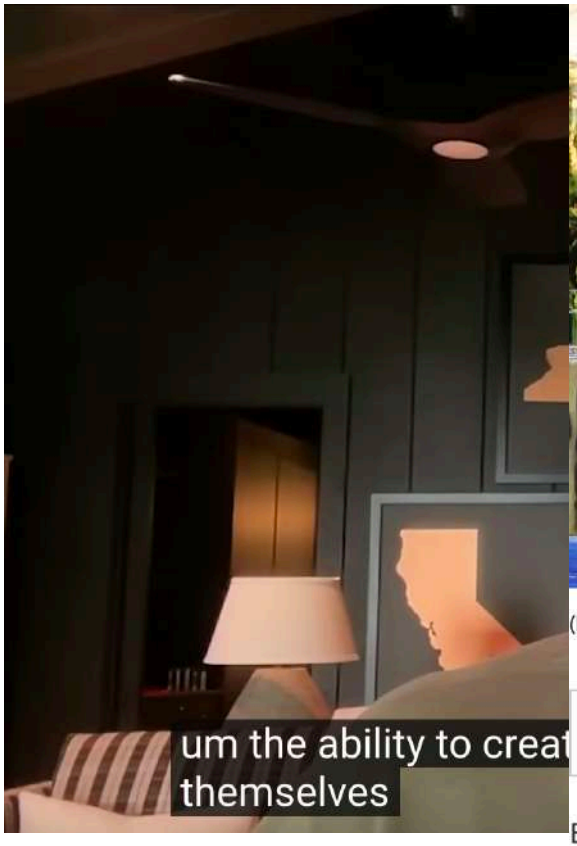


Capability	Benchmark	ERNIE-4.5 -0.3B-Base	Qwen3 -30B-A3B-Base	ERNIE-4.5 -21B-A3B-Base	DeepSeek-V3 -671B-A37B-Base	ERNIE-4.5 -300B-A47B-Base
General	C-Eval	40.7	87.2	**88.0**	90.2	**91.5**
	CMMLU	39.8	86.0	**86.2**	88.2	**91.2**
	MMCU	37.2	88.8	**94.0**	94.0	**95.9**
	AGIEVAL	28.5	**72.8**	68.4	75.8	**78.4**
	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**
	CMATH	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!)
- Distributed computing on demand

Image generated by ChatGPT

STRUCTURE OF APPLE INTELLIGENCE

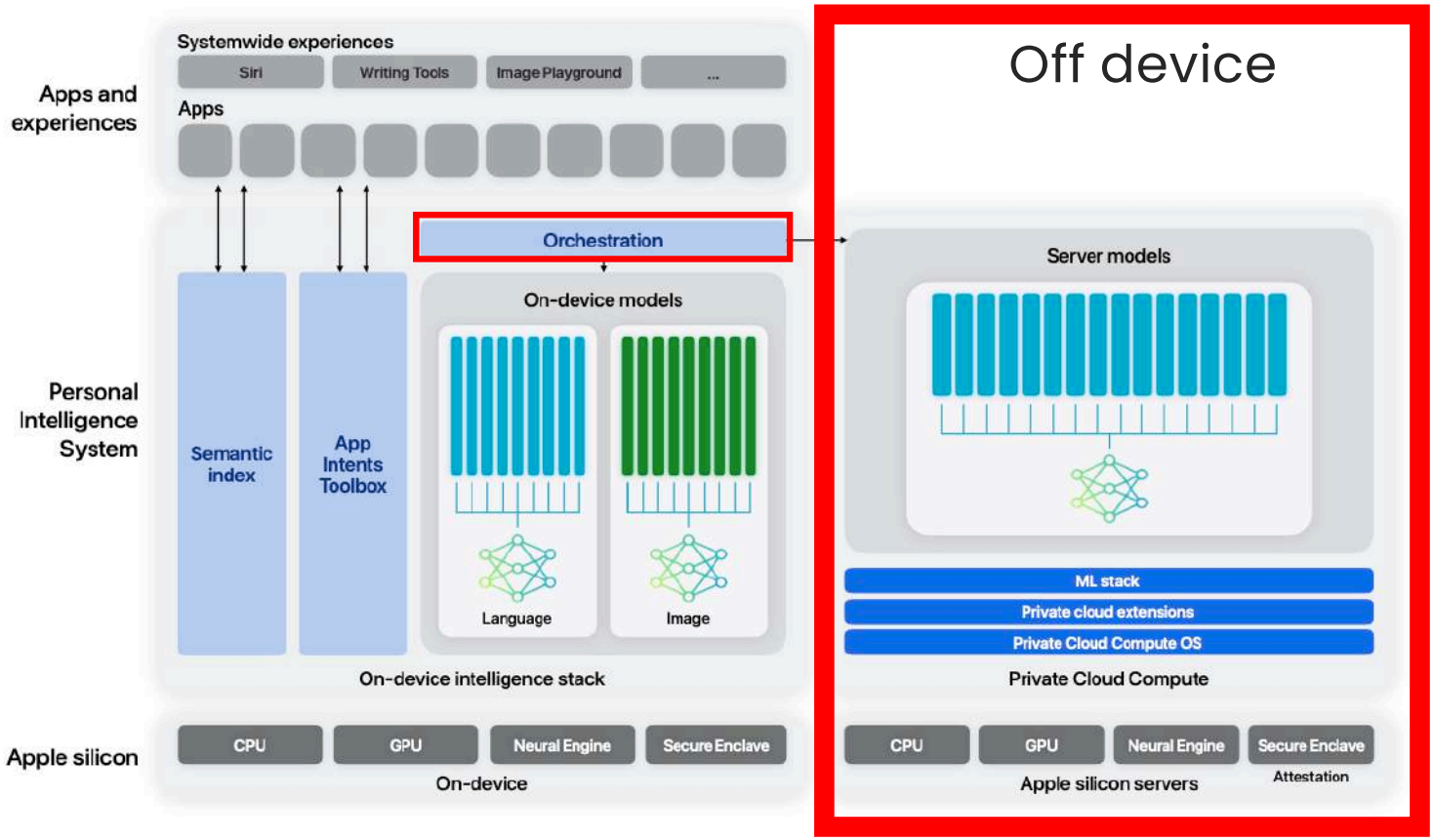


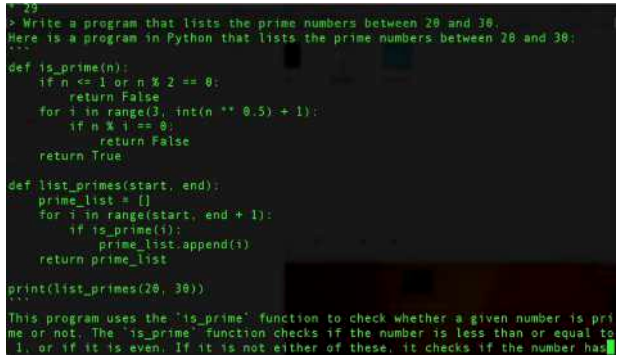
Figure 2: Architecture of Apple Intelligence with adapters for the language on-device and server models and the image models. In this report we are only describing the text models.

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

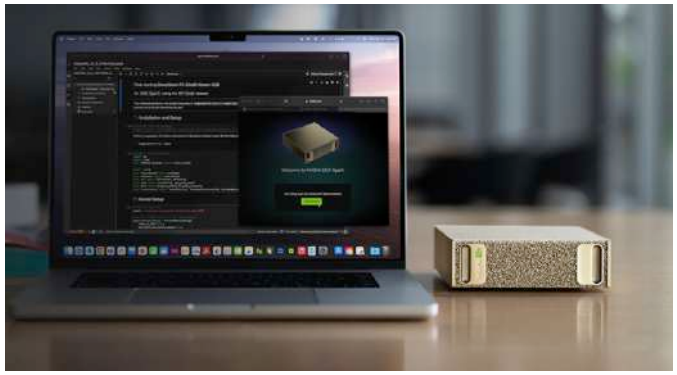
GENERATIVE AI IN THE COMPUTING CONTINUUM…

(US) AI data center: xxxx MW

LLM running locally on Mac mini: about 20W

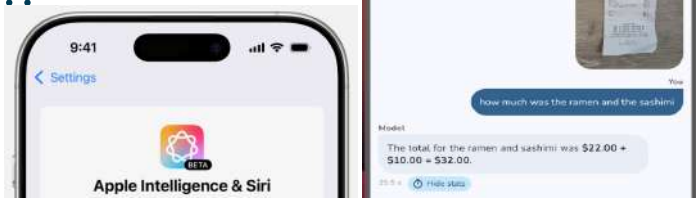


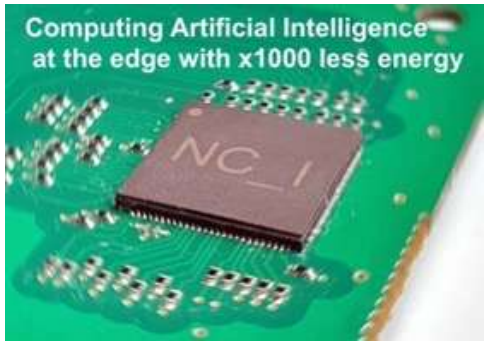
IoT LLM box: 1-2 W



NVidia DGX Spark: 150W ???

AI smartphones: 5-10W



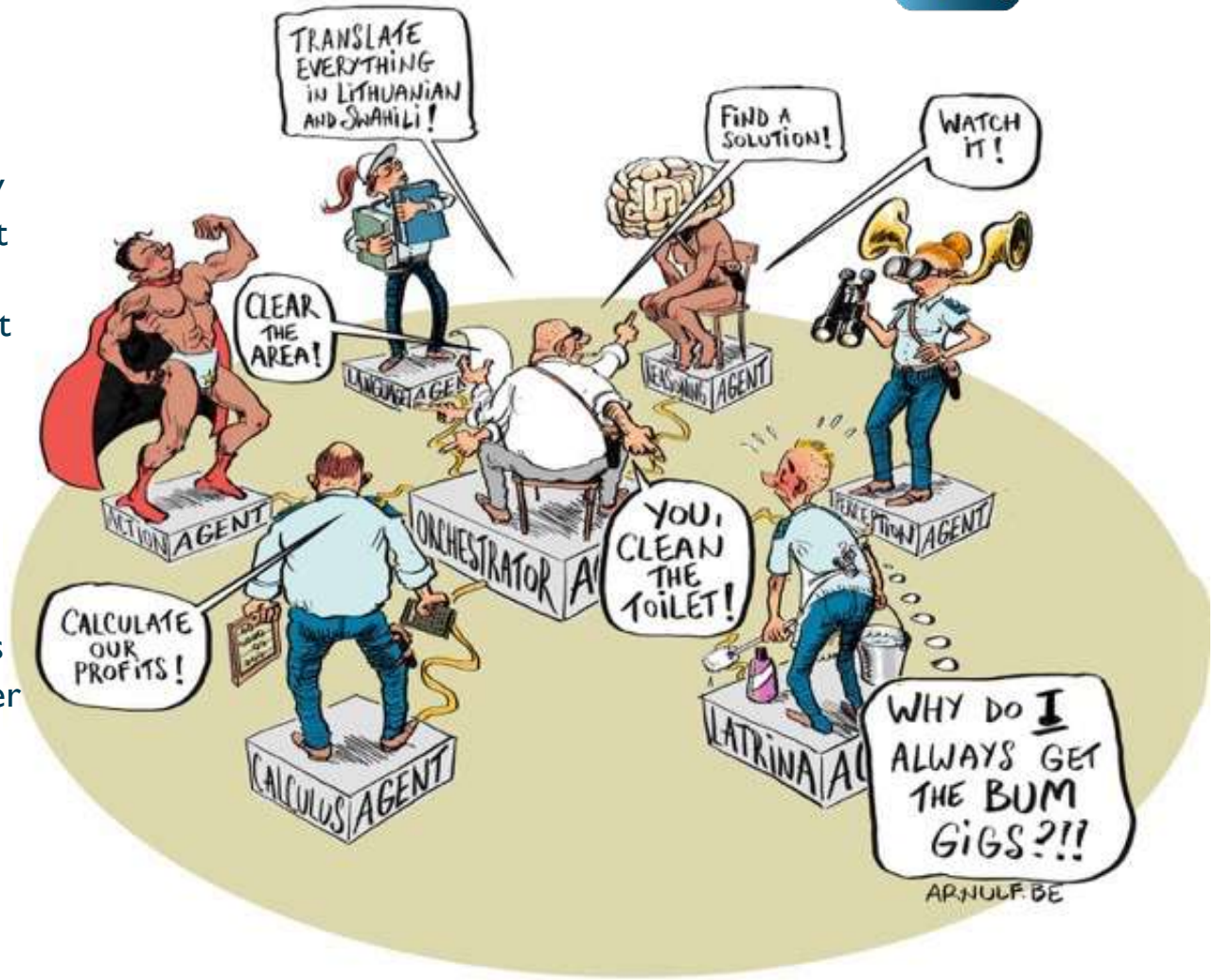


Object detection on
HD images at 30FPS for 23mW

A small thought experiment:

- 500 M (5.10^8) smartphones in Europe
- Typical mid-range 2025 phones deliver roughly (≈ 10–25 TOPS) (10^{13}) (will be most of market in 3-5 years from now)
- 40% of them recharging at night (so they might be remotely available for computing)
- Therefore, if we can use all of them into a super distributed compute fabric in 2030, the potential compute power will be of **2 ZOPS** (2.10^{21})!!!

- But most of the use of AI by European citizens could be done more or less locally with smaller models orchestrated together



INTEROPERABILITY PROTOCOLS FOR FEDERATED AGENTIC AI

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

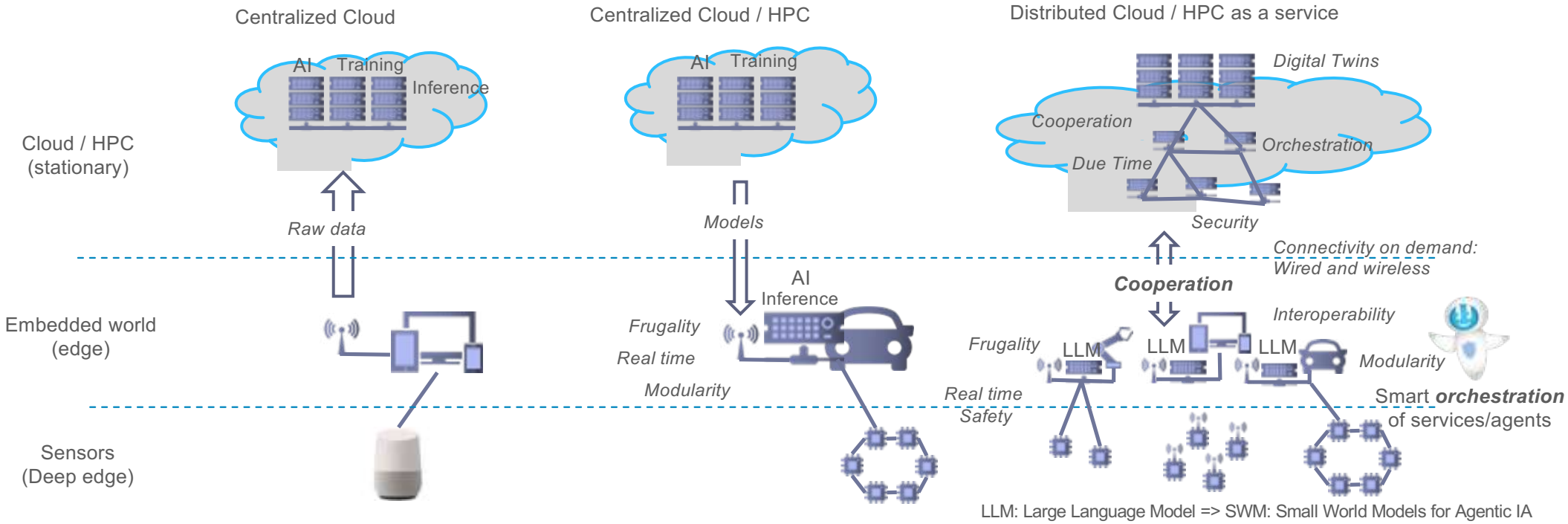
- rely not only on functional requirements (like MCP*, A2A, …)
- but also on non-functional requirements (providing enough information such as the orchestrator can selects the right services, e.g. according to criteria such as **response time, potential level of hallucinations, cost**, localization,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

* Model Context Protocol from Anthropic
** for example Agent Payments Protocol from Google

EVOLUTION OF COMPUTING: CLOUD, CPS, IOT, AI → NEXT COMPUTING PARADIGM



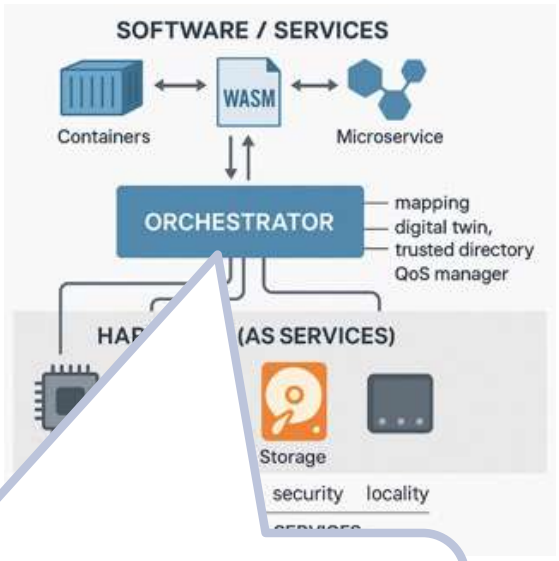
Cloud computing → Intelligent Cyber-Physical Systems → Distributed and cooperative systems

From Denis Dutoit, inspired from HiPEAC 2023

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**
  - Code (*not only data*) can migrate from one location to another
- They are **orchestrated by an intelligent system** which selects the services from your own specific requirements.
- **Taking into account non-functional properties** like time, latency, localization, privacy, cost, security, …..



Various devices can be aggregated to form a super device, with resource sharing and fast/low latency connectivity, for improved capabilities. Services can be seamlessly transferred to the most suitable device.

PARTIAL CONCLUSION*

Europe should ***not copy solutions*** that are not fit for it due to its fragmentation

- But working **together with common goals** and interests to avoid “Brownian noise”

Defining a **common European approach to create an ecosystem for chiplets** and fast design of interposers is key

- Create fast and flexible solutions fitting European market (flexibility, diversity, fast TTM instead of big quantities)

Concentrating compute power into multiple GigaWatt data centers is perhaps not the best approach for offering AI benefits to European citizens (*except for supremacy goals*)

- Create **smaller and diverse compute fabrics** that can be on premises (from cites, companies, home, car, edge,..) based mainly on European chiplets ecosystem
- **Working on interoperability across technologies (from chiplets to services)** is important for the future of Europe technologies, leveraging on its diversity and creating active ecosystems from start-ups, SMEs to big companies
- Interoperability, the notion that **everything** (including hardware) **is a service** will facilitate the building of a heterogeneous computing fabric where Europe has leverage: open ISAs and accelerators, chiplet-based modularity, memory-safe system software, robust toolchains and more modular/maintanable software that make performance, energy and «green» efficiency, reliability, security and safety first-class citizens.

It is the right moment due to this disruptive period due to AI : "A New Golden Age for Computer Architecture”**

- but Europe need to react fast…

** not including software, Open Source, etc…

** Quote from Dave Patterson, cf for example https://www.youtube.com/watch?v=aA5pqklkkvI


“The best way to predict the future is to invent it.”
Alan Kay