

EPI Forum

Barcelona, 9–10.10.2024.





SIPEARL

Rhea GPP technology, key features of Rhea1,
targeting HPC and AI workloads with best-in-class energy-efficiency

Exascale system level EPI view

October 2024

Contents

- 04 Introduction
- 09 The evolving landscape of HPC
- 18 AI: LLMs overview
- 26 Rhea1 key features



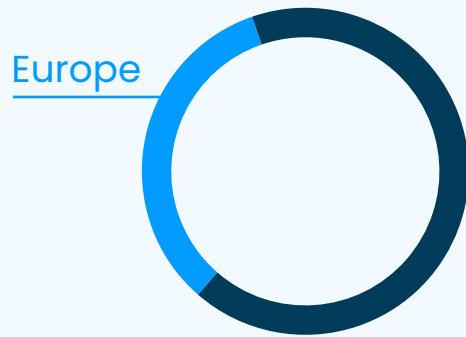
Introduction

A bit of history

Supercomputing industry: EU behind the curve

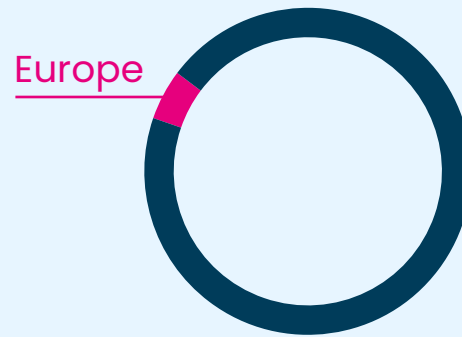
1/3

of global supercomputing
ressources are consumed
by Europe⁽¹⁾.



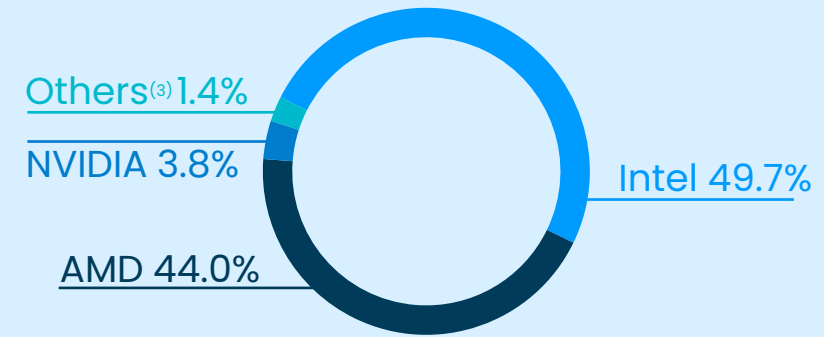
1/20

of global supercomputing
equipment are supplied
by Europe⁽¹⁾.



0%

of the processors
powering Europe supercomputers
are European⁽²⁾.



This lack of homegrown technology has serious implications
on Europe sovereignty, intellectual property and security.

The European Union response (1/2)

Launch of EuroHPC JU & European Processor Initiative

September 2018



Launch of the EuroHPC JU to deploy in Europe a world class exascale supercomputing infrastructure.

December 2018



Launch of a call for proposals in 2017 to design and implement a roadmap for a new family of high-end European processors

- Target: high-performance and energy-efficiency

Coordinated by Bull (an Eviden company), the EPI consortium won this call for proposals.

6 years of hand-in-hand work involving over 200 engineers from 30 partners

- Scientists: research institutes, universities and supercomputing centres
- Industry: European leaders, IT, electronics and automotive specialists



The European Union response (2/2)

Creation of the private company to bring to life the European high-performance low-power processor



Incorporated

In June 2019



Arm architecture

Energy-efficiency quick time to market, proven ecosystem



Funded

By the European Union



Key partnerships

Joint-offering with

AMD intel NVIDIA
EVIDEN Hewlett Packard Enterprise



Financing

Series-A to date: €113m
(€105m equity + €8m bank loans)



Identified customers

Server manufacturers based on user specifications: First, EuroHPC ecosystem before going global.

+200

Employees

from



7 locations

Maisons-Laffitte (HQ), Barcelona, Bologna, Duisburg, Grenoble, Massy, Sophia Antipolis

Arm technology

An historic and well-considered choice

Arm: from smartphone to supercomputer

- Performance
- High-energy efficiency

The Arm logo, consisting of the word "arm" in a lowercase, blue, sans-serif font.

Mont-Blanc projects, an initiative launched in 2011 with the EU support⁽¹⁾

3 initial partners

- Bull
- Barcelona Supercomputing
- Arm

To design a new type of computer architecture around Arm technology, an architecture capable of delivering a new level of performance/energy ratio for HPC applications

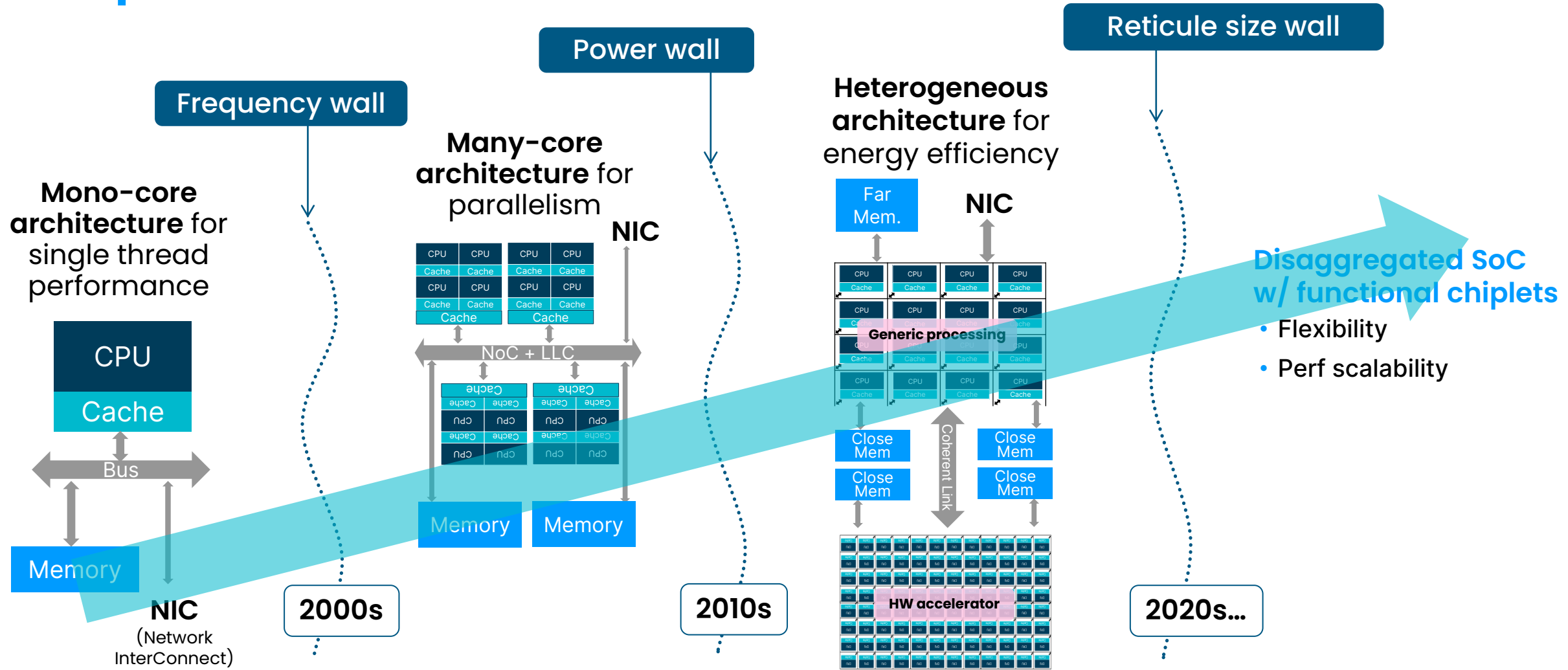
The Mont-Blanc logo, featuring the words "MONT" and "BLANC" in a stylized, blue, sans-serif font. The "M" is larger and more prominent, with the "ONT" and "BLANC" stacked to its right.

April 2020: Licence agreement for Arm® Neoverse V1 platform to design Rheal



The evolving landscape of HPC

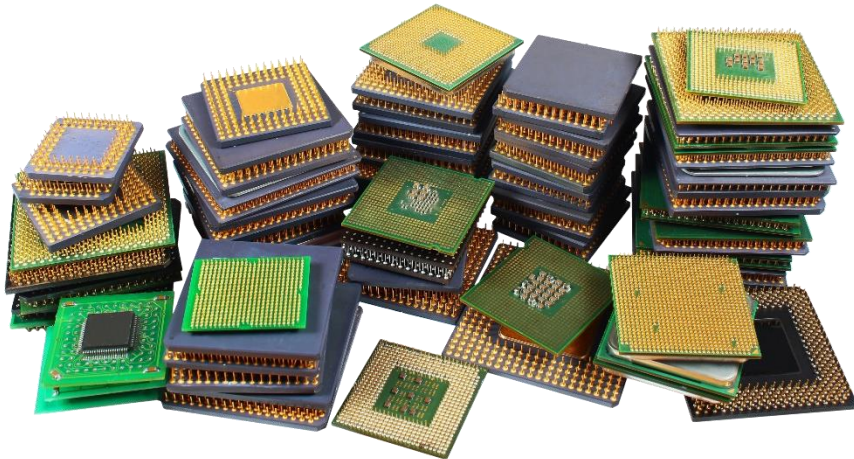
Compute Element Evolution



The early days: CPU based systems

70s & 80s

- System compute was primarily purpose-built CPUs.
- Limited or no use of accelerators.



Early supercomputers Cray 1



The cluster era

Since 90's

90s

- Systems built with standard CPUs & clusters introduction
- Top 500 created based on Linpack benchmark.



2000s

- Frequency wall: the trend of increasing processor clock speed migrated to multicore architectures



2010s

Increased deployment of accelerators

Today

Typical supercomputers are hybrid solutions with:

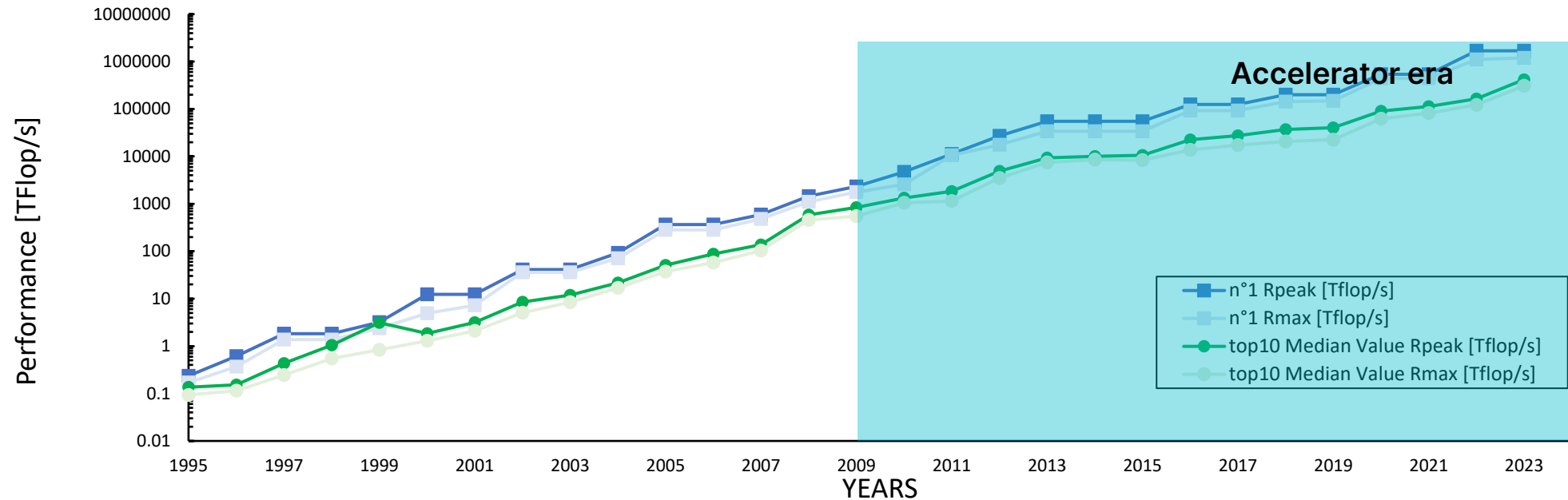
- Multicore processor nodes, cache and RAM shared by all cores
- xPU accelerator to offload specific types of computation



HPC Performance (1/4)

HPL: Measures how fast a computer system can solve a linear system of equation

Good for core performance but not representative of real workload nowadays

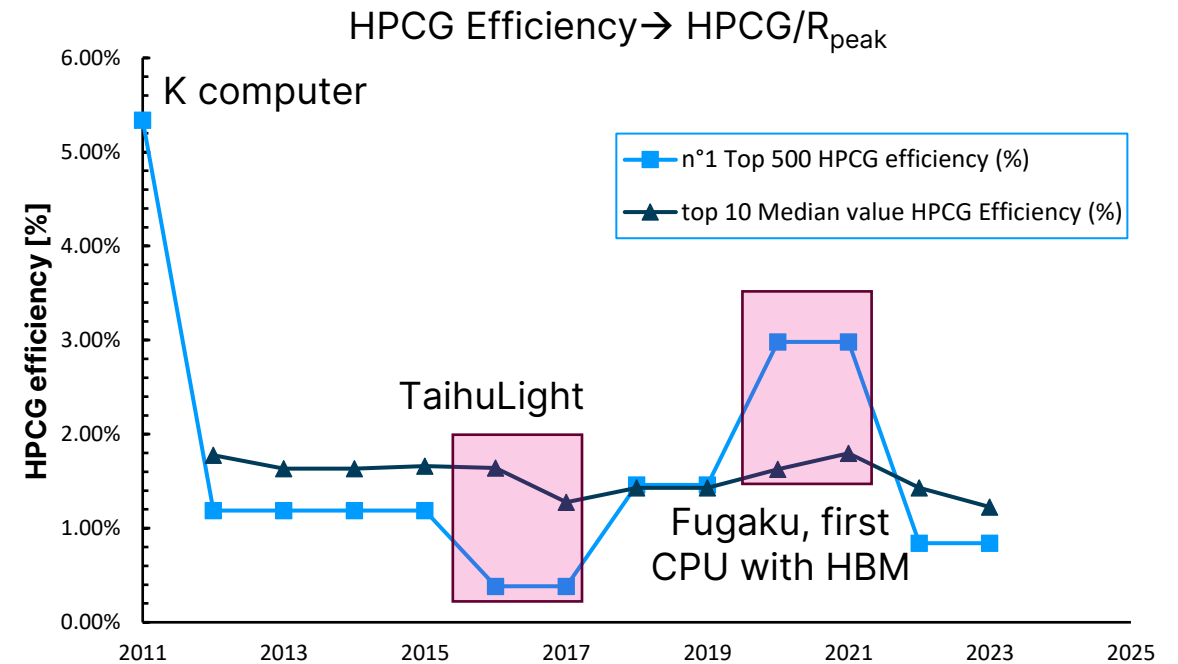
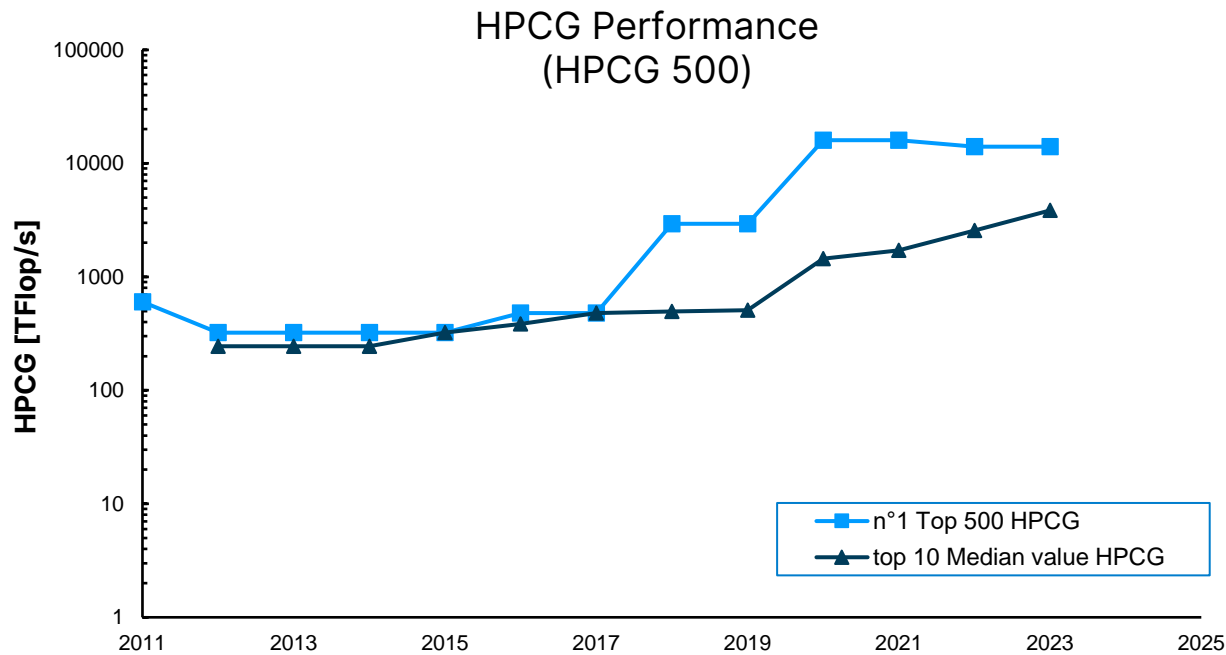


GPUs brought significant core performance gains

HPC Performance (2/4)

HPCG: It is intended to test the effect of limitations of the memory subsystem and internal interconnect of the supercomputer on its computing performance

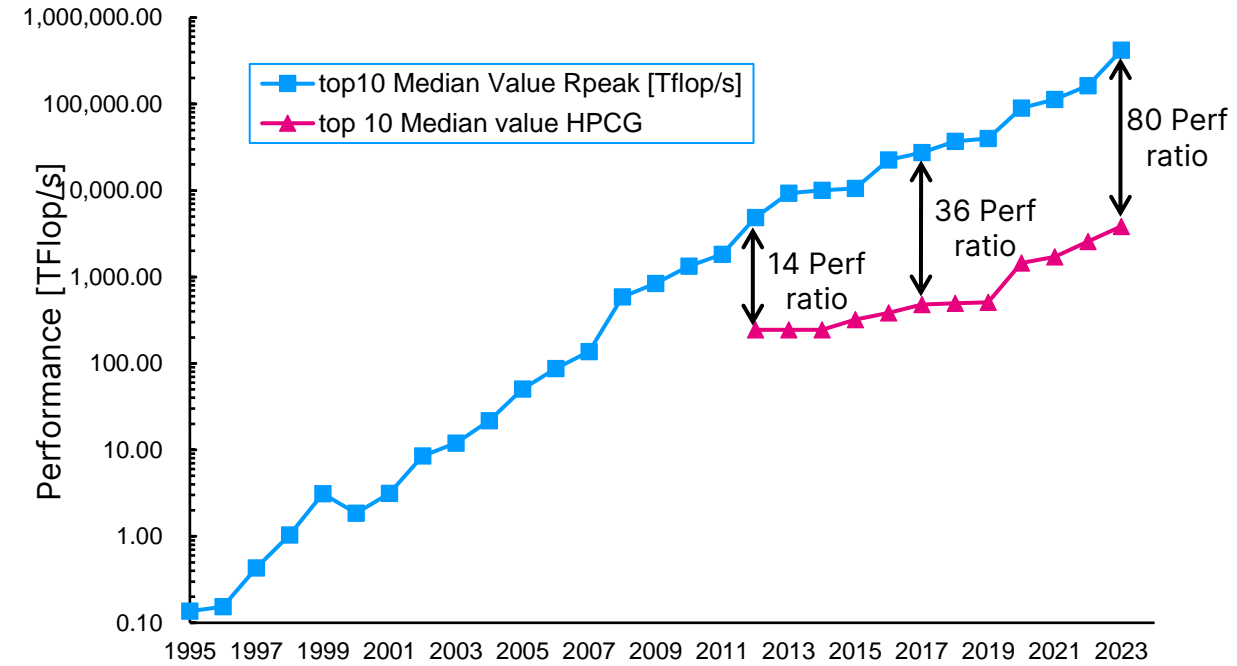
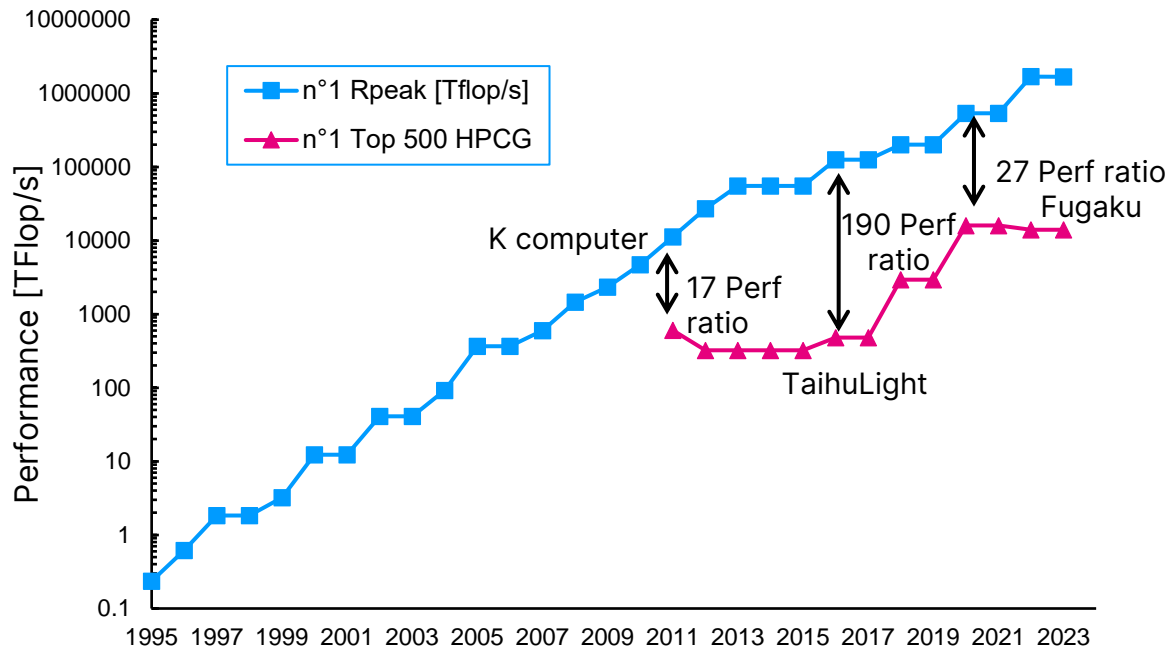
- Tests compute core performance and memory bandwidth
- Better represents real workload performance (vs HPL)



HPCG does not show the same trend as HPL

HPC Performance (3/4)

HPL vs HPCG (No data available before 2011)

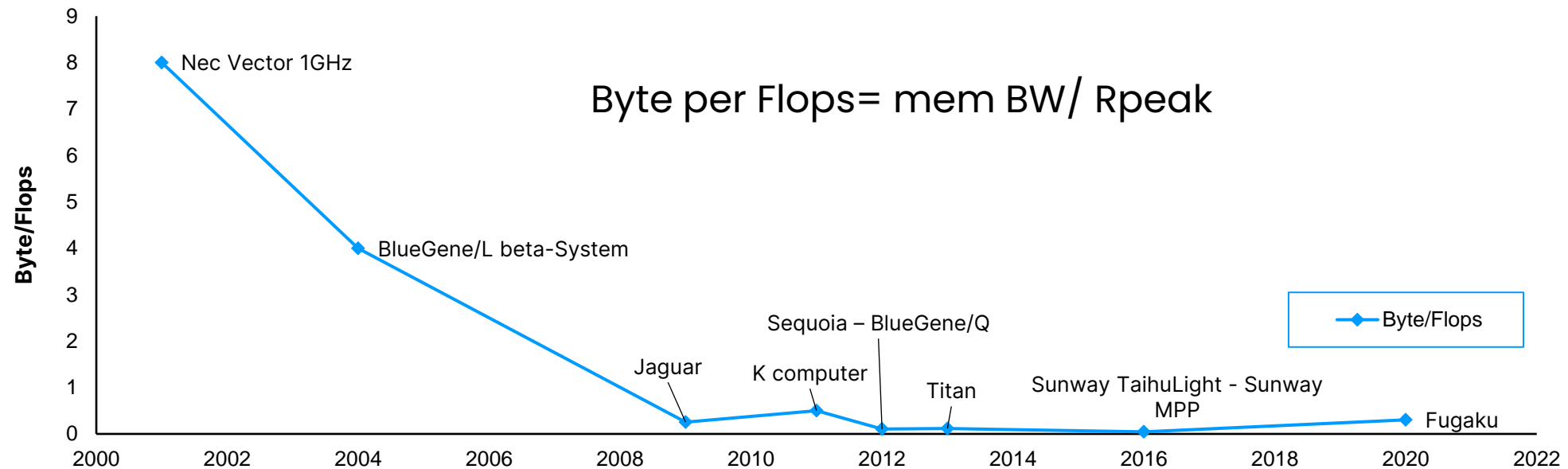


Median performance ratio HPL/HPCG increases

HPC Performance (4/4)

Byte per Flops: core calculation intensity before reaching the data bandwidth limit.

Test compute core and memory bandwidth

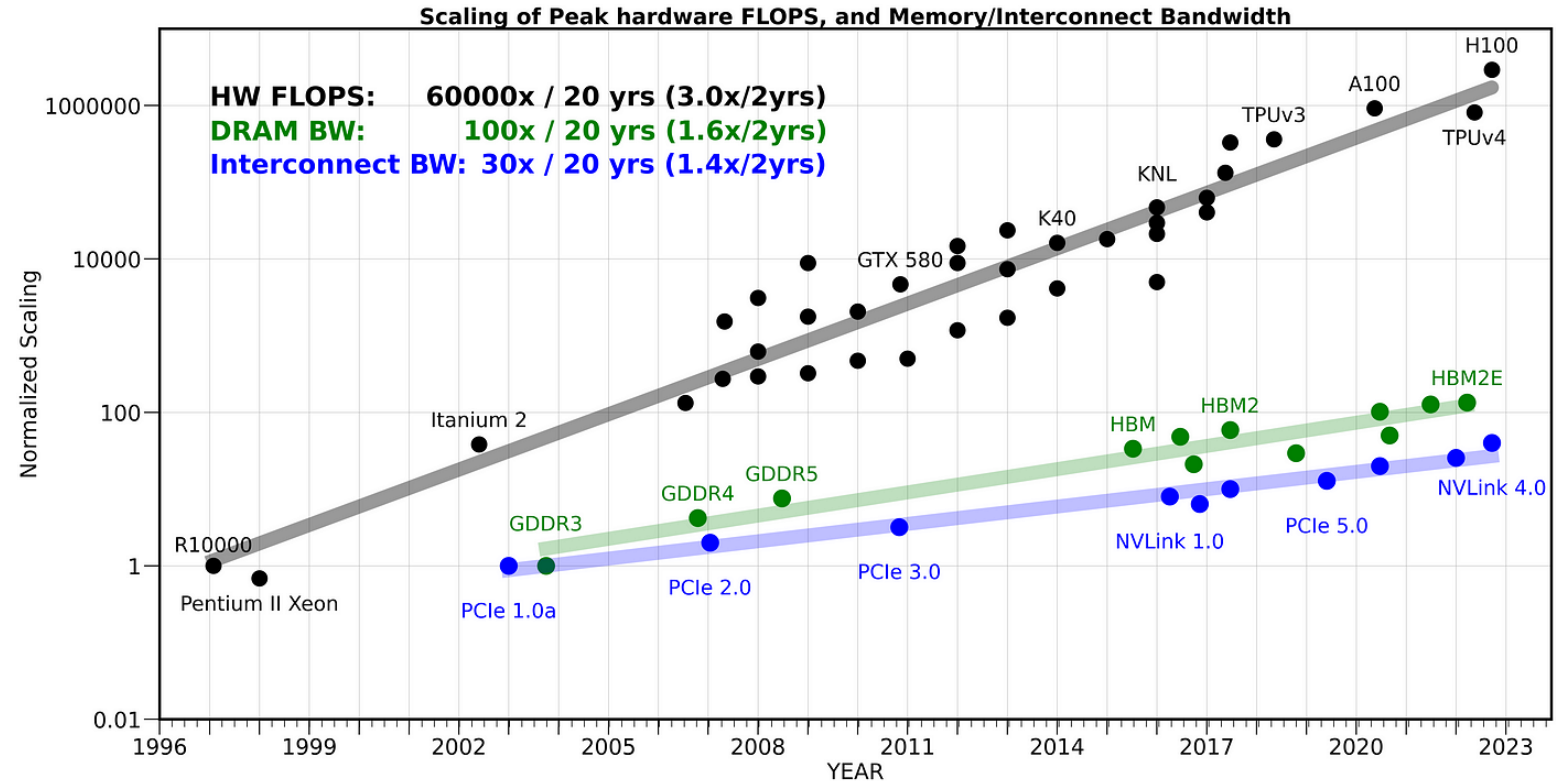


Byte per Flops, like HPCG efficiency, does not show the same trend as HPL

The memory speed bottleneck: A shifting paradigm

Linpack measure
only core compute power

Linpack performance
increased, but memory &
network bandwidth did not
follow at the same pace.



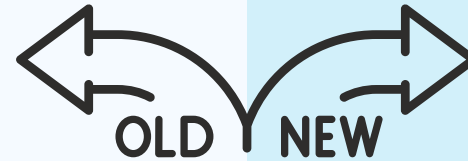
Source: <https://medium.com/riselab/ai-and-memory-wall-2cb4265cb0b8>

Linpack does not represent the range of real workloads

Emerging requirements for modern HPC processor



Mflops
Compute power



Sovereignty

Ensuring data security and control within geographical boundaries.



Energy-efficiency

Reducing operational costs, the environmental impact and optimizing the TCO.



Data Transfer

Addressing the bottleneck for optimal performance.

Benchmark

HPL

HPCG
Stream
Bytes/Flop
GF/Watt



AI: LLM overview

What are LLMs?

Large Language Models (LLMs) are a category of foundation models trained on immense data sets. They have the potential to improve productivity across industries and academia to solve the world's toughest problems.



Healthcare,
Energy



Finance, Law,
Education



Scientific
Research



Security



Natural Language
Processing



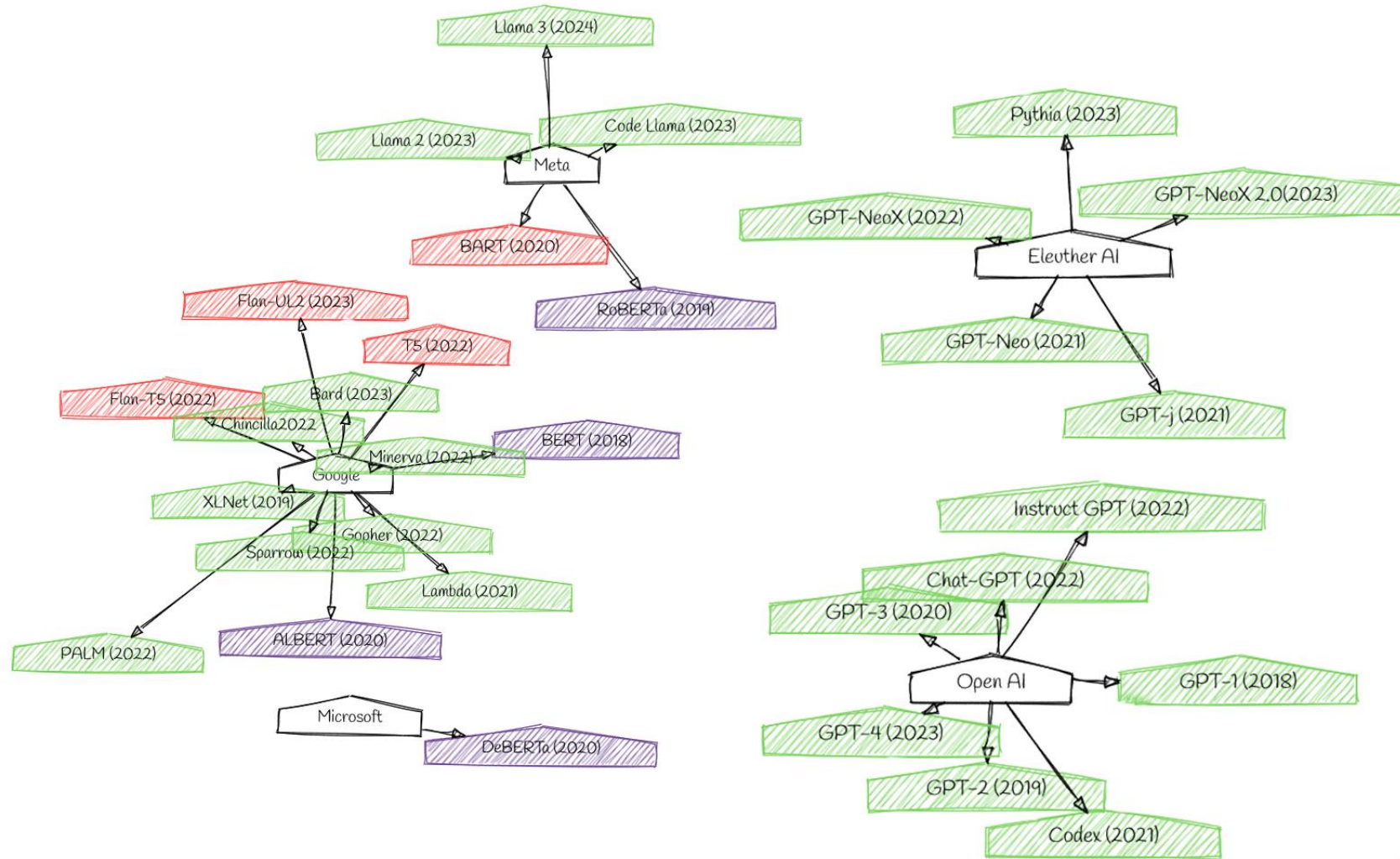
Arts, etc.

LLMs can equal or surpass human performance

- Coding tasks
- Complex dataset analysis
- Information retrieval tasks
- Text translation
- Educational tutoring
- Text summarization
- Questions answering
- Sentiment analysis
- Clinical Note Summarization
- Scenario generation & recommended measures
- Mental Health support
- Scientific explanation
- ...

**LLMs exist as a subset of deep learning models,
which are a subset of machine learning models**

A variety of LLMs driven by hyperscalers



Legend:

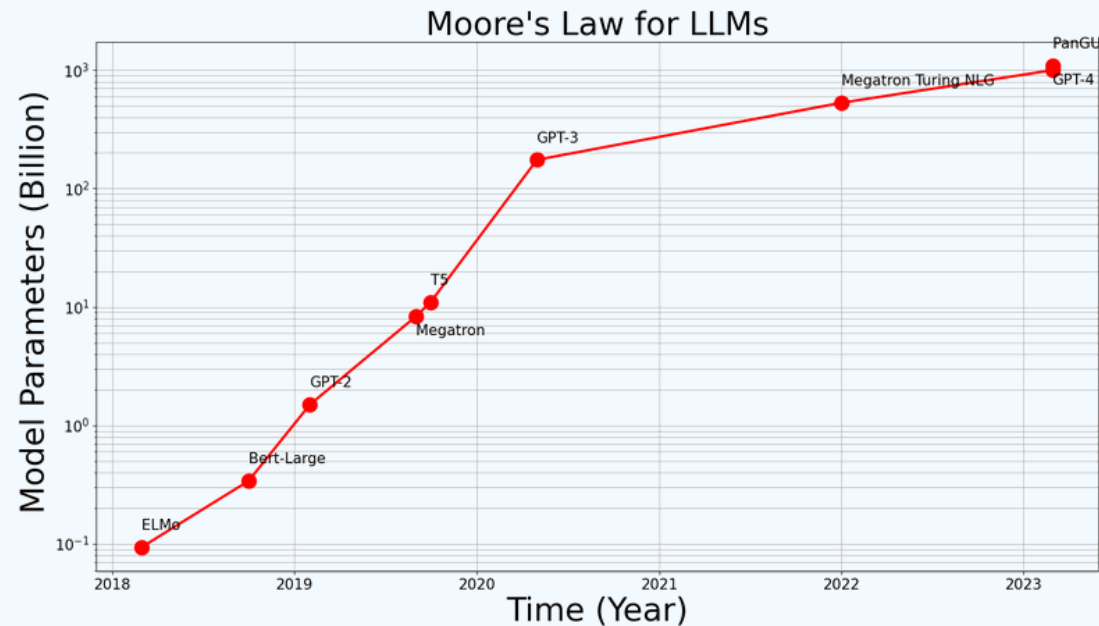
Encoder Only

Encoder Decoder

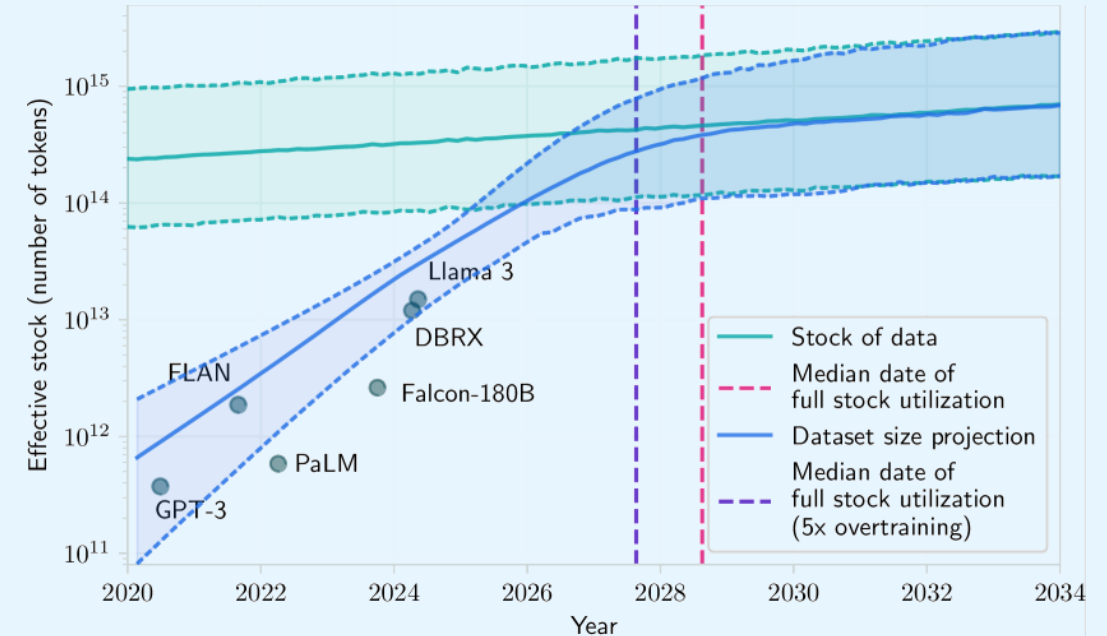
Decoder Only

Evolution: trends & future of LLMs parameters

Models are growing exponentially in size with time, even faster than the rate of the original Moore's law



Source: ChatGPT in the Age of Generative AI and Large Language Models: A Concise Survey

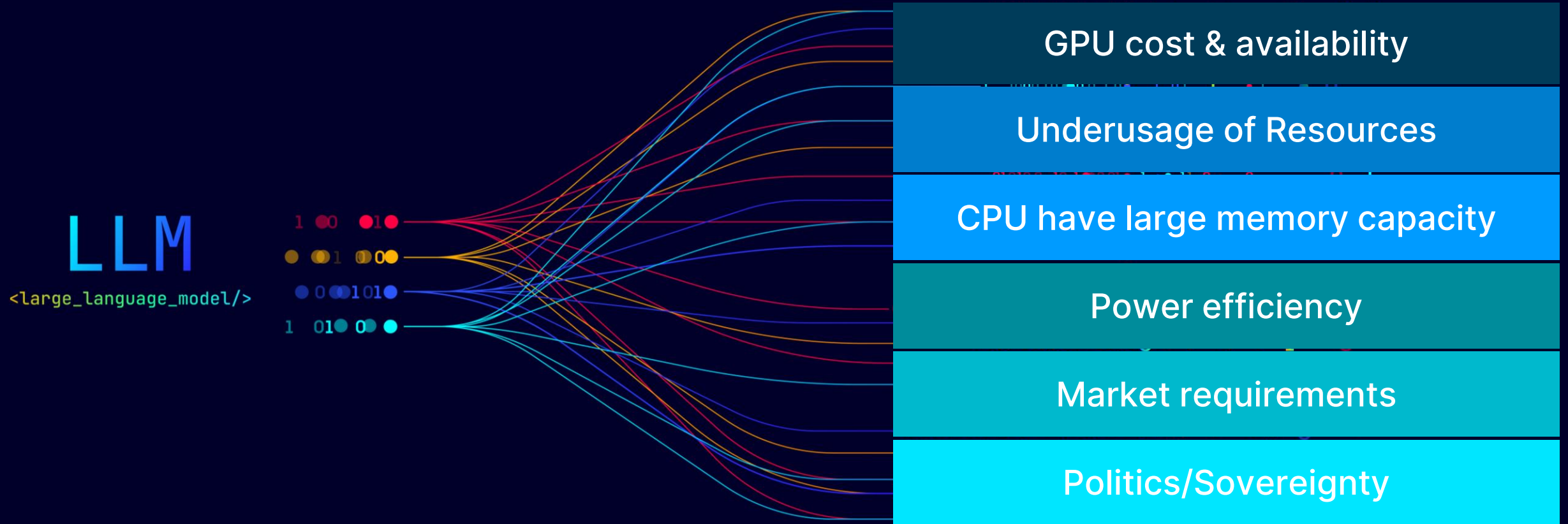


Source: Will we run out of data? Limits of LLM scaling based on human-generated data

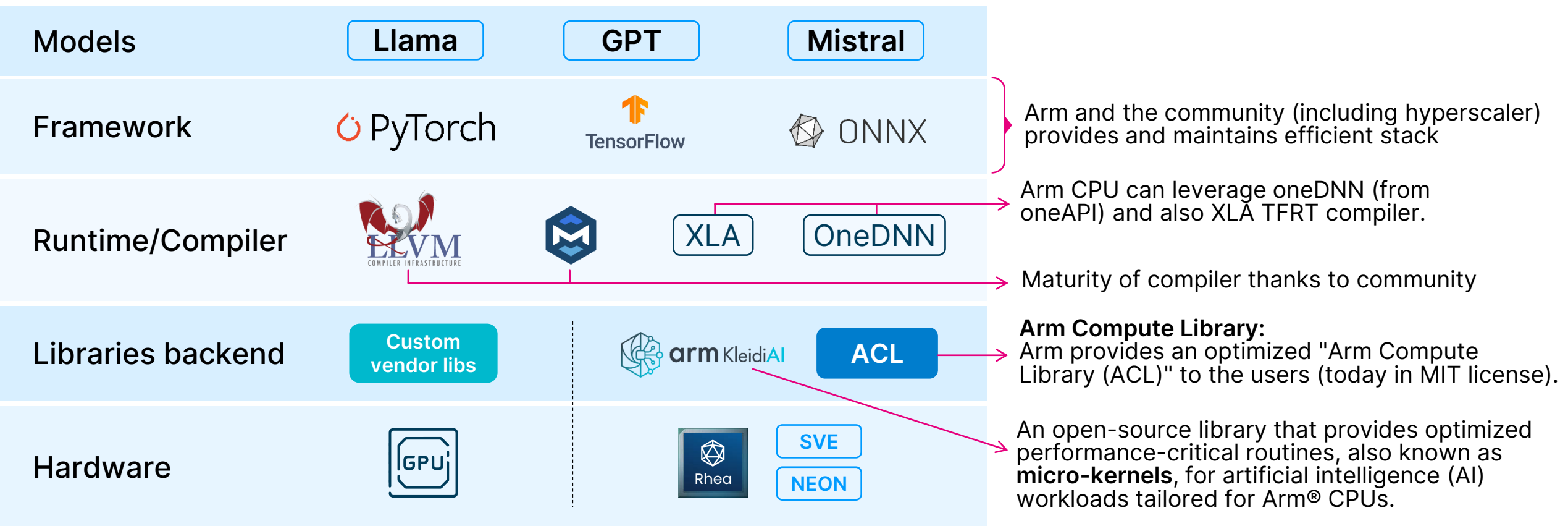
A variety of Language Models

Feature	Small Language Models	Large Language Models
Number of Parameters	Millions to (a few dozen) Billions	Billions to Trillions
Operational Requirements	Lower (faster, less memory/power)	Higher (slower, more memory/power)
Training Data	Smaller, more specific datasets	Massive, diverse datasets
Performance on Simple Tasks	Good performance	Good to excellent performance
Domain Expertise	Can be fine-tuned for specific domains	More general across multiple domains
Perf. on Complex Tasks / Generalization	Lower capability / Limited generalization	Higher capability
Example of Use Cases	Chatbots, simple text generation, domain-specific NLP	Creative writing, question answering, general NLP
Examples	ALBERT, DistilBERT, TinyBERT, Phi-3	Less GPT-3, BERT, T5

Key considerations



Maturity of Software stack for AI on ARM



Covers Inference, Finetuning, Quantization and Training

Common characteristics of HPC and AI/LLM workloads

Challenges are similar for AI/LLM and HPC

AI tends to be more demanding in latency

HPC tends to be more demanding in Bandwidth

CPU & GPU

Memory Latency

Memory
Bandwidth

Power efficiency

Core Performance
(FLOPS)

Interconnect
bandwidth

Interconnect
latency



Rheal key features

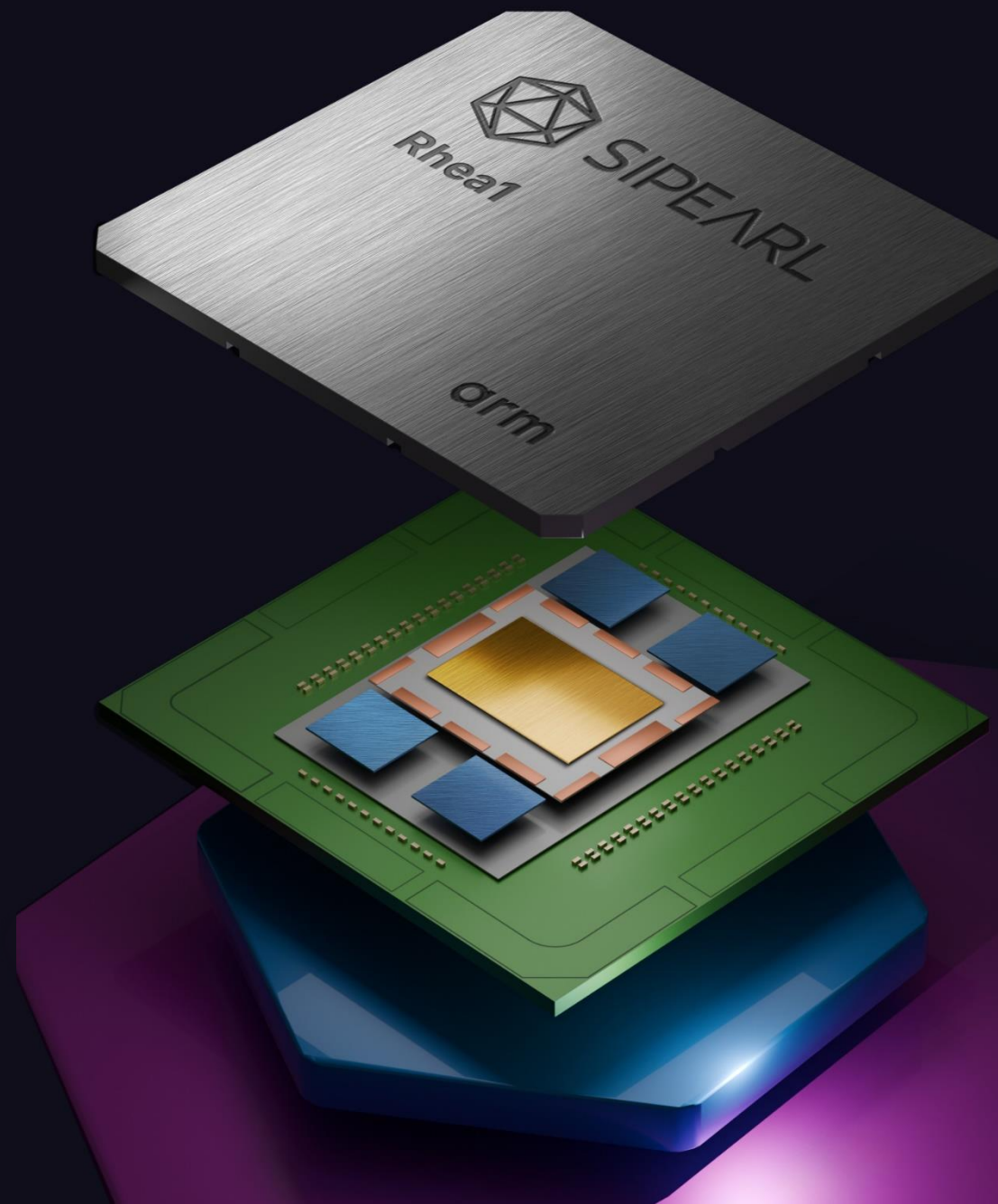
RHEA1

HPC and AI microprocessor

80 arm® Neoverse V1 cores
with 2 x 256 SVE each

4 x HBM

4 x DDR5 interfaces



Rhea1, our 1st generation microprocessor

High performance per watt: Arm ISA power efficiency

- Arm cores have ~30% smaller area vs equivalent performance x86 cores

Very high memory bandwidth

Built-in HBM

- Ideal performances for Generative AI

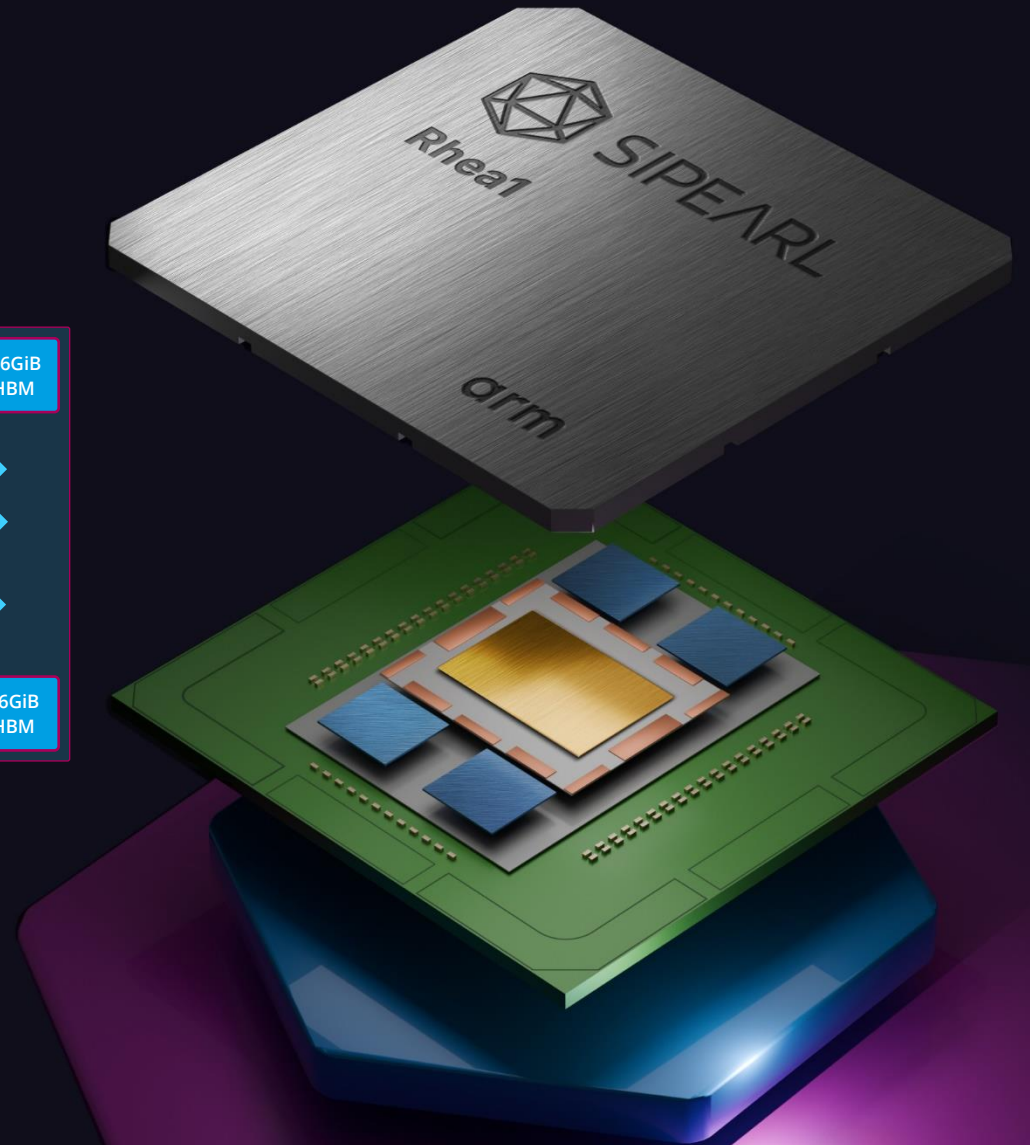
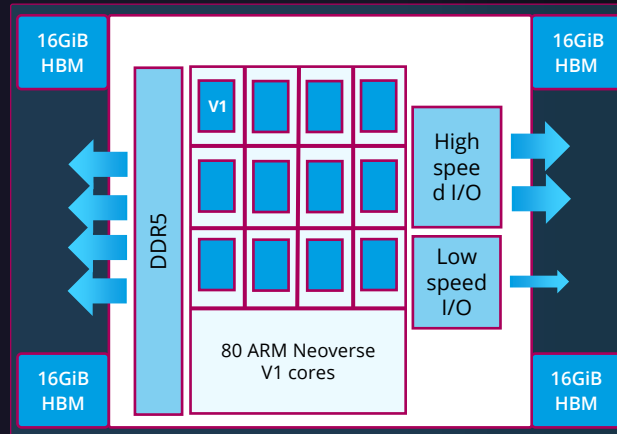
Unique memory architecture: High Byte/Flop

Openness

- Arm ecosystem from IoT/edge to HPC and cloud

Fully auditable – backdoor-free

Rhea1 will deliver extraordinary performance and efficiency with an unmatched Byte/Flop ratio.



Rhea1 performance targets

Addressing the requirements of HPC & AI at exascale level



Data Transfer

Memory BW $\gg 1.6$ TB/s
Stream > 1.6 TB/s



Mflops

Compute power

HPL > 2 TF



Energy-efficiency

HPCG Efficiency $> 5\%$
Byte/Flop (Rpeak) > 0.5
Byte/Flop (Rmax) > 0.4

Rhea1 in 5 key benefits



High Performance

To surpass the performance of 10,000,000 desktop computers.



Energy-efficiency

To halve power consumption for equivalent computing power.



Flexibility

Designed to work with any third-party accelerator (GPU, artificial intelligence, quantum).



Backdoor-free security

To protect data with secure end-to-end network transmission.



Sovereignty

To further Europe's technological leadership and independence.



About... SiPearl

SiPearl is building the European high-performance low-power microprocessor dedicated to supercomputing and artificial intelligence. This new generation of microprocessors will first target EuroHPC Joint Undertaking ecosystem, which is deploying world-class supercomputing infrastructures in Europe for solving strategic sovereign challenges in medical research, generative AI, security, energy management and climate with a reduced environmental footprint.

SiPearl is working in close collaboration with its 30 partners from the European Processor Initiative (EPI) consortium - leading names from the scientific community, supercomputing centres and industry - which are its stakeholders, future clients and end-users.

SiPearl employs more than 200 people in France (Maisons-Laffitte, Grenoble, Massy, Sophia Antipolis), Italy (Bologna), Germany (Duisburg) and Spain (Barcelona).

SiPearl is part of French Tech 120 program 2024 class.

