SIPEARL

Leveraging EPI Outcomes for an Open Source Cloud-Based Services Ecosystem

The Role of SiPearl's ARM-Based CPUs

Roberto MOSTALLINO

Novembre 2024

Agenda

#1 Market Opportunities for ARM CPUs in the Cloud

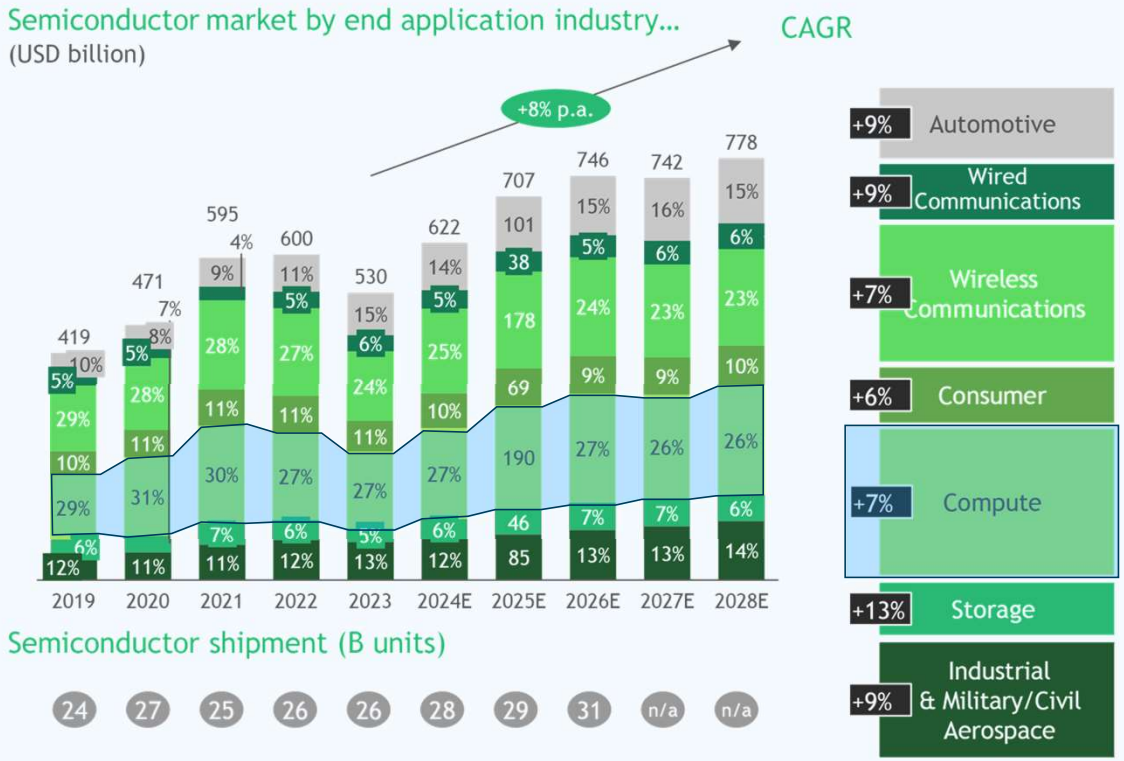
#2 SiPearl in the ARM ecosystem

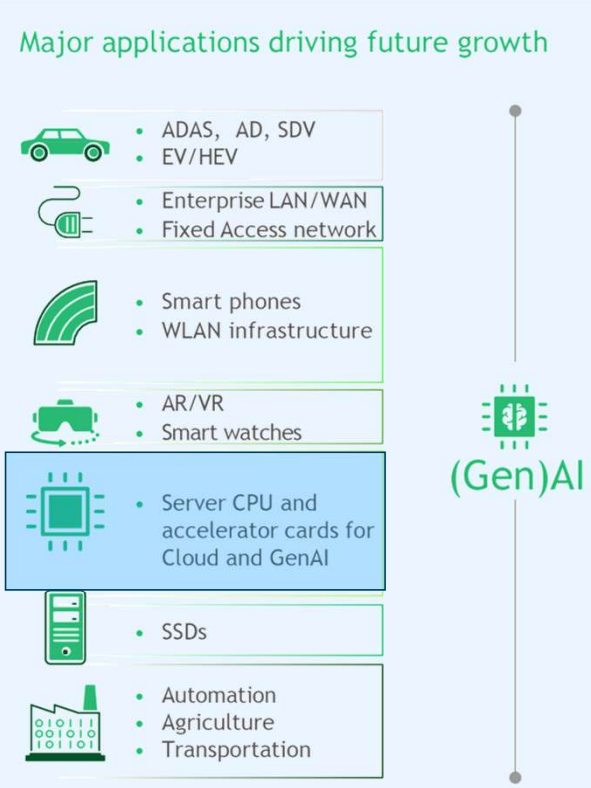
#3 Overview of SiPearl contribution to RISER Project

#4 SiPearl's ARM-Based CPU & Server Blades Seine platform


SIPEARL
Market Opportunities
for ARM CPUs in the Cloud

Semiconductor Market is expected to growth at 8% CAGR until 2028

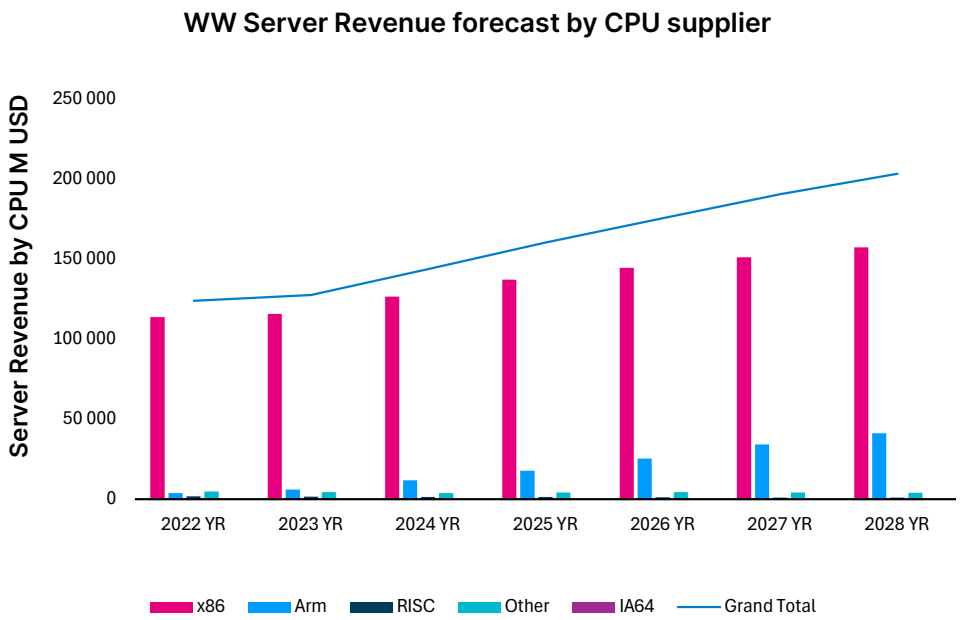


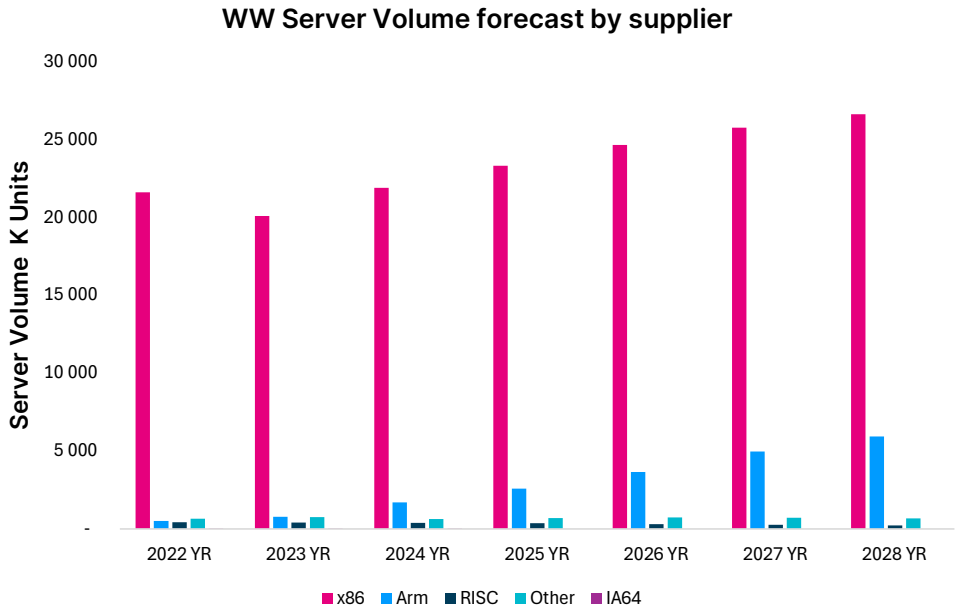


Source: Gartner semiconductor report and BCG analysis

Server Market Forecast WW

Forecast 2022-2028 of Server Revenue by CPU

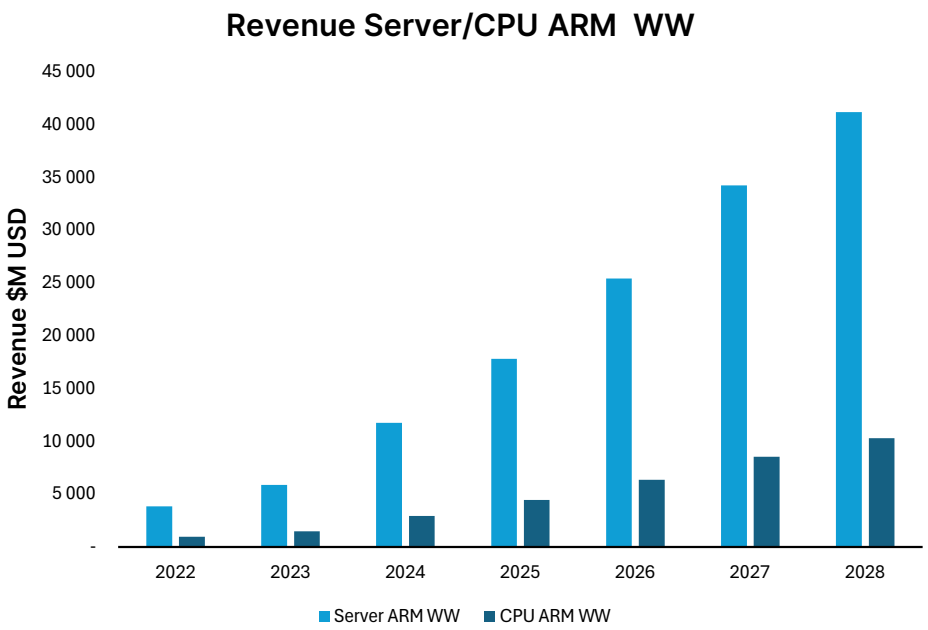


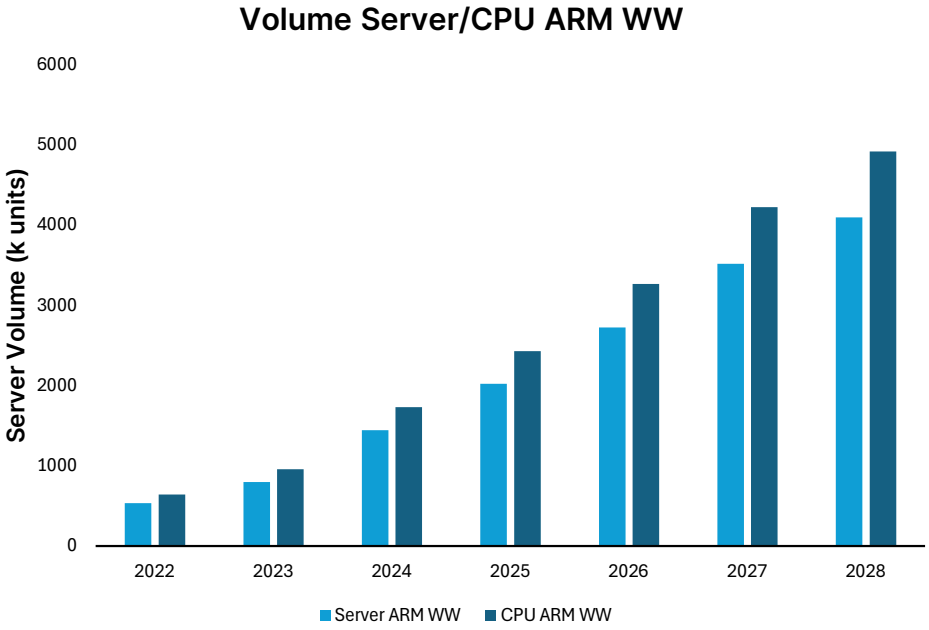


X86 shows a linear progress and ARM based server is the second biggest market

Source: Gartner Servers_Forecast_1Q24 (808654)

Server/CPU Forecast for ARM in WW

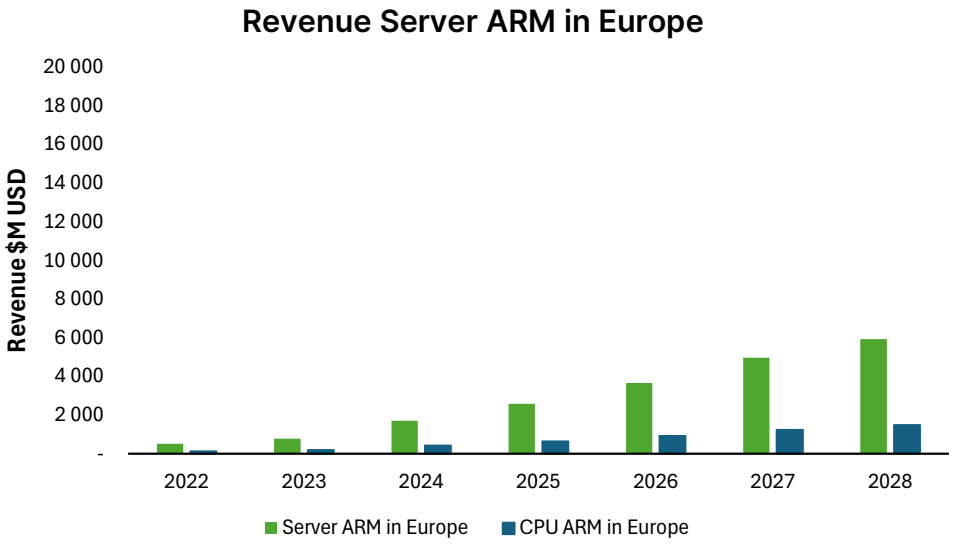


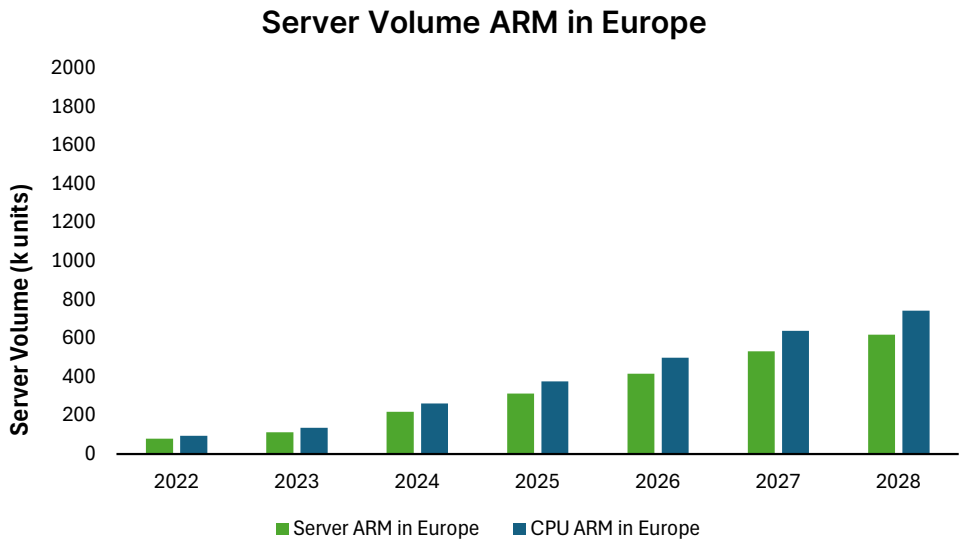


The WW market for ARM CPU will increase by a factor of 3 in terms of shipment and revenue in four years.

Source: Gartner Servers_Forecast_1Q24 (808654)

Server/CPU Forecast for ARM in **Europe**



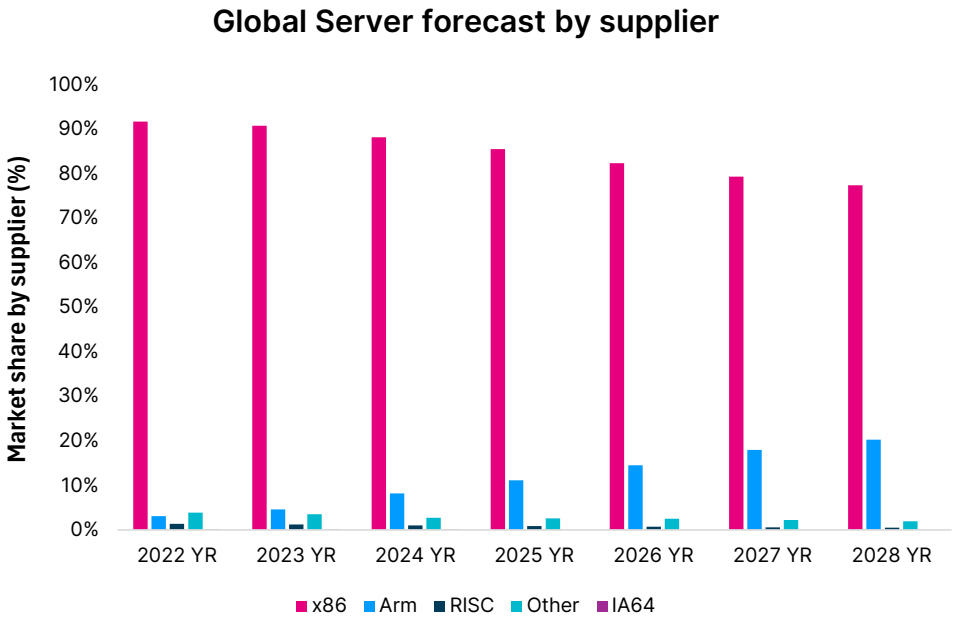


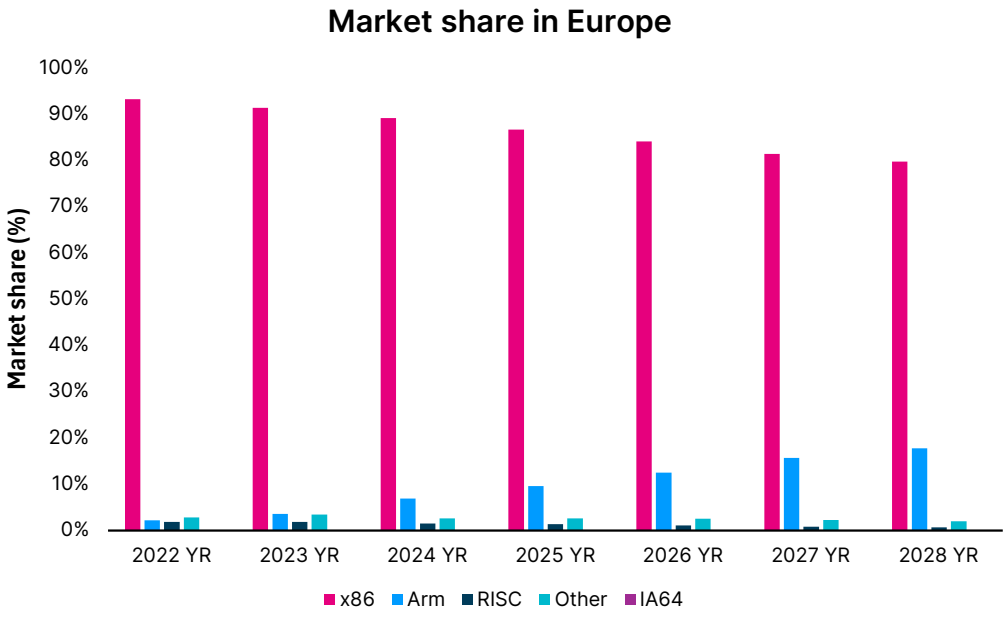
The European market for ARM CPU will increase by a factor of 3 in terms of shipment and revenue in 4 years.

Source: Gartner Servers_Forecast_1Q24 (808654)

Server Market Share Forecast by **CPU**

Global server forecast by Instruction Set Architecture (ISA)





Although x86 will maintain its dominant position in the WW market, the ARM ecosystem will see a 2,5x increase in market share in 4 years.

Source: Gartner Servers_Forecast_1Q24 (808654)

SiPearl in the ARM ecosystem

Competitive positioning Highlights

Market analysts predict a shift of market share, with ARM gaining traction at the cost of x86

Present

ARM & x86

- Ampere design CPU ARM for datacenter, with focus on storage
- Nvidia design ARM-based server CPU, mostly coupled with proprietary GPU.
- Type 1 Hyperscalers such as AWS, Google and Microsoft produce CPUs for internal needs

Future

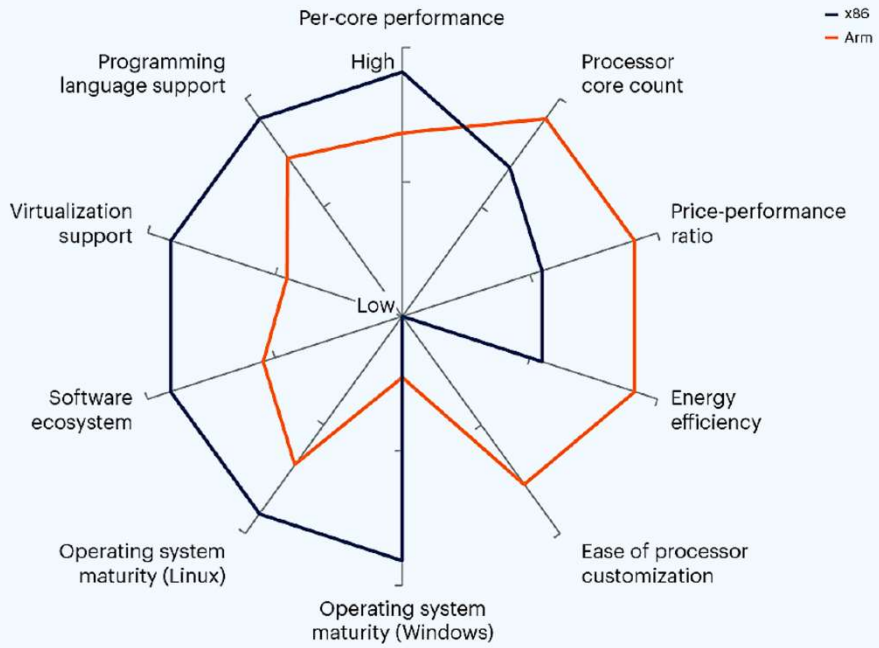
SiPearl differentiator

- Leader in performance for memory bound workload.
- Independent backdoor-free CPU design, fully auditable.
- Unique position to win European AI market.
- Cost-effectiveness compared to x86 CPU.

Software Ecosystem

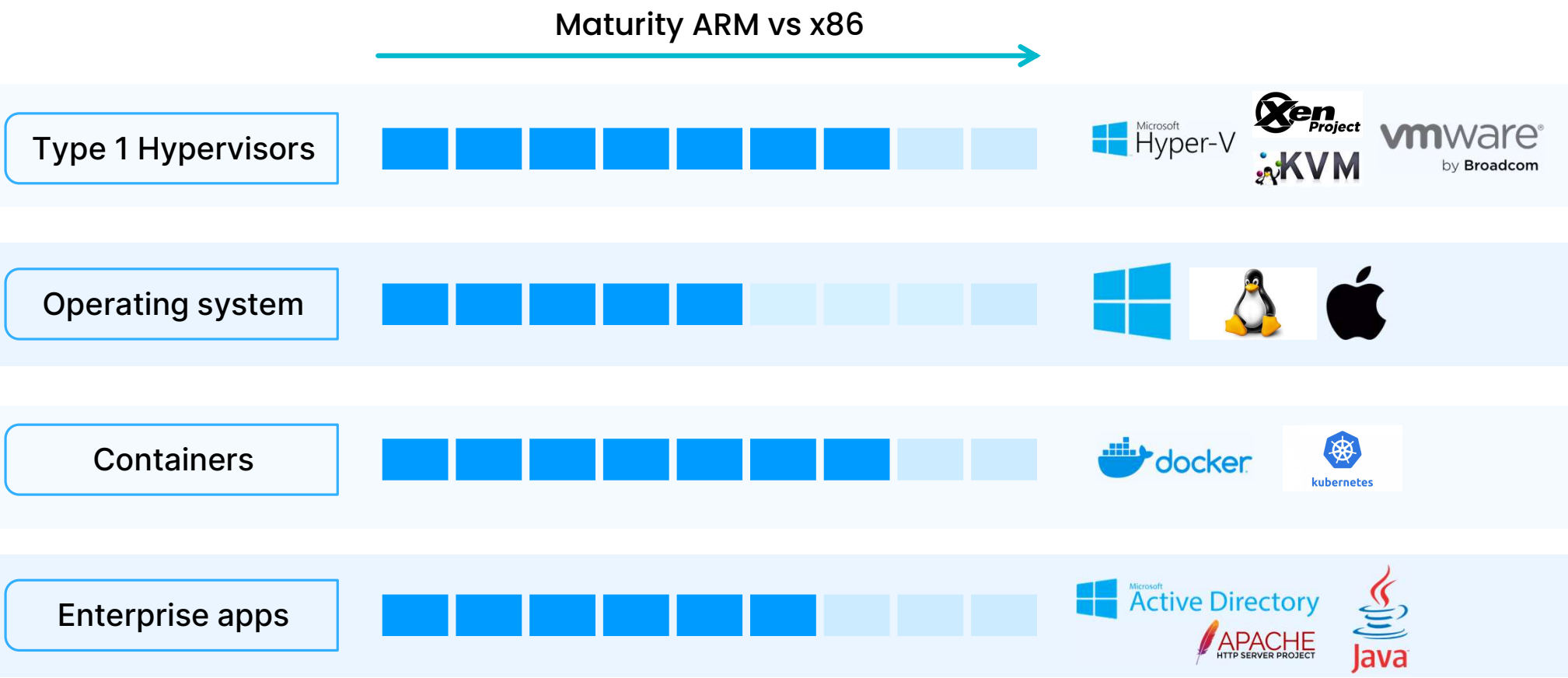
Since ARM's entry into the server CPU market in 2008, its superior price-performance ratio and energy efficiency has fueled robust growth year after year.

Comparison of Arm and x86
Illustration of factors assessed in this research



Source: Gartner
805675_C

Datacenter Software Ecosystem maturity



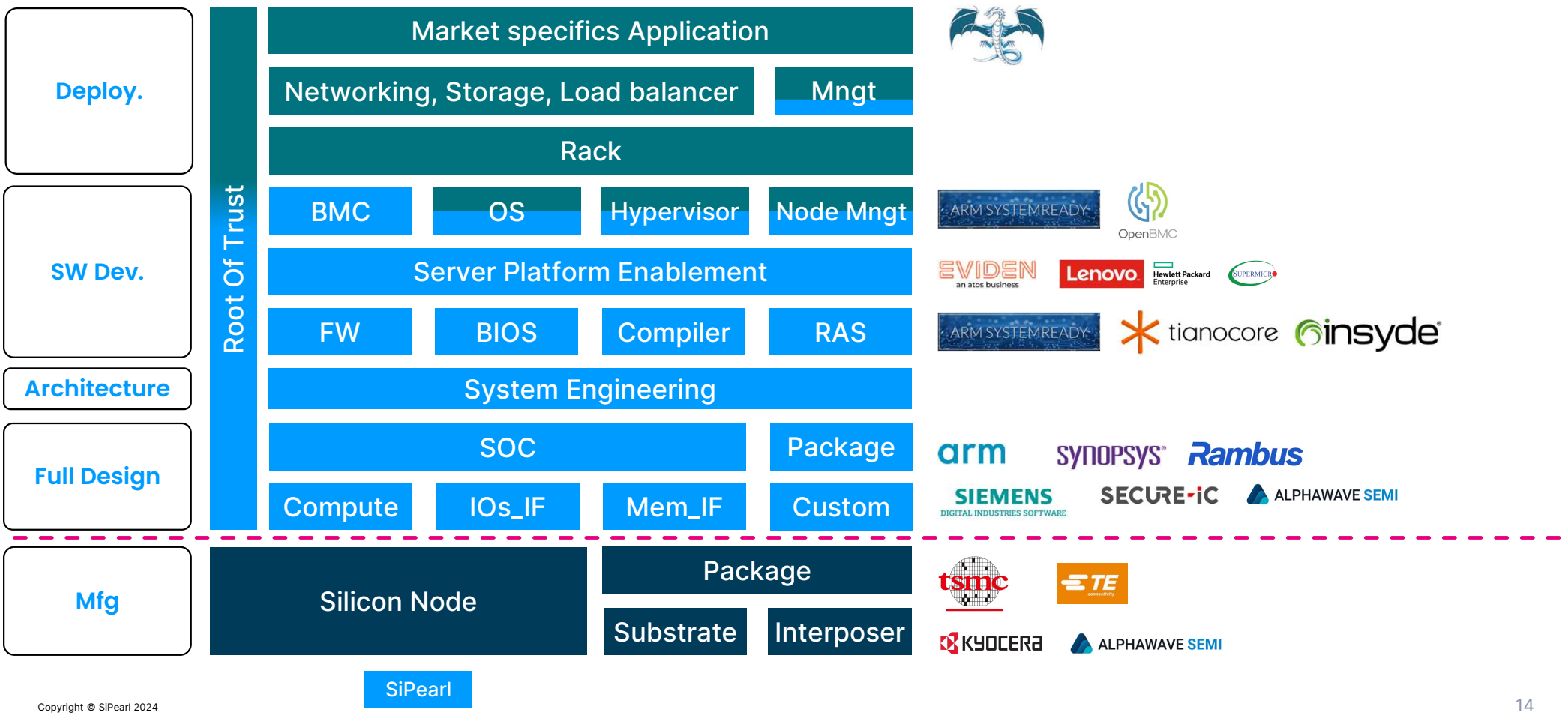
Server Based CPU landscape

ODM & OEM

CPU & GPU

Foundries

SiPearl in the value chain



Overview of SiPearl contribution to RISER Project

RISER – project summary and SiPearl role

Scope

- Leverage the upstream of EPI and EUPilot projects to develop the first all-European RISC-V cloud server infrastructure.
- Europe's open strategic autonomy in the semiconductor technology market

SiPearl contribution

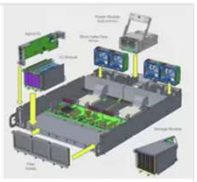
- Seine Platform based on RHEA processor and integrated with the PCIe acceleration board
- Requirement and use case definition for Riser prototype

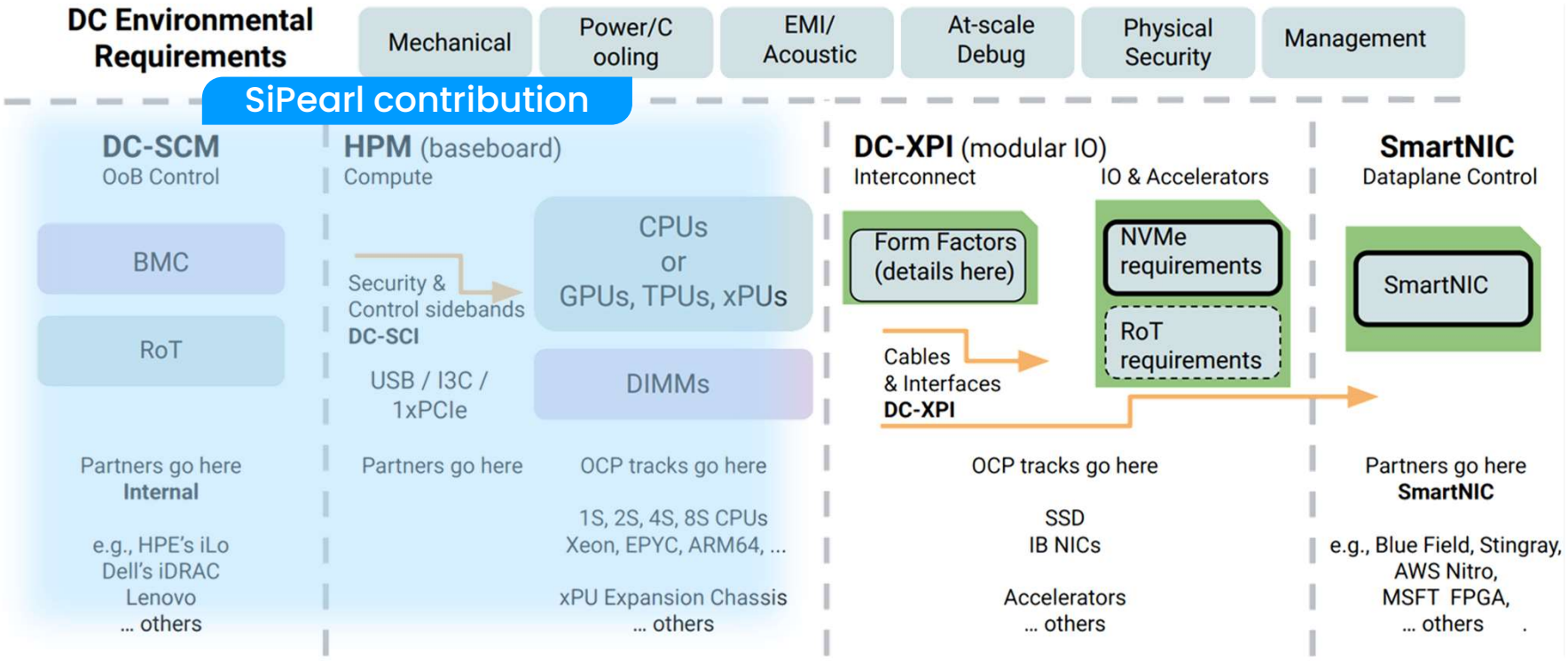
Future

- Usage of Seine Platform for Cloud and AI applications
- Bring up of the demonstration Riser platform
- Development of drivers for Riser accelerator
- Evaluation of different use cases

Hyperstack Hardware Modules

Logical Blocks overlaid on Physical Blocks for
a *Datacenter-ready Integrated System (DC-Stack)*





After Riser: Higher

Open Compute Project OCP

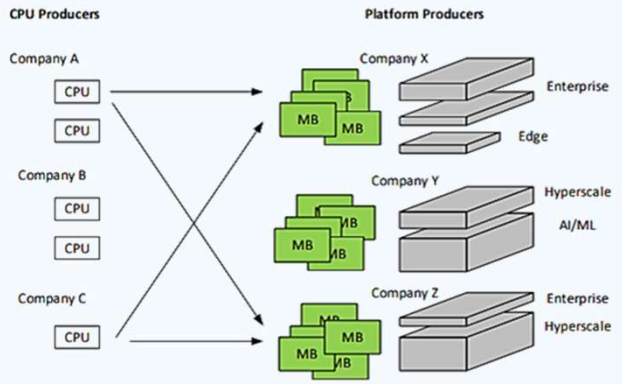
The OCP focuses on redesigning hardware technology to efficiently support the growing demands on compute infrastructure

OCP Server Project: provides standardized server system specifications for scale computing

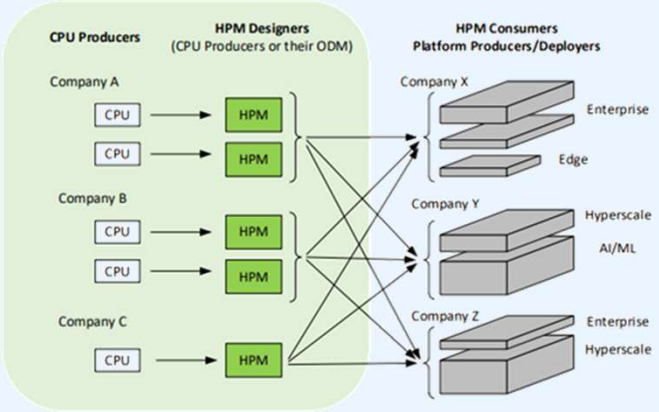
DC-MHS: Data Center –Modular Hardware System

- Interoperability between key elements of datacenter, by providing consistent interfaces and form factors among modular building blocks

Today's Model



New Model w / DC-MHS



HPM stands for Host Processor Module

SiPearl role in Higher

- **Typical Higher DC-MHS server composed of:**

DC-SCM (Security & Control Module)

Interchangeable Host Processor Modules based on European processors technology: HPM-Rhea2, HPM-EUPILOT

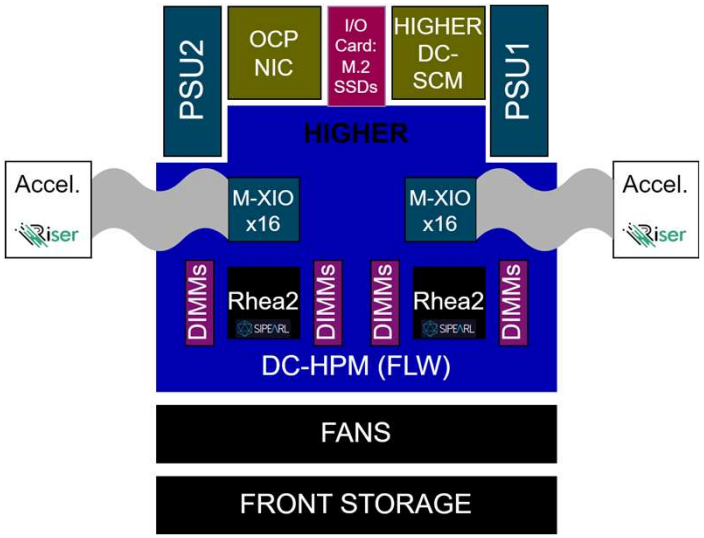
OCP Compliant Network Interface Card (NIC)

PCIe Acceleration board inherited from Riser1

- **In this context:**

SiPearl to develop a Rhea2 based HPM (dual socket)

Exapsys and SiPearl develop the EPI-based DC-SCM (BMC board)



SiPearl's ARM-Based CPU & Server Blades Seine platform

RHEA1

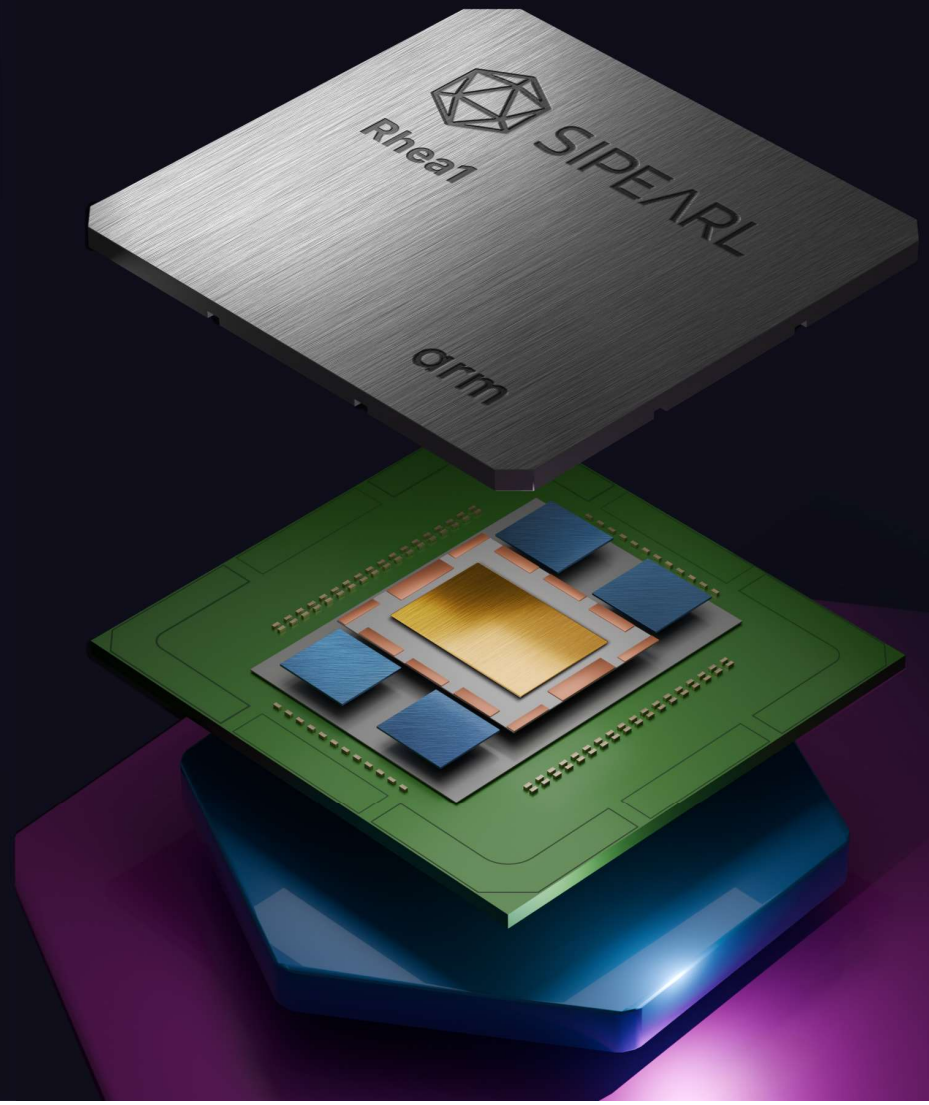
HPC and AI processor

Designed with

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

4 x HBM

4 x DDR5 interfaces



Rhea1, our 1st generation processor

High performance per watt: Arm ISA power efficiency

Very high memory bandwidth

Built-in HBM

- Ideal performances for AI inference

Unique memory architecture: High Byte/Flops

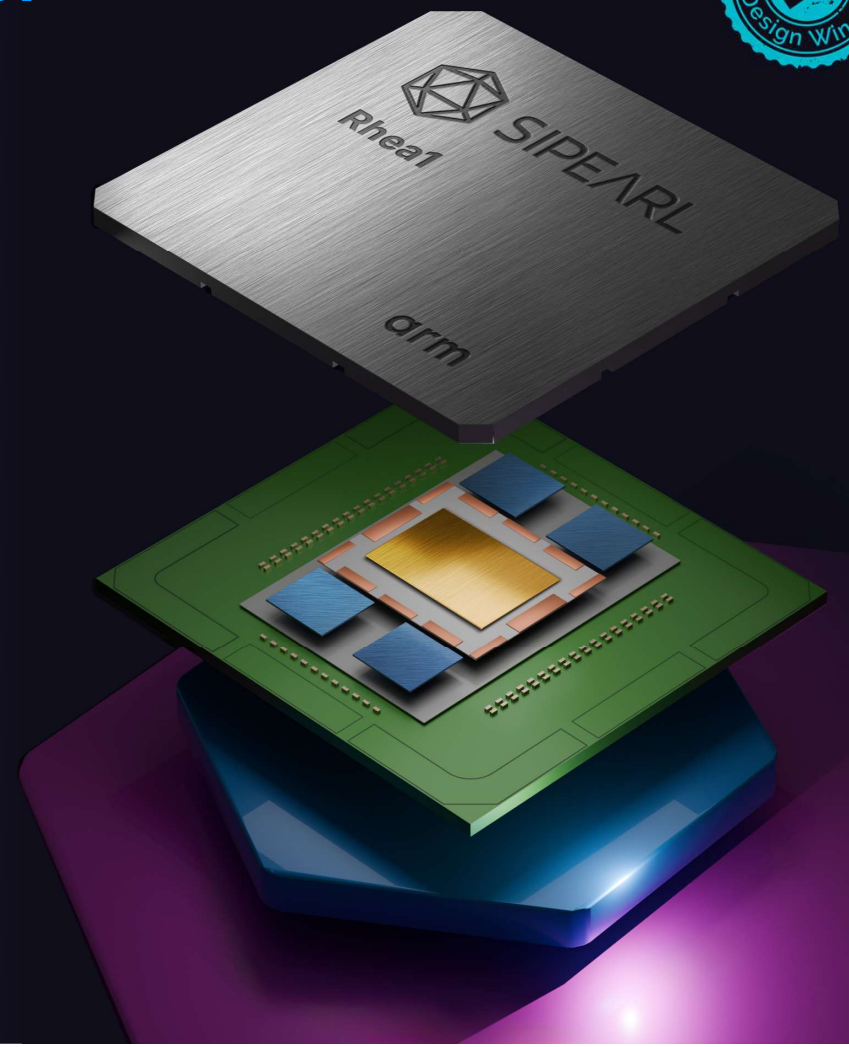
Openess

- Arm ecosystem from IoT/edge to HPC and cloud

Fully auditable & backdoor-free

Pre-integration with proven accelerator (AMD, Intel, Nvidia)

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



Rhea1 CPU performances

Data Access

Highest performance for memory access il literature

Memory BW >1.6 TB/s
Stream ~1 TB/s

Compute power

TOPs (int 8) >50

Energy-efficiency

Highest value of HPCG eff and Byte/Flop

(*)HPCG Efficiency >5%
Byte/Flop >0,5

(*)CPU with high HPCG efficiency can perform the same amount of work using less electrical power

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

Scientific Research

Natural Language Processing

Finance, Law, Education

Security

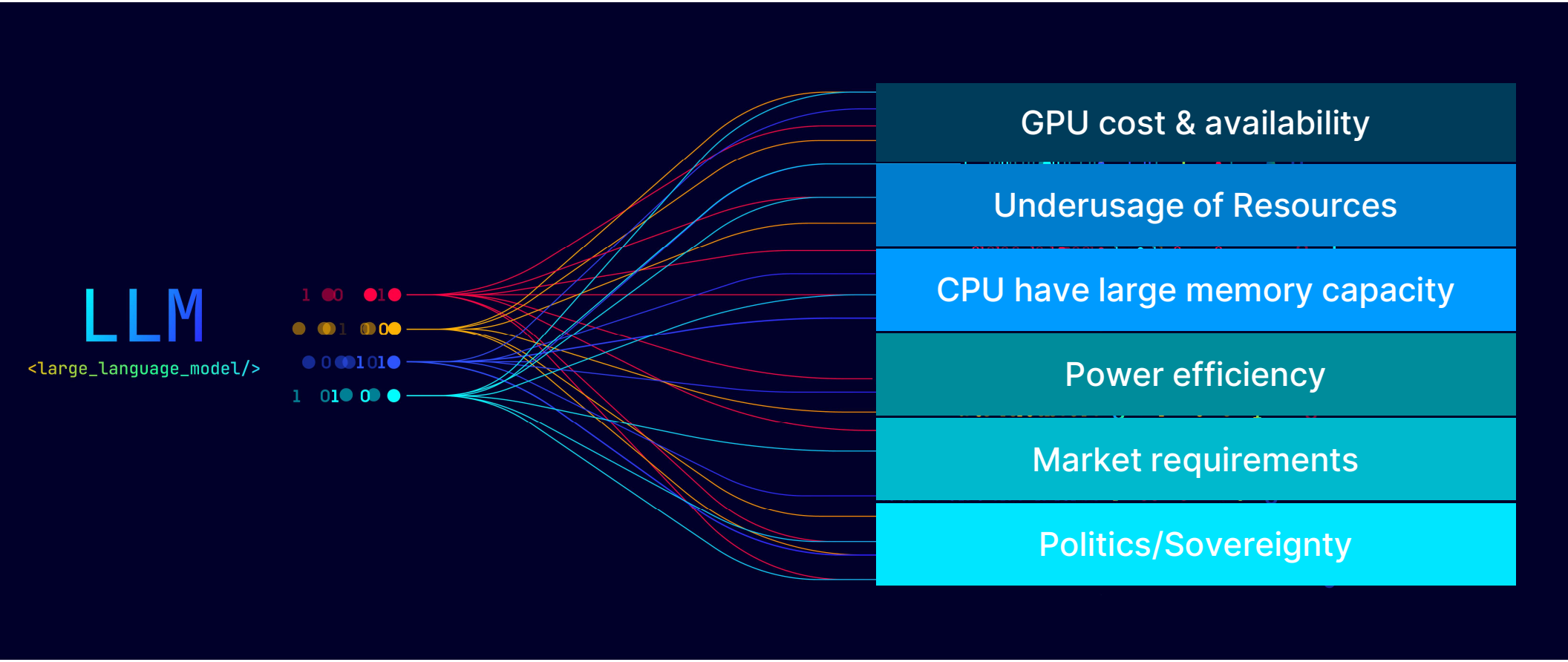
Arts, etc.

LLMs can equal or surpass human performance

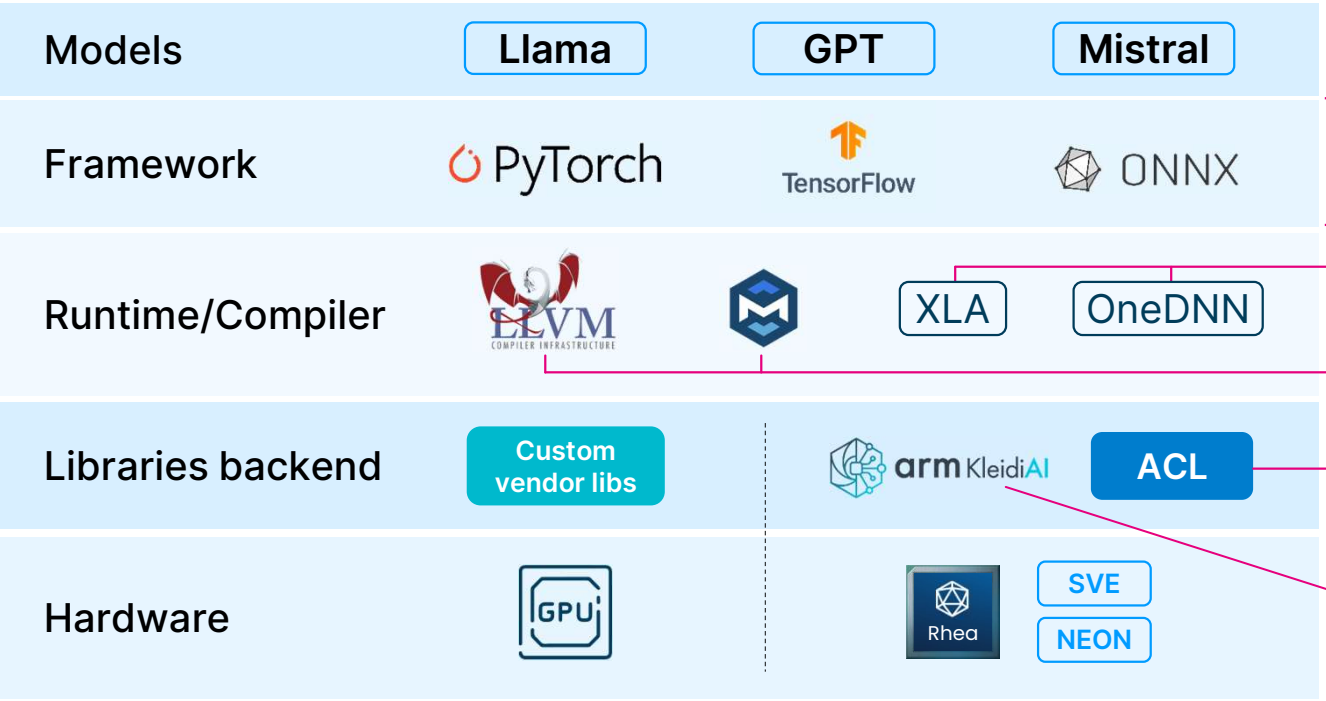
- Coding tasks
- Complex dataset analysis
- Information retrieval tasks
- Text translation
- Educational tutoring
- Text summarization
- Questions answering
- Sentiment analysis
- Clinical Note Summarization
- Scenario generation & recommended measures
- Mental Health support
- Scientific explanation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

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

Key considerations



Maturity of Software stack for AI on ARM



Arm and the community (including hyperscaler) provides and maintains efficient stack

Arm CPU can leverage oneDNN (from oneAPI) and also XLA TFRT compiler.

Maturity of compiler thanks to community

Arm Compute Library:
Arm provides an optimized "Arm Compute Library (ACL)" to the users (today in MIT license).

An open-source library that provides optimized performance-critical routines, also known as **micro-kernels**, for artificial intelligence (AI) workloads tailored for Arm® CPUs.

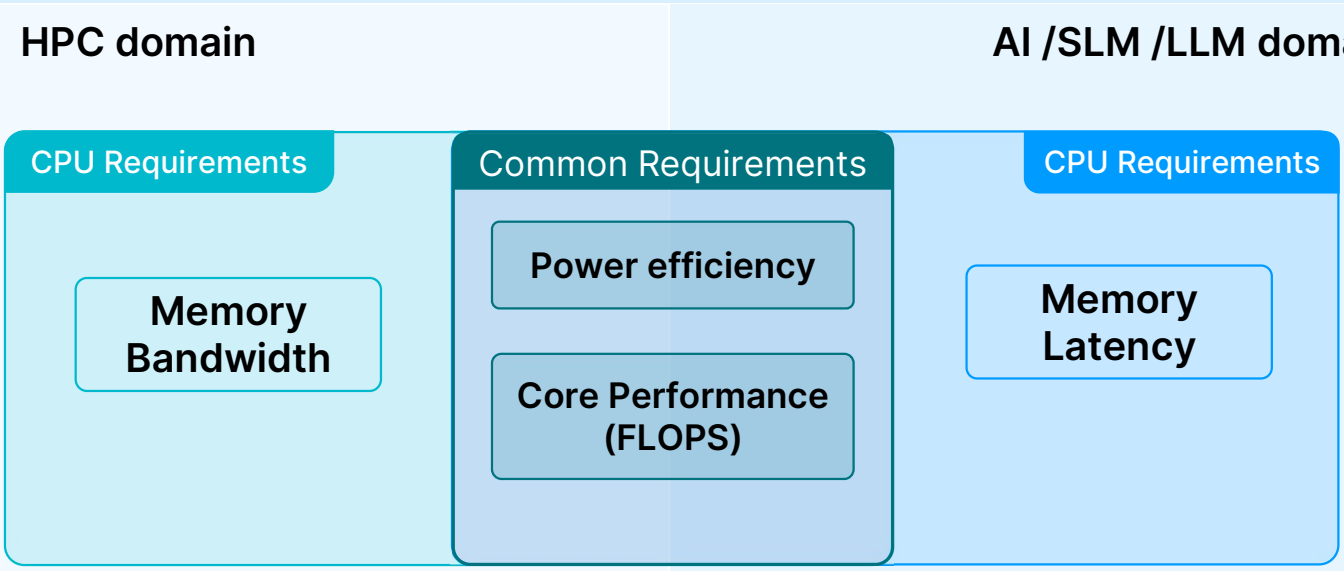
Covers Inference, Finetuning, Quantization and Training

Common characteristics of HPC and AI /SLM /LLM workloads

Challenges are similar for AI/LLM and HPC

AI tends to be more demanding in latency

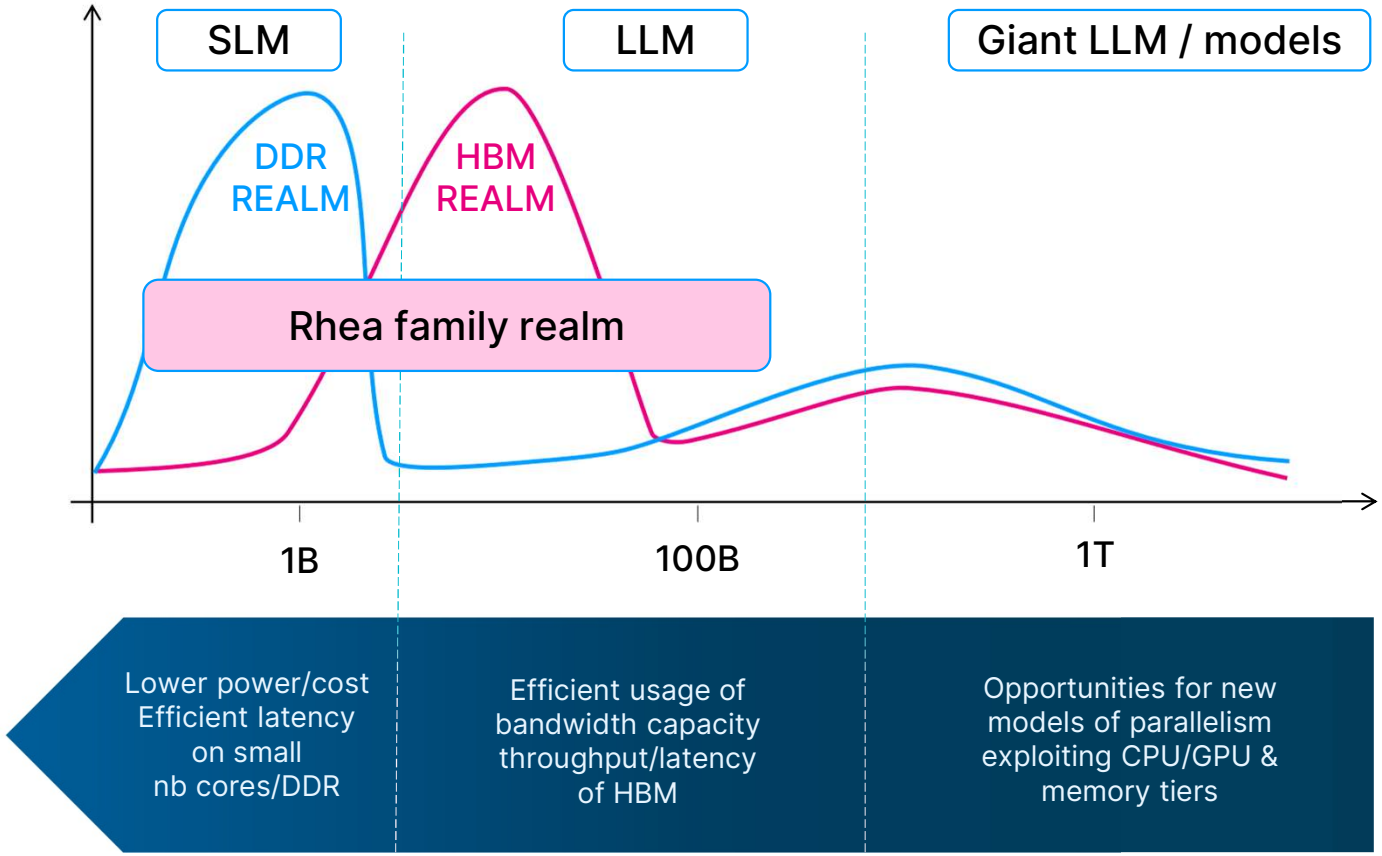
HPC tends to be more demanding in Bandwidth



Rhea1 supporting AI

HBM allows state of the art models to be tackled with lower cost and lower power while minimizing latency.

Rhea supports state of the art precision types (int8, bfloat16) for LLM



Server board design as reference design

A key step in the development of high-performance microprocessors

- Internal testing and characterization of silicon
- Software development and integration
- Demos

An essential prerequisite for the server design of our direct customers

- Reduced design costs

A platform to be used by our partners, EU & local projects, customers

- Software development and integration
- Performance evaluation

A design meeting all the functional requirements of different uses

#1 Reference design

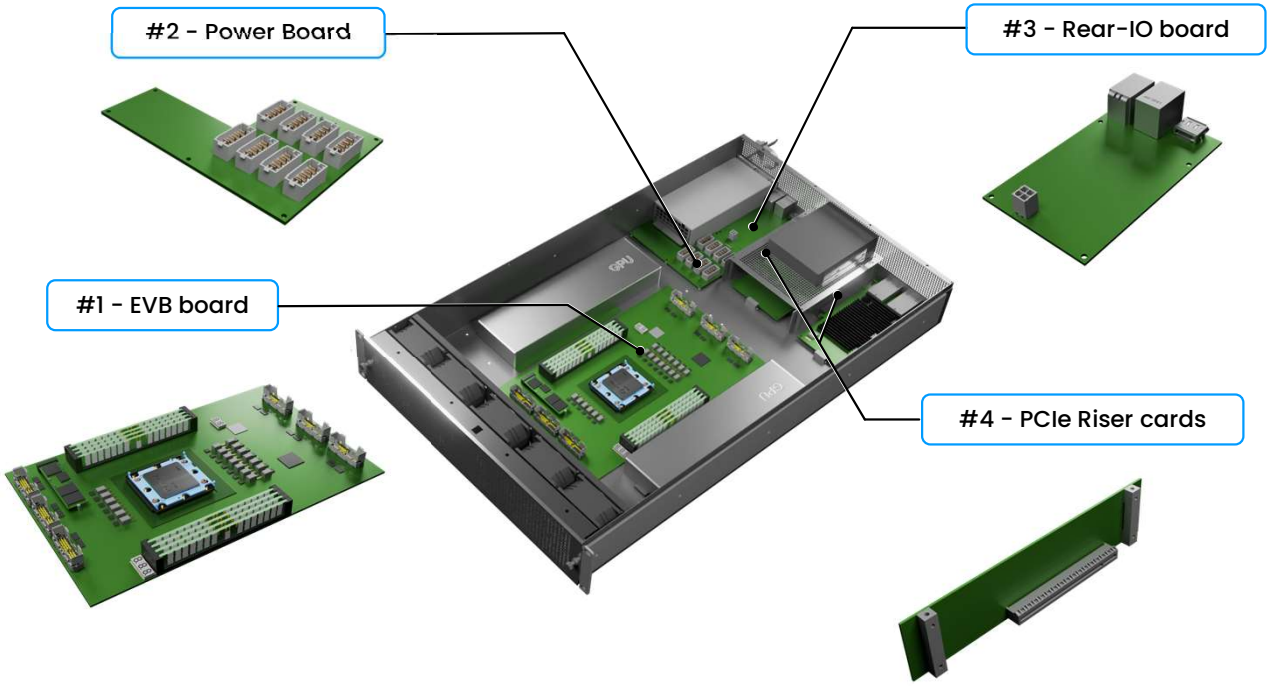
#2 Validation & testing

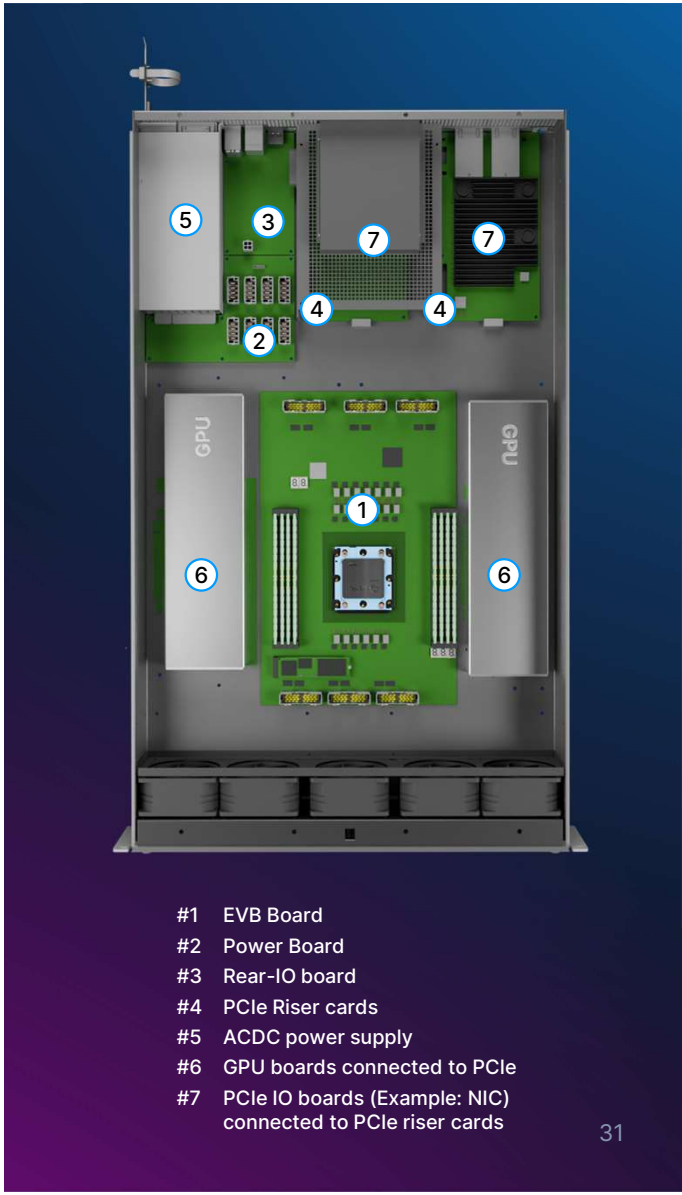
#3 Multiple demos

A multifonctional, flexible & versatile platform

In response: a single modular reference server solution

Overview of Seine Platform



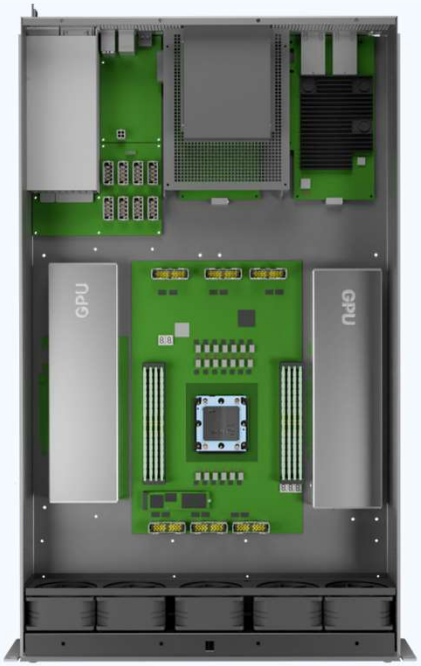


#1 EVB Board
#2 Power Board
#3 Rear-IO board
#4 PCIe Riser cards
#5 ACDC power supply
#6 GPU boards connected to PCIe
#7 PCIe IO boards (Example: NIC) connected to PCIe riser cards

A modular solution for every application (1/3)

Config #1

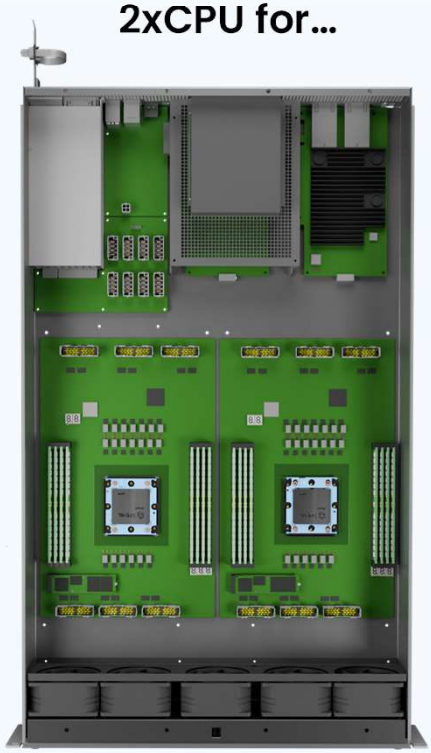
Single-CPU with 2xGPU for…



A modular solution for every application (2/3)

Config #2

2xCPU for…

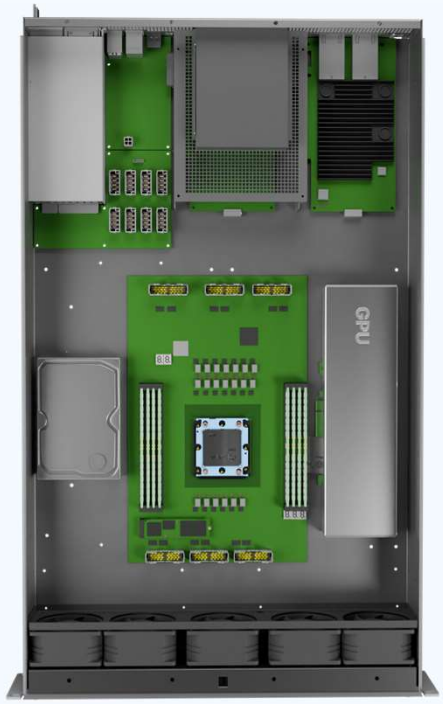


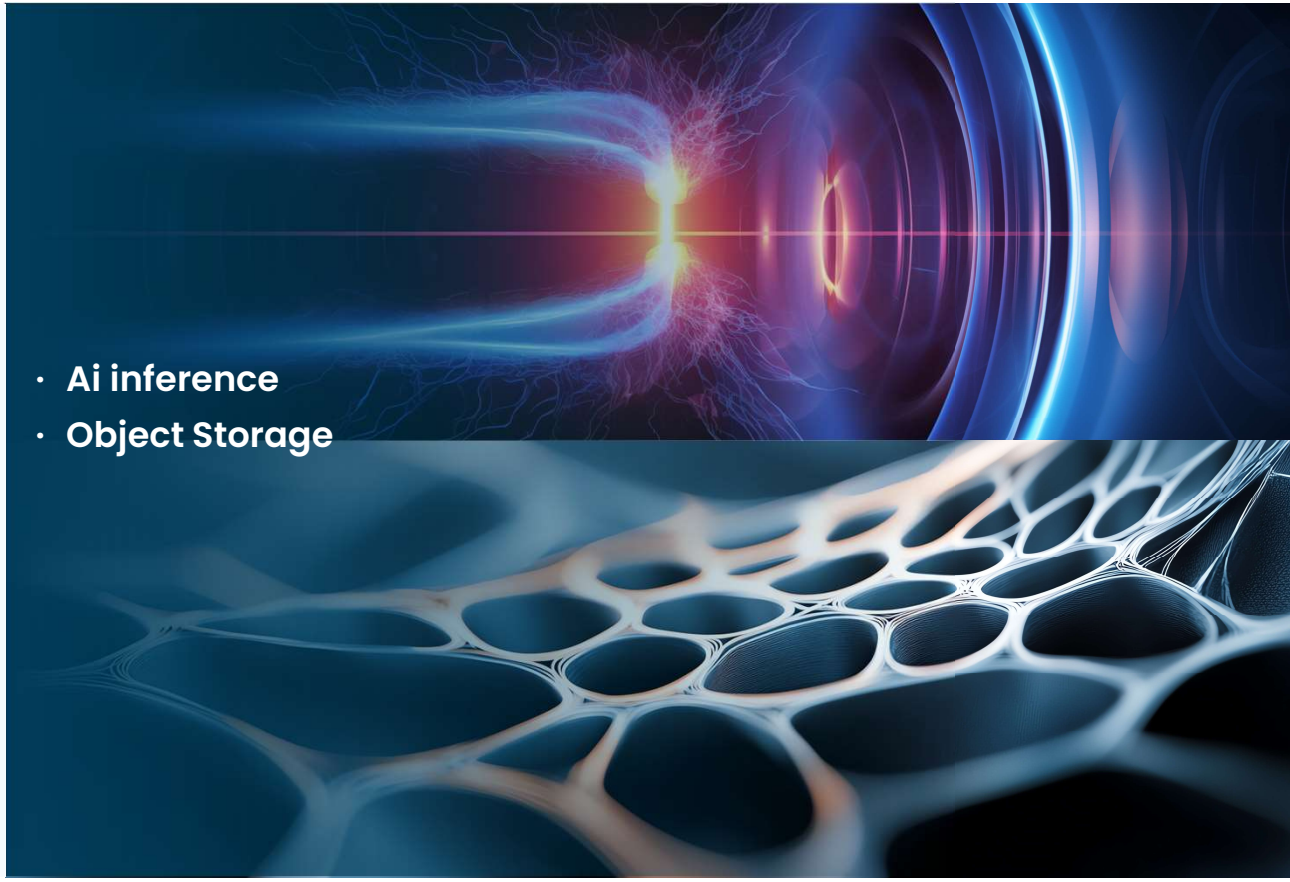
- **Ai inference**
- **Virtualization**

A modular solution for every application (3/3)

Config #3

Single-CPU with 1xGPU & 1xSATA





Seine server is now available and ready for Rhea1