Dr. Jens Krüger
Fraunhofer ITWM

The European Processor Initiative

Agenda

1. Short Introduction and Background
2. The European Processor Initiative (EPI) and related projects
3. Is souvereignity possible?
4. The Stencil- und Tensor Accelerator (STX)

Disclaimer: vermutlich sehr subjektive Sichtweisen des Autors

Short Introduction and Background

The Fraunhofer-Gesellschaft

At a Glance

Application-oriented research with a focus on future-relevant key technologies and the utilization of the results in business and industry. Guide and source of inspiration for innovative developments.

> 20 000
Employees

3,4 Billion €
Budget

76 Institutes and Research Facilities

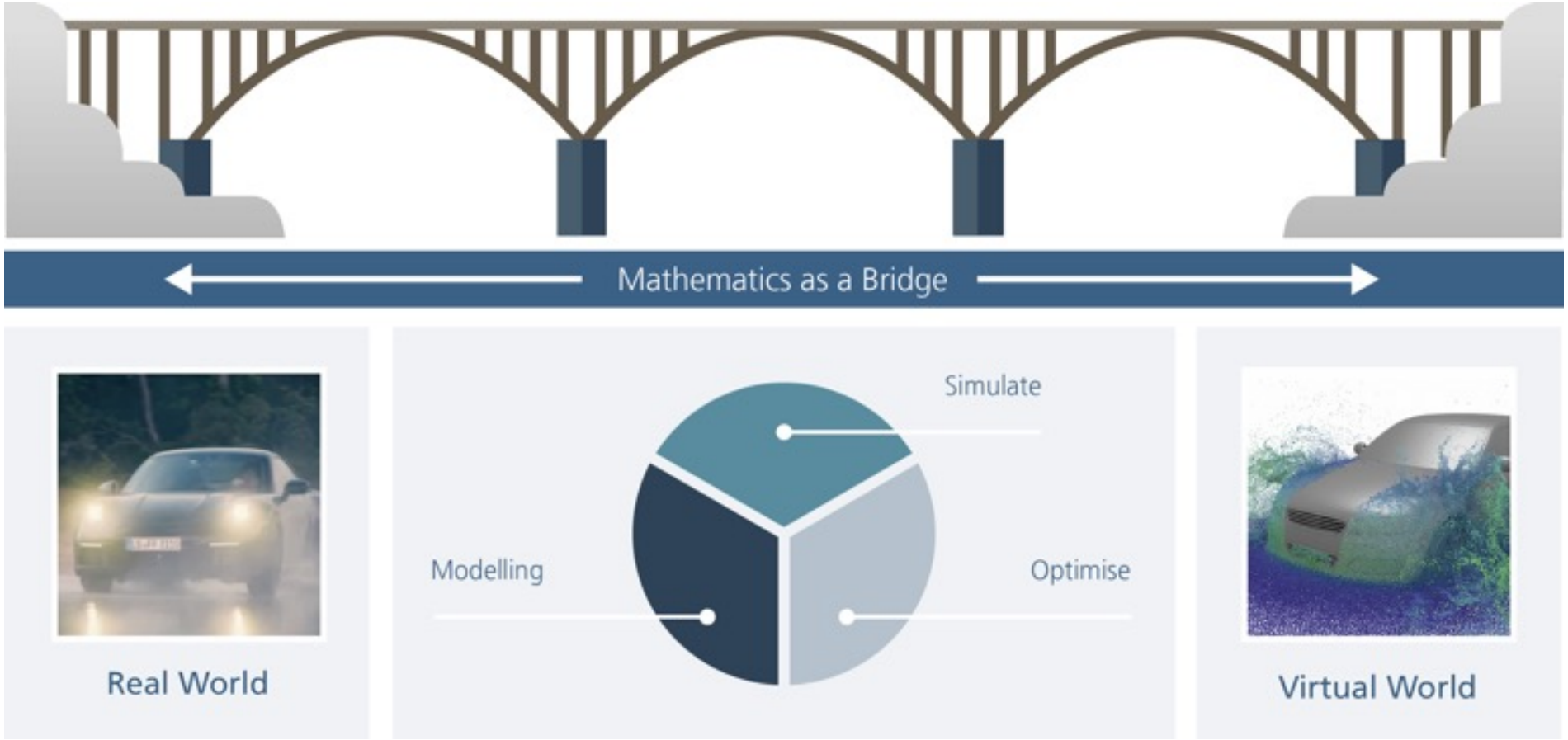
9 Associations:

- Energy Technologies & Climate Protection
- Health
- Innovation Research
- **IUK-Technology**
- Light & Surfaces
- Microelectronics
- Production
- Resource Technologies and Bioeconomy
- Materials and Components

- Fraunhofer Defense, Prevention and Security Division VVS

Fraunhofer ITWM

Mathematics is a Key Technology

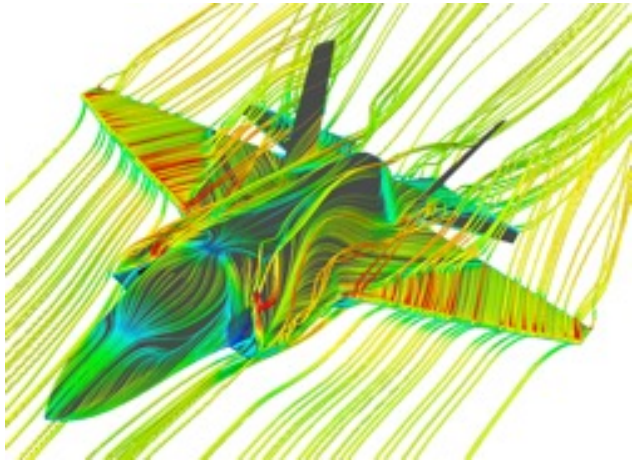


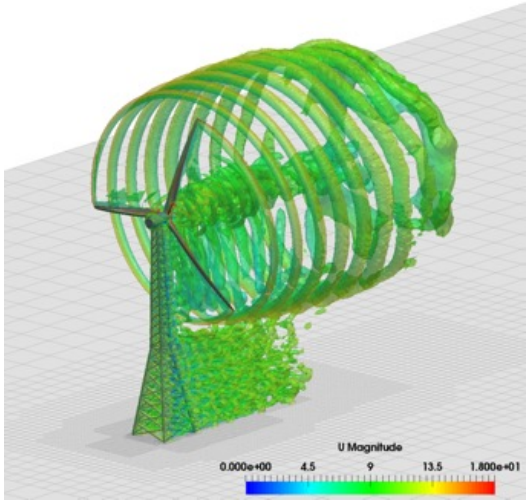
From real life experiments …

https://www.deutschlandfunk.de/atomwaffen-warum-es-immer-noch-13-400-atombomben-gibt-100.html

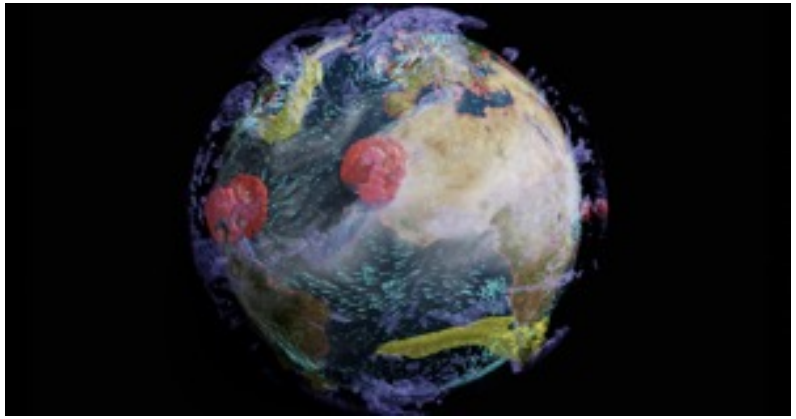
https://i.imgur.com/YhAR4yg.jpg

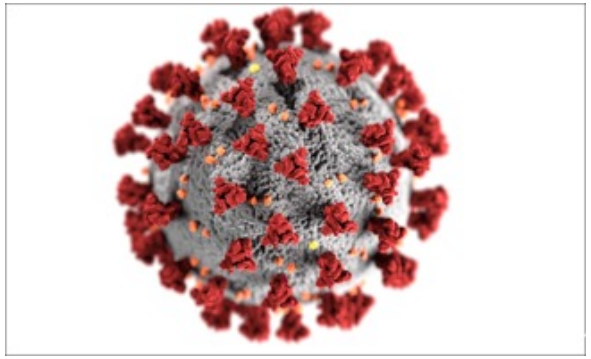
to model based simulations

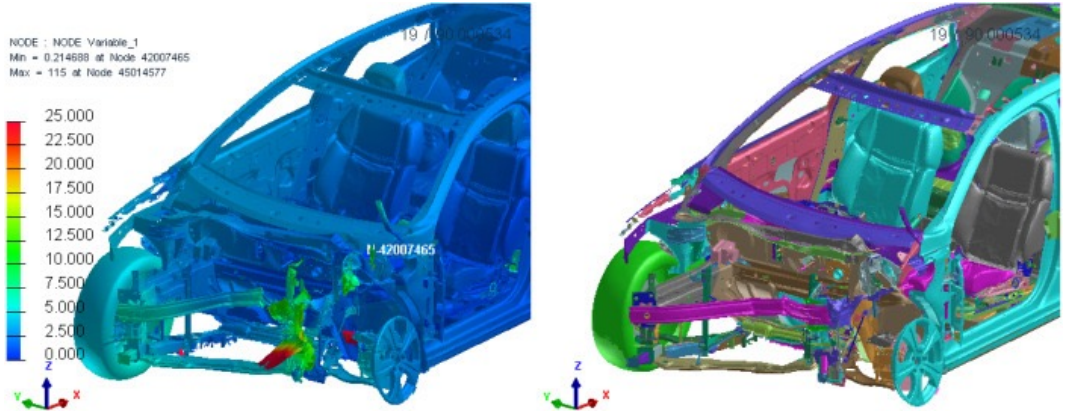




https://blogs.nvidia.com/blog/2021/11/12/earth-2-supercomputer/

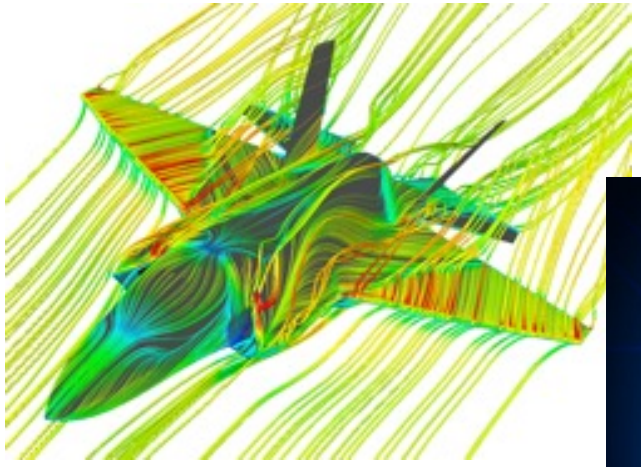


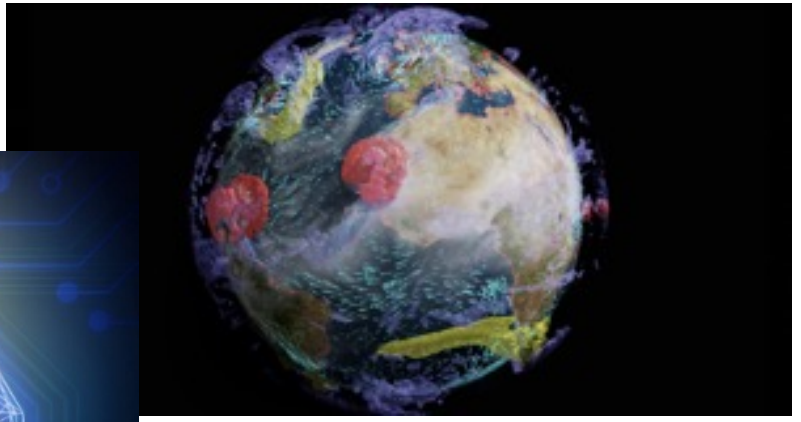


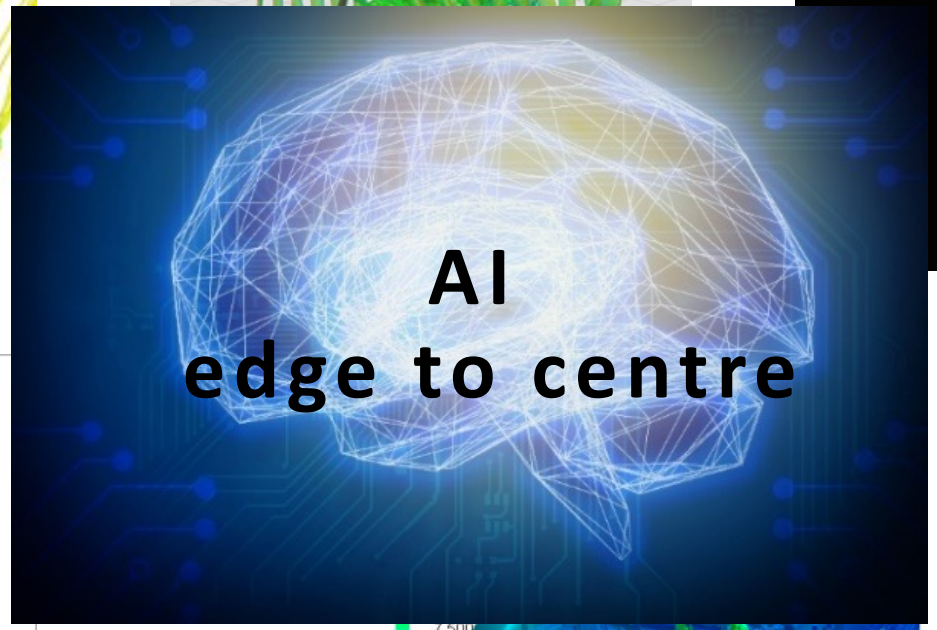


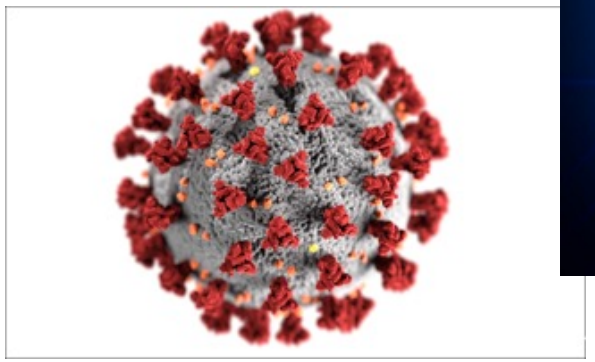
to data driven simulation and modelling

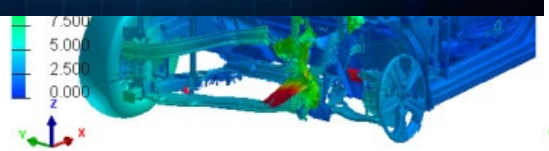
https://blogs.nvidia.com/blog/2021/11/12/earth-2-supercomputer/

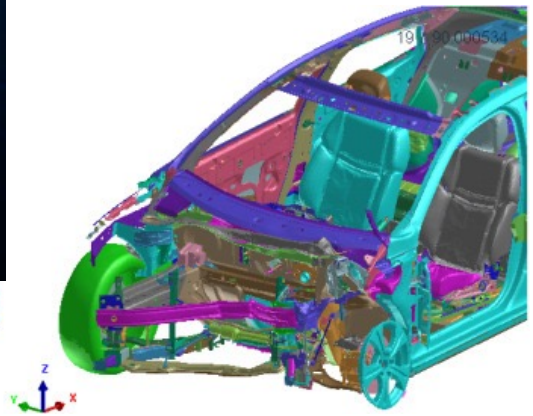








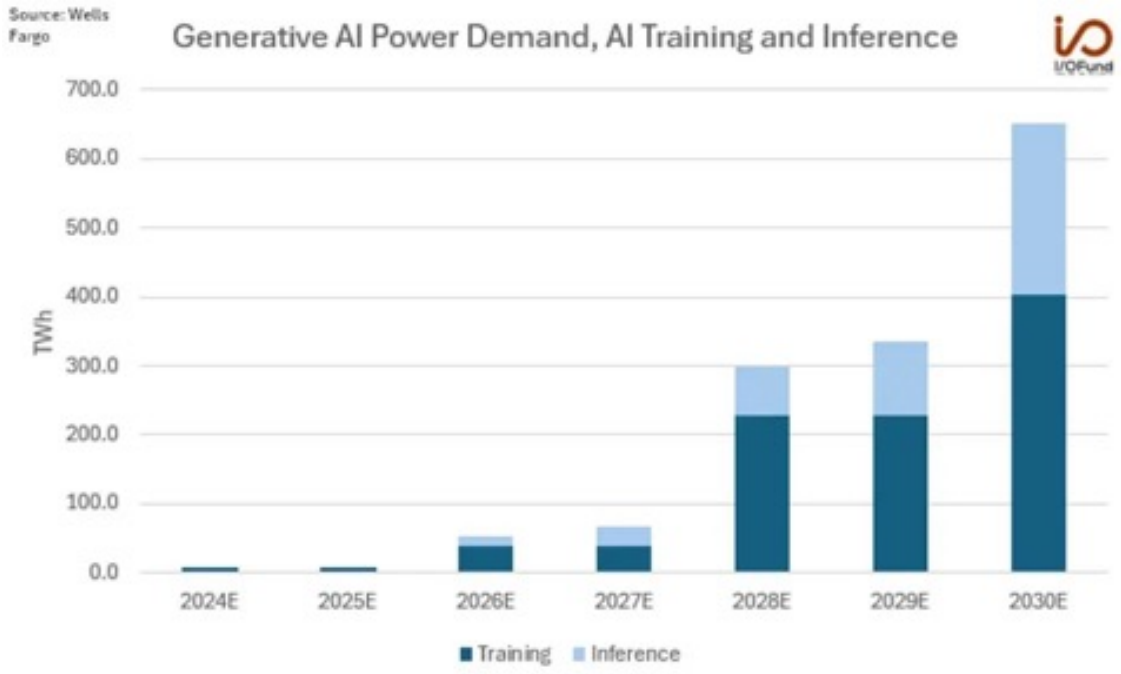




Current trends are not sustainable

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**El Capitan** - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS, HPE DOE/NNSA/LLNL United States	11,039,616	1,742.00	2,746.38	29,581
2	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Cray OS, HPE DOE/SC/Oak Ridge National Laboratory United States	9,066,176	1,353.00	2,055.72	24,607
3	**Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	9,264,128	1,012.00	1,980.01	38,698

Source: Top500.org



Source: https://www.forbes.com/sites/bethkindig/2024/06/20/ai-power-consumption-rapidly-becoming-mission-critical/

Estimate assume ~200MW peak for the xAI Collossus

Innovation gefährdet durch zu wenig Anschlussleistung?

Startseite › Lokales › Erding › Wartenberg

Wartenberg: Für Großrechenzentrum fehlt der Strom

23.06.2025, 12:00 Uhr
Von: Markus Schwarzkugler

Kommentare Drucken Teilen

DARSTELLUNGEN

MITTELDEUTSCHE REGION

Netzkapazität wird KI-Rechenzentrumsbau ausbremsen

In Brandenburg, Sachsen-Anhalt und Sachsen sollen über 60 KI-Rechenzentren entstehen. Doch die Anschlusskapazität von 13 Gigawatt ist nicht vorhanden.

in Pocket speichern merken

24. Juni 2025, 18:57 Uhr, Achim Sawall

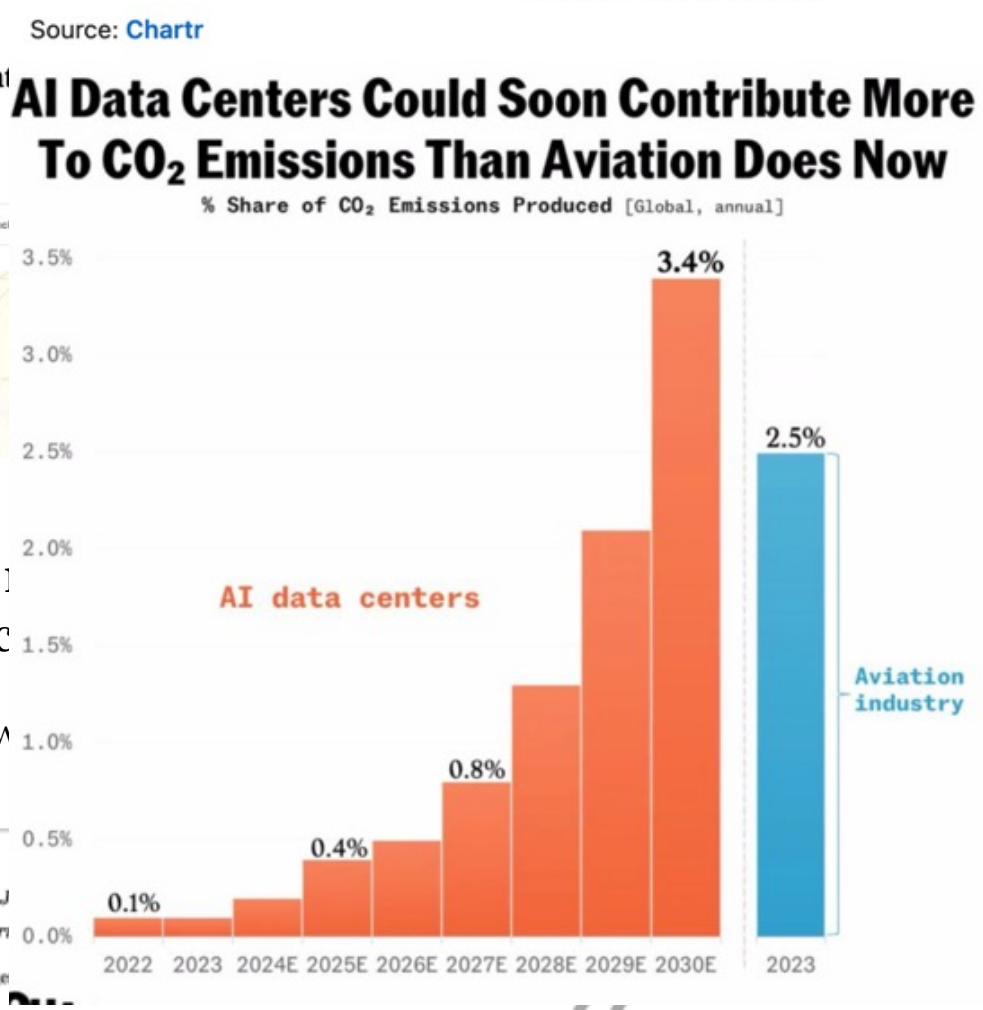
Pröbst berichtete von Gesprächen mit Tennet und Bayernwerk: „300 Megawatt bräuchten wir."

„Demnächst rüsten sie zwar um 700 Megawatt auf, aber das ist alles schon vergeben."
Dass wirklich nichts zu machen ist

> *Zehn Jahre Lieferzeit für Strom. Da muss ich schon sagen: Gute Nacht, Bayern.*
>
> - Bürgermeister Christian Pröbst (CSU) -

Bereits vor gut zwei Jahren hatte der Marktrat beschlossen, den

Innovation gefährdet durch zu wenig Anschlussleistung?



KI-Rechenzentrumsbau

und Sachsen sollen über 60 KI-Rechenzentren
pazität von 13 Gigawatt ist nicht vorhanden.

24. Juni 2025, 18:57 Uhr, Achim Sawall

Pröbst berichtete von Gesprächen
Bayernwerk: „300 Megawatt bräuc

„Demnächst rüsten sie zwar um 700 Megaw
Dass wirklich nichts zu machen ist

Zehn J
Bayern
- Bürge

Bereits vor gut zwei Jahren hatte der Marktrat beschlossen, den

Fraunhofer ITWM

Innovation gefährdet durch zu wenig Anschlussleistung?

Source: Chartr

Startseite › Lokales › Erding › Wartenberg

Warte
fehlt d

23.06.2025, 12:0
Von: Markus Sch

Kommenta

DARSTELLUNGEN

NVIDIA

Home | AI | Data Center | Driving | Gaming | Pro Graphics | Robotics | Healthcare | Startups | AI Podcast | NVID

Gearing Up for the Gigawatt Data Center Age

Inside the AI factories powering the trillion-parameter era — and why the network matters more than ever.

August 21, 2025 by Gilad Shainer

KI-Rechenzentrumsbau

…sen sollen über 60 KI-Rechenzentren
…n 13 Gigawatt ist nicht vorhanden.

24. Juni 2025, 18:57 Uhr, Achim Sawall

Pröbst berichtete
Bayernwerk: „30c

„Demnächst rüsten si
Dass wirklich nichts z

Source: https://blogs.nvidia.com/blog/networking-matters-more-than-ever/ (last checked: 24.8.25)

Bereits vor gut zwei Jahren hatte der Marktrat beschlossen, den

Fraunhofer ITWM

Next Generation Computing

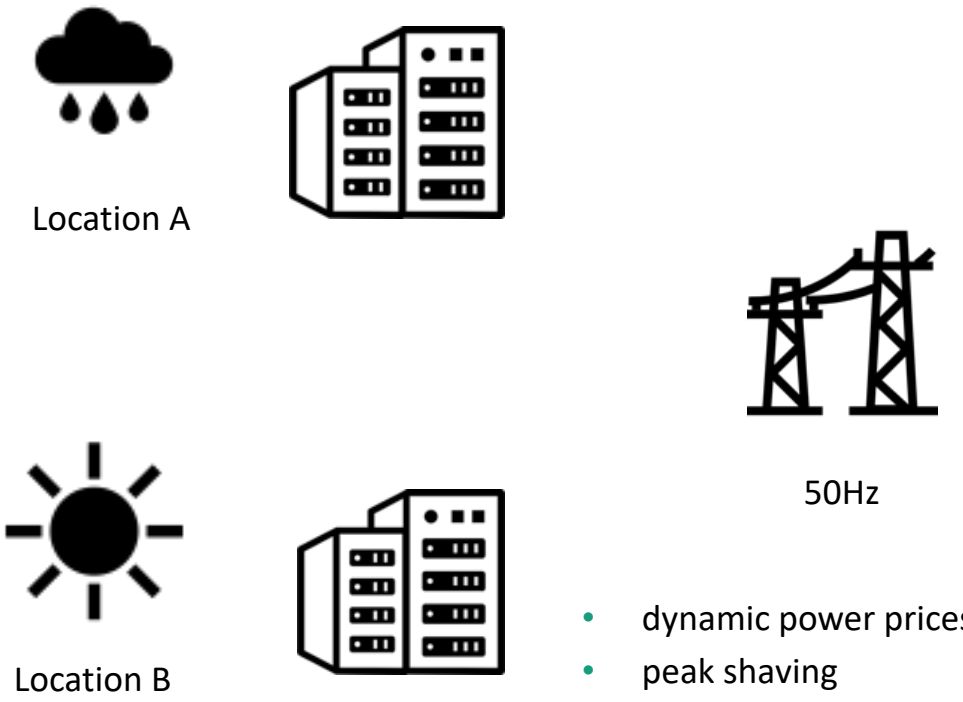
- What if we could solve problems up to 100.000x more energy efficient than today?
- What if we could solve problems in seconds that would normally take decades?
- What if we could have secured supply chains and trusted hardware in Europe?

Fraunhofer ITWM

Holistic Optimization for Efficiency

- Co-Design System architectures for best Total Cost of Ownership (TCO) – Multi-centre to IoT
- Fraunhofer consulting to find the best combination of computing resources and optimized software
- Holistic approach:
  - Algorithmic optimization
  - Hardware specific optimization
  - Efficient Software/Tools
  - Data access optimization
  - Scalability optimization
- Next generation Computing Systems:
  - Neuromorphic Computing
  - Quantum Computing
  - Novel Efficient Classical Computing

Smart Application Scheduling



- dynamic power prices
- peak shaving
- grid frequency stabilization
- CO2 footprint / cost reduction
- hardware as a flexibility

European Processor Initiative

European Processor Initiative

Framework Partnership Agreement In European Low-power Microprocessor Technologies

EPI TIMELINE

SIPEARL

EU Exascale Systems 

EUPEX Pre-Exascale Pilots EUPILOT

DARE

Phase 1:
GPP & Accelerator test chips

Phase 2:
GPP V1 Go to Market
Accelerator Demonstrator

Next project(s)
GPP V2 Go to Market
RISC-V Accelerators Go to Market

Mont-Blanc projects
Arm HPC ecosystem & HW design space exploration

Centres of Excellence in HPC Applications

2011 – 2014 | 2015 – 2018 | 2019 – 2021 | 2022 - 2025 | 2025 -

Nov 2018
EuroHPC JU

Sept 2020. State of the Union
- Europe's digital sovereignty - investment of 8 billion euros in the next generation of supercomputers
- European industry to develop our own next-generation microprocessor

Sept 2023.
Digital sovereignty: European Chips Act enters into force

Nov 2023.
Chips JU launched

PROJECT FAMILY OVERVIEW

- European Processor Initiative (EPI)
  - Framework Partnership Agreement (FPA) 6 years
  - Specific Grant Agreements (SGA1 3years 130M€, SGA2 3years 70M€)
  - Until end of 2025
- The European Pilot (30M€) / EUPEX (40M€)
  - Project funded by EuroHPC JU
  - 2022 - 2026
- Digital Autonomy with RISC-V in Europe – DARE
  - EuroHPC FPA 6 years
  - AI
  - SGA1 - 3years (240M€)

EPI PROJECT FACTSHEET

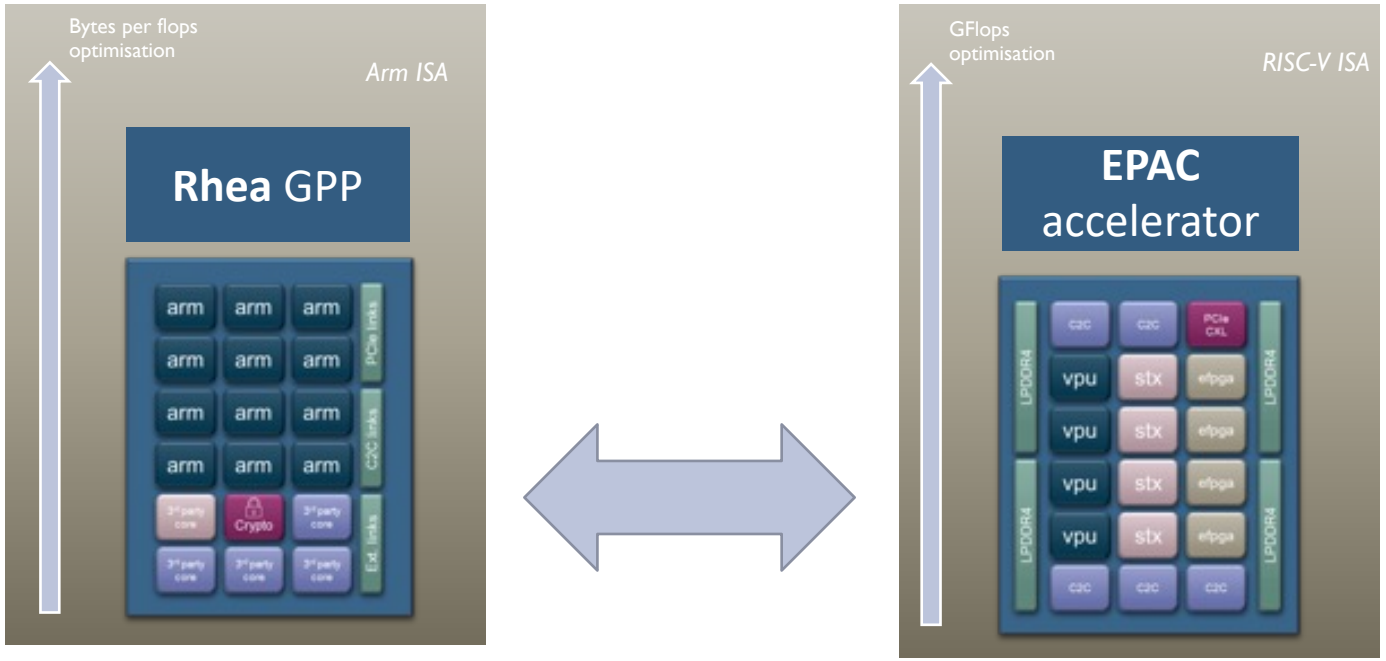
- Currently in Phase 2 (2022-2025)
- Consortium of 27 strategically chosen key European academic and industrial partners
- Funded by EuroHPC JU (50%)
  - and co-funded by Croatia, France, Germany, Greece, Italy, the Netherlands, Portugal, Spain, Sweden and Switzerland

EPI & PILOTS : MAIN OBJECTIVES

EU chips fit for HPC usage - at Exascale level

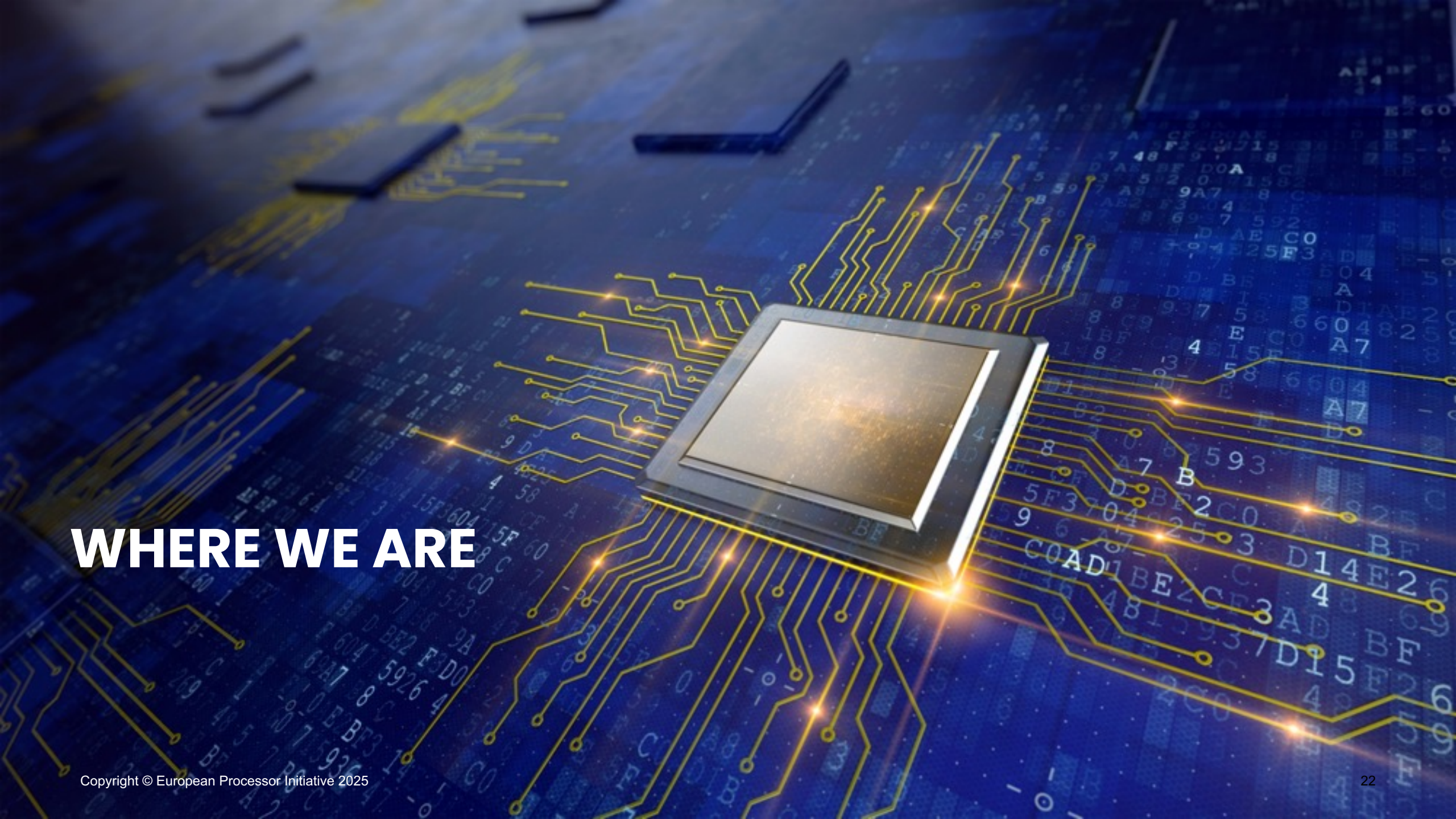
General Purpose Processors:
Legacy & programmability

Accelerators:
Computing force



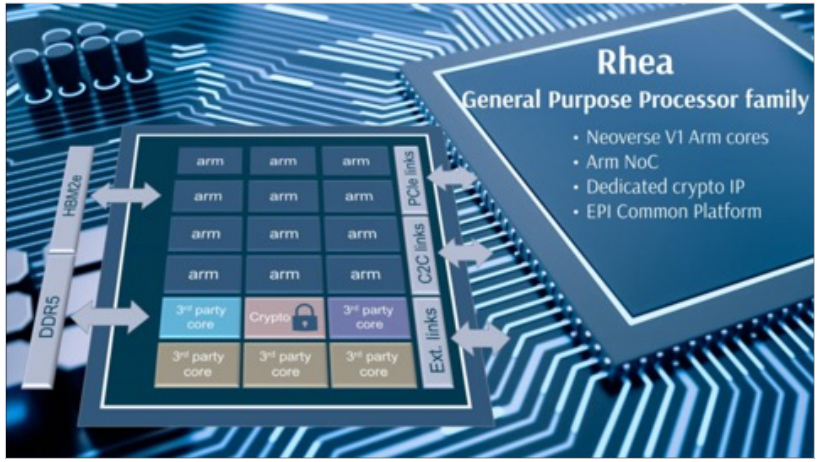
EPI – EUROPE'S AMBITION

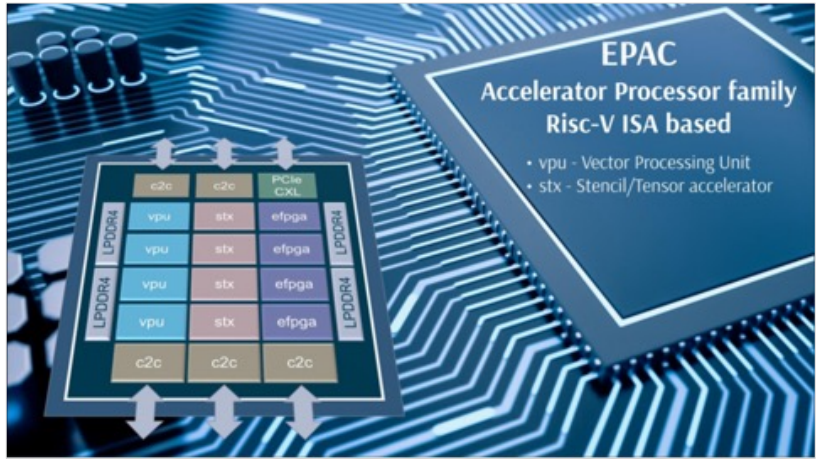
- In Phase 2 (SGA2):
  - Finalize the development of the first generation of low-power processor units (GPP) developed in Phase 1
  - Develop the second generation of low-power accelerators and related software stack
  - Develop the second generation of GPP by applying technological enhancements concerning the GPP (Rhea)
  - Focus on the industrialisation & commercialisation path of the developed IPs and innovations & Enable the long-term economic sustainability and scalability
  - Develop an open Hardware and Software Common Platform standard



WHERE WE ARE

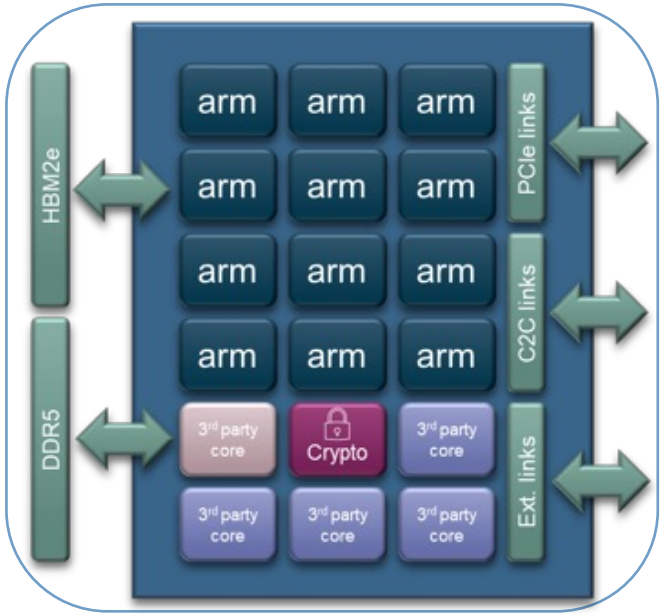
EPI: TWO PROCESSOR FAMILIES





RHEA1: GENERAL PURPOSE PROCESSOR

- **Chip**
  - 6 nm
  - integrating both CPU and accelerator parts
  - Network on Chip (NoC): mesh linking all components
- **Core**
  - Arm ISA
  - ARM Neoverse V1 core design – 80 cores
  - Arm tile with 2 SVE256 (Scalable Vector Extension) units
- **Memory hierarchy**
  - Cache subsystem: L1, L2, SLC (System Level Cache)
  - 4 HBM: High Bandwidth Memory
  - 4 DDR5 interfaces (Standard Memory)
- **High speed I/O**
- **Security block – dedicated crypto IP**



Design targets:

- High Byte/Flop ratio
- Compute performance and efficiency for real-applications
- Ideal for HPC & AI inference

RHEA DEVELOPMENT STATUS

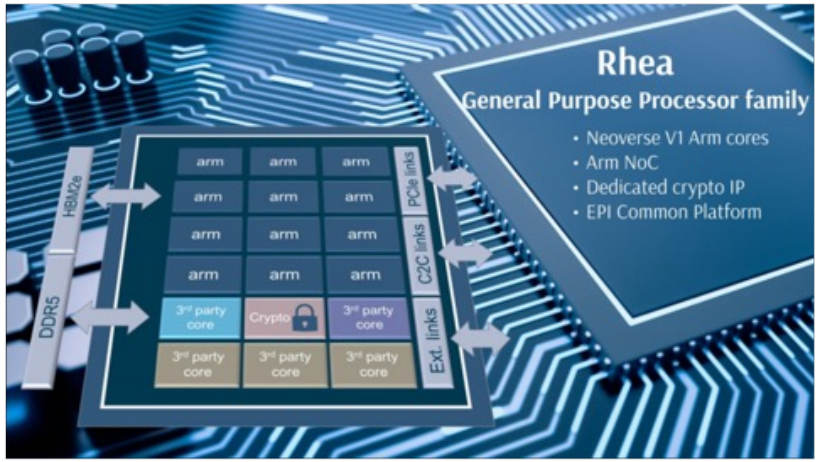
SiPearl Serie A funding completion

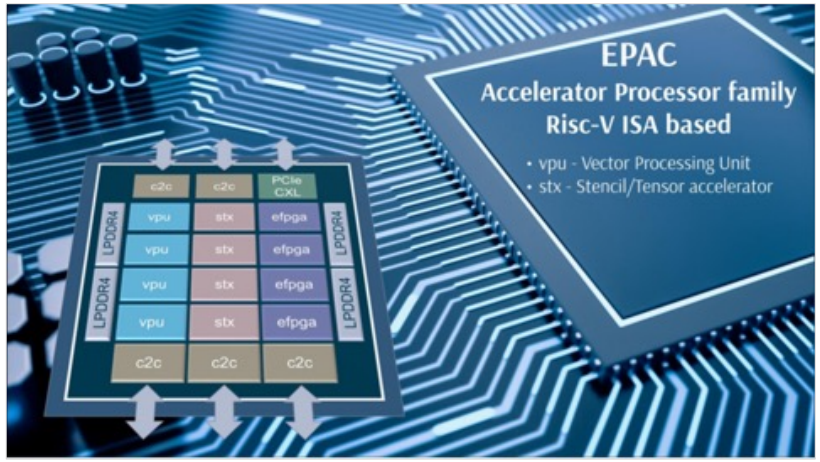
Chip announced @ ISC

Rhea1 Design finished

Tests & emulation (booting Linux on dual socket)

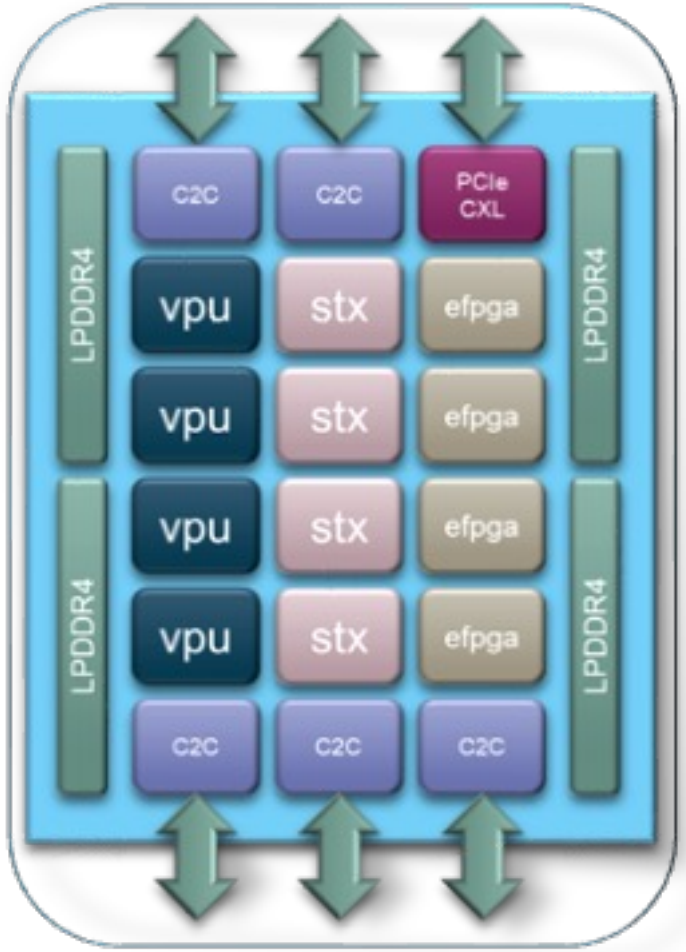
EPI: TWO PROCESSOR FAMILIES





EPAC VISION AND CONTRIBUTIONS

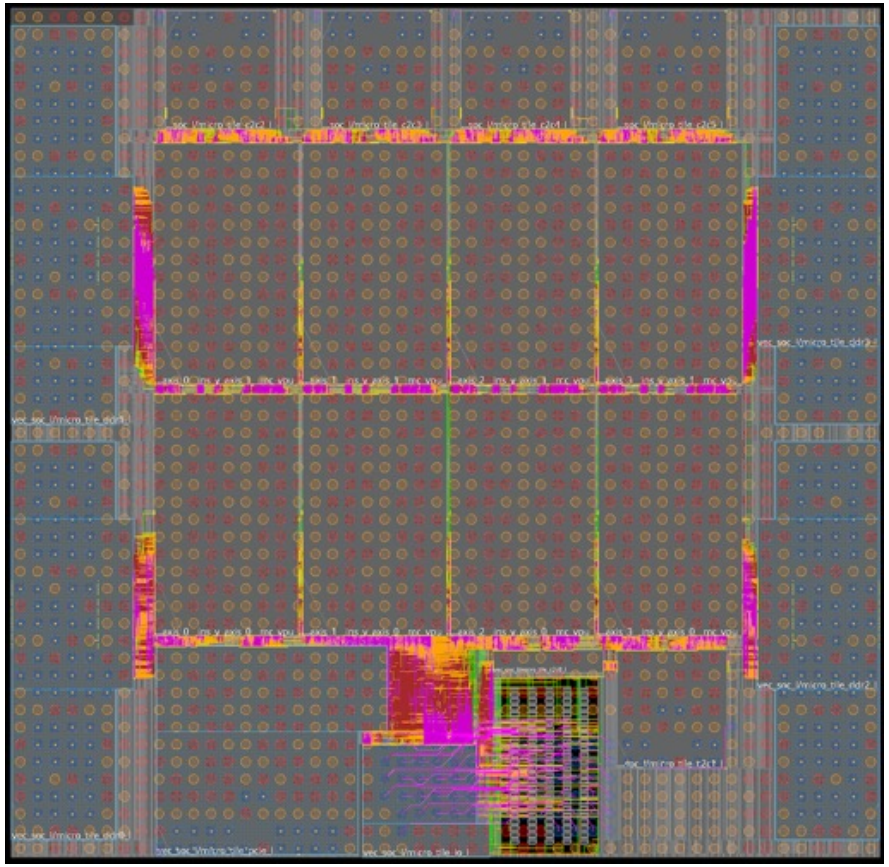
- **VEC -** Self-hosted RISC-V CPU + wide VPU (256 double elements) supporting RVV 0.7.1 / 1.0
- **STX -** RISC-V CPU + specific cores for stencil and neural network computation
- **VRP -** RISC-V CPU with support for variable precision arithmetic (data size up to 512 bit)
- **eFPGA -** On-chip reconfigurable logic
- **Ziptillion -** IP compressing/decompressing data to/from the main memory
- **KVX -** FPGA demonstrator of the Kalray RISC-V CPU targeting HPC and ML



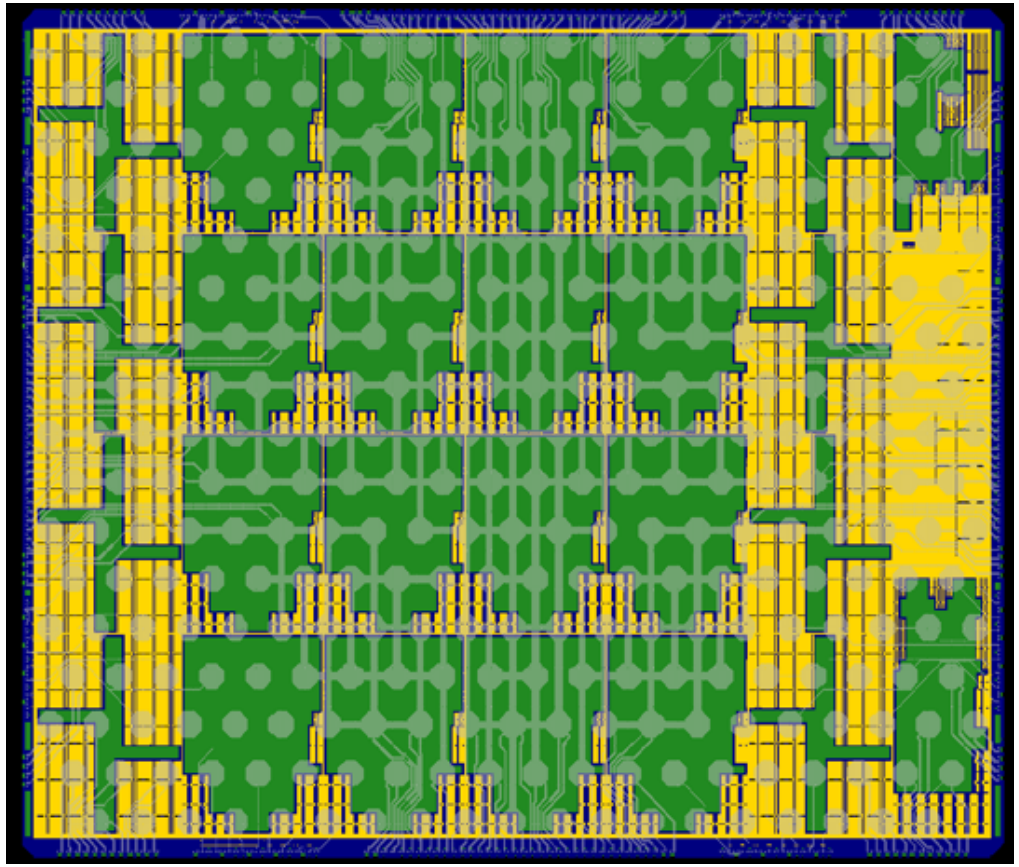
BREAKING NEWS: 21.8.25

THE EUROPEAN PILOT TAPEOUTS

VEC – full SoC (GF12LPP)



MLS/STX – compute only (7nm TSMC)

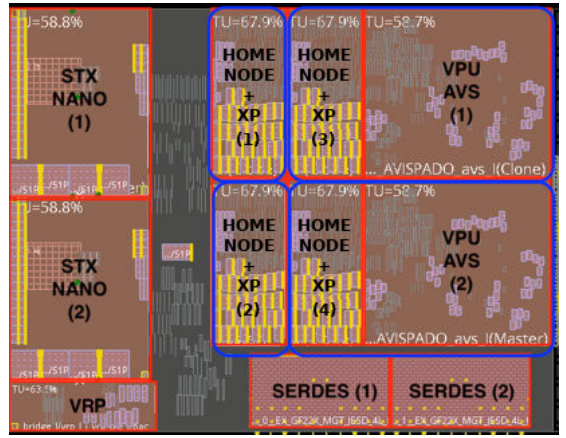


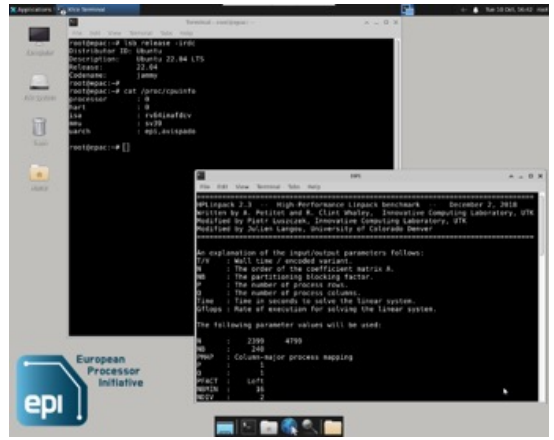
EPAC DEVELOPMENT STATUS

EPAC 1.5 realised and running well (demoed …)

Sw Development Vehicle: training performed (FZJ, HiPEAC …)

IPs consolidation for further steps (EUPILOT & DARE …)





CO-DESIGN AND VALIDATION

Hardware platforms

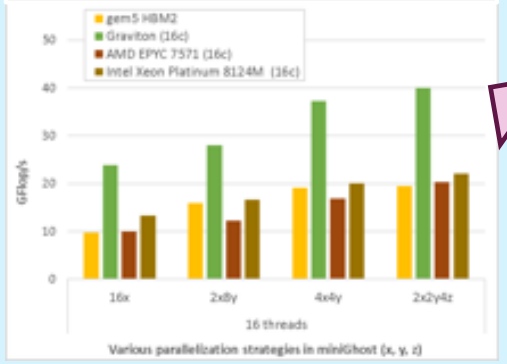
Refer. Platforms

SDVs

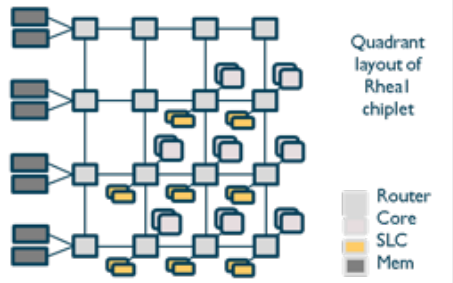
EPI Prototype

Validation and Co-design

Validation



Codesign



Rhea architects

EPAC architects

Future Chips

EPI TIMELINE

original timeline (2022)

EPI 1

2019-2021

2022

Rhea platform

EPAC V1.5 sent for manufacturing

- KVX RISC-V based accelerator architecture definition
- Menta FPGA chiplet architecture definition

Rhea Family - Gen1 GPP

EPI Common Platform
Arm & Risc-V (STX, VRP, ..)
Arm Neoverse V1 Core – N6
External IP's

- EPAC V1.5 multi-node demonstration cluster
  EPAC V2.0 sent for manufacturing
- Menta FPGA chiplet sent for maufacturing
- Kalray RISC-V SDK release

2023

2024

Rhea2 Family - Gen2 GPP

Dual chiplet implementation
First EU Exascale system with Rhea processors
FPGA Die-to-Die Demonstrator of the HW Common Platform

- RISC-V KVX FPGA emulator
- EPAC V2.0 Platform

Cronos Family - Gen3 GPP

EU Exascale system with Rhea2 processors

- Hurricane PCIe Acceleration card based on EPAC technology

2025

EPI TIMELINE

PAGE CONTENTS

- JUPITER
- LUMI
- LEONARDO
- MARENOSTRUM 5
- MELUXINA
- KAROLINA
- DISCOVERER
- VEGA
- DEUCALION
- DAEDALUS

JUPITER

JUPITER is located at the Forschungszentrum Jülich campus in Germany and operated by the Jülich Supercomputing Centre. It is based on Eviden's BullSequana XH3000 direct liquid cooled architecture. Soon to be capable of delivering 1 ExaFLOP of computing power, JUPITER will become Europe's first Exascale supercomputer.

https://www.eurohpc-ju.europa.eu/supercomputers/our-supercomputers_en

793.40 petaflops	930.00 petaflops
Sustained performance	**Peak performance**

Compute partitions:	Booster Module (highly-scalable GPU accelerated) Cluster Module (general-purpose, high memory bandwidth)
Central Processing Unit (CPU):	The Cluster Module utilises the SiPearl Rhea1 processor (ARM, HBM), integrated into the BullS… platform.
Graphics	

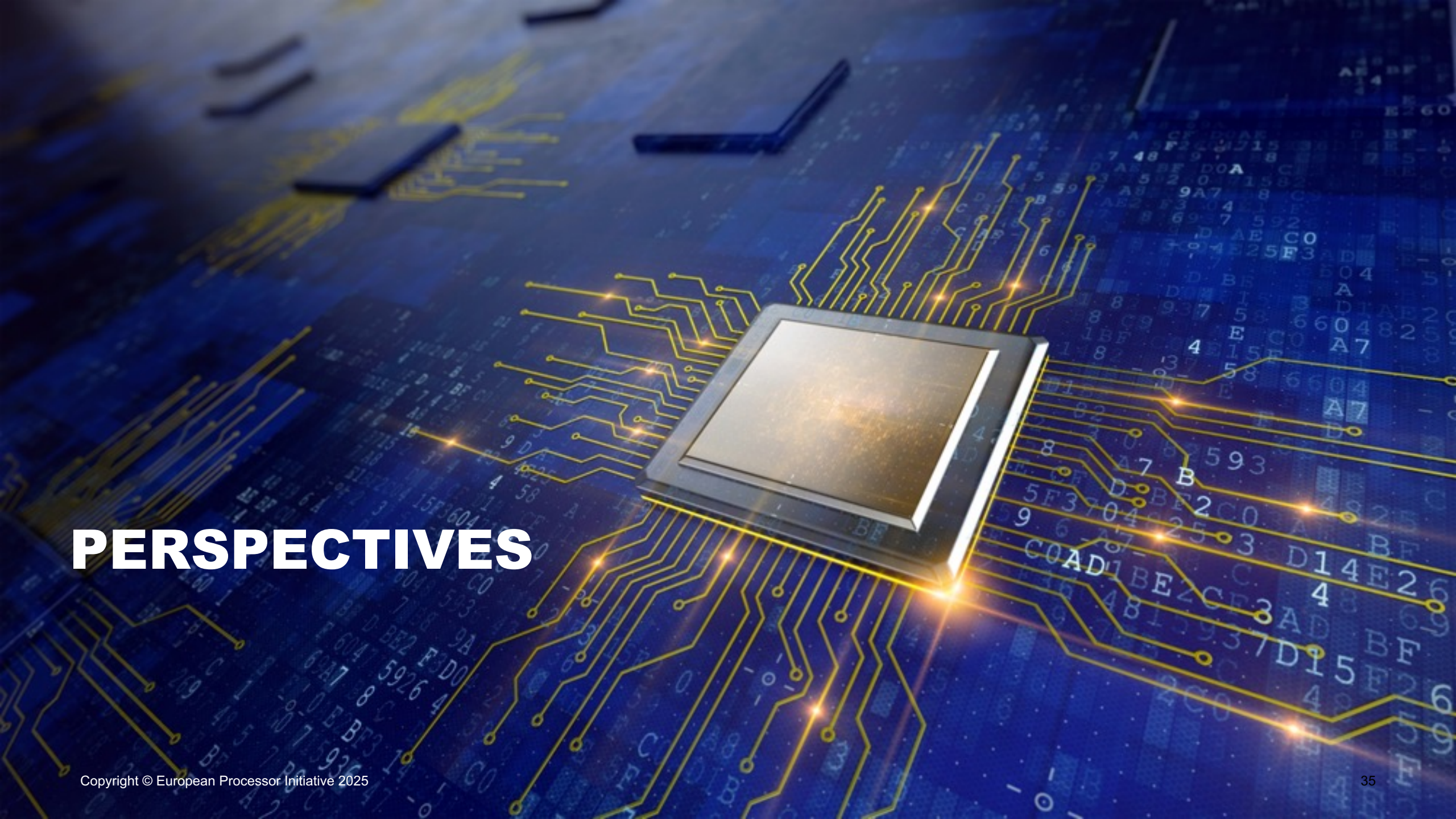
SIPEARL RHEA-1

Tape-out des ersten EU-Prozessors bei TSMC

Die ersten Test-CPUs sollen bereits in einigen Monaten an Partner verschickt werden. Ziel ist ein Supercomputer mit Prozessoren aus Europa.

15. Juli 2025 um 12:27 Uhr / Martin Böckmann

https://www.golem.de/news/sipearl-rhea-1-tape-out-des-ersten-eu-prozessors-bei-tsmc-2507-198121.html



TAKEAWAY

- We are progressing and delivering
- However, it’s a long road
- Collaboration is key

A pathfinder towards an autonomous European HPC supply chain

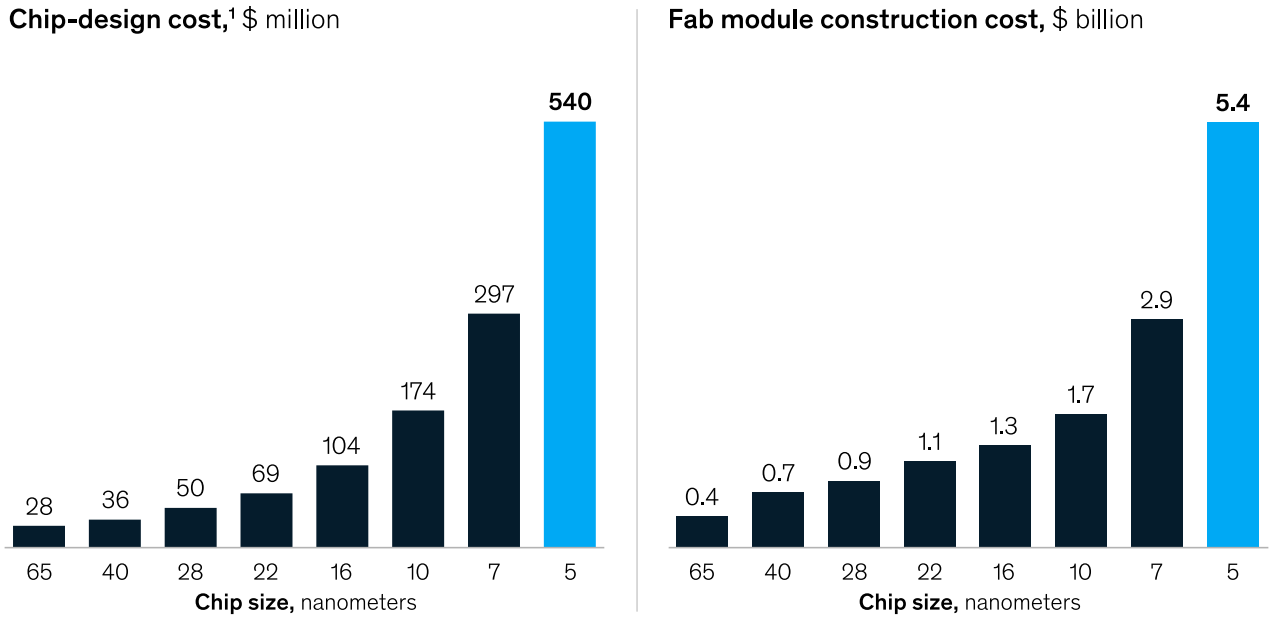
Is sovereignty possible?

Constraints

- **Energy Efficient - T**o fit into the Green Deal agenda
- **Independent** – To support European sovereignity
- **Competitive** – To be able to exploit the results commercially
- and
- **Cheap** – To fit into the limited budget available

Development Costs

R&D for chips and fab module construction costs are soaring.



[1]Major components include IP qualification, architecture, verification, physical, software, prototyping, and validation.
Source: IBS; McKinsey

McKinsey & Company

Quelle:
Semiconductor design and manufacturing: Achieving leading-edge capabilities
August 20, 2020 | Article

https://www.mckinsey.com/industries/industrials-and-electronics/our-insights/semiconductor-design-and-manufacturing-achieving-leading-edge-capabilities (checked: 25.8.25)

Chips tend to become complex

Typ	Release	Takt (GHz)	Stack	pro Stack (1024 bit)	
				Kapazität (2^{30} Byte)	Datenrate (GByte/s)
HBM 1	Okt. 2013	0,5	8× 128 bit	1×4 = 4	128
HBM 2	Jan. 2016	1,0…1,2		1×8 = 8	256…307
HBM 2E	Aug. 2019	1,6…1,8		2×8 = 16	409…461
HBM 3	Okt. 2021	2,8…3,2	16× 64 bit	2×12 = 24	717…819
HBM 3E	2023	4,0…4,8		3×16 = 48	1024…1229
HBM 4	2026	5,6		2×16 = 32	1434[4]

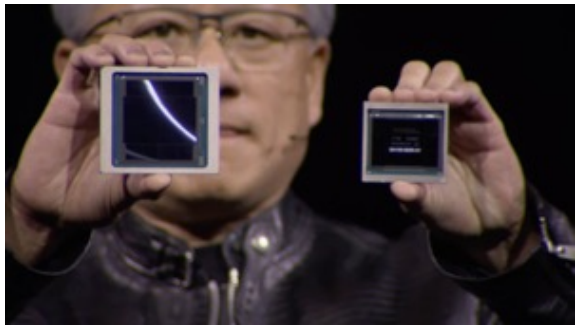
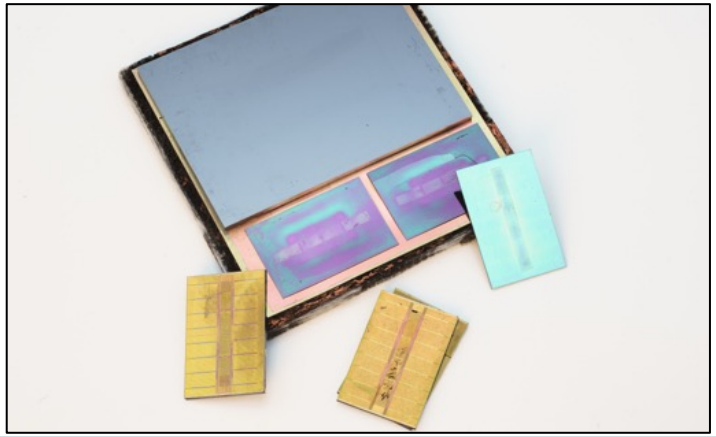


Bild: Nvidia



Source: https://de.wikipedia.org/wiki/High_Bandwidth_Memory

(CC0 1.0) Fritzchens Fritz; https://www.flickr.com/photos/130561288@N04/

Historic Example of European Success

France, Germany, England – 2 main companies

European Initiatives

CALL FOR PROPOSALS | Closed

Framework Partnership Agreement (FPA) for developing a large-scale European initiative for High Performance Computing (HPC) ecosystem based on RISC-V

The aim is to establish a partnership between the EuroHPC JU and a consortium of industry,

Participants

(999655520) - COORDINATOR
(996058081) - BENEFICIARY
(999843409) - BENEFICIARY
(999645820) - BENEFICIARY
(999984059) - BENEFICIARY
(999981149) - BENEFICIARY
(999980470) - BENEFICIARY
(999993953) - BENEFICIARY
(999992401) - BENEFICIARY
(968741911) - BENEFICIARY
(999980373) - BENEFICIARY
(950337810) - BENEFICIARY
(999995893) - BENEFICIARY
(999876486) - BENEFICIARY
(902999288) - BENEFICIARY
(885311241) - BENEFICIARY
(999613325) - BENEFICIARY
(897072006) - BENEFICIARY
(900064941) - BENEFICIARY
(882661880) - BENEFICIARY
(901837907) - BENEFICIARY
(952408275) - BENEFICIARY
(899161677) - BENEFICIARY
(999868144) - BENEFICIARY
(998322837) - BENEFICIARY
(972434119) - BENEFICIARY
(999990946) - BENEFICIARY
(999844573) - BENEFICIARY
(999654356) - BENEFICIARY
(999874546) - BENEFICIARY
(999864846) - BENEFICIARY
(999916741) - BENEFICIARY
(999613422) - BENEFICIARY
(999547074) - BENEFICIARY
(998627417) - BENEFICIARY
(999751938) - BENEFICIARY
(999977463) - BENEFICIARY

- + more

European Initiatives

Participants

(999655520) - COORDINATOR
(996058081) - BENEFICIARY
(999843409) - BENEFICIARY
(999645820) - BENEFICIARY
(999984059) - BENEFICIARY
(999981149) - BENEFICIARY
(999980470) - BENEFICIARY

(999864846) - BENEFICIARY
(999916741) - BENEFICIARY
(999613422) - BENEFICIARY
(999547074) - BENEFICIARY
(998627417) - BENEFICIARY
(999751938) - BENEFICIARY
(999977463) - BENEFICIARY

- + more

Fraunhofer ITWM

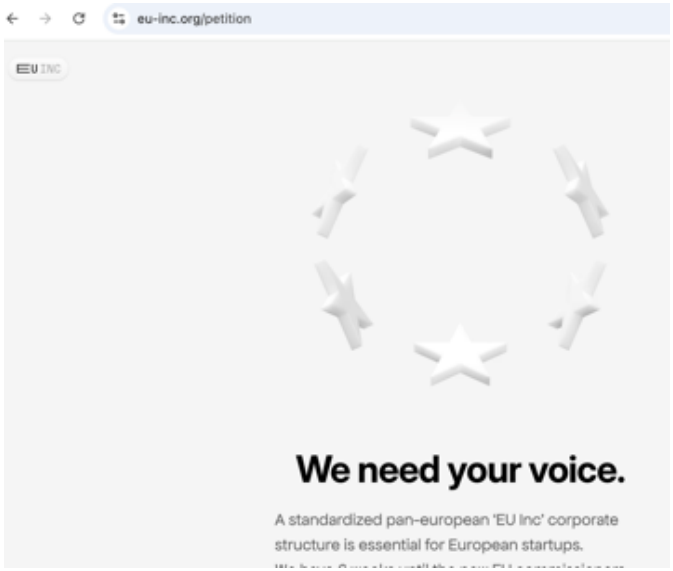
Markets

Entity	Population (2023)	Area (Km²)	Density (P/Km²)	World Population Share	GDP (trillion $ 2023)
China	**1,422,584,933**	9,388,211	152	17.6 %	$ 19.37
E.U.	**450,657,522**	3,997,838	113	5.6 %	$ 17.82
U.S.A.	**343,477,335**	9,147,420	38	4.2 %	$ 26.85
Japan	**124,370,947**	364,555	341	1.5 %	$ 4.41

https://www.worldometers.info/population/china-eu-usa-japan-comparison/

Freepik License: Free

Ideas?



Building a true single market

The multitude of countries and cultures in Europe is its unfair advantage. But because of that, our startup scene is fragmented. Legal and regulatory compliance is a burden, and cross-border collaboration is rare. While US angels invest coast-to-coast, European capital sticks within national borders. The results: stifled momentum, unrealized potential, and an artificial limit on our startups' chances of success.

Despite these challenges, Europe stands at the threshold of an exciting future. We're outflanking the US in new tech founders. We're a global leader in climate tech investment. Our deep tech scene is thriving, and European startups create jobs three times faster than other sectors.

We are at a unique inflection point of talent and technology. What we need is the framework to unleash it.

A single pan-European startup entity

Created under the 28th regime, this new entity will establish the institutional foundations for Europe's future competitiveness, yielding exponential returns. The new entity would:

- ✓ Standardize investment processes to enable genuine pan-European investments.
- ✓ Establish a unified employee stock options program to share startup success more widely.

+ commitment

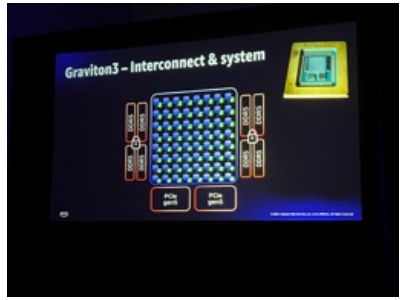
Is it possible to be succesfull?

- Open Source Hardware
- Open Source Software Stack
- Open Source EDA (?)
- Open Source IPs (?)

- Hardware Software Co-Design
- Domain Specific Acceleration

© Apple

© Alibaba



© AWS

© Google

Fraunhofer ITWM

The STX Accelerator for Efficient Simulations

HPC „Made in Germany“

- Open Source Hardware
- Open Source Software Stack
- PCIe x16 Gen5
- HBM Memory
- ~170mm^2 GF12LPP SoC
- Up tp 3x better energy efficiency than leading competitors
- funded by EU, BMBF and Fraunhofer

- Hardware Software Co-Design
- Domain Specific Acceleration
- SpinOff: UNEEC Systems GmbH

CC BY-NC-ND 2.0; flickr.com

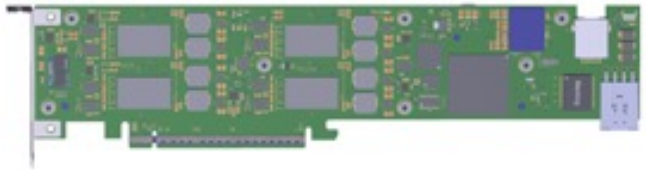
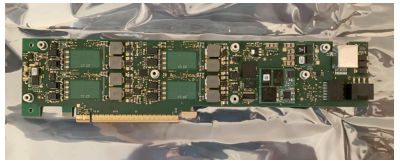


Abb.: STX PCIe Gen5 x16 Beschleunigerkarte



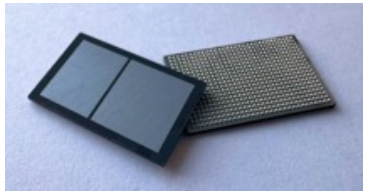
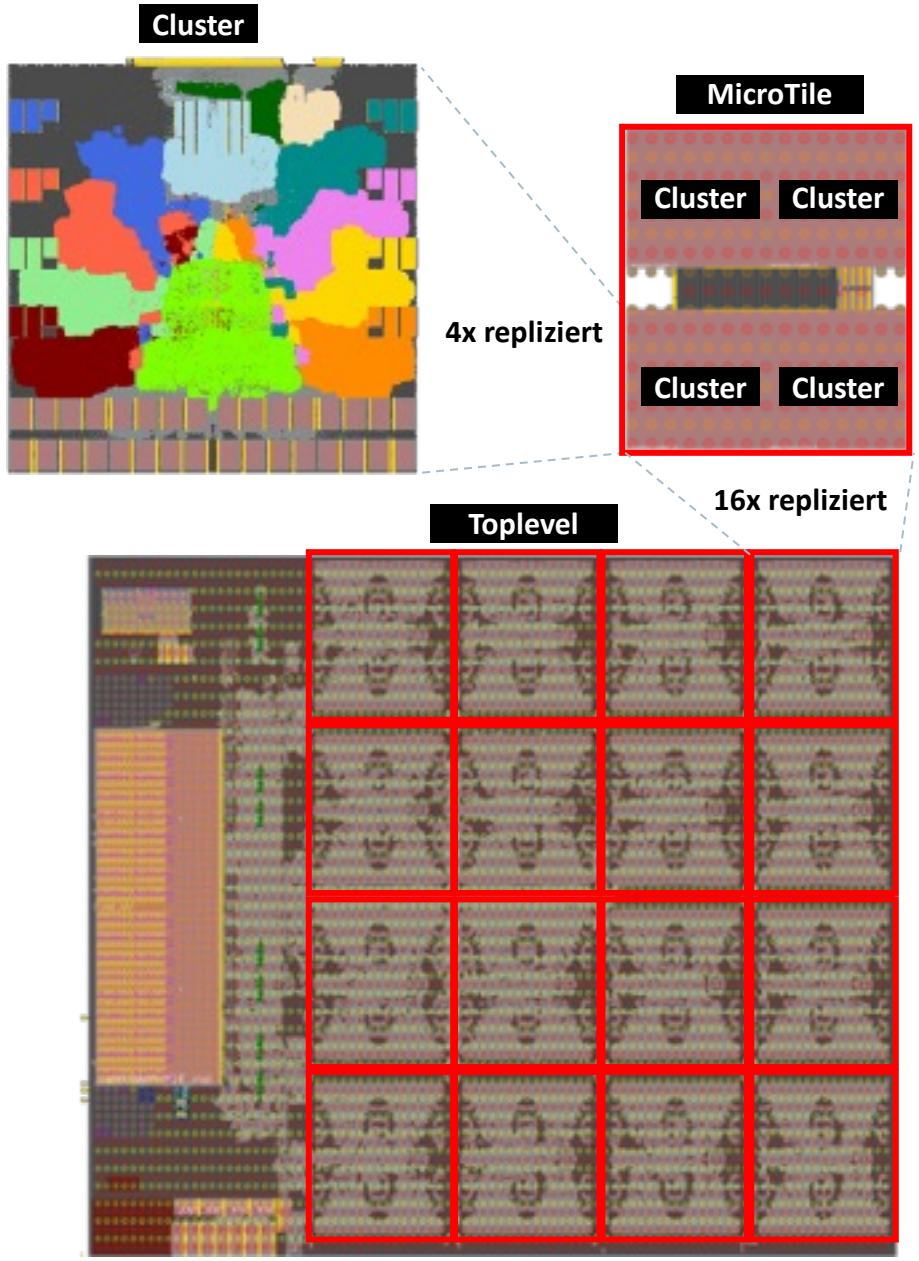


Abb.: Test Interposer AVT Test mit dummy HBM und STX Die

STXMOD ASIC Backend

Partitions

- 64x internal compute clusters with 8 RISC-V32 Cores + Stencil extensions and scratchpad memories
- 2D AXI4/5 network on chip
- Multistage network for HBM access
- PCIe Gen5 Ctrl
- CVA6 RISC-V64 with MMU zu run Linux
- Peripherals

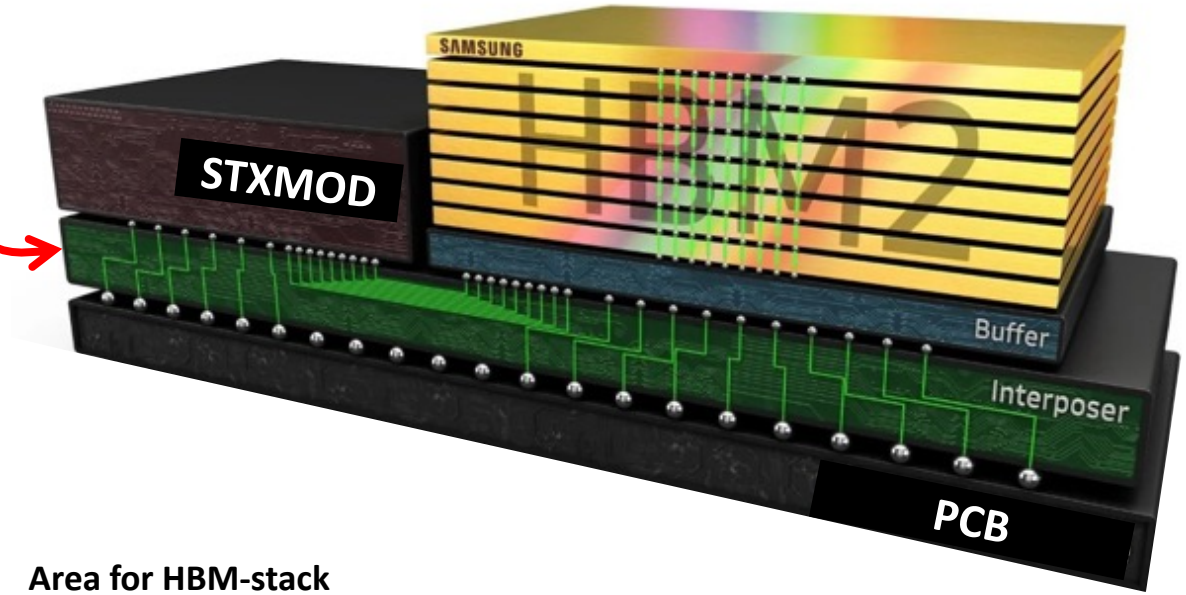


Interposer

2.5D integration

- Additional Silicon (reduced production process)
- Additional IC-design step incl. routing and bump placement
- Additional Signal- and power-integrity analysis and design
- Additional level of functional verification

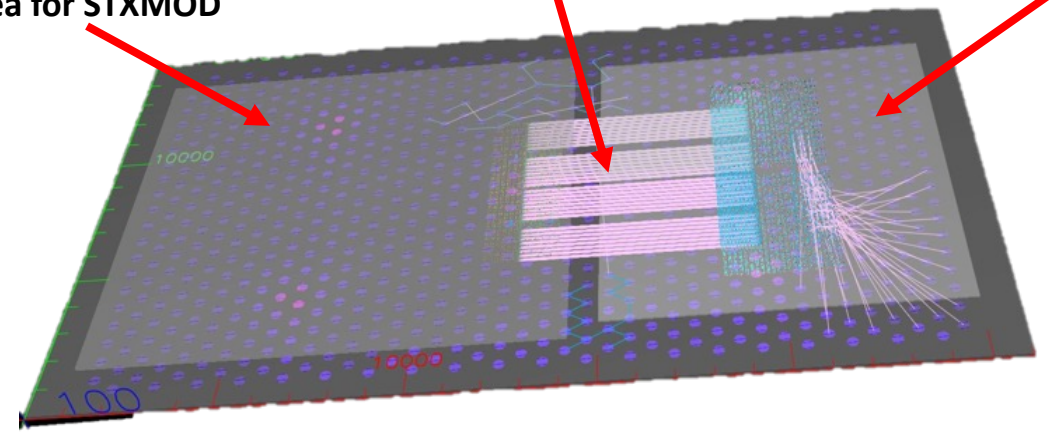
Interposer example



>2k connections between STXMOD and HBM-stack

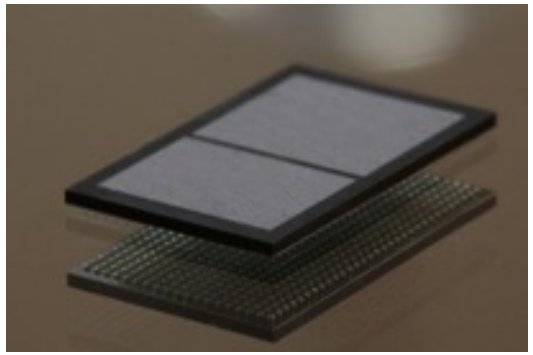
Area for STXMOD

Area for HBM-stack

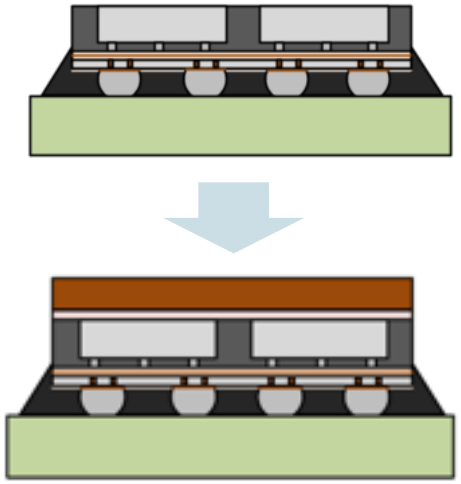


STX System Assembly

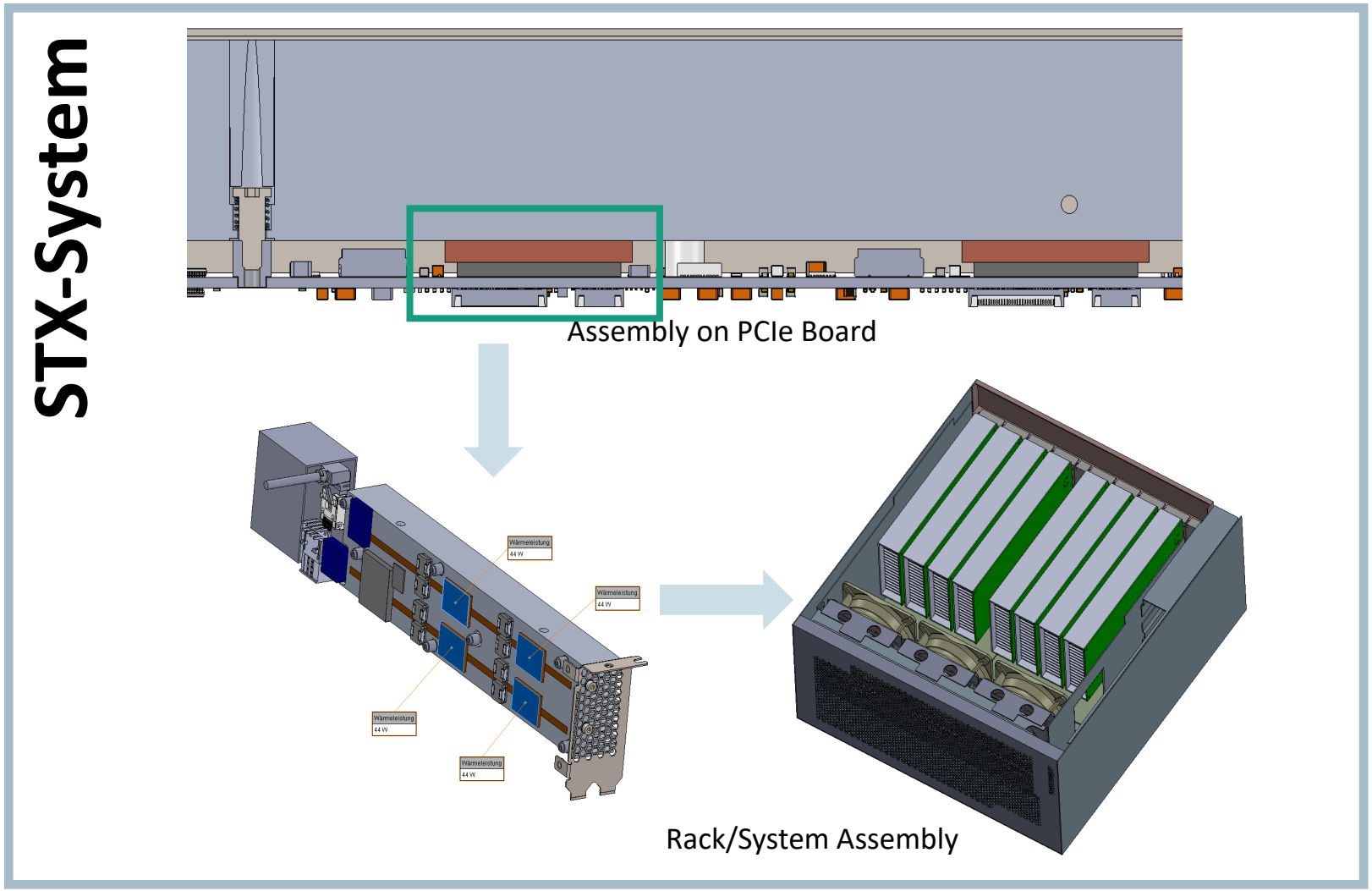
Heatsink and Assembly on Board



- Package with solderballs



- TIM and Heatsink

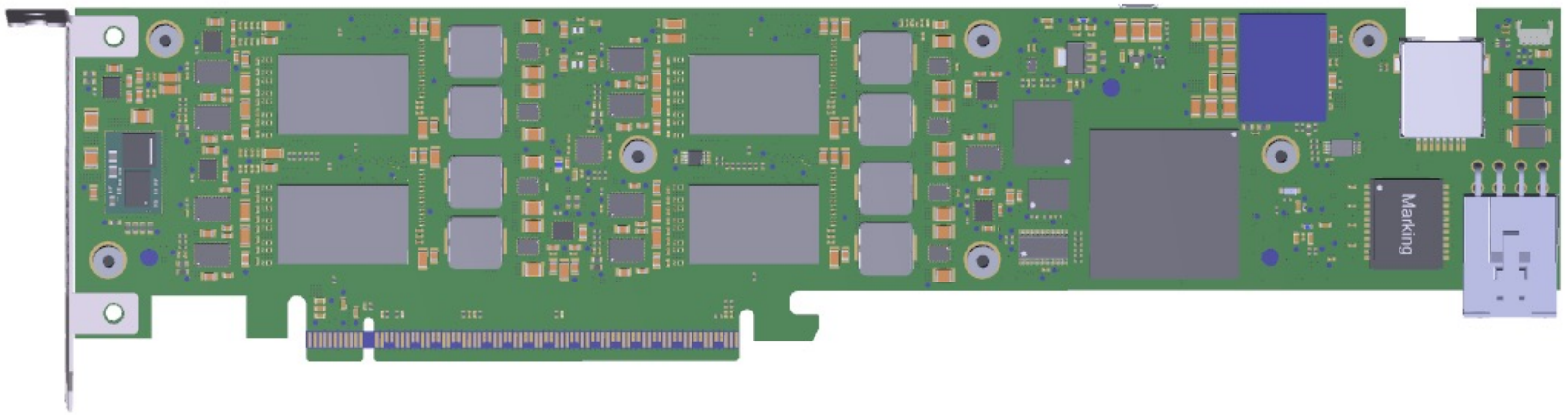


It does not get boring …

Further challenges:

- Software (on-Chip, PCI board, system, applications)
- Business Strategy (Markets, Domains, …)
- Venture Capital
- Generation2 (+AI) and 3

STX
Sovereignty?



- Open Source Hardware from Europe
- Frontend + Physical Design done in Europe (~170mm^2)
- Interposer Design & package assembly done in Germany
- but major parts not from the EU: IPs, EDA tools, Fab (GF12LPP)

Conclusions

1. Can Europe catch up in manufacturing high end chips and systems?

- Yes – but major efforts necessary
- Europe needs to commit to European technology

2. Can Europe be fully independent?

- No – but it can be key not to be fully dependent

3. Is it exciting?

- definitely

Contact: jens.krueger@itwm.fraunhofer.de

