

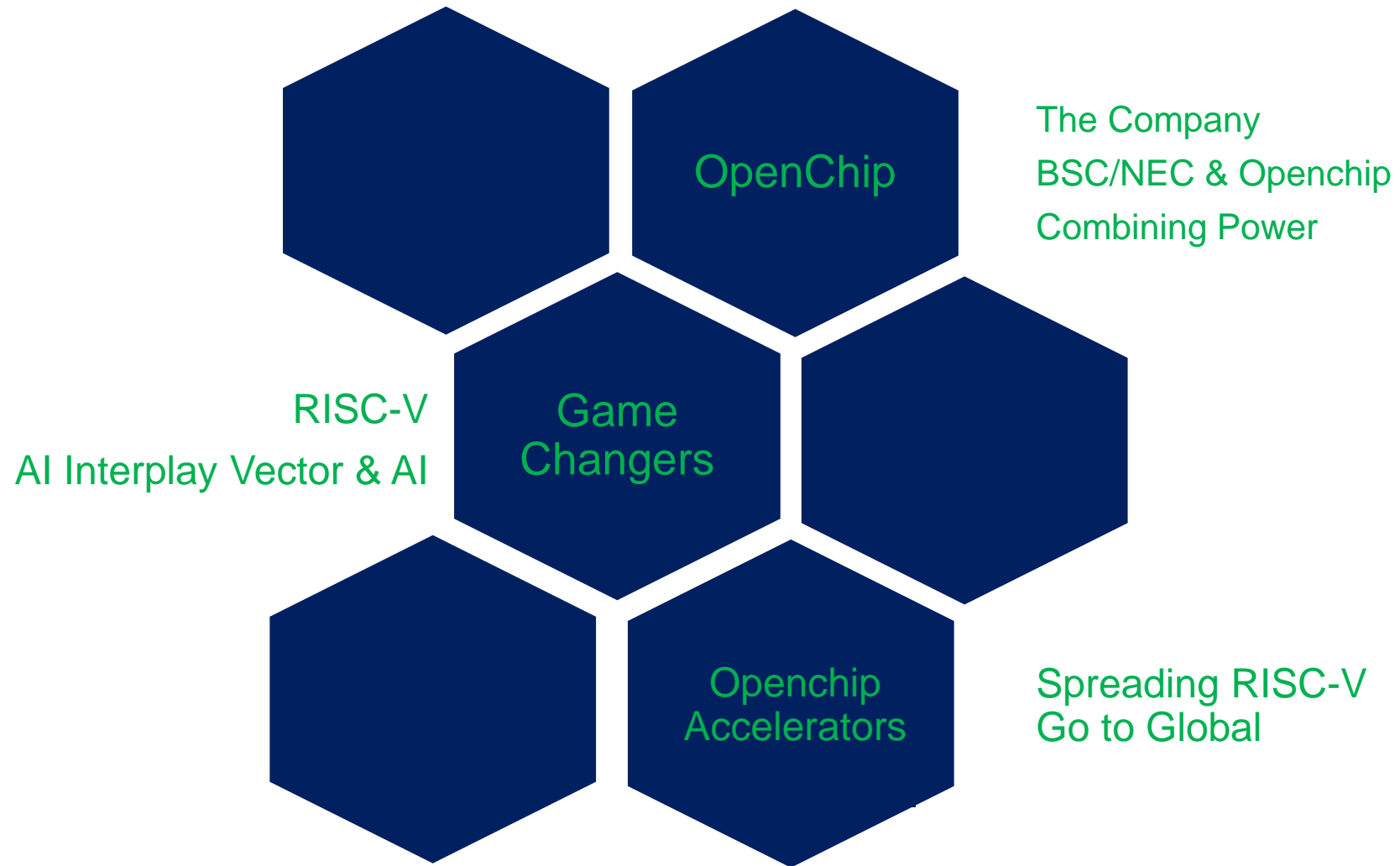


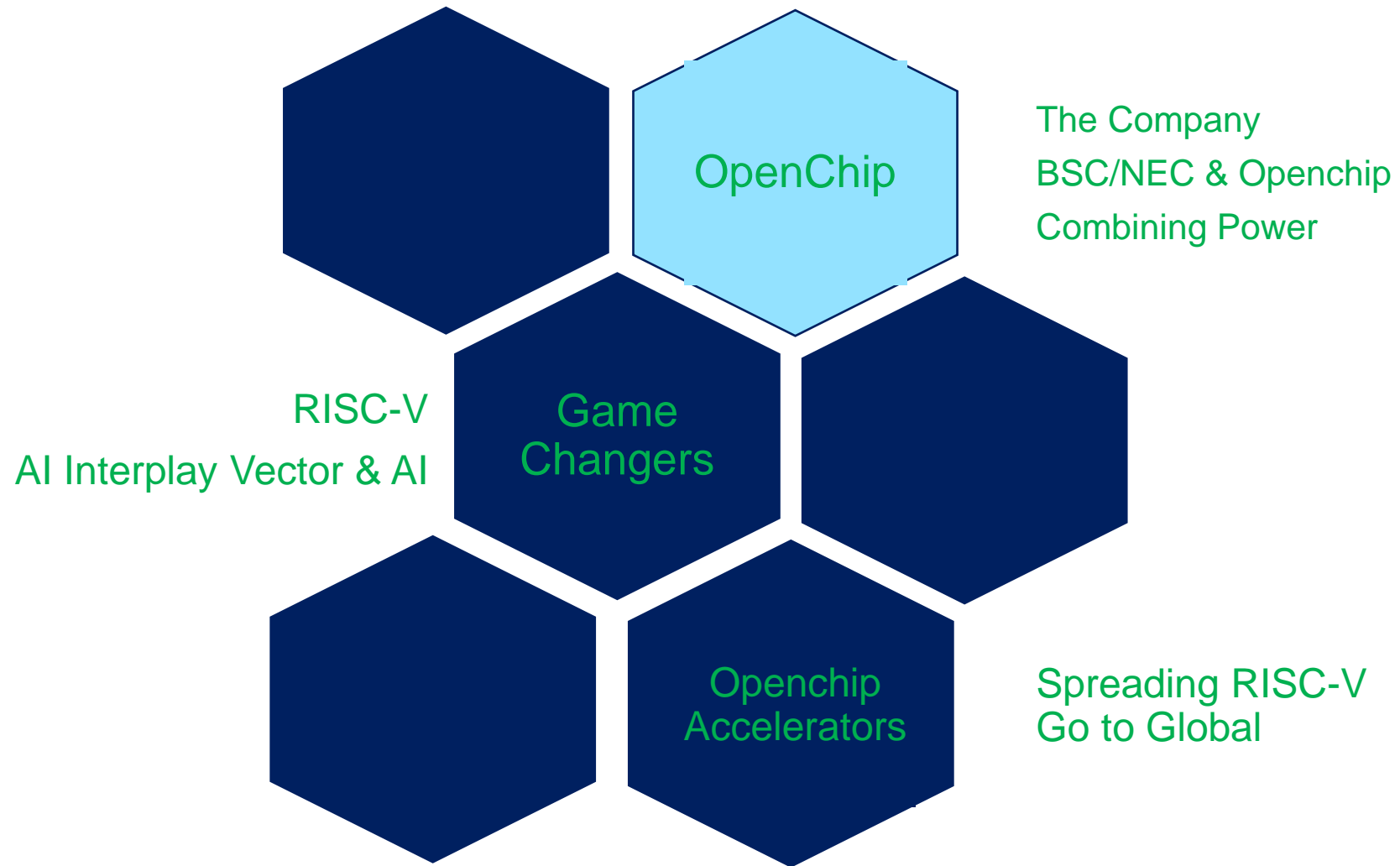
# Challenges in HPC, AI/ML/DL for RISC-V accelerator chips

Akira Tsukamoto

Openchip

Security and Ecosystem Architect Fellow







Openchip & Software Technologies, S.L.  
Established in 2024  
as a private company in Barcelona

IPCEI & DARE Programs



### Clients in 15 Countries - More than 1.500 high-tech projects

GTD works across some of the most demanding industries, providing software, systems, and services for safety, mission, and business-critical applications. We provide our clients with secure, reliable technology around the world. Out in space, our software orbits the Earth 24/7, 365 days a year. Closer to home, our software keeps aircraft flying high, makes vehicles safer, powers smart meter networks, and does much more. For over **34 years**, we've been transforming the way the world uses technology.



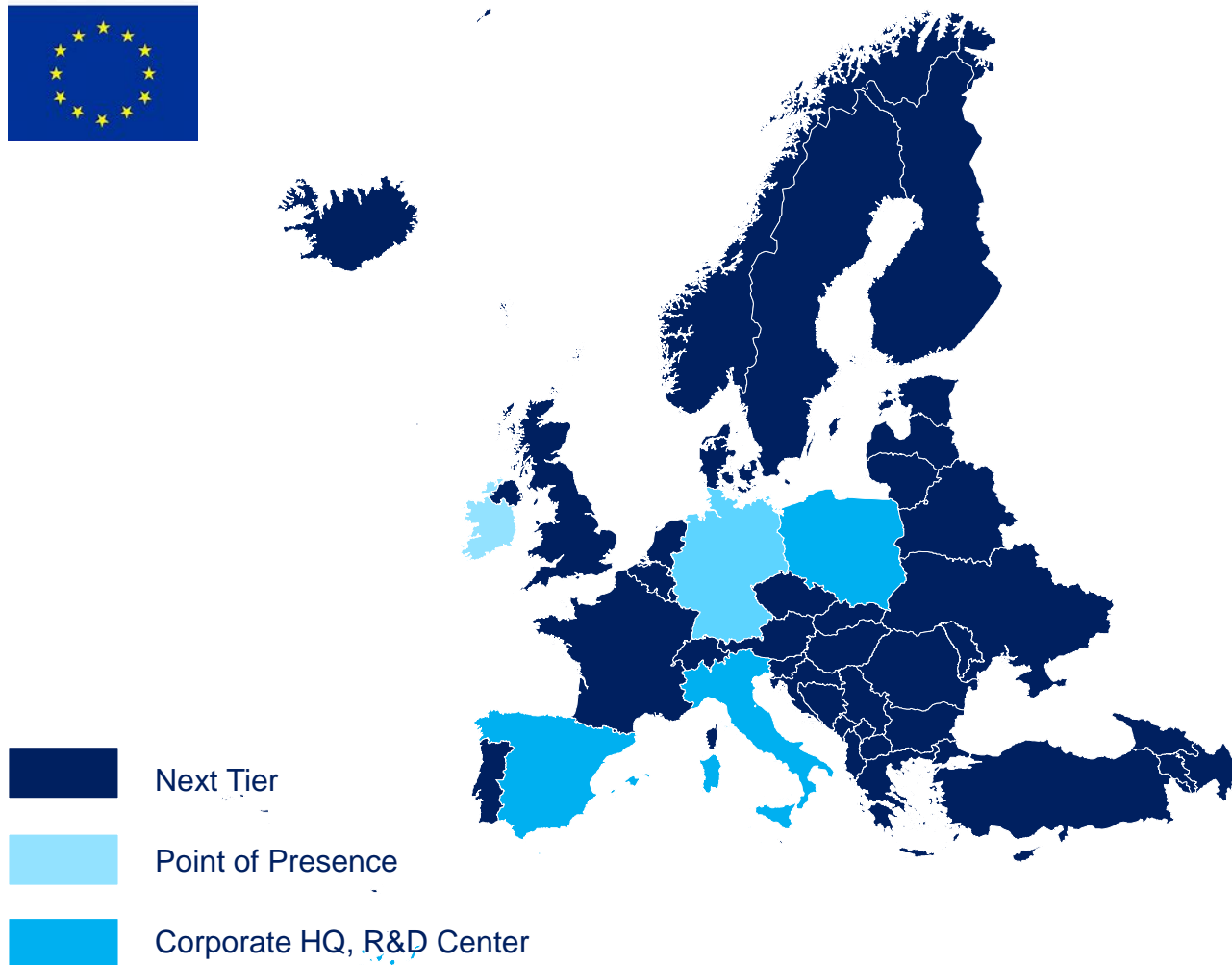
**EuroHPC**  
Joint Undertaking



Una manera de hacer Europa



# Openchip Strategy to Develop a Global Market



## Strategy for European Union Expansion

- Openchip is a Spanish Project truly multinational and multicultural: Europe is in our DNA.
- +120 employees from 15 nationalities: Spain, Italy, Poland, Germany, France, UK, USA, Romania, Armenia, India, Pakistan, Ukraine, Japan, ...
- We are determined to promote the collaboration within EU countries: Hiring a talented team, distributing R&D facilities and having a point of presence next to our priority customers.
- Company HQ is located in Barcelona, Spain, a global hub for business.
- We aim to lead the efforts of **European Union for Digital Sovereignty** and adoption of **RISC-V as the ISA for Supercomputing and Artificial Intelligence**.

## Chief Executive Office & C-Level Staff



### Cesc Guim – CEO

PhD HPC – 60 publications +530 Patents  
5 Years at Barcelona Supercomputing Center  
17 Years Silicon and System Design at Intel



### Marc Fernández - CFO

Bachelor in BA, MBA ESADE, Postgraduate degree in Law  
+25 years in Finance, Consulting and General Management



### Gaspar Mora - CTO

PhD Computer Architecture - 14 Patents  
15 Years Silicon and System Architecture at Intel (Xeon, HPC's Intel Omnipath) and Nvidia (GPU memory system)



### Ingacio Astilleros – CSO

Bachelor Physics, MBA IE. Master in Economy & Innovation. 4 patents.  
+25 years Sales, BizDev. & Management.  
INDRA, Sun Microsystems, Huawei, Intel.



### Tommaso Vali – Chief HW Eng.

Master Degree in Electronic engineering – 15 publications, 263 Patents, >30 Tapeouts, 35 Years of Silicon and System Design at Texas Instruments(10Y), Micron and Intel (25Y)



### Violante Moschiano – Chief HW Arch

Master Degree in Electronic Engineering  
+350 Patents, 10 publications, ISSCC ITC,  
20 Years of DTCD, System Design & Architecture . Micron and Intel



### Edgar González – Chief SW & AI Off.

PhD AI – 27 publications, 4 patents  
20 Years AI, ML, and NLP  
12 Years as production SW team lead (Google Research, Cloud & Assistant)



### Ivan Rodero – Chief of Innovation

PhD HPC – over 180 publications  
5 Years at Barcelona Supercomputing Center  
20 Years High-Performance Computing and world-class Advanced Cyberinfrastructure

## VPs, Fellows and Senior Fellows



### Akira Tsukamoto – Security and Ecosystem Architect Fellow

Master of Science Degree in Computer Science  
RISC-V Ambassador, Co-Chair at IETF  
40 years of Computer experience on B.E., SCE, NEC, AIST



### Erich Focht – HPC SW Chief Architect & Fellow

PhD HPC – 70 publications  
27 years at NEC R&D, contributing to tens of Top500 HPC systems, system designs and algorithms.



### Product Chief Officer -- VP

EMBA from London Business School and a Meng in Telcom. University of Cantabria.  
25 years in Semiconductor industry industries at Alcatel Bel, Lucent Microelectronics, Nokia,



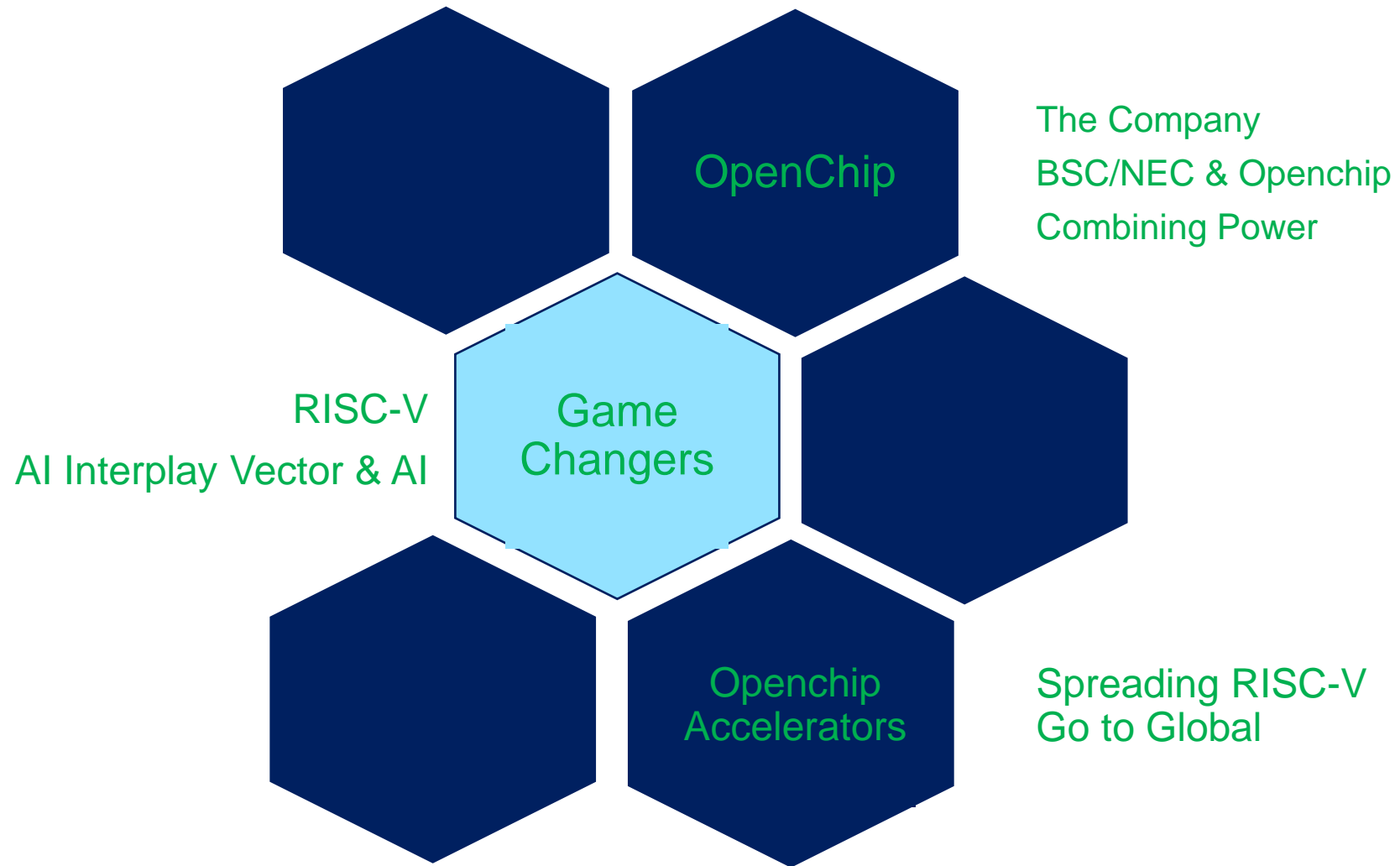
### Satoru Tagaya – HPC Chief Architect & Senior Fellow

MS in Computer Science 10 Patents  
30 Years of Vector Computer System Architecture and Microarchitecture (SX series and SX Aurora) at NEC Corporation in Tokyo.



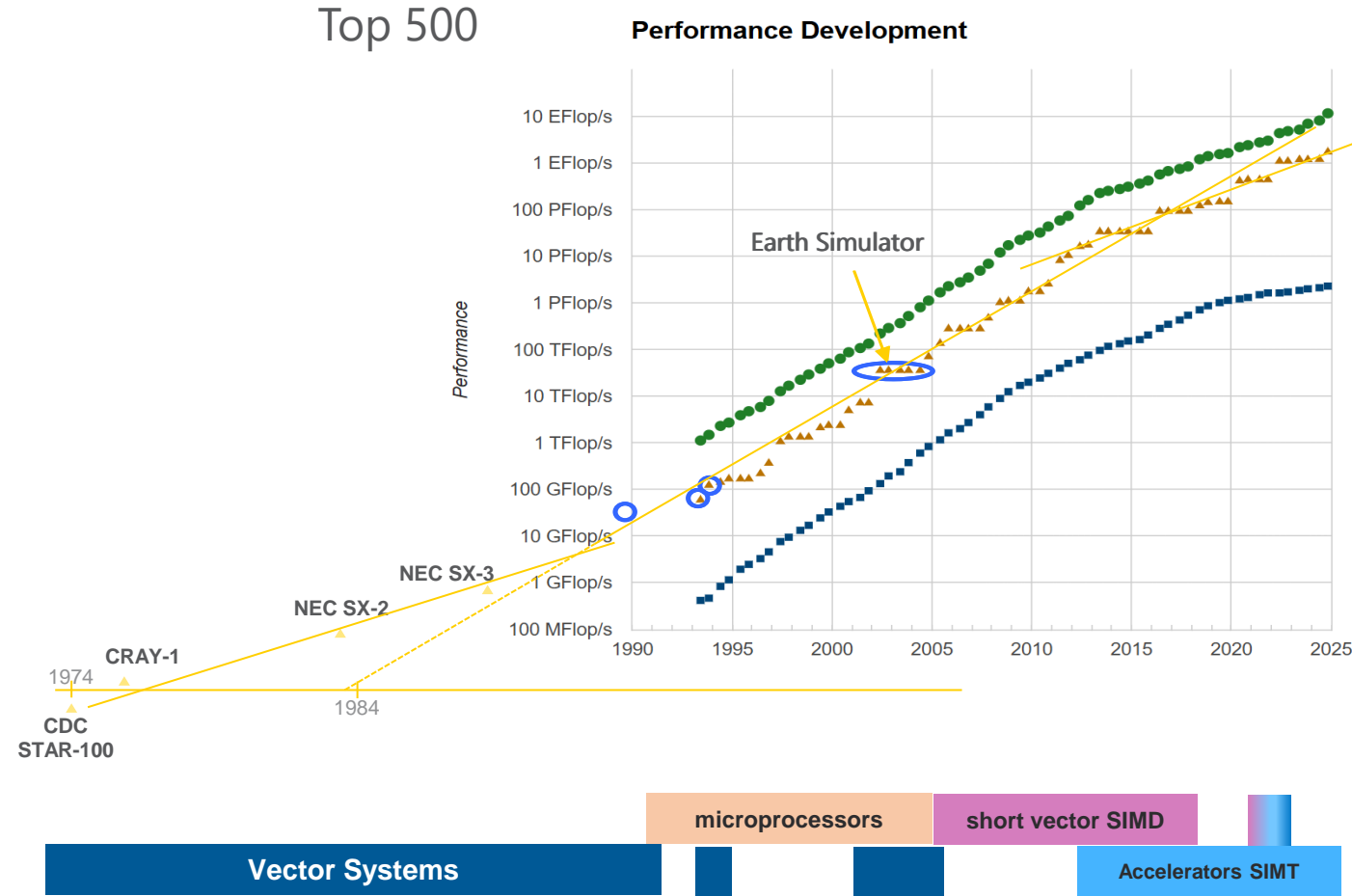
### Therese Jamaa, VP Industry Alliances (Marketing, Communications, RSC)

Information Security, Business Administration, Marketing, Business Intelligence. Coach.  
Schlumberger, Vodafone, Qualcomm, GSMA, Huawei. +25 years of experience



# HISTORY OF SUPERCOMPUTING

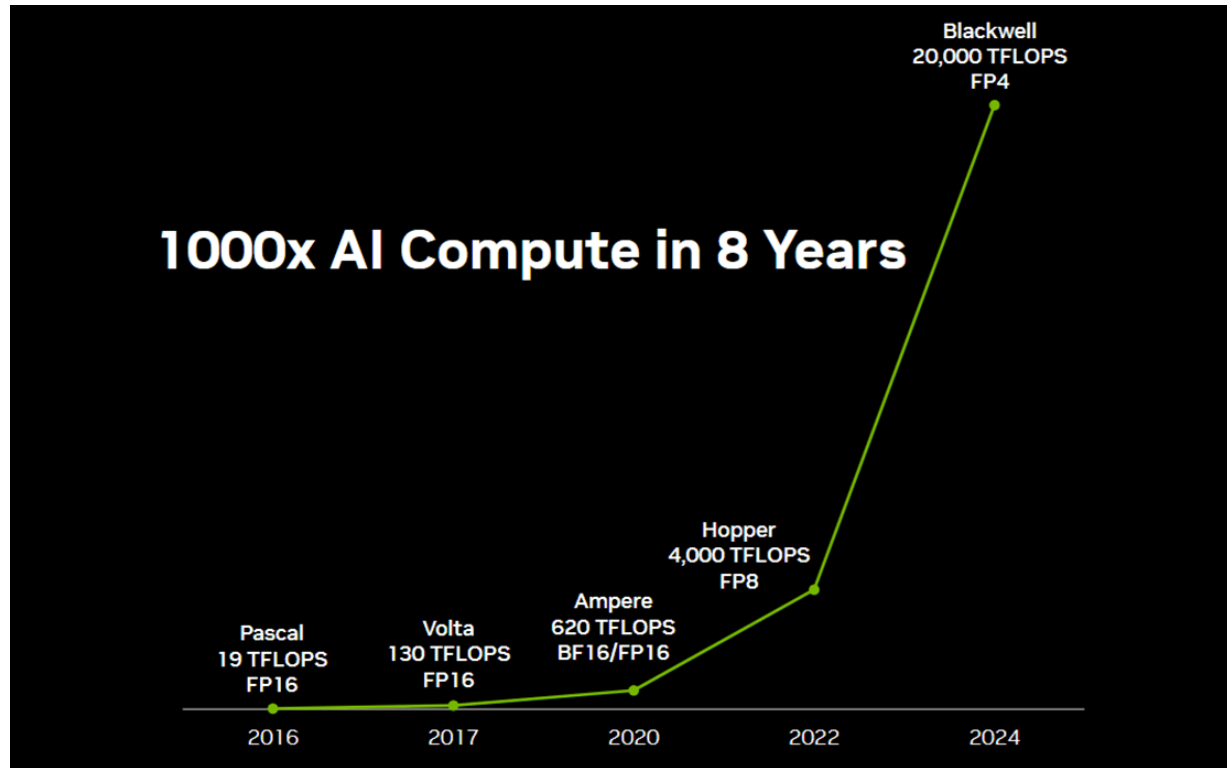
Slide courtesy of Dr. Erich Focht



- ✓ Vector systems were the general purpose supercomputers from the 1970s to beginning of 2000s.
- ✓ Dennard scaling and Moore's law pushed general purpose microprocessors to the top for ~15-20 years.
- ✓ Short vector SIMD: return of vector in general purpose processors: until now.
- ✓ End of Moore's law: shift to accelerators, SIMD (Xeon PHI) and SIMT (GPUs) + increasingly higher power.
- ✓ SIMT can be mapped to SIMD (vector)



# END OF MOORE'S LAW VS. AI MARKETING MACHINE



- ✓ Progress comes from
- Technology node
  - Tensor units
  - Numeric precision reduction
  - Structural (fake) sparsity
  - Blackwell: block scaled numeric formats

# LONG VECTOR SYSTEMS EVOLUTION

Slide courtesy of Dr. Erich Focht




**CDC STAR-100**  
First commercial vector  
1974 – 100 MFLOPS  
Streaming from memory  
to memory  
16 Bytes/FLOP



**CRAY-1**  
Commercially successful vector!  
1976 – 160 MFLOPS

- Vector registers!
- Vector length: 64 elements
- 16 Bytes/FLOP
- Excellent scalar performance
- Pipelined vector ops
- Chained operations

History of SX Vector Supercomputer	<b>SX-2</b> 1983		Technology: Bipolar CPU Frequency: 166 MHz CPU Performance: 1.3 GFlops CPU Memory Bandwidth: 10.7 GB/sec
	<b>SX-3</b> 1989		Technology: Bipolar CPU Frequency: 340 MHz CPU Performance: 5.5 GFlops CPU Memory Bandwidth: 12.8 GB/sec
	<b>SX-4</b> 1994		Technology: 350 nm CPU Frequency: 125 MHz CPU Performance: 2.0 GFlops CPU Memory Bandwidth: 16.0 GB/sec
	<b>SX-5</b> 1998		Technology: 250 nm CPU Frequency: 250 MHz CPU Performance: 8.0 GFlops CPU Memory Bandwidth: 64.0 GB/sec
	<b>SX-6</b> 2001		Technology: 150 nm CPU Frequency: 500 MHz CPU Performance: 8.0 GFlops CPU Memory Bandwidth: 32.0 GB/sec
	<b>SX-7</b> 2002		Technology: 150 nm CPU Frequency: 552 MHz CPU Performance: 8.8 GFlops CPU Memory Bandwidth: 35.3 GB/sec
	<b>SX-8</b> 2004		Technology: 90 nm CPU Frequency: 1.0 GHz CPU Performance: 16.0 GFlops CPU Memory Bandwidth: 64.0 GB/sec
	<b>SX-9</b> 2007		Technology: 65 nm CPU Frequency: 3.2 GHz CPU Performance: 102.4 GFlops CPU Memory Bandwidth: 256.0 GB/sec
	<b>SX-ACE®</b> 2013		Technology: 28 nm CPU Frequency: 1.0 GHz CPU Performance: 256.0 GFlops CPU Memory Bandwidth: 256.0 GB/sec

NEC Innovations:  
Multi-lane pipelines, vector cache, multi-core, ...



**NEC SX-Aurora Tsubasa**  
First vector accelerator card  
2018 – 2.45 TFLOPS (DP)

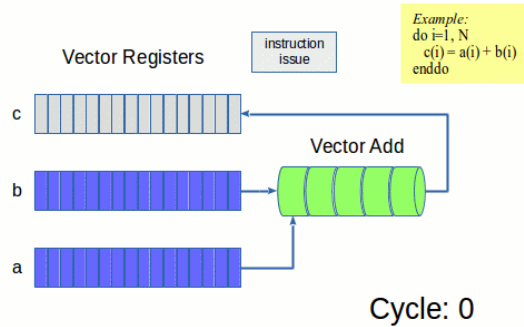
- 16nm
- 8 – 10 cores
- 6 x HBM2 (48 GB)
- 1.2 – 1.55 TB/s mem BW
- VL = 256 DP elements
- 32 vector lanes

2022 – 4.9 TFLOPS (DP)

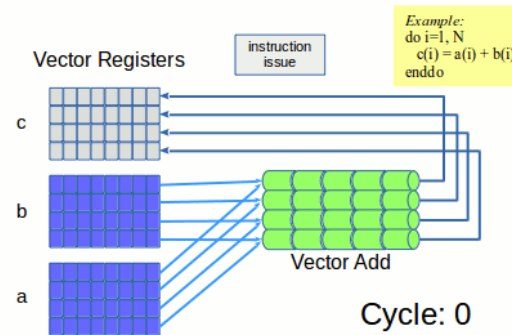
- 7nm
- 16 cores
- 6 x HBM2e (96 GB)
- 2.4 TB/s mem BW

# FROM VECTOR TO SIMD AND BACK

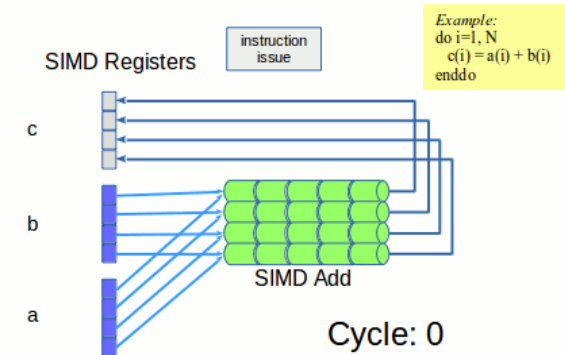
Slide courtesy of Dr. Erich Focht



CRAY-1 style  
vector pipeline  
**Long Vector**



NEC SX-4 style  
parallel vector pipeline  
**Long Vector**



AVX2 style  
SIMD pipeline  
**Short Vector**

- ✓ Classical long vectors are excellent in handling data level parallelism (DLP)
  - One instruction keeps the pipeline(s) busy for long time
  - Hide latency of memory access through long vectors, temporal execution
  - Less sensitive to occasional latency increase (like NoC congestion)
  - Parallelism explicit, in the ISA!

- ✓ SIMD & short vectors
  - Cache/prefetch/OoO to hide latency
  - More difficult to keep pipelines full
  - Moderate DLP handling
  - Sensitive to latency increase
  - Good with short vectors

# VECTOR IS THE BETTER ISA



RISC-V decided for a VECTOR ISA

- Not a SIMD ISA!
- **Variable vector length**
  - Vector length register like SX Aurora
- Vector register size not prescribed
  - VLEN = VREG size in bits, **is not fixed!**
  - Left to the implementation
  - Must be power of 2
  - $ELEN \geq 8$  max element size in bits
  - $ELEN \leq VLEN \leq 65536$

***Code can be VLEN agnostic!***

*Runs on any implementation of RISC-V Vector.*

<https://github.com/riscv/riscv-v-spec>

NOTE: ARM SVE (Scalable Vector Extension) allows implementing 128-2048bit SIMD units.

- ✓ Can implement SIMD, short vector, long vector with same ISA!
- ✓ Data parallelism → Loop vectorization
- ✓ Code ported to RISC-V long vector runs on short vector cores, too.
- ✓ Investment is protected!

## RISC-V Vector Implementations:

- Long vectors: Vitruvius, Ara, (Hwacha)
- Short vectors: Spatz, Saturn, commercial



# Processing Vector and GPGPU with RISC-V



## **Vector: Small number of decently parallel operations**

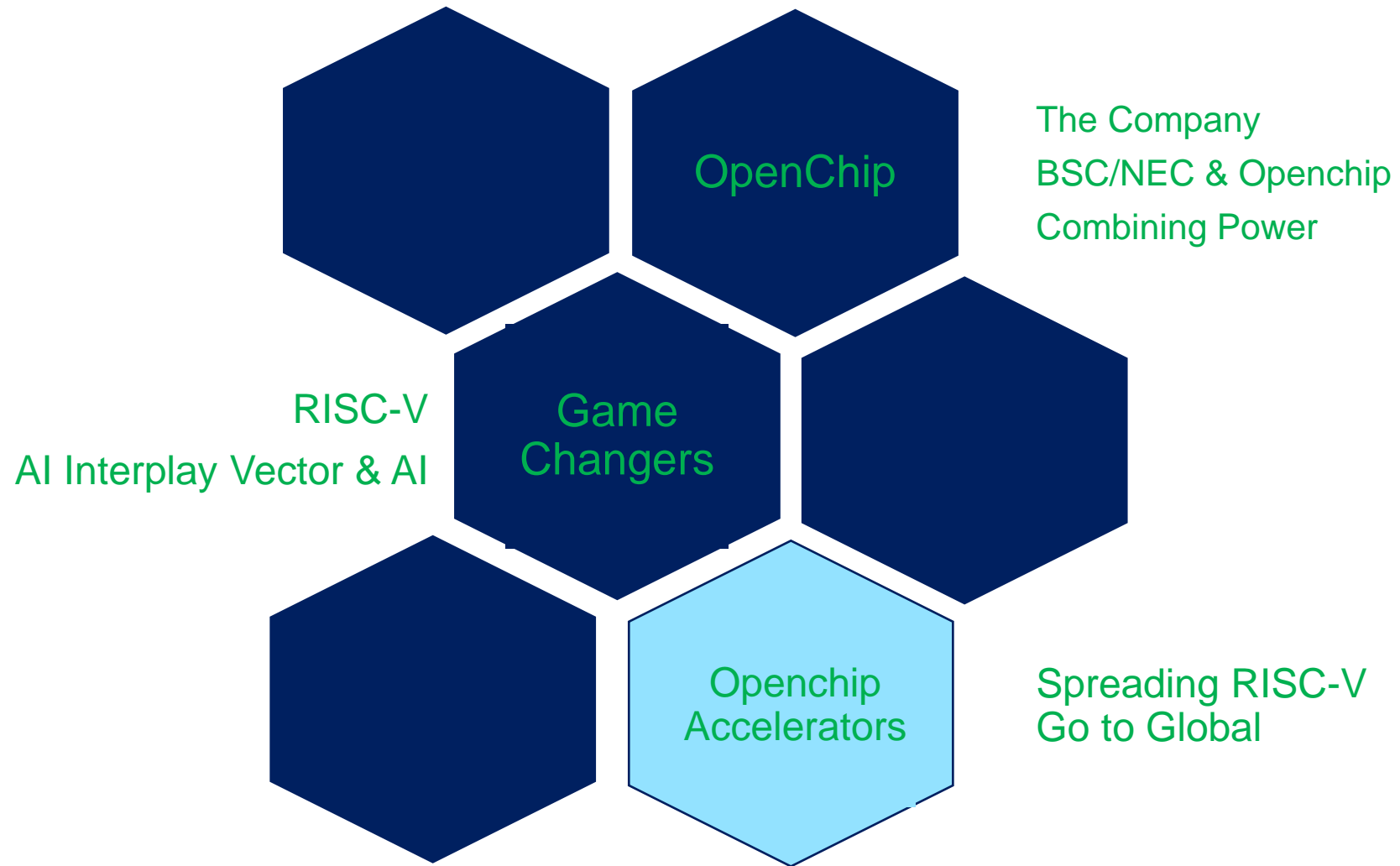
- Flexible use model
- More power efficient:
  - Single control for certain volume of computation



## **GPU: A huge number of single threads**

- Generally Inflexible use model
- Less power efficient:
  - Control for each computation thread required





## What do we deliver?

## HPC Applications

## OpenChip provides:

Application Optimization Services  
Modeling and system design

Software  
(Management, SDK, etc.)

## OpenChip provides:

Baseline SDKS  
Advanced Management and Observability tools



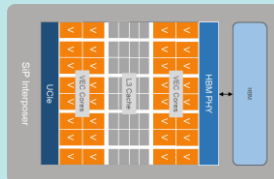
## Host, HW Platform & Chassis

## OpenChip provides:

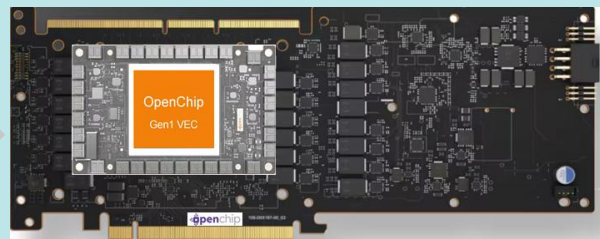
Reference Desings

List of OEMS/ODMS validated implementations

## Silicon Design



## PCB (PCIE etc.)



## OpenChip provides:

Silicon & Accelerators (e.g. PCIe)  
HPC Vector and AI Accelerators

## HPC

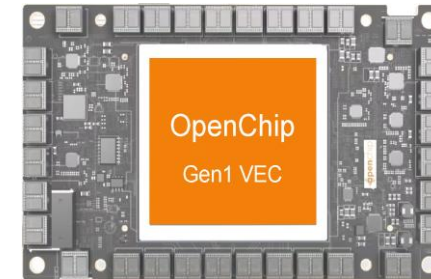
### Vector & AI

Our strategic relationship with Barcelona Supercomputing Center and other Industry leaders like the Japanese NEC and other ecosystem partner, will help us build SoC's for the most demanding environments of supercomputing, focusing on Vector processing and Artificial Intelligence.

## HPC-DC-Cloud

### Secure & Efficient AI

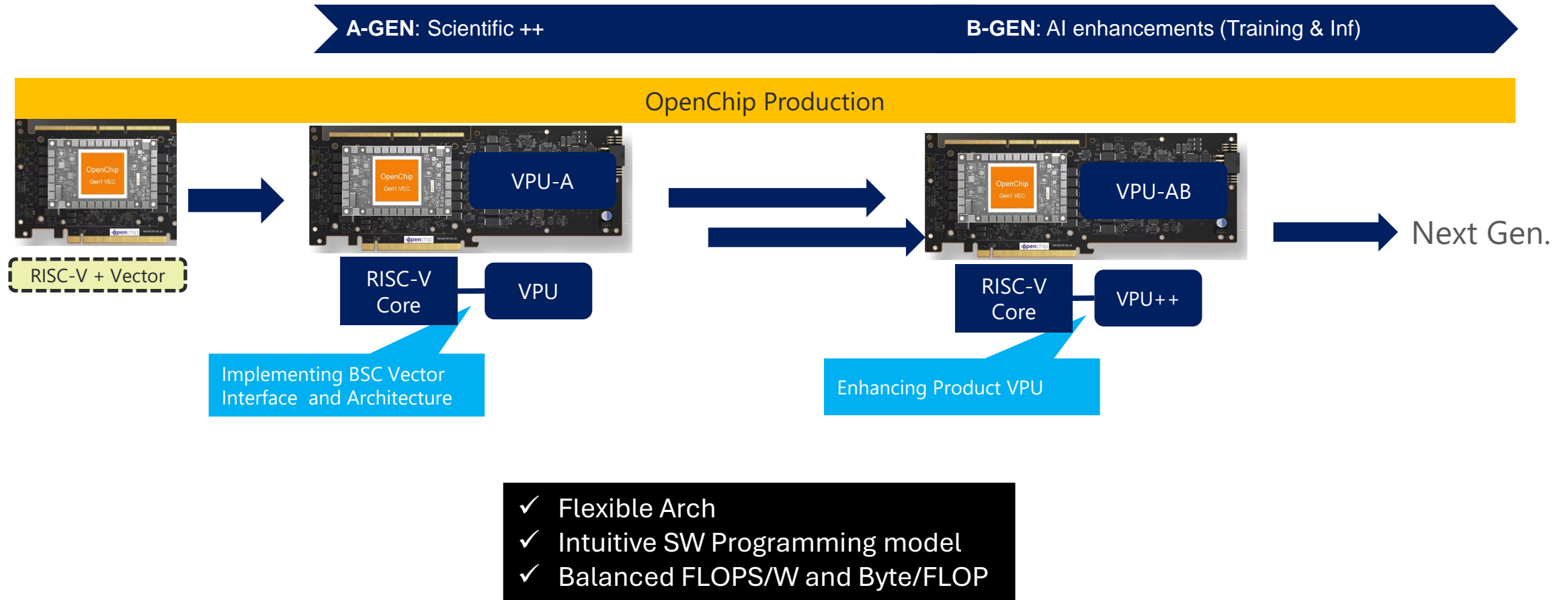
Better performance per watt. Drastic reduction of carbon footprint. Secure by design, our SoC's will provide a full stack of technologies to secure the cloud and guarantee the privacy of the data and the digital sovereignty. Artificial Intelligence LLM (GPT) modes: Inference and training.



Openchip's ability to combine in a SOC multiple chiplets with **open interconnection** (UCIe) can help to provide flexibility and shorter time to market of innovation. The **adoption of RISC-V** cores based on an instruction-set free of license can help to reduce costs, provide energy savings and access a wide ecosystem of applications that is being massively adopted in multiple industries. Openchip, being **AI centric**, will design their products by combining RISC-V with AI accelerators, Security Processors, Vector Processors and other accelerated chiplets. Born with an open philosophy, Openchip is developing strategic agreements with global silicon industry leaders and will develop in-house some accelerators in cooperation with Spanish & European R&D centers and Academia.

# PRODUCTION STRATEGY

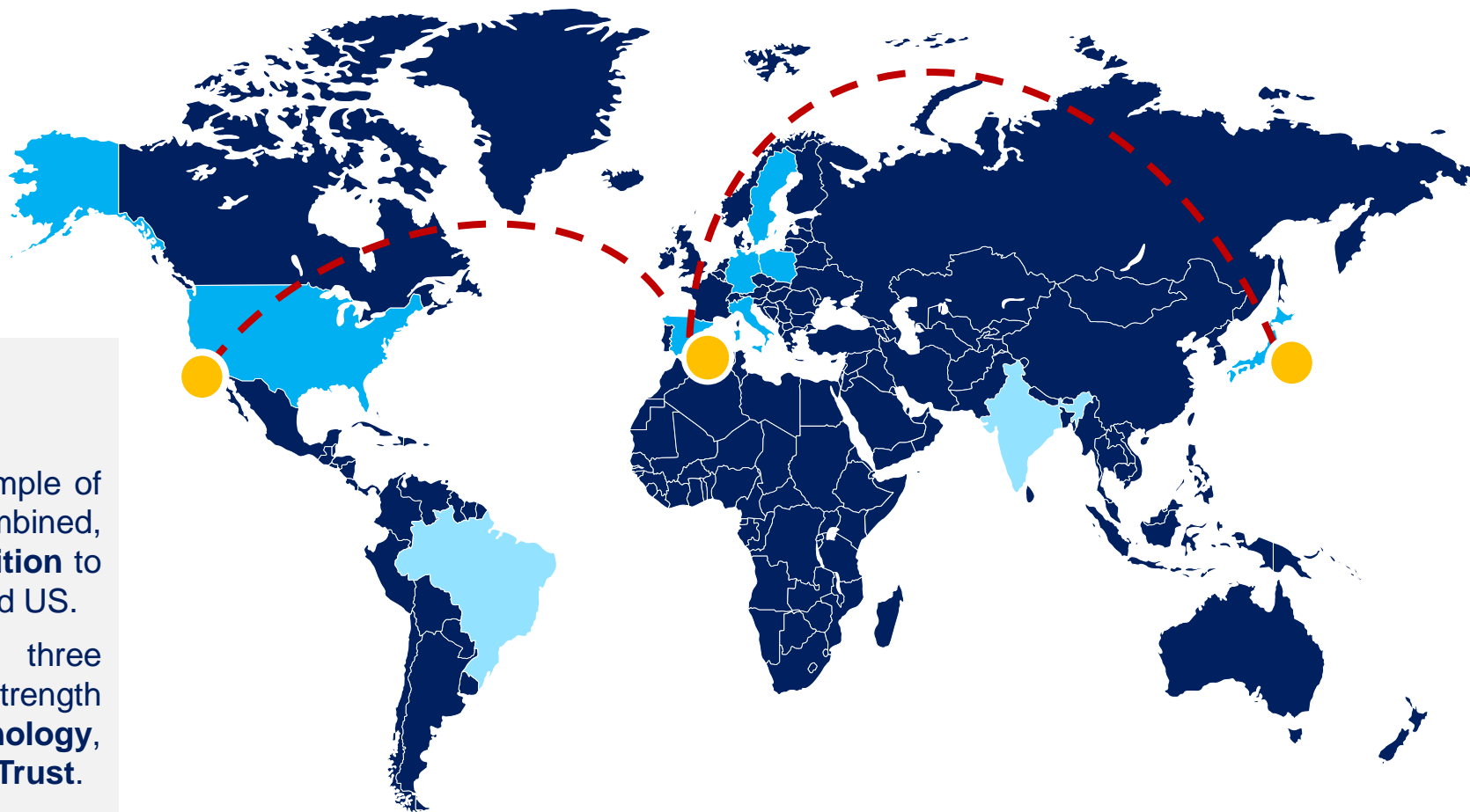
COMBINING ADVANCED R&D WITH HIGH-END PRODUCTS



# Openchip: Cultivating a Global Iron Triangle

## On becoming a Global Leader

- Openchip & NEC represent an example of competitive technologies that, combined, could bring a **unique value proposition** to address multiple markets: EU, JP and US.
- The alliance between these three innovative companies is, in fact, a strength that beyond **Innovation** and **Technology**, brings the values of **Resilience** and **Trust**.





Thanks!