Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

HPC Perspectives and Challenges for Europe

Prof. Mateo Valero
BSC Director

02/2019

Instituto Politécnico Nacional

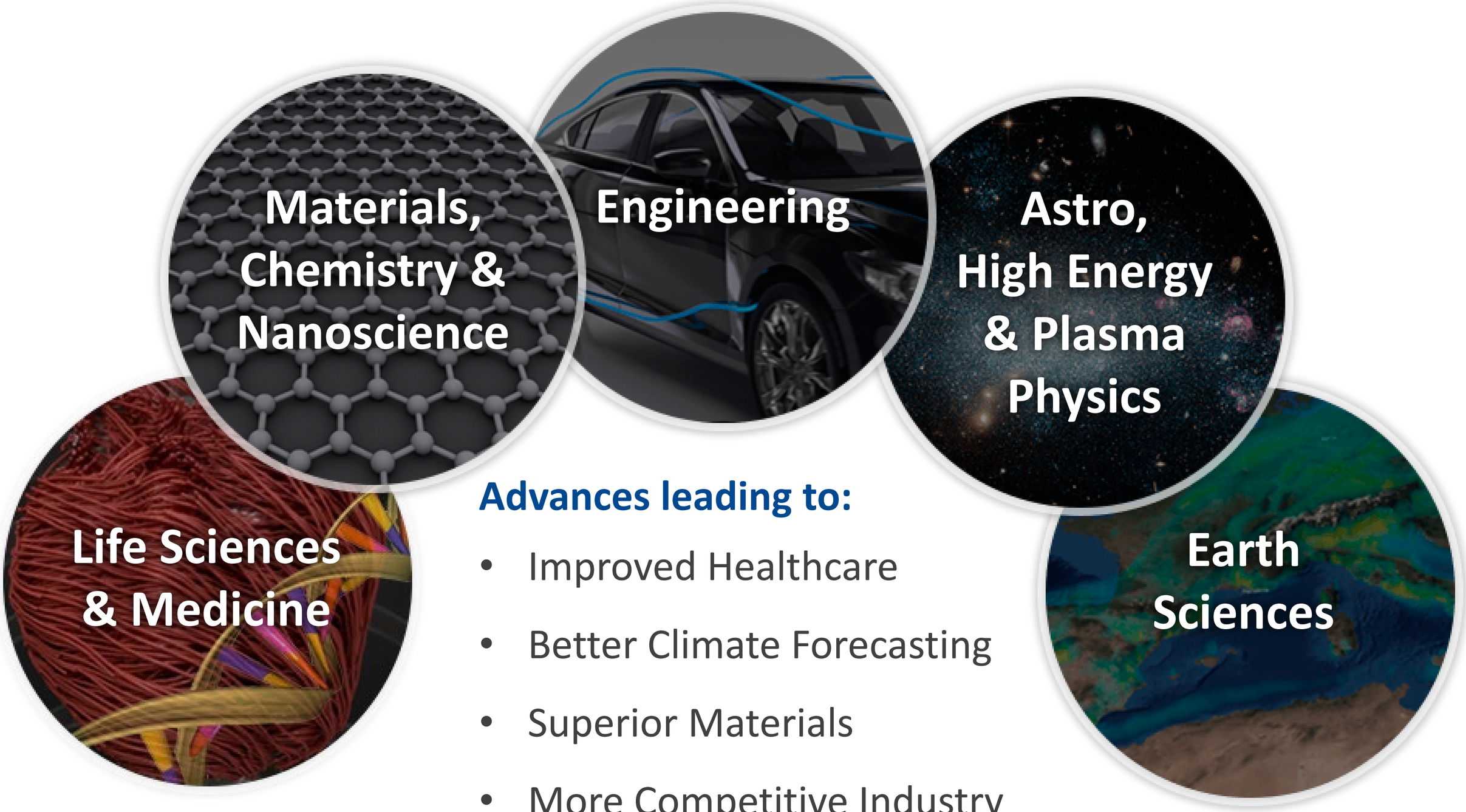
Centro de Investigación en Computación
Instituto Politécnico Nacional

The Evolution of the Research Paradigm

Numerical Simulation and Big Data Analysis

- Reduce expense
- Avoid suffering
- Help to build knowledge where experiments are impossible or not affordable

HPC: An enabler for all scientific fields



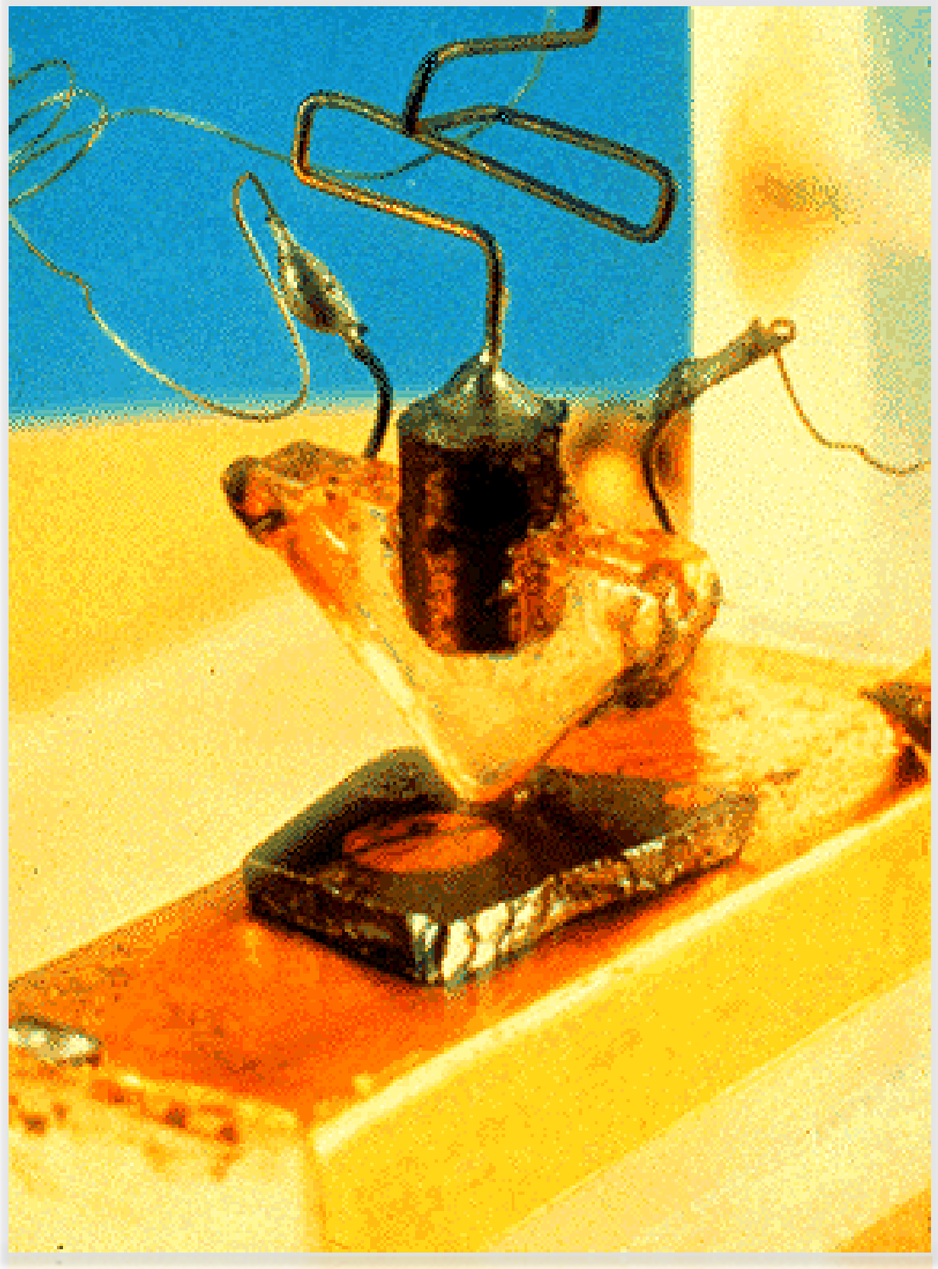
Advances leading to:

- Improved Healthcare
- Better Climate Forecasting
- Superior Materials
- More Competitive Industry

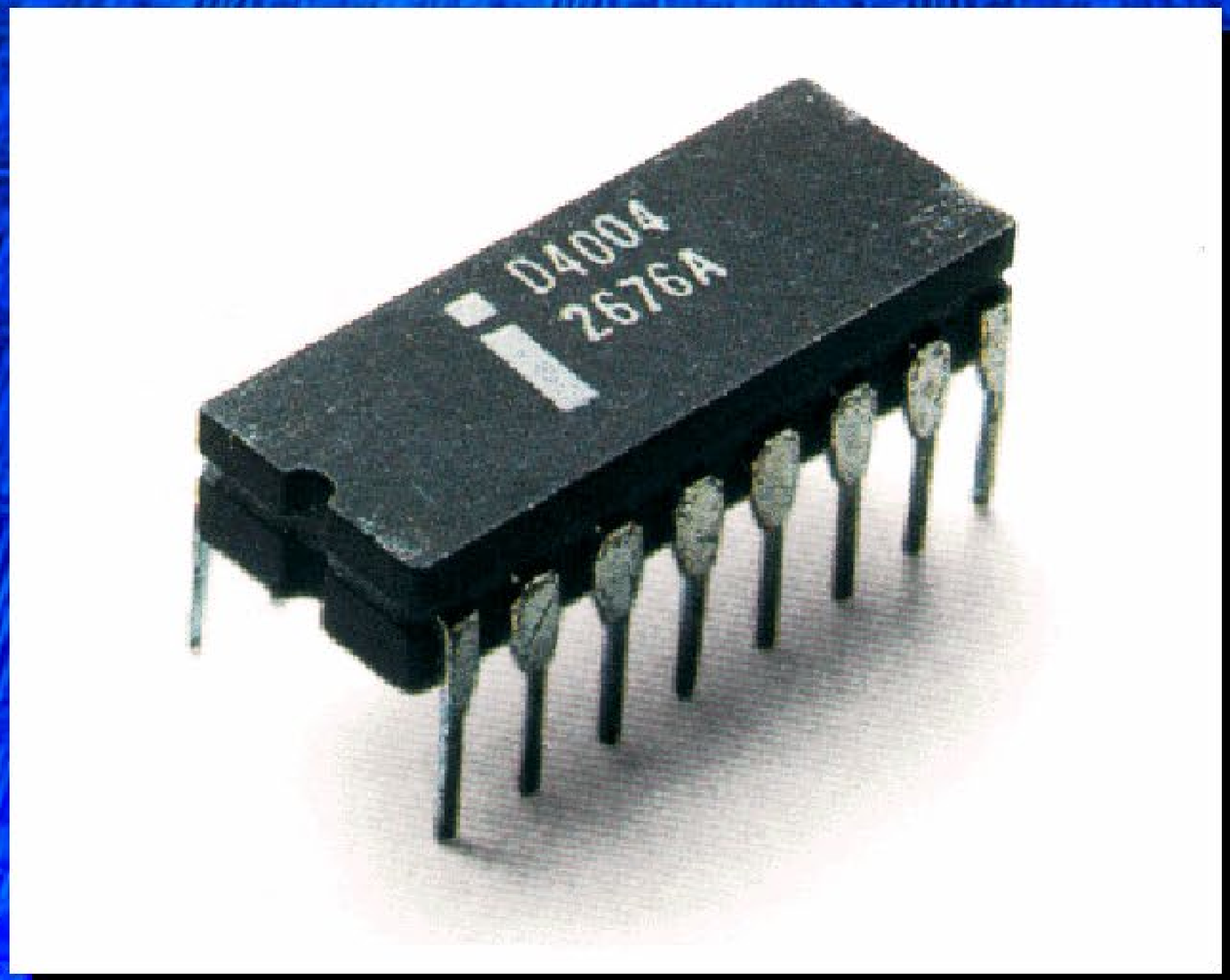
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

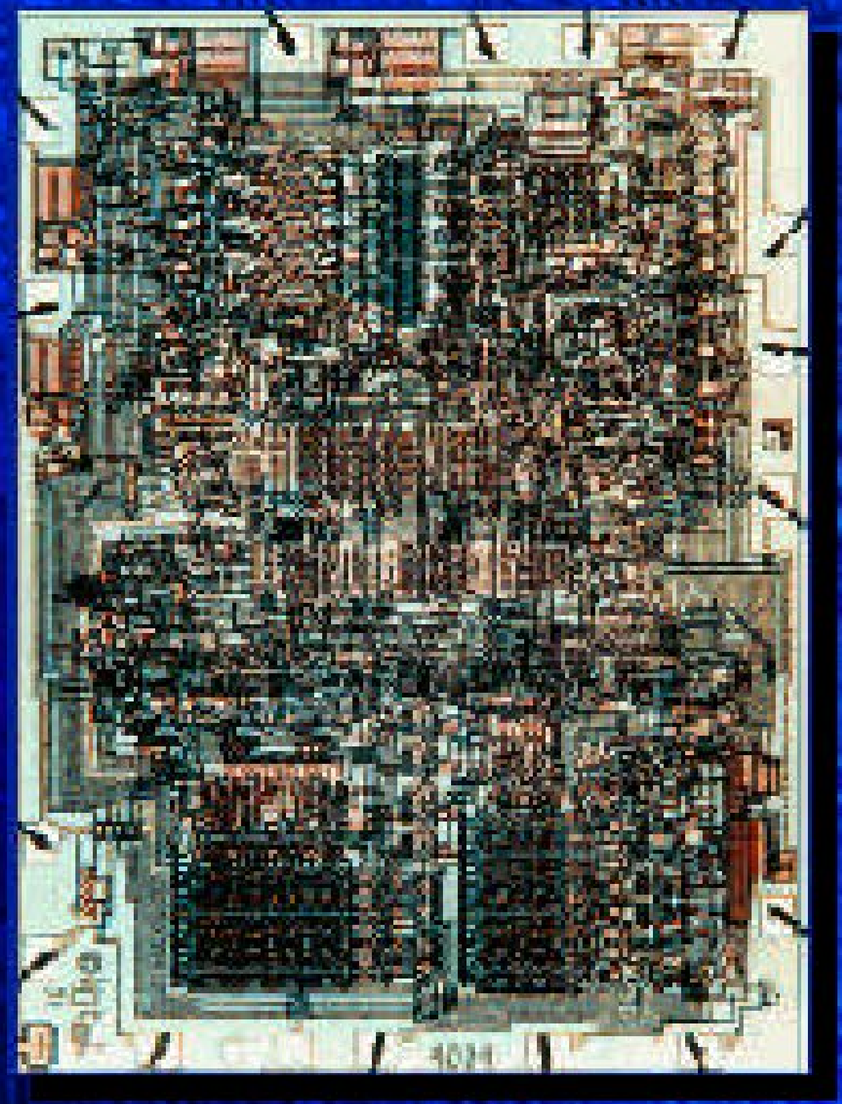
Technological Achievements

- Transistor (Bell Labs, 1947)
  - DEC PDP-1 (1957)
  - IBM 7090 (1960)
- Integrated circuit (1958)
  - IBM System 360 (1965)
  - DEC PDP-8 (1965)
- Microprocessor (1971)
  - Intel 4004



Birth of the Revolution -- The Intel 4004





Introduced November 15, 1971

108 KHz, 50 KIPs , 2300 10μ transistors

ANNOUNCING TESLA V100

GIANT LEAP FOR AI & HPC
VOLTA WITH NEW TENSOR CORE

21B xtors | TSMC 12nm FFN | 815mm^2

5,120 CUDA cores

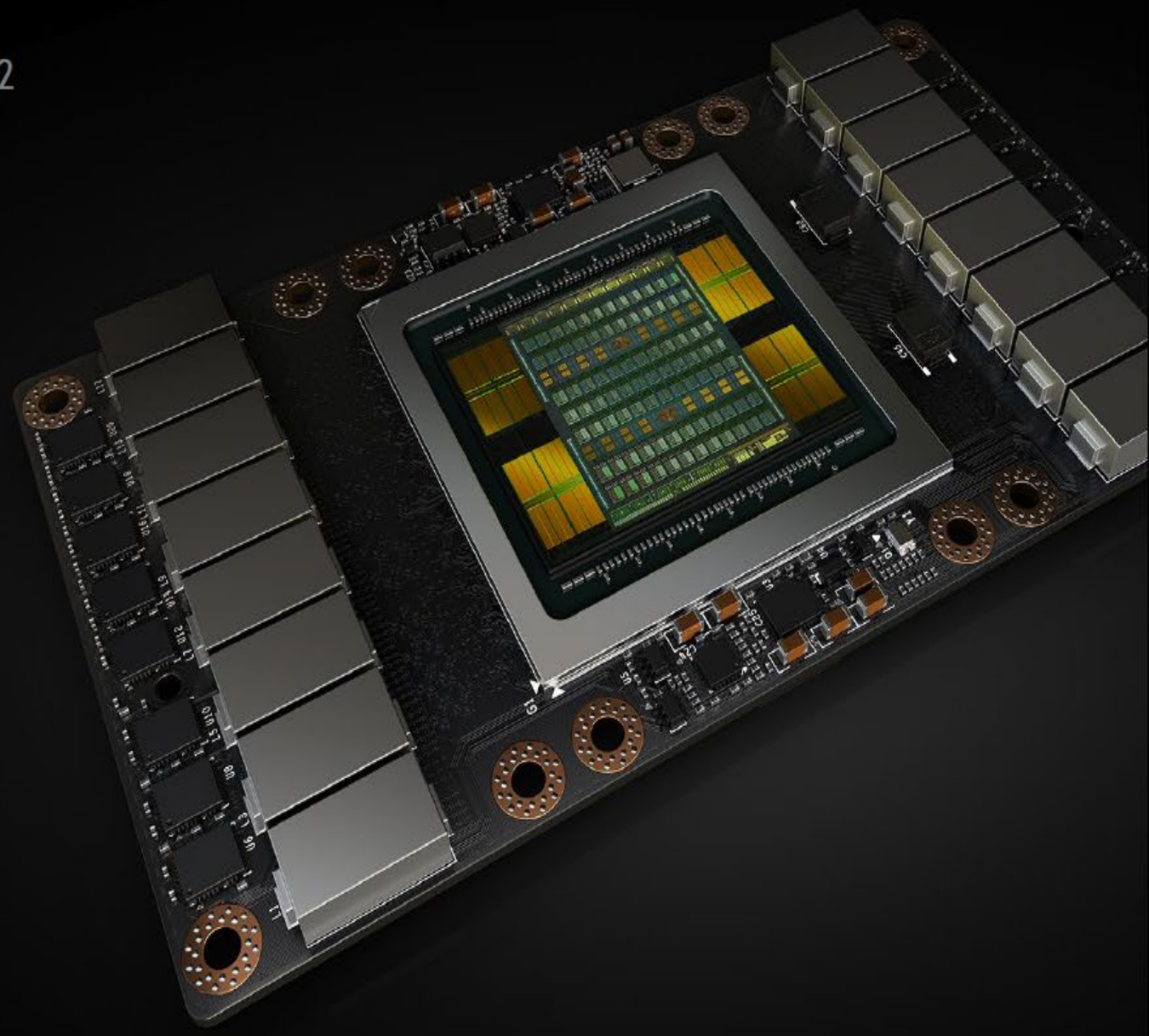
7.5 FP64 TFLOPS | 15 FP32 TFLOPS

NEW 120 Tensor TFLOPS

20MB SM RF | 16MB Cache

16GB HBM2 @ 900 GB/s

300 GB/s NVLink



Top-500, Nov 2018

Rank	Name	Site	Manufacturer	Country	Cores	Accelerators	Rmax [TFlop/s]	Rpeak [TFlop/s]	GFlops /Watts
1	Summit	DOE/SC/Oak Ridge National Laboratory	IBM	US	2,397,824	2,196,480	143,500	200,795	14.67
2	Sierra	DOE/NNSA/Lawrence Livermore National Lab.	IBM/NVIDIA	US	1,572,480	1,382,400	94,640	125,712	12.72
3	Sunway TaihuLight	National Supercomputing Center in Wuxi	NRCPC	China	10,649,600		93,015	125,436	6.05
4	Tianhe-2ª	National Super Computer Center in Guangzhou	NUDT	China	4,981,760	4,554,752	61,445	100,679	3.33
5	Piz Daint	Swiss National Supercomputing Centre	Cray Inc.	Switz	387,872	319,424	21,230	27.154.3	8.90
6	Trinity	DOE/NNSA/LANL/SNL	Cray Inc.	US	979,072		20,158,7	41,461	2.66
7	AI Bridging Cloud Inf.	National Inst. of Adv Industrial Science & Tech.	Fujitsu	Japan	391,680	348,160	19,880	32,577	12.05
8	SuperMUC-NG	Leibniz Rechenzentrum	Lenovo	Germany	305,856		19,477	28,872,86	
9	Titan	DOE/SC/Oak Ridge National Laboratory	Cray Inc.	US	560,640	261,632	17,590	27,113	2.14
10	Sequoia	DOE/NNSA/Lawrence Livermore National Lab.	IBM	US	1,572,864		17,173	20,133	2.18
25	MareNostrum	Barcelona Supercomputing Center	Lenovo	Spain	153,216		6,471	10,296	3.97

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Pre-Exascale: ORNL Summit System

#1 in World - top500.org

System Performance

- Peak of 200 Petaflops (FP_{64}) for modeling & simulation
- Peak of 3.3 ExaOps (FP_{16}) for data analytics and artificial intelligence
- Max power 13 MW

The system includes

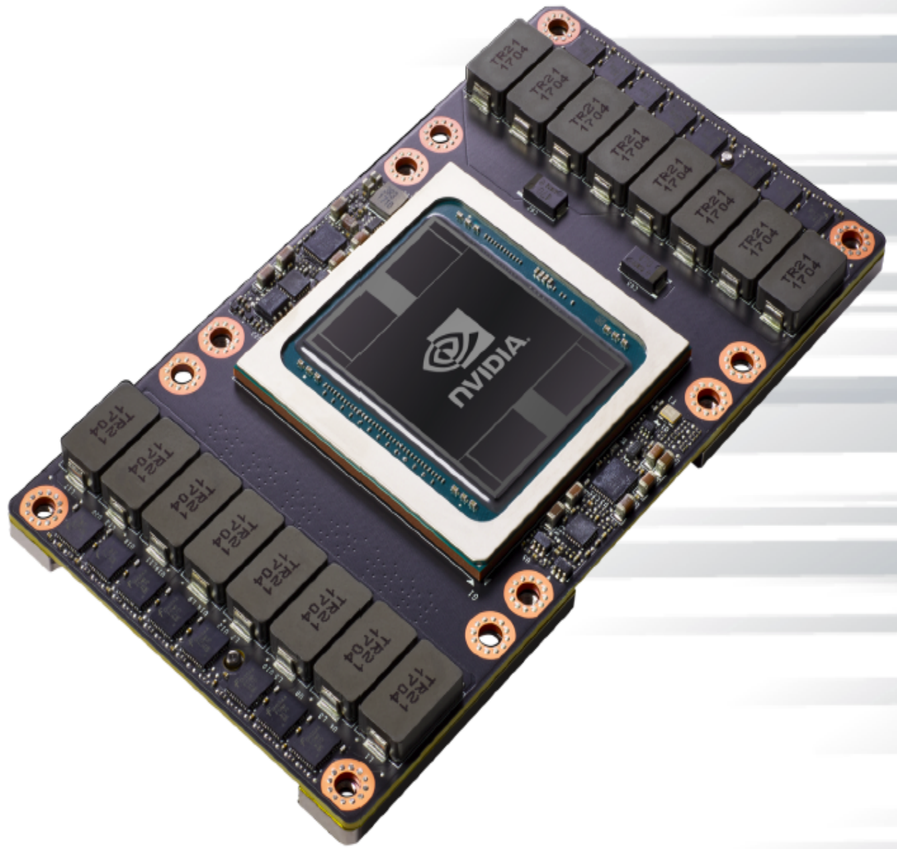
- 4,608 nodes
- Dual-rail Mellanox EDR InfiniBand network
- 250 PB IBM file system transferring data at 2.5 TB/s

Each node has

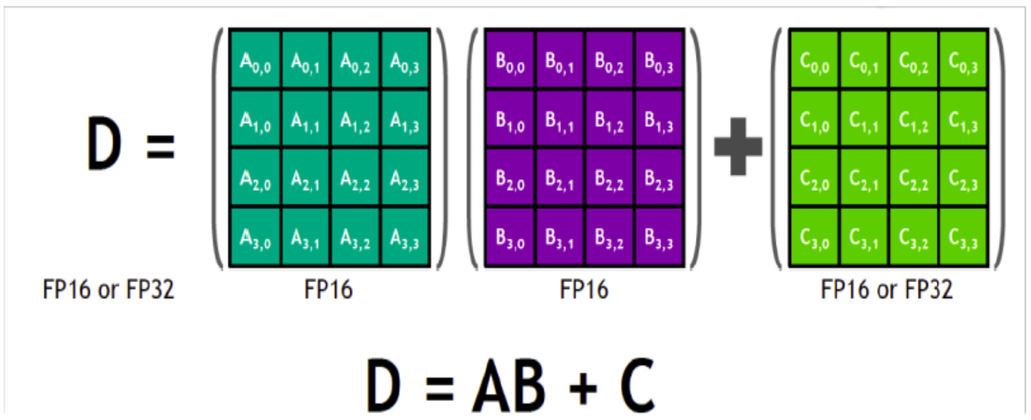
- 2 IBM POWER9 processors
- 6 NVIDIA Tesla V100 GPUs
- 608 GB of fast memory (96 GB HBM2 + 512 GB DDR4)
- 1.6 TB of NV memory

NVIDIA's tesla v100

- 5,120 CUDA cores (64 on each of 80 SMs)
- 640 NEW Tensor cores (8 on each of 80 SMs)
- 20MB SM RF | 16MB Cache | 16GB HBM2 @ 900 GB/s
- 300 GB/s NVLink
- 7.5 FP64 TFLOPS | 15 FP32 TFLOPS | 120 Tensor TFLOPS
- ~57 times faster in 64-bit peak floating point performance than the CM-5 we worked on 25 years ago
- >27K of these coming on ORNL's Summit system!



- Mixed precision matrix math 4x4 matrices



Type	Size	Range	$u = 2^{-t}$
half	16 bits	$10^{\pm 5}$	$2^{-11} \approx 4.9 \times 10^{-4}$
single	32 bits	$10^{\pm 38}$	$2^{-24} \approx 6.0 \times 10^{-8}$
double	64 bits	$10^{\pm 308}$	$2^{-53} \approx 1.1 \times 10^{-16}$
quadruple	128 bits	$10^{\pm 4932}$	$2^{-113} \approx 9.6 \times 10^{-35}$

- The M&S community must figure how out to "cheat" and utilize mixed / reduced precisions
- Ex: Jack Dongarra shows he can get 4x FP64 peak for 64bit LU on V100 with iterative mixed precision (using GMRES!)

Design of Superscalar Processors

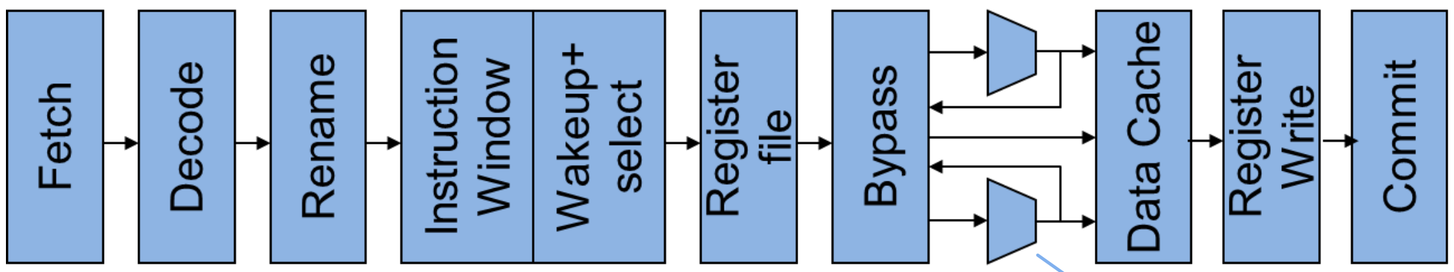
Decoupled from the software stack

Applications

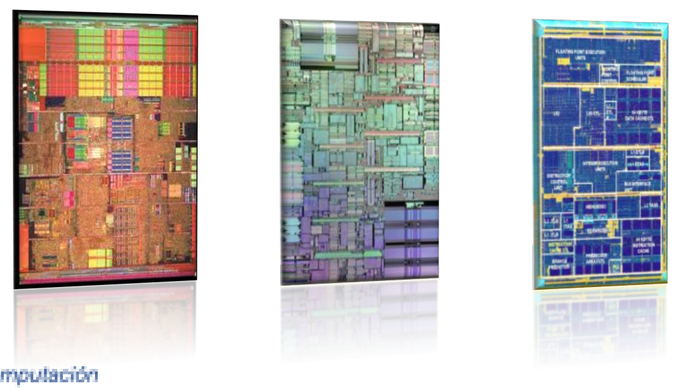
Programs
“decoupled”
from hardware

ISA

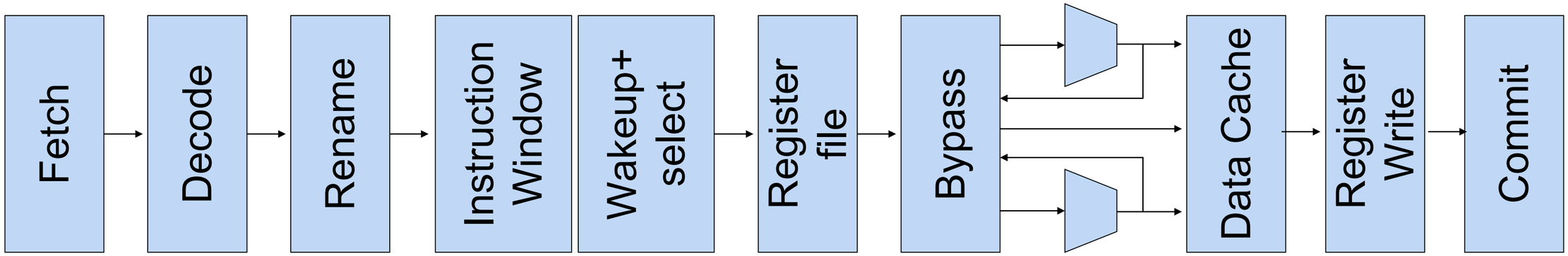
Simple interface
Sequential
program



ILP



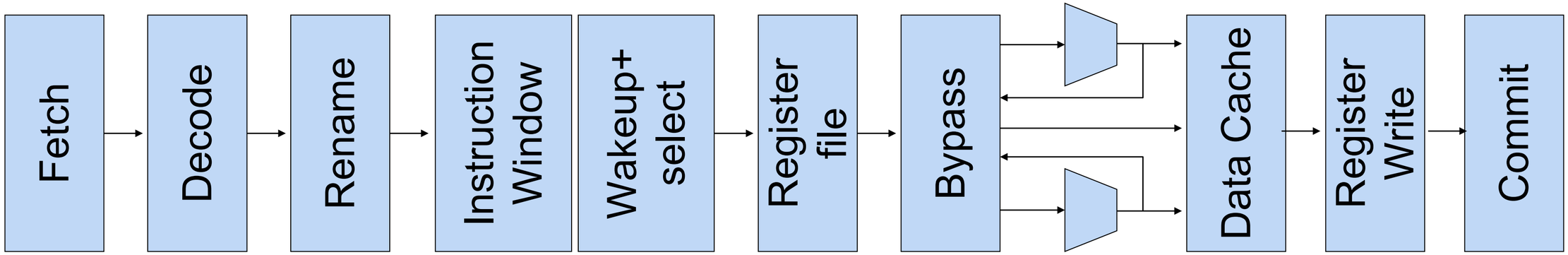
Latency Has Been a Problem from the Beginning... ☹



- Feeding the pipeline with the right instructions:
  - Software trace cache (ICS'99)
  - Prophet/Critic Hybrid Branch Predictor (ISCA'04)
- Locality/reuse
  - Cache Memory with Hybrid Mapping (IASTED87). Victim Cache ☺
  - Dual Data Cache (ICS'95)
- A novel renaming mechanism that boosts software prefetching (ICS'01)
- Virtual-Physical Registers (HPCA'98)
- Kilo Instruction Processors (ISHPC03,HPCA'06, ISCA'08)

Memory Wall

… and the Power Wall Appeared Later ☹☹☹

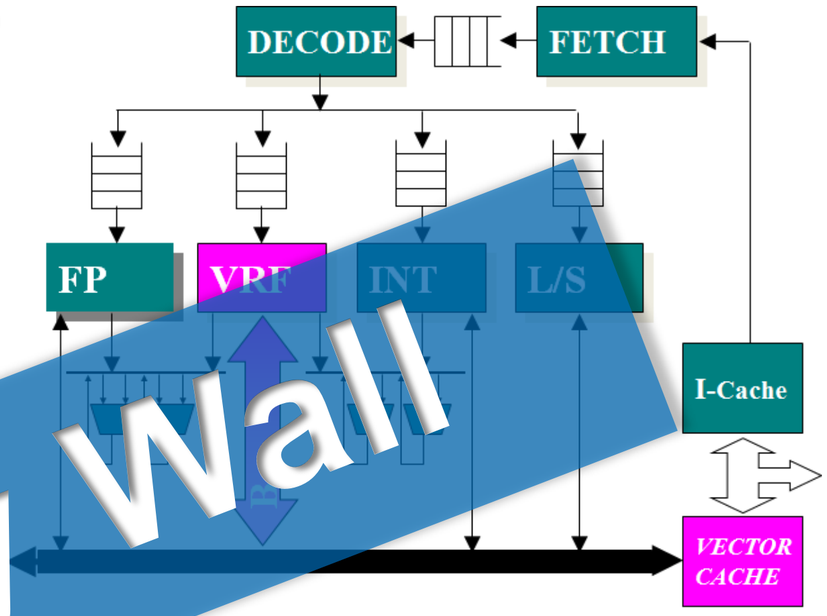


- Better Technologies
- Two-level organization (Locality Exploitation)
  - Register file for Superscalar (ISCA'00)
  - Instruction queues (ICCD'05)
  - Load/Store Queues (ISCA'08)
- Direct Wakeup, Pointer-based Instruction Queue Design (ICCD'04, ICCD'05)
- Content-aware register file (ISCA'09)
- Fuzzy computation (ICS'01, IEEE CAL'02, IEEE-TC'05). Currently known as Approximate Computing ☺

Power Wall

Vector Architectures… Memory Latency and Power ☺☺☺

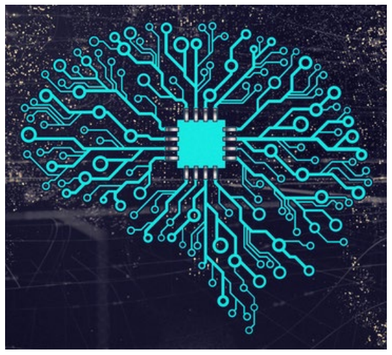
- Out-of-Order Access to Vectors (ISCA 1992, ISCA 1995)
- Command Memory Vector (PACT 1998)
  - In-memory computation
- Decoupling Vector Architectures (HPCA 1996)
  - Cray SX1
- Out-of-order Vector Architectures (Micro 1996)
- Multithreaded Vector Architectures (HPCA 1997)
- SMT Vector Architectures (HICS 1997, IEEE [illegible])
- Vector register-file organization (PACT 1997)
- Vector Microprocessors ([illegible])
- Architectures with Short Vectors (PACT 1997, ICS 1998)
  - [illegible], Knights Corner
- Vector Architectures for Multimedia (HPCA 2001, Micro 2002)
- High-Speed Buffers Routers (Micro 2003, IEEE TC 2006)
- Vector Architectures for Data-Base (Micro 2012, HPCA2015,ISCA2016)



Power, Latency Wall

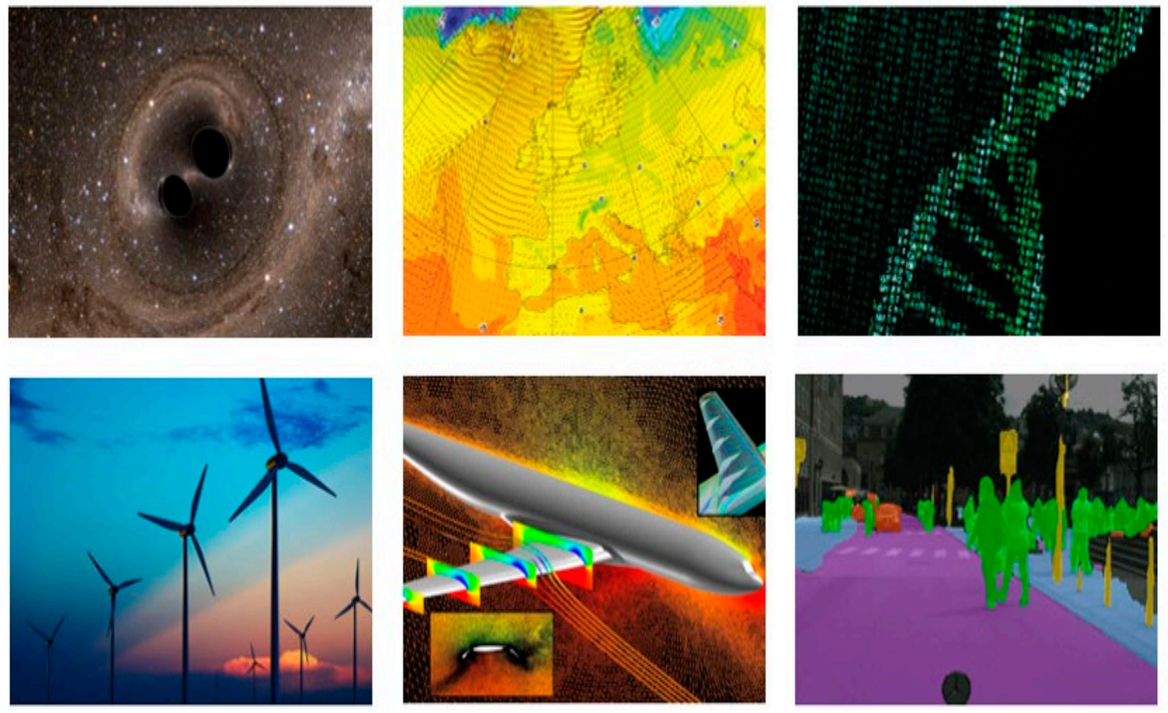
Some Context

- I was kindly invited by Roberto Viola to give a talk at ICT in Lisbon in October 2015
- I tried to emphasise the importance of HPC and of producing European technology
- A lot has happened since then…



Some Societal Challenges

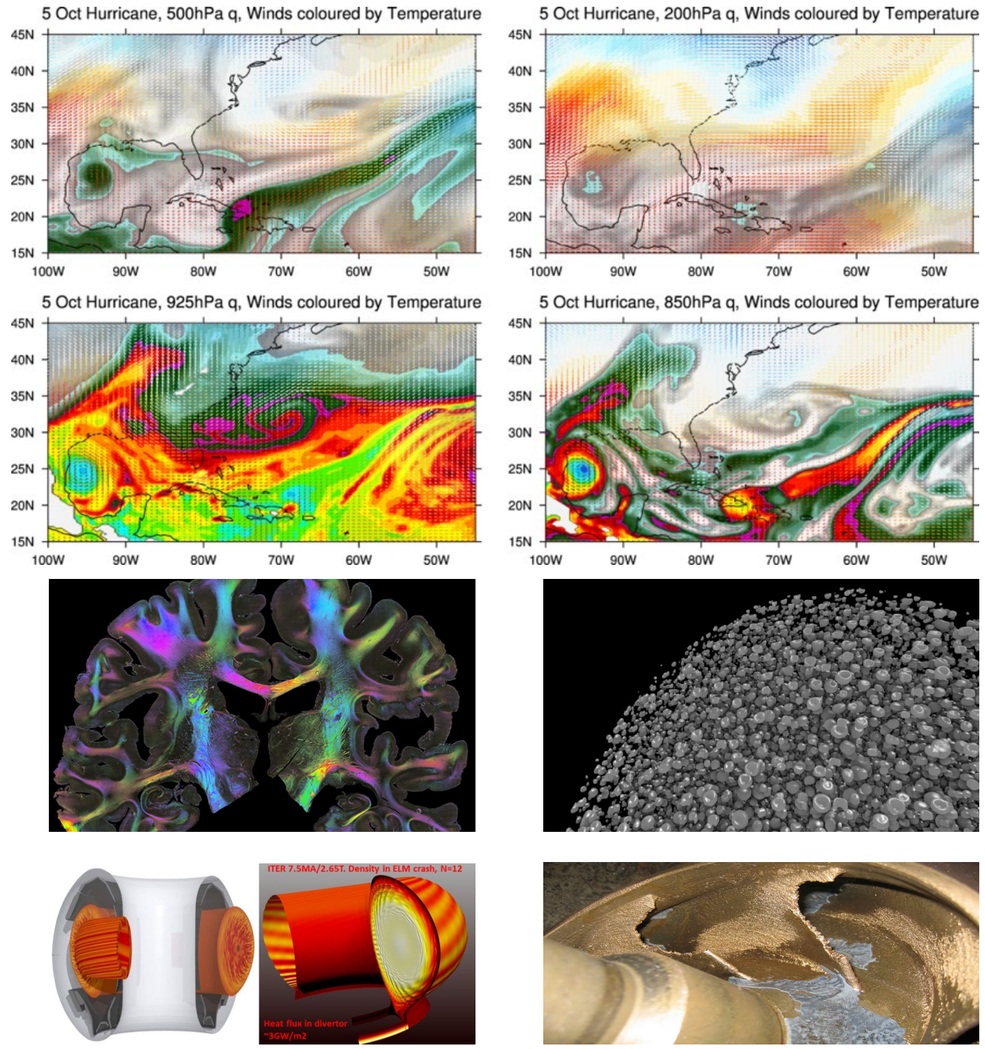
- Aging population
- Climate change
- Cybersecurity
- Increasing energy needs
- Intensifying global competition



Images courtesy of The PRACE Scientific Steering Committee, “The Scientific Case for Computing in Europe 2018-2026”

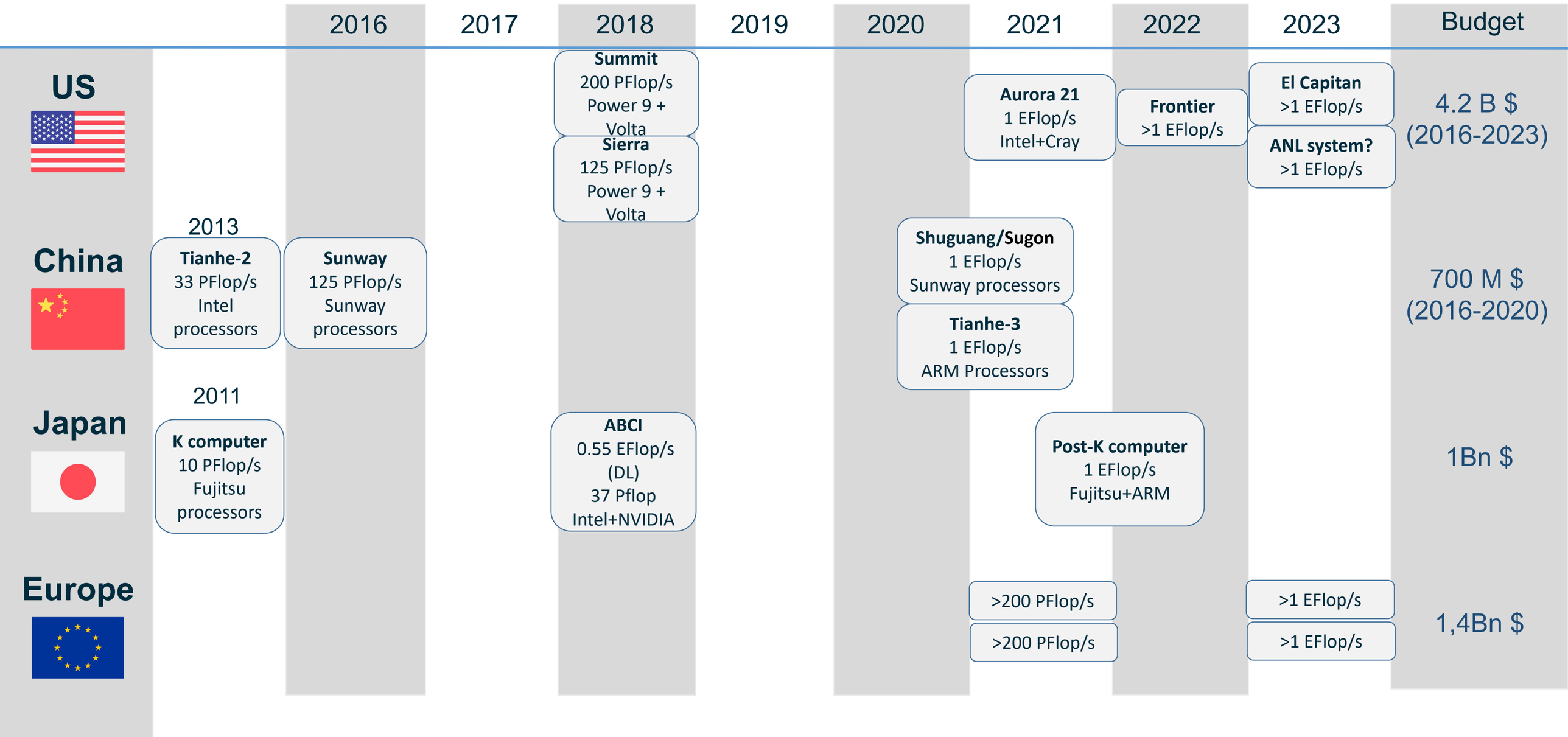
Why is HPC Needed?

- Will save billions by helping us to adapt to climate change
- Will improve human health by enabling personalized medicine
- Will improve fuel efficiency of aircraft & help design better wind turbines
- Will help us to understand how the human brain works



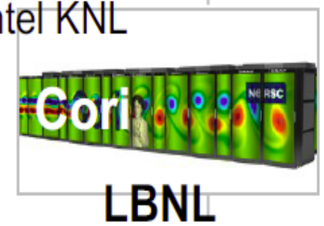
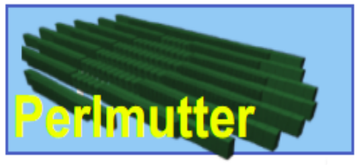
Images courtesy of The PRACE Scientific Steering Committee, "The Scientific Case for Computing in Europe 2018-2026"

Update on the Race for Exascale



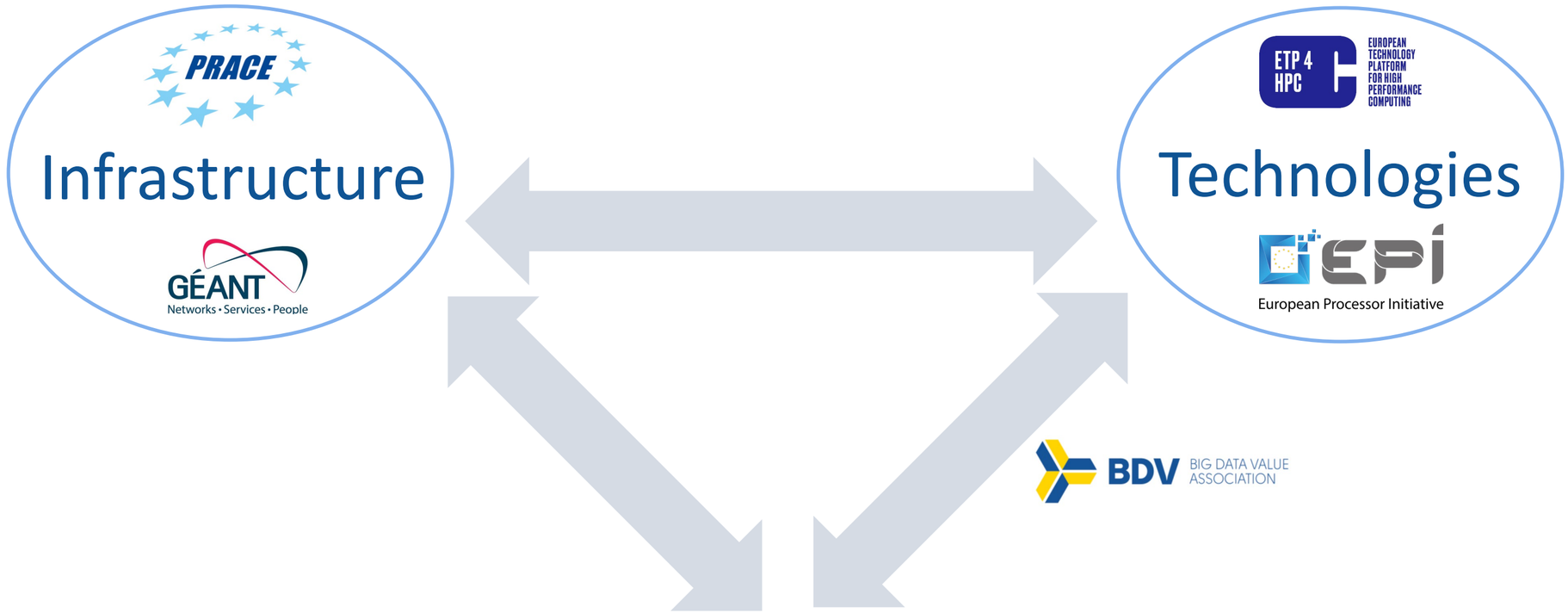
Department of Energy Roadmap to Exascale Systems
portfolio includes a variety of vendors and architectures

First U.S. Exascale Computers

2012	2016	2018	2020	2021-2022
 **ORNL** Cray/AMD/NVIDIA		 **ORNL** IBM/NVIDIA		 **ORNL** TBD
 **ANL** IBM BG/Q	 **ANL** Cray/Intel KNL	 **LBNL** Cray/Intel Xeon/KNL	 **LBNL** Cray/AMD/NVIDIA	 **ANL** Intel/Cray
 **LLNL** IBM BG/Q	 **LANL/SNL** Cray/Intel Xeon/KNL	 **LLNL** IBM/NVIDIA	 **LANL/SNL** TBD	DOE Classified Computing  **LLNL** TBD

EXASCALE COMPUTING PROJECT

Current European HPC Ecosystem



Applications

esiwace
CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

EXCELLERAT

MAX

ChEESE

bioexcel

HiDALGO

CompBioMed

POP

EoCoE

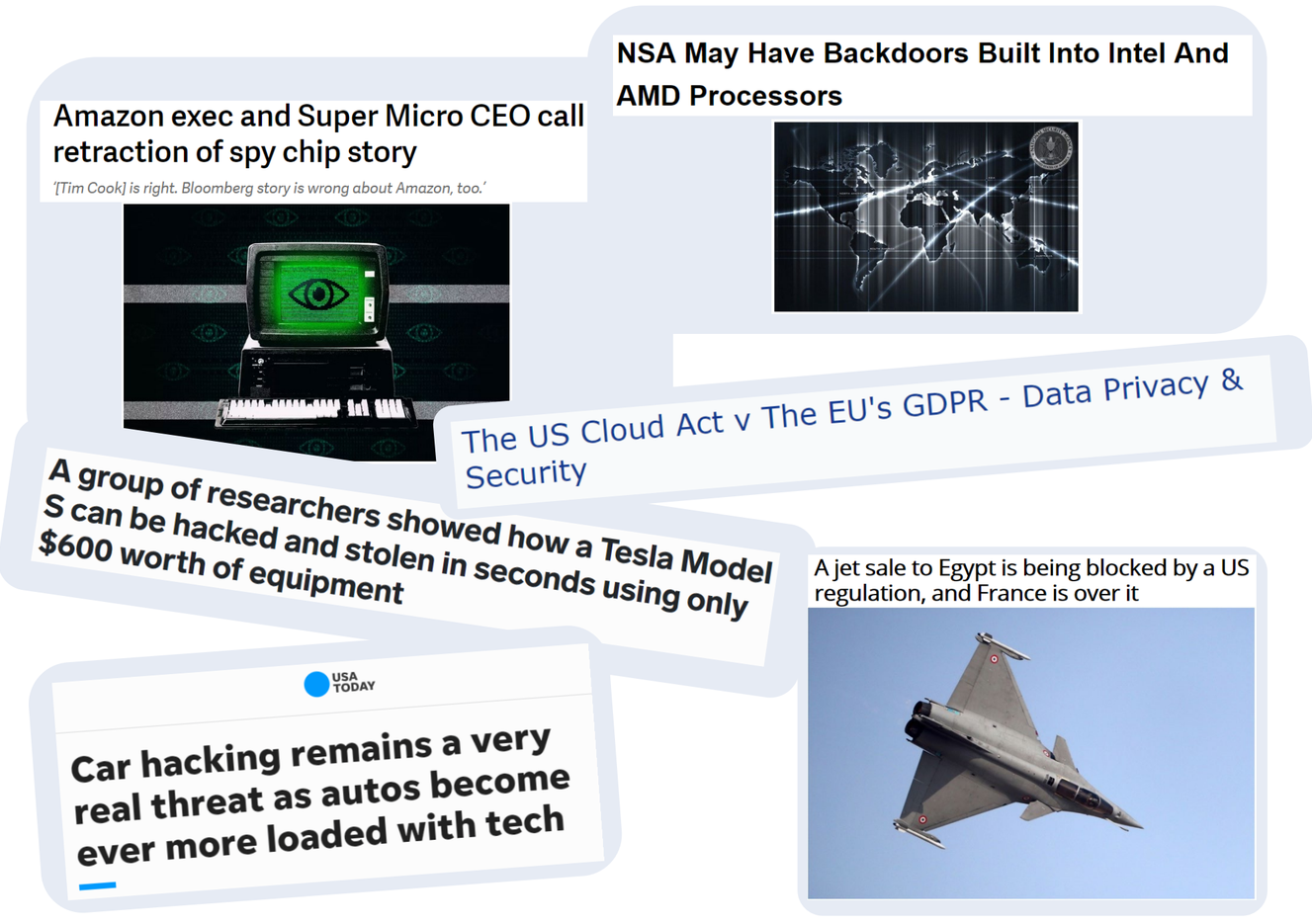
Where Europe needs to be stronger

- Only 1 of the 10 most powerful HPC systems is in the EU
- HPC codes must be upgraded
- Vital HPC hardware elements are missing: General Purpose Processor and Accelerators

- EU needs its own source of as many of the system elements as possible

Why Europe needs its own Processors

- Processors now control almost every aspect of our lives
- **Security** (back doors etc.)
- Possible **future restrictions on exports to EU** due to increasing protectionism
- **A competitive EU supply chain** for HPC technologies will create jobs and growth in Europe



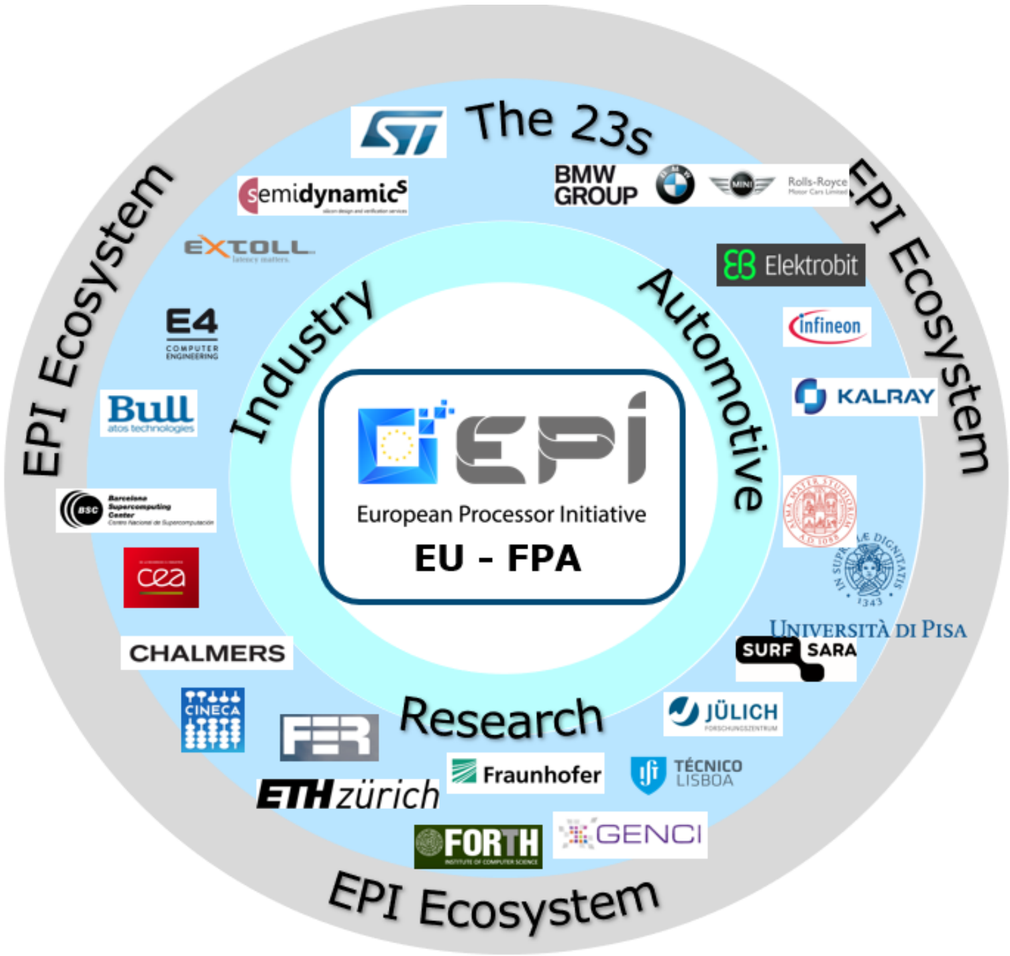
Images courtesy of European Processor Initiative

How EuroHPC will help to make us stronger

- Developing a new European supercomputing ecosystem: HPC systems, network, software, applications, access through the cloud
- Making HPC resources available to public and private users, including SMEs.
- Stimulating a technology supply industry

EuroHPC & EPI (European Processor Initiative)

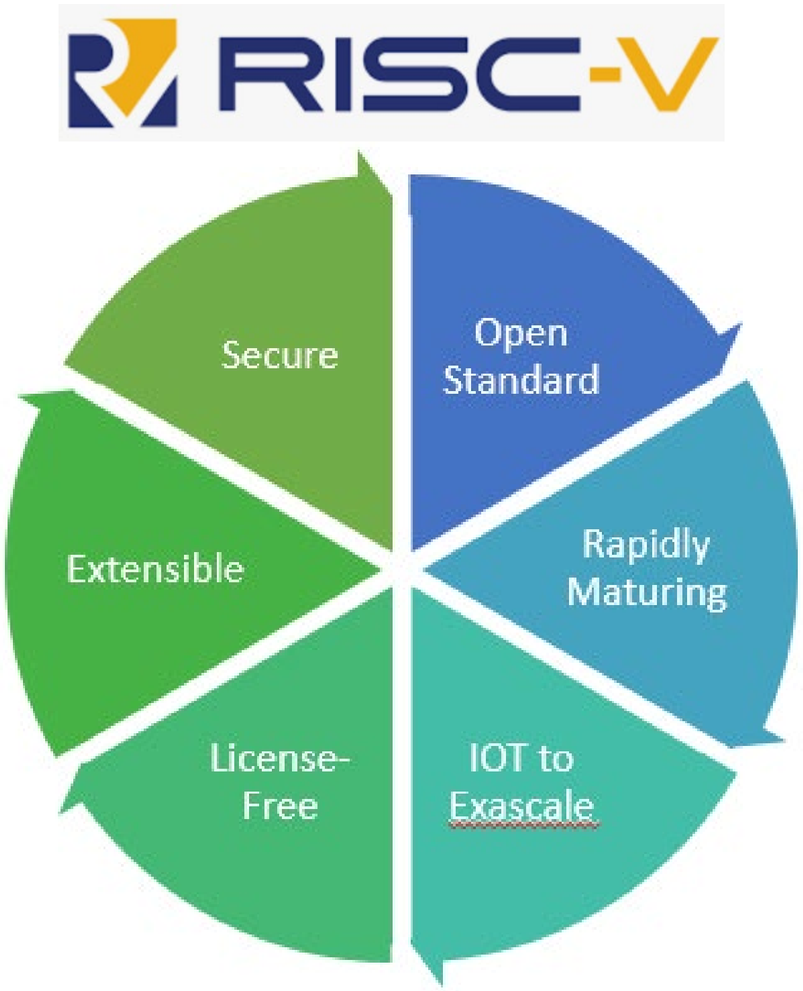
- High Performance General Purpose Processor for HPC
- High-performance RISC-V based accelerator
- Computing platform for autonomous cars
- Will also target the AI, Big Data and other markets in order to be economically sustainable



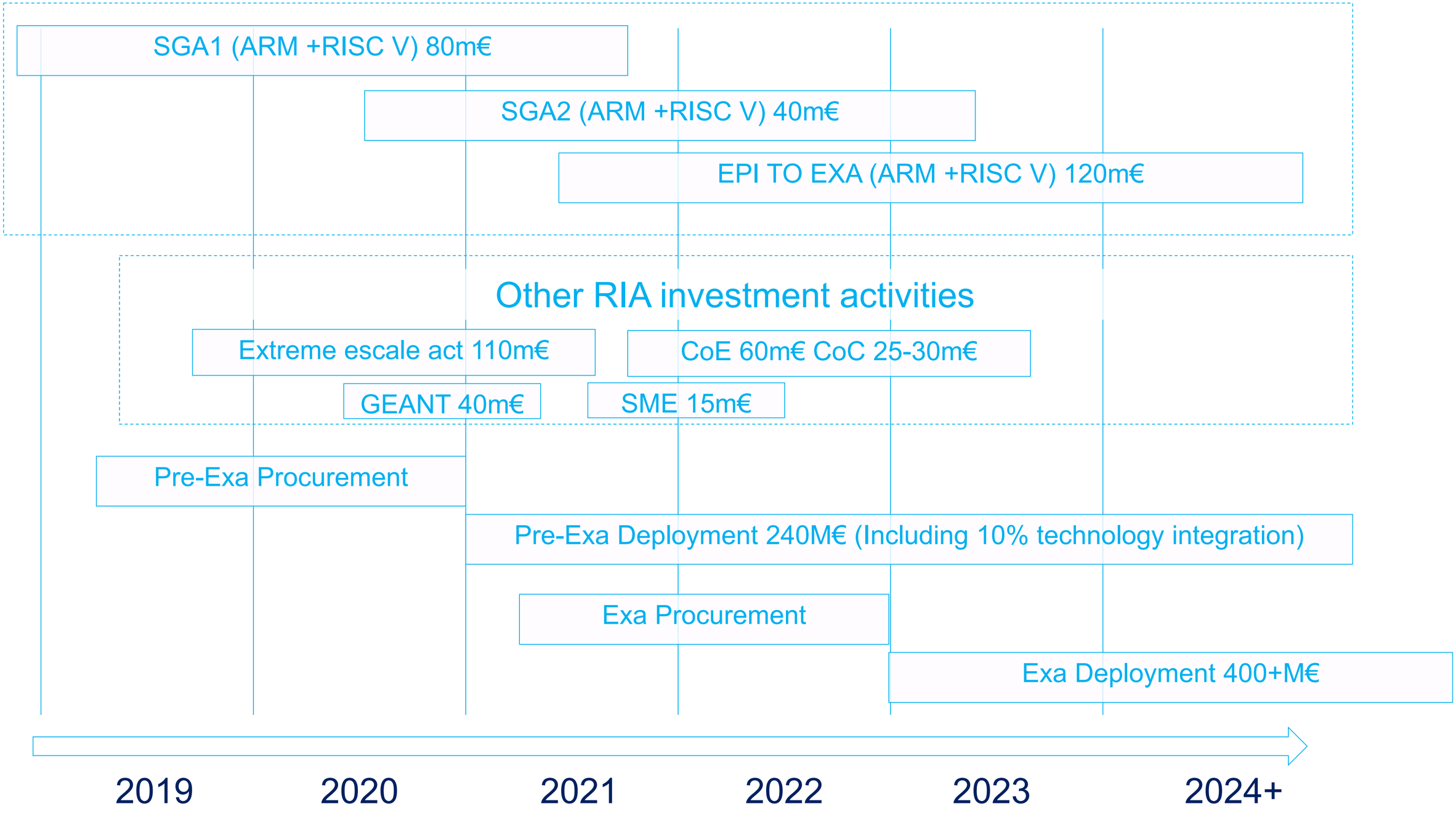
Images courtesy of European Processor Initiative

The Open-Source Hardware Opportunity

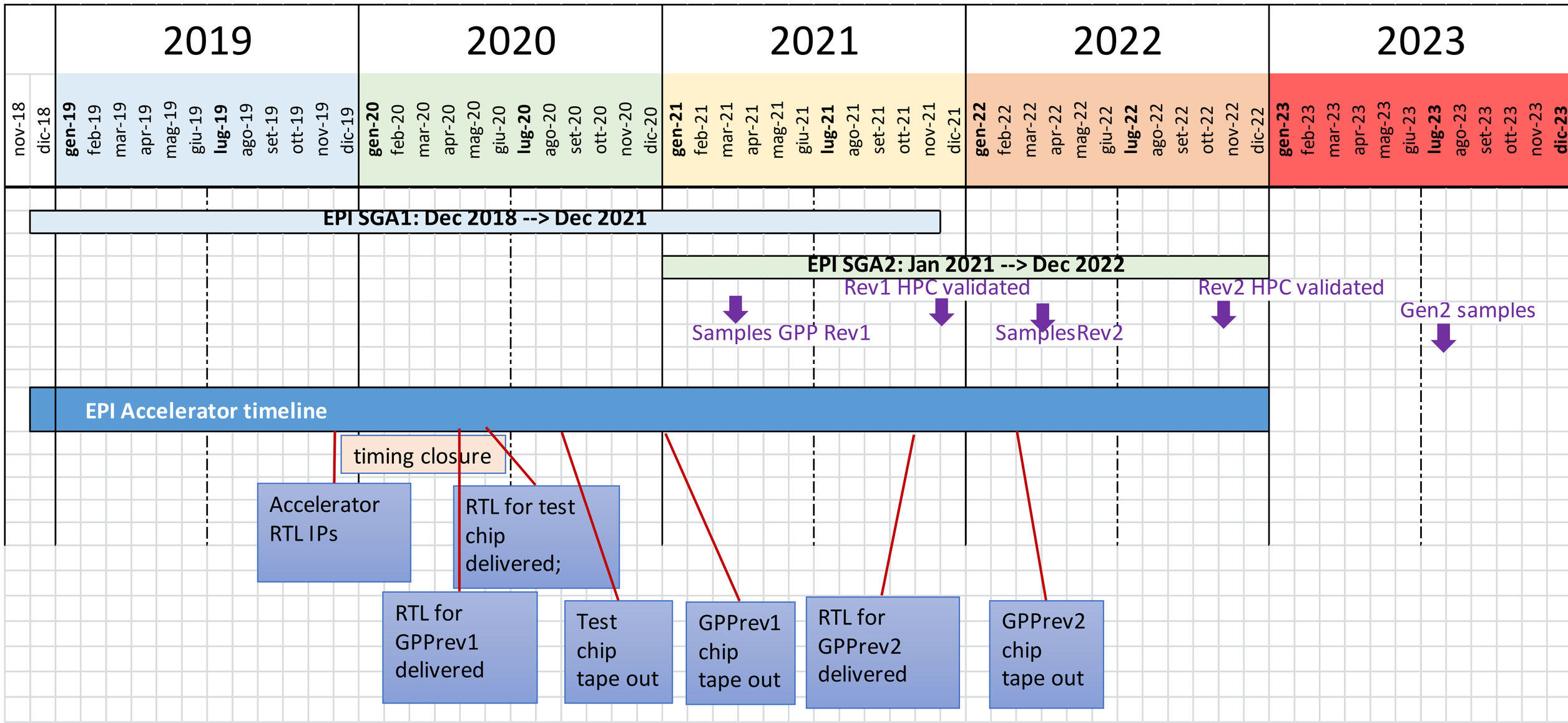
- In 2015 I said I believed a European Supercomputer based on ARM was possible (Mont-Blanc).
- Even though ARM is no longer European, it can form part of the short-term solution
- The fastest-growing movement in computing at the moment is Open-Source and is called RISC-V
- The future is Open and RISC-V is democratising chip-design



HPC Roadmap


SGA1 (ARM +RISC V) 80m€
SGA2 (ARM +RISC V) 40m€
EPI TO EXA (ARM +RISC V) 120m€
Other RIA investment activities
Extreme escale act 110m€
CoE 60m€ CoC 25-30m€
GEANT 40m€
SME 15m€
Pre-Exa Procurement
Pre-Exa Deployment 240M€ (Including 10% technology integration)
Exa Procurement
Exa Deployment 400+M€
2019
2020
2021
2022
2023
2024+

Accelerator Timeline

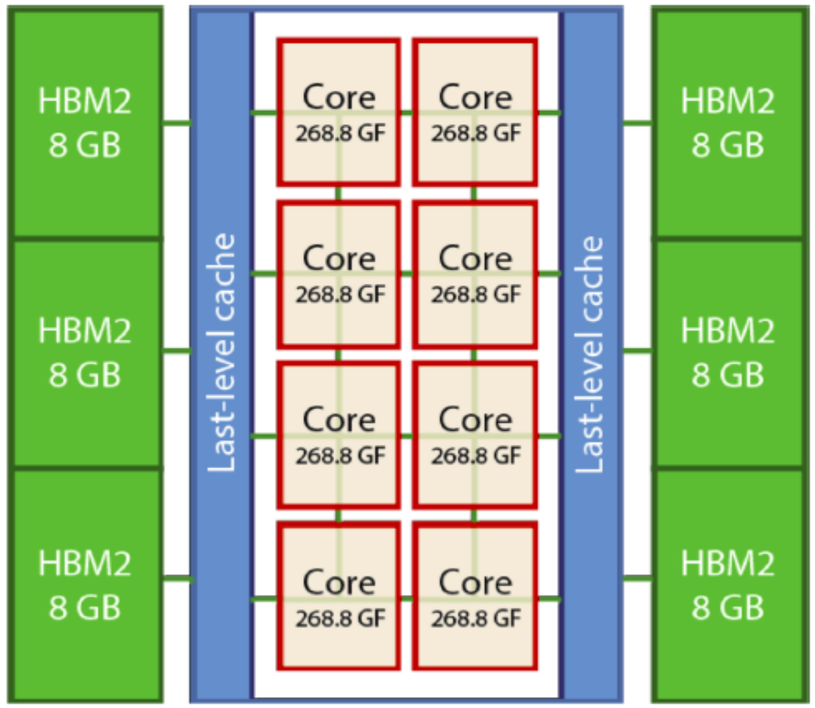


Related chips

Provider	Fujitsu	Aurora	V100	GPP	Intel Xeon Gold 6152
Techinogy	7nm	16nm	12nm	7nm	14nm
Size			815mm^2		
Transistors			21,1 billon		
ISA	ARM V8.2 + SVE	NEC	Cuda	ARM V8.4+SVE	X64-AVX512
Cores	48+4 (4 clusters)	8	80 SM (each divide internaly in 4)	36 (max 8x8)	22
Vector registers	32	64	256KB (mem size)	32	32
Vector Length	512 bits (16 DP)	16384bits (256 DP)		256bits (8 DP)	512bits (16DP)
Ops per cycle (double precision)	2X512bits 64 flops (32FMA)	192 flops (96FMA)	64 flops (32 FMA)	2X256 bits flops	2x512 64 flops
Flops core	57 GF/s	409 GF/s	98 GF/s	24 to 48 GF/scal	141 GF/s
Flops chip	2,7 Tflops/s	3,2 TF/s	7,8 TF/s	0,87 to 1,7 Tflops/s	3,1 TF/s
Memory	4xHBM2 32GB	6xHBM2 48GB	4 HBM2 32 GB	4xHBM2 32GB	6xDDR4 2666 Up to 768 GB
Memory bandwidth	1024GB/s	1,2TB/s	900 GB/s		120 GB/s
L1 Cache	64Kb-4way 230+115 GB/s 2X512 loads sustained throughput Combined gather operation	32 KB only scalar (each, inst and data)	Up to 96 KB, but in the fig of the white paper shows 128KB.	64Kb-4way	32KB
L2	32 MB (8 per cluster) Bandwidth (3,6 according to the slide, 2,2 according my calculations) Seems shared in the cluster	256 KB unified for scalar.	6 MB	1 MB per core (local)	!MB per core
LLC		16 MB shared, 3TB/s		64 MB shared, distributed	30 MB

Specification of SX-Aurora TSUBASA

- Memory bandwidth
  - 1.22 TB/s world's highest memory bandwidth
    - Six HBM2 memory modules integration
  - 3.0 TB/s LLC bandwidth
    - LLC is connected to cores via 2D mesh network
- Computational capability
  - 2.15 Tflop/s@1.4 GHz
    - 8 powerful vector cores
    - 16 nm FINFET process technology
    - 4.8 billion transistors
    - 14.96 mm x 33.00 mm



Block diagram of a vector processor

3

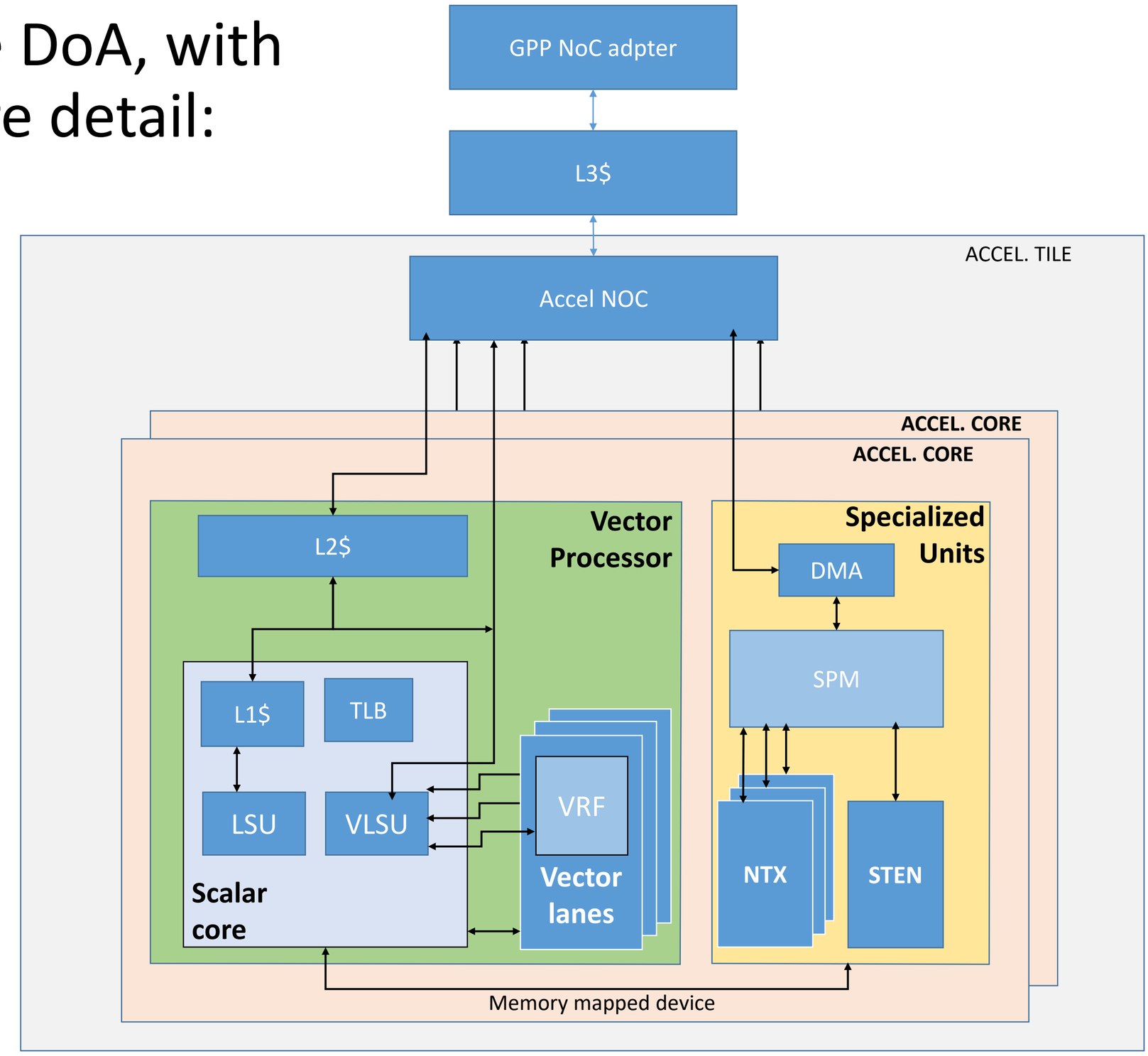
CHICHÉN ITZÁ ACCELERATOR

Chichén Itzá -
Teotihuacán

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Architecture (with GPP integration)

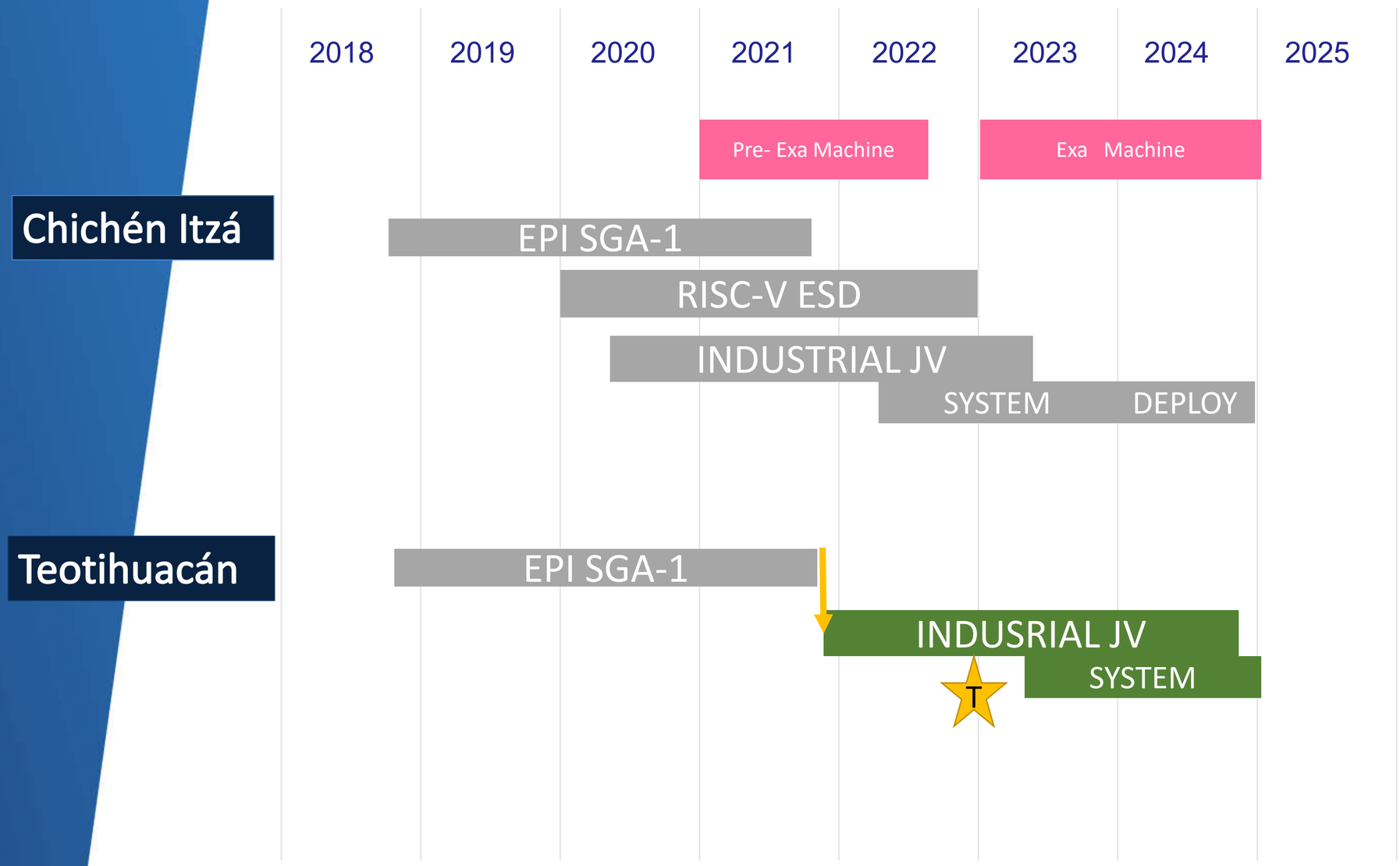
- As per the DoA, with some more detail:



Vector processor

- Vector processor:
  - RISC-V vector ISA
    - (+ extensions in second phase ?)
  - Vector lanes per core: 8-16
  - Vector length per lane: 64 elements, DP
  - Physical Vector Register File per lane: 20 KB
  - Vector load-store data-path can bypass L2 cache
- Target clock frequency:
  - Test chip (22nm) 1 GHz
  - GPP (7 nm) 2 GHz
- FLOP/cycle per vector core: 16-32 DP, 32-64 SP
- Connectivity
  - Lanes-to-NoC interconnect width, per core: 512 bits
- Lanes-to-NoC bandwidth, per core: 64 GB/s @ 1 GHz

FULL SYSTEM ROADMAP



Chichén Itzá - Teotihuacán

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

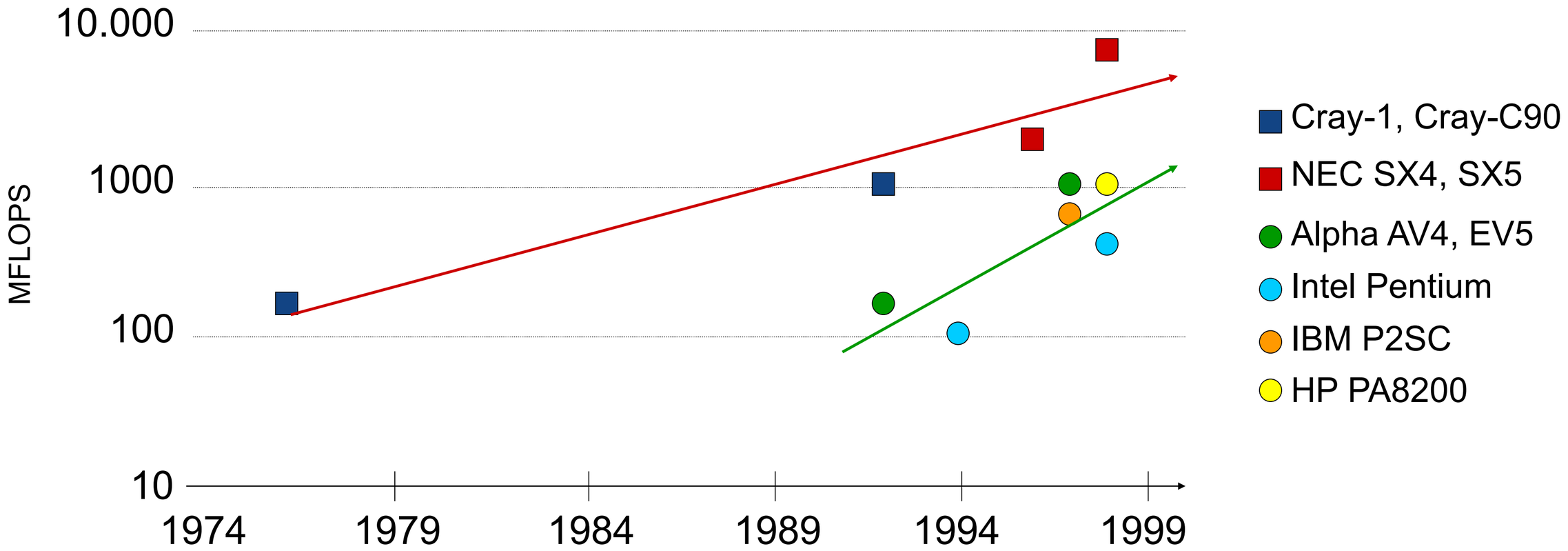
4

TEOTIHUACÁN
GPP

Chichén Itzá -
Teotihuacán

Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

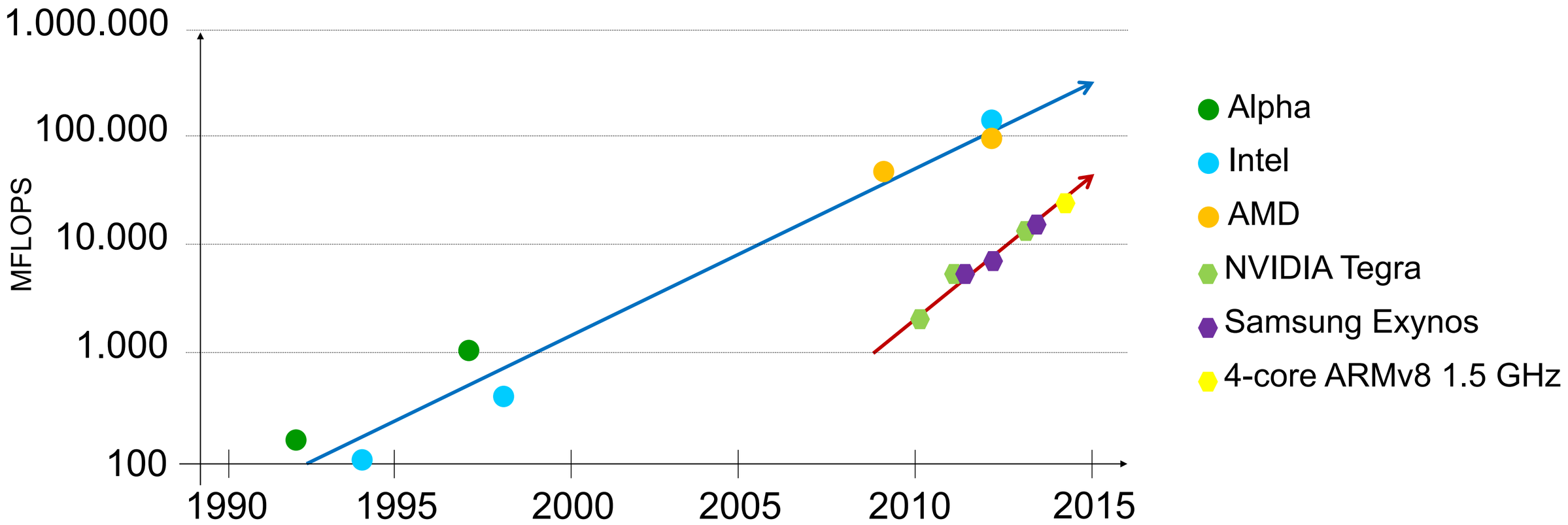
“Killer microprocessors”



- Microprocessors killed the Vector supercomputers
  - They were not faster ...
  - ... but they were significantly **cheaper** and **greener**
- 10 microprocessors approx. 1 Vector CPU
  - SIMD vs. MIMD programming paradigms

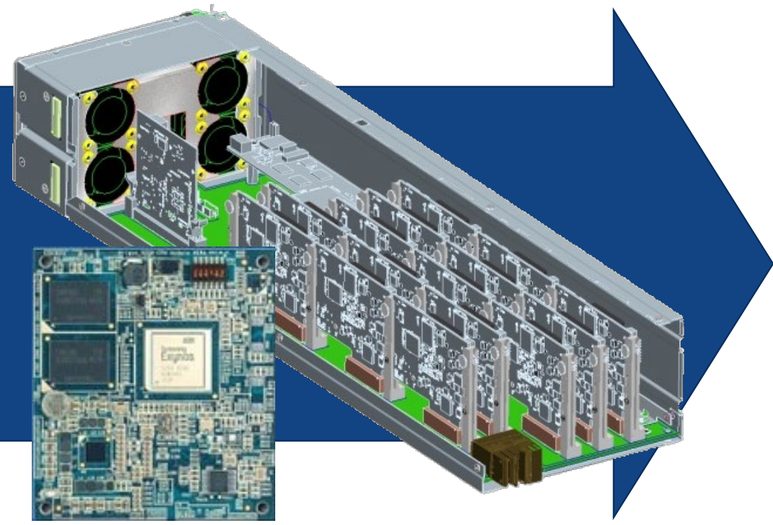
M. Valero. “Vector Architectures: Past, Present and Future”. Keynote talk. ICS-11. IEEE-ACM. Melbourne, 1998

The killer mobile processors™



- Microprocessors killed the Vector supercomputers
  - They were not faster ...
  - ... but they were significantly cheaper and greener

- History may be about to repeat itself …
  - Mobile processor are not faster …
  - … but they are significantly cheaper



2011

Tibidabo

ARM multicore

2012

KAYLA

ARM + GPU
CUDA on ARM

2013

Pedraforca

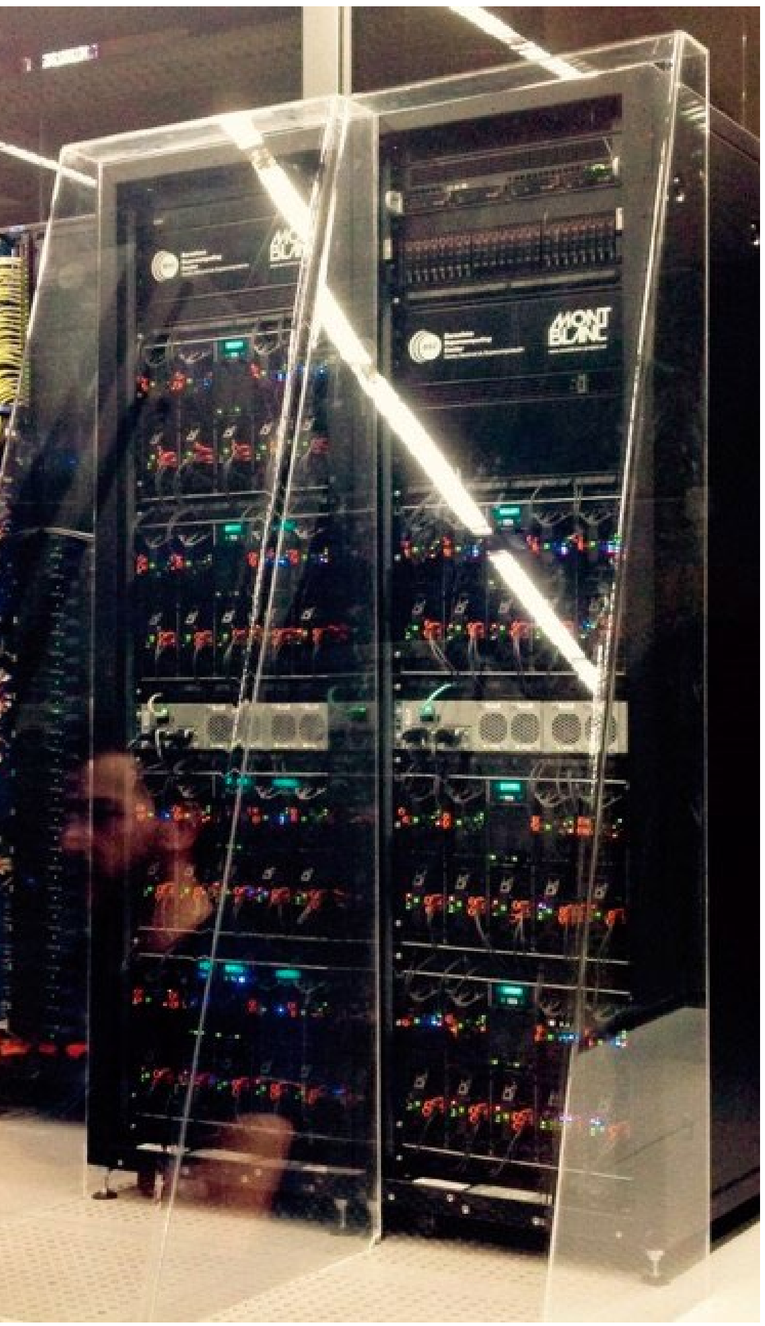
ARM + GPU
Inifinband
RDMA

2014

Mont-Blanc

Single chip ARM+GPU
OpenCL on ARM GPU

Mont-Blanc HPC Stack for ARM



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Industrial applications

Pharmacelera
MUREX
Termo Fluids
DASSAULT AVIATION
AVL
Cenaero
Excellence in Simulation Technologies for Aeronautics
ROLLS ROYCE

Applications

GENCI
CINECA
UNI GRAZ
HLRS
University of BRISTOL

System software

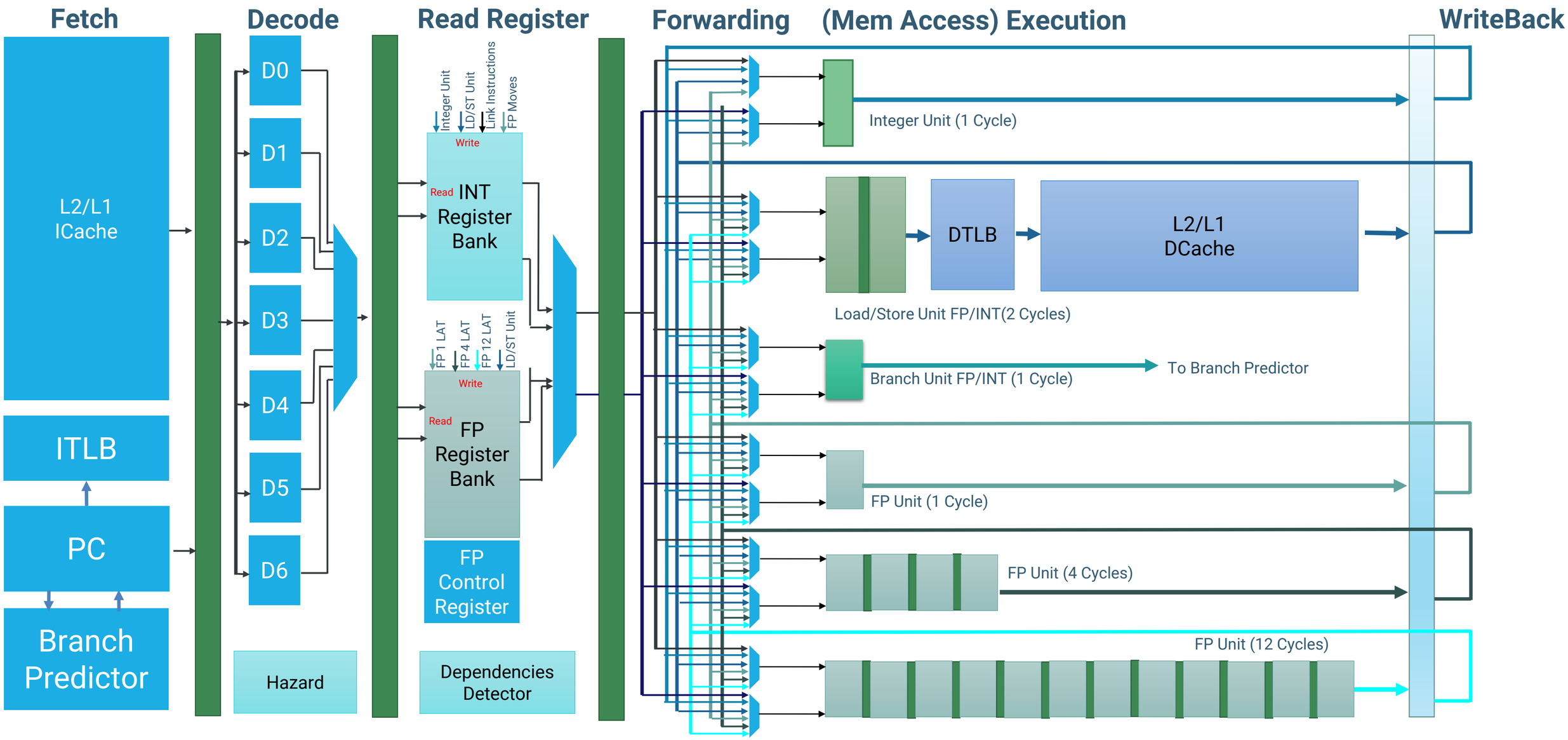
ARM
ETH zürich
Informatiques mathématiques Inria
Barcelona Supercomputing Center
Centro Nacional de Supercomputación
JÜLICH FORSCHUNGSZENTRUM
Bull atos technologies

Hardware

ARM
Barcelona Supercomputing Center
Centro Nacional de Supercomputación
Bull atos technologies
cea
UC UNIVERSIDAD DE CANTABRIA

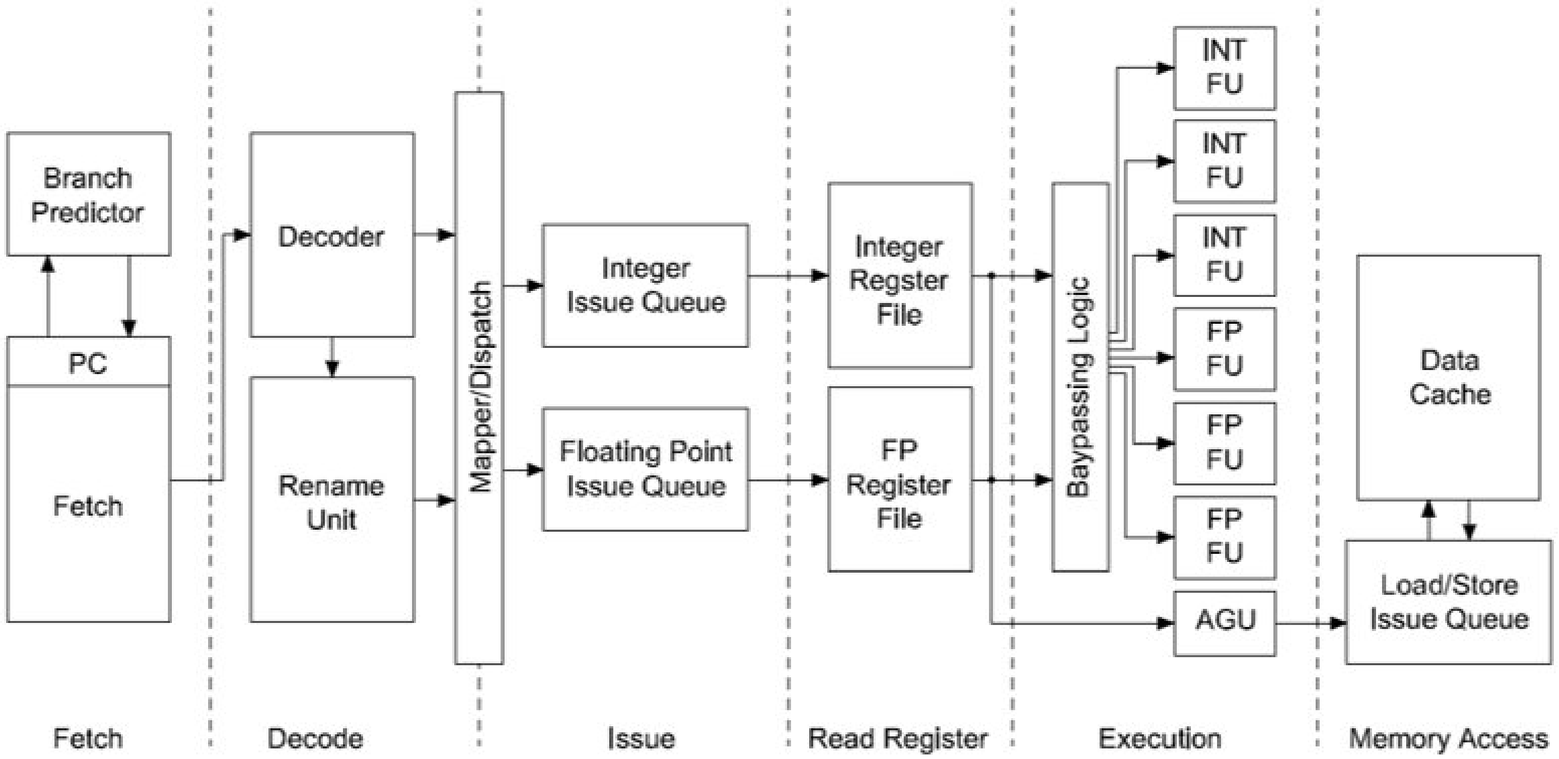
Lagarto I

Microarchitecture



Microarquitectura

Procesador Lagarto II



7 etapas de procesamiento
Predictor de saltos tipo G-Share

ISA RV64I (62 instrucciones)
Extensión M (13 instrucciones)
ALU/ADD (2)
MULT/DIV/REM (1)
Extensión F (30 instrucciones)
ADD/MUL/DIV/SQRT
Extensión D (32 instrucciones)
ADD/MUL/DIV/SQRT
Extensión A (9 instrucciones)
Atómicas
Manejador de Excepciones (privilegiadas)

Ejecución fuera de orden de 2-vías
5 instrucciones ejecutadas por ciclo de reloj
2 instrucciones graduadas

128-entradas para el ROB
128 registros físicos de tipo entero (INT)
128 registros físicos de punto flotante (FP)

32-entradas para la cola de enteros
32-entradas para la cola de punto flotante
32-entradas para la cola de acceso a memoria

Módulos patentados/Trámite

Procesador Lagarto II

Mecanismo de recuperación de excepciones en procesadores con ejecución fuera de orden.

Mecanismo de planificación dinámica para procesadores de alto desempeño y bajo consumo de energía

Unidad de traducción de direcciones virtuales a direcciones físicas de memoria para procesadores de alto desempeño y bajo consumo de energía

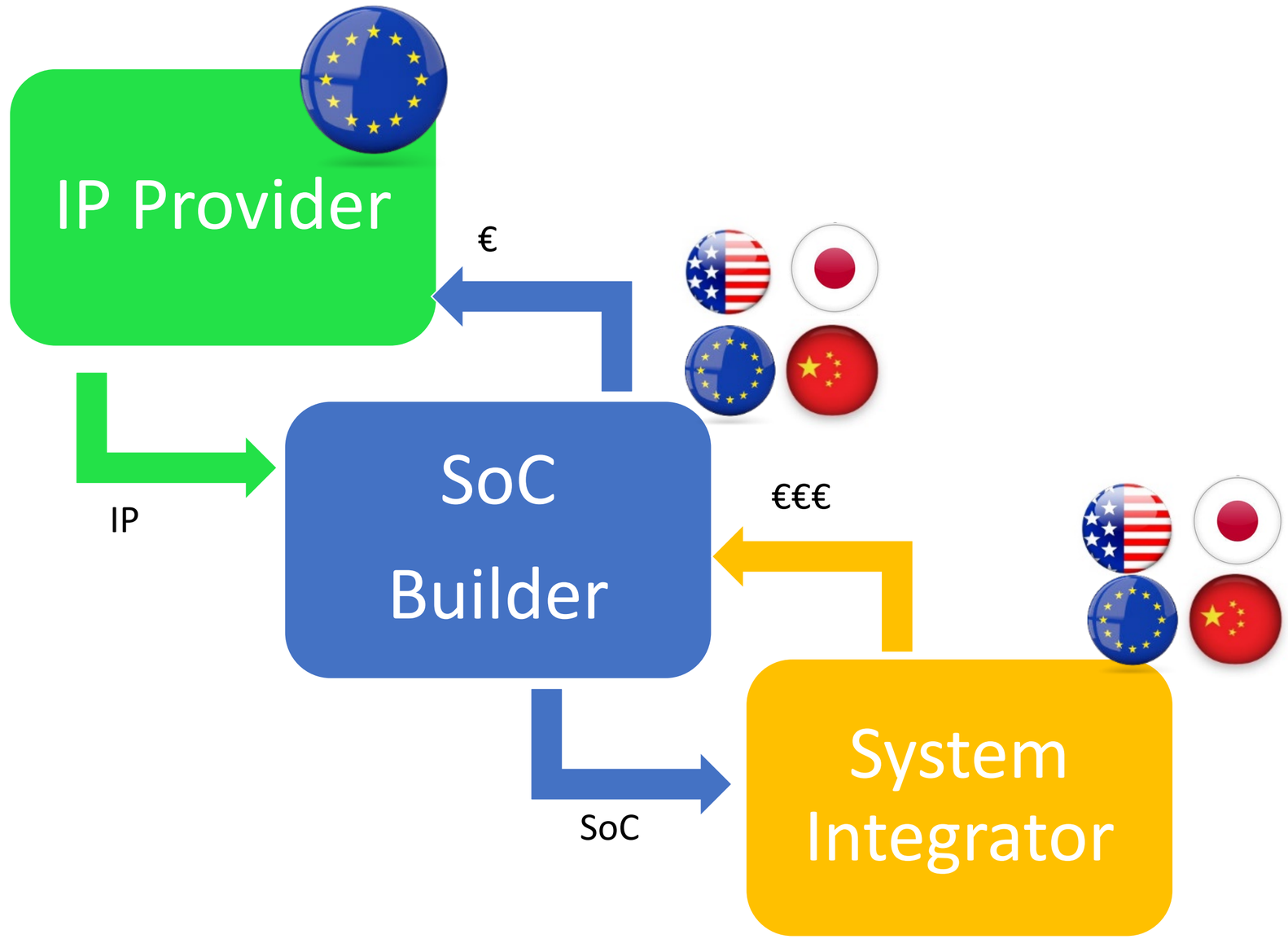
Unidad de renombrado de registros lógicos con detección de registros viejos prematuros

Máquina de ejecución de instrucciones de punto flotante compatible con formato IEEE 754 para la ejecución de operaciones escalares o vectoriales

Coprocesador para la aceleración de rutinas de los servicios de excepción en procesadores de propósito general

Mecanismo distribuido de reordenamiento de instrucciones de bajo consumo de energía para procesadores con emisión fuera de orden

JV PARTNERS



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

RISC-V ECOSYSTEM

SoC Builder

E2E

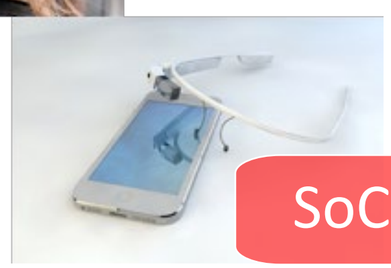
ExaScale

IP Provider

Supercomputing Centers

SoC Builder

Entertainment



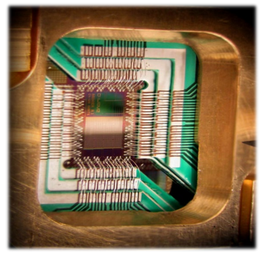
SoC Builder

System Integrator

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

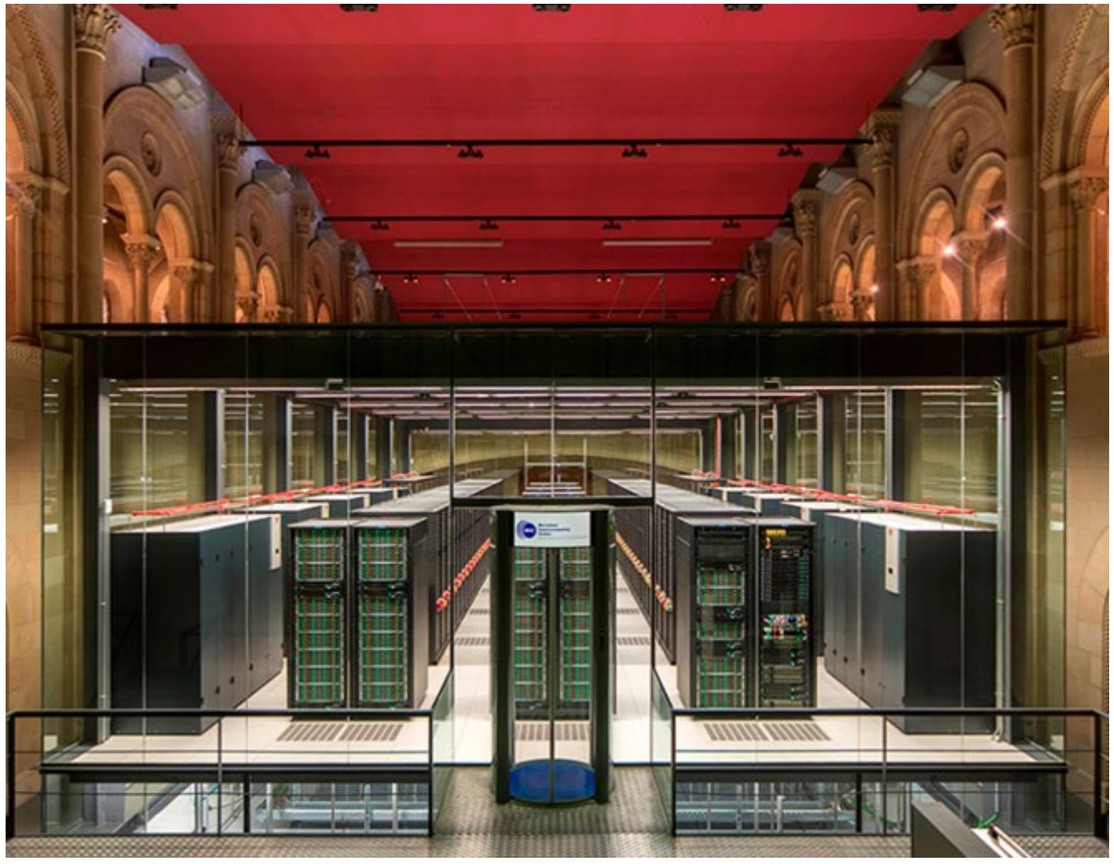
New Paradigms for Exascale

- Increasingly complex and heterogeneous hardware, including AI, Neuromorphic, Quantum and other accelerators. New memory systems and hierarchies
- Will require a monumental effort to redesign system software and applications for Exascale systems
- AI will be essential to harness the complexity of future Exascale systems



Conclusion

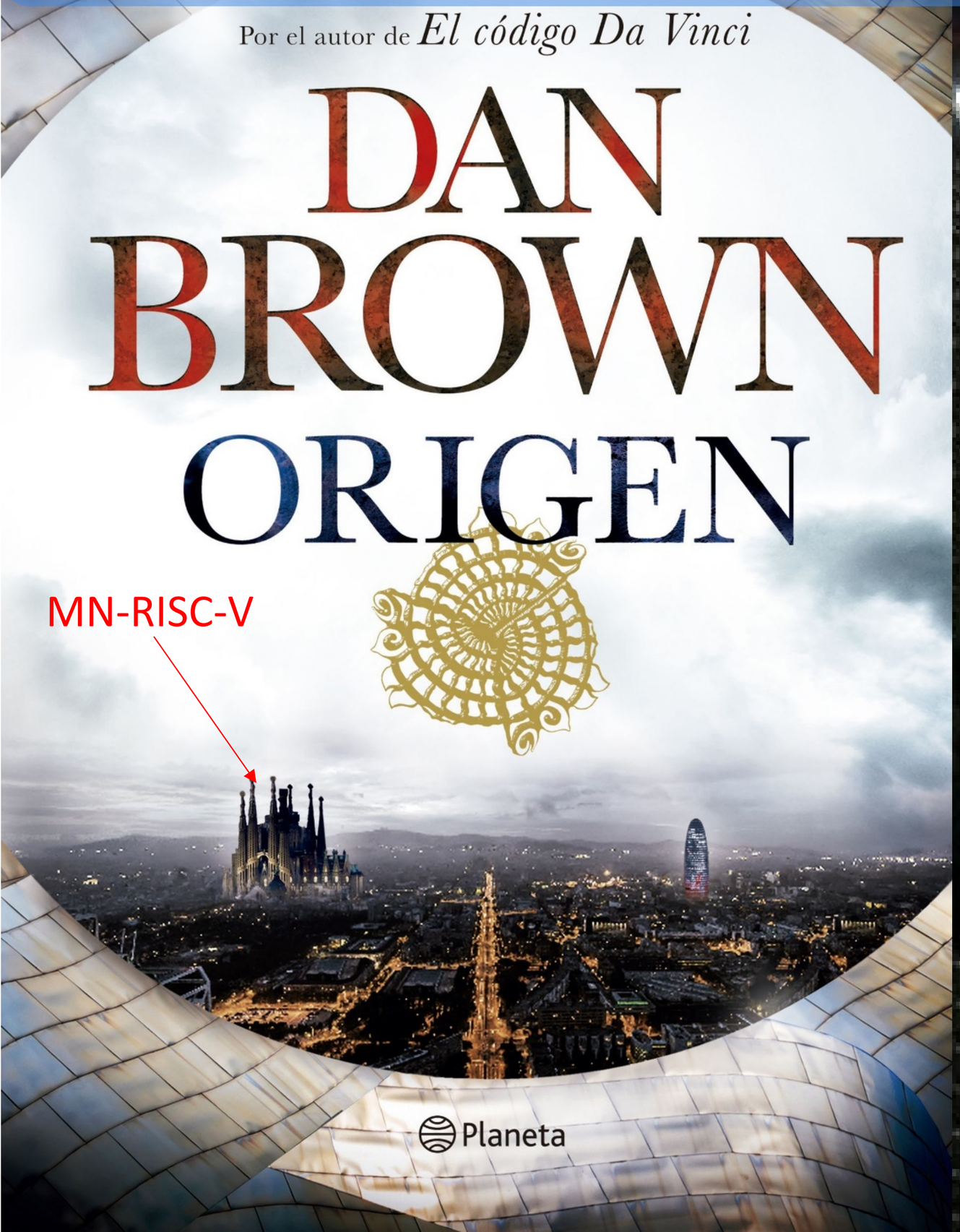
- HPC is crucial to resolve societal challenges and preserve European competitiveness
- Europe is going in the right direction with EuroHPC. This must be sustained in the long-term
- The chip design effort must continue for the EU's security and competitiveness, and should create a processor ecosystem covering IoT, servers, cloud, autonomous connected vehicles and HPC



EuroHPC opens a window of opportunity to create the Airbus/Galileo of HPC

EuroHPC
Joint Undertaking

Mare Nostrum RISC-V inauguration 202X



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Thank you

mateo.valero@bsc.es

02/2019