KEYNOTE

Forum TERATEC 23
Unlock the future

Adressing data challenges of the second wave of AI

Animée par James Coomer
14h15 – 14h30

James Coomer
Senior VP for products, **DDN**

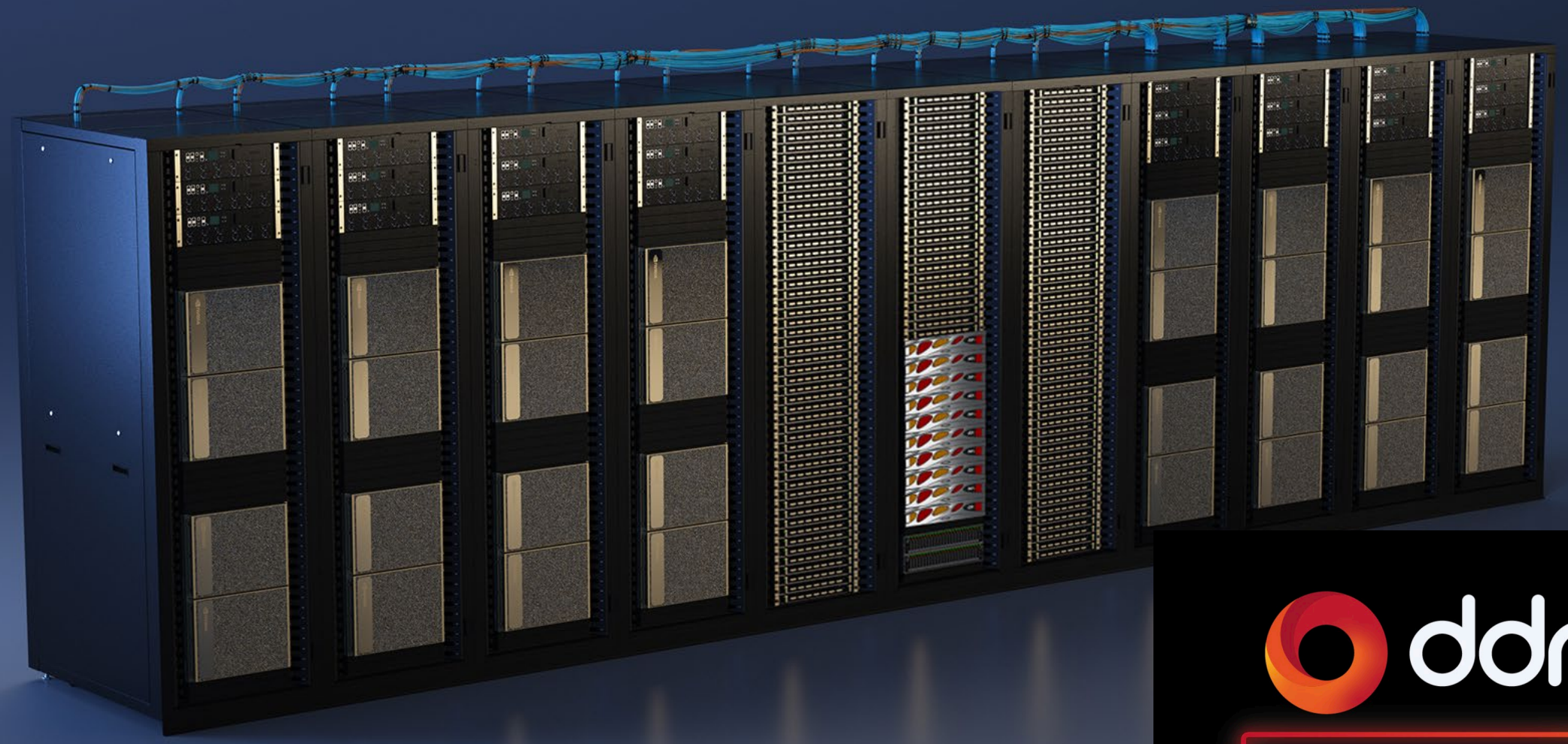
Forum TERATEC 23

Addressing the Data Challenges of the Second Wave of AI

ddn

The AI Data Company

DDN builds the Data Storage to drive Advances in AI, HPC, LifeSciences, Autonomous Driving and Finance



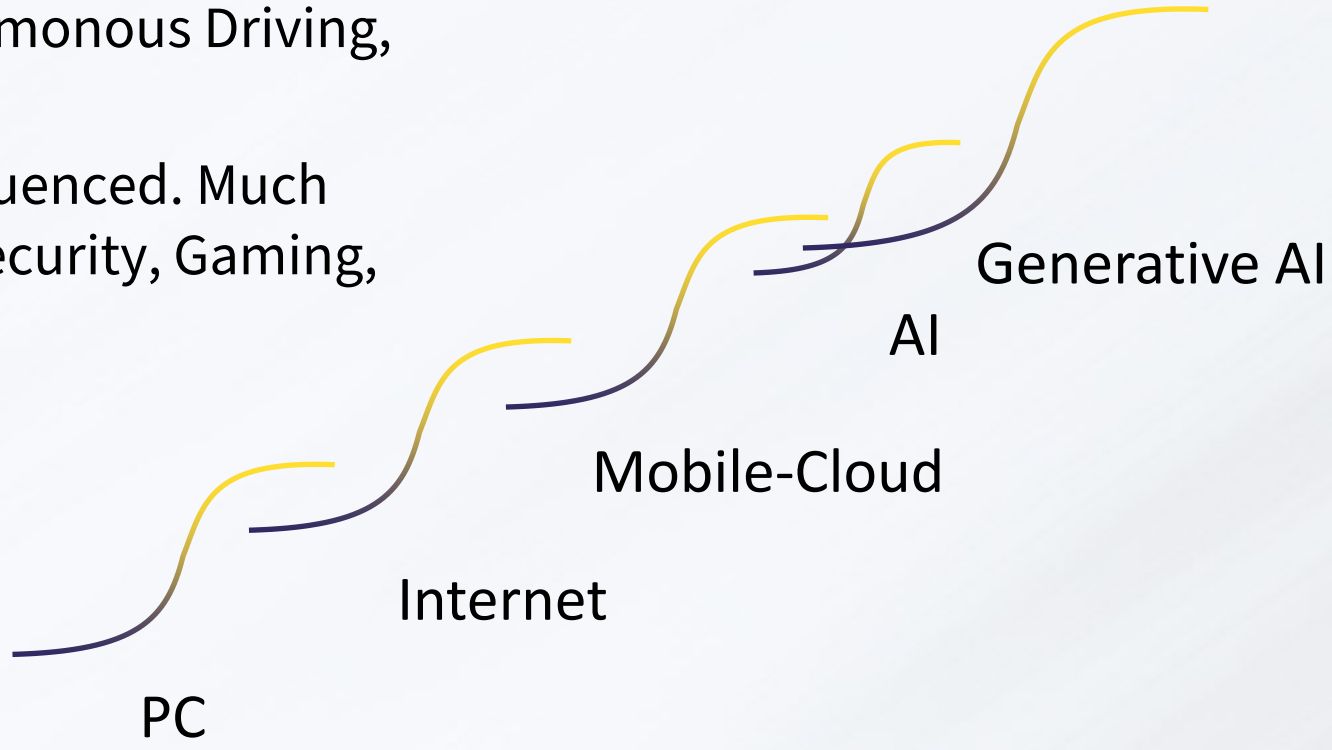
ddn

THE AI DATA COMPANY

What is the Second Wave?

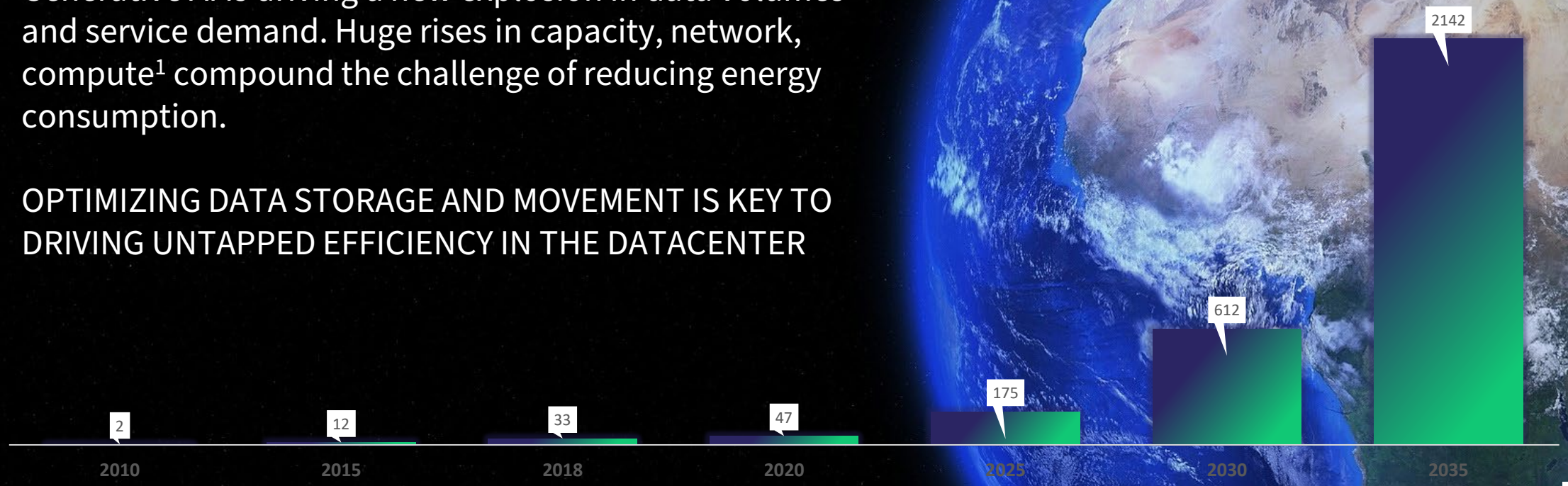
The Second Wave of AI

- ***1st Wave*** – NLP, Image and Video, Automonous Driving, Life Sciences
- ***2nd wave*** – Generative AI. ChatGPT influenced. Much Broader Take Up – All the Above plus Security, Gaming, Manufacturing, defence…



Datacenters Need to Become MUCH More Efficient

- Datacenters use 1% of electricity worldwide
- Generative AI is driving a new explosion in data volumes and service demand. Huge rises in capacity, network, compute[1] compound the challenge of reducing energy consumption.
- OPTIMIZING DATA STORAGE AND MOVEMENT IS KEY TO DRIVING UNTAPPED EFFICIENCY IN THE DATACENTER



Global Data Growth

[1]https://www.science.org/doi/10.1126/science.aba3758

What's the Difference in the second Wave?

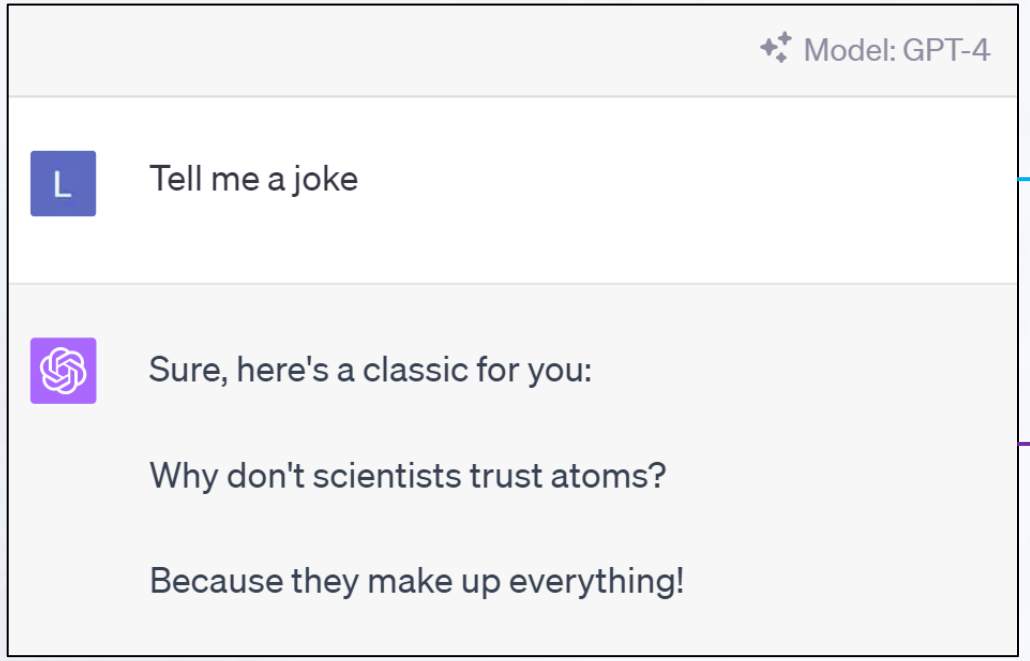
1. New Large Language Model methods creating many new opportunities for all Enterprises that use Data

2. Established **Accelerated Computing** creating economically viable ways of accessing the potential of AI with **MUCH better Efficiency** through
   1. New Levels of Parallelisation (GPU, Network, Storage)
   2. Full Stack Integration (App ← → Device)

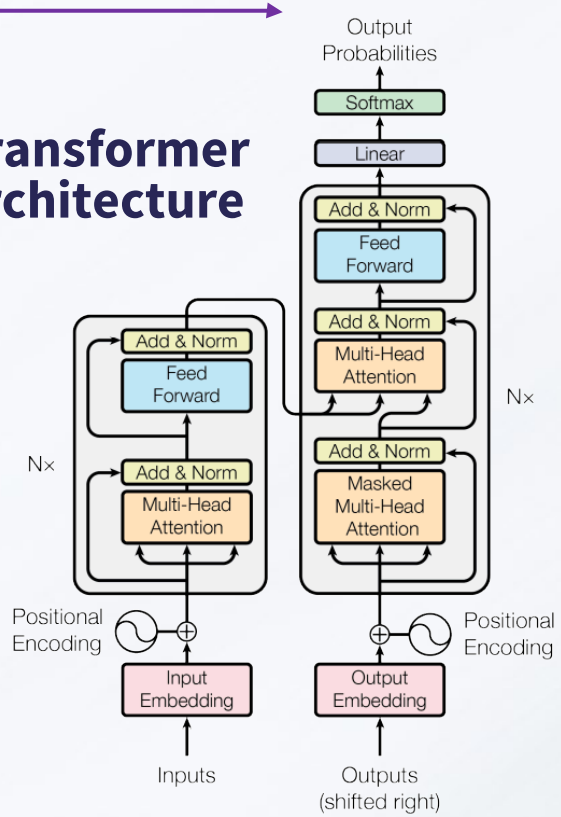
Challenges of Generative AI

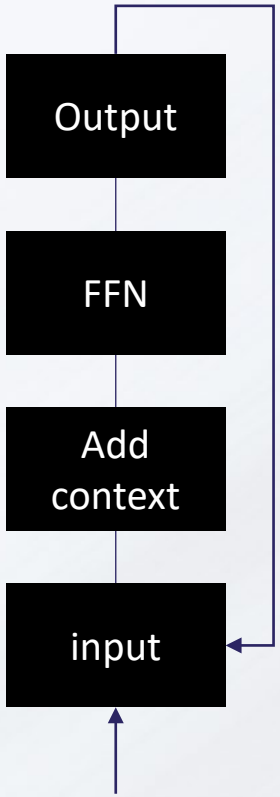
Natural Language Processing is driven by Transformers

ChatGPT



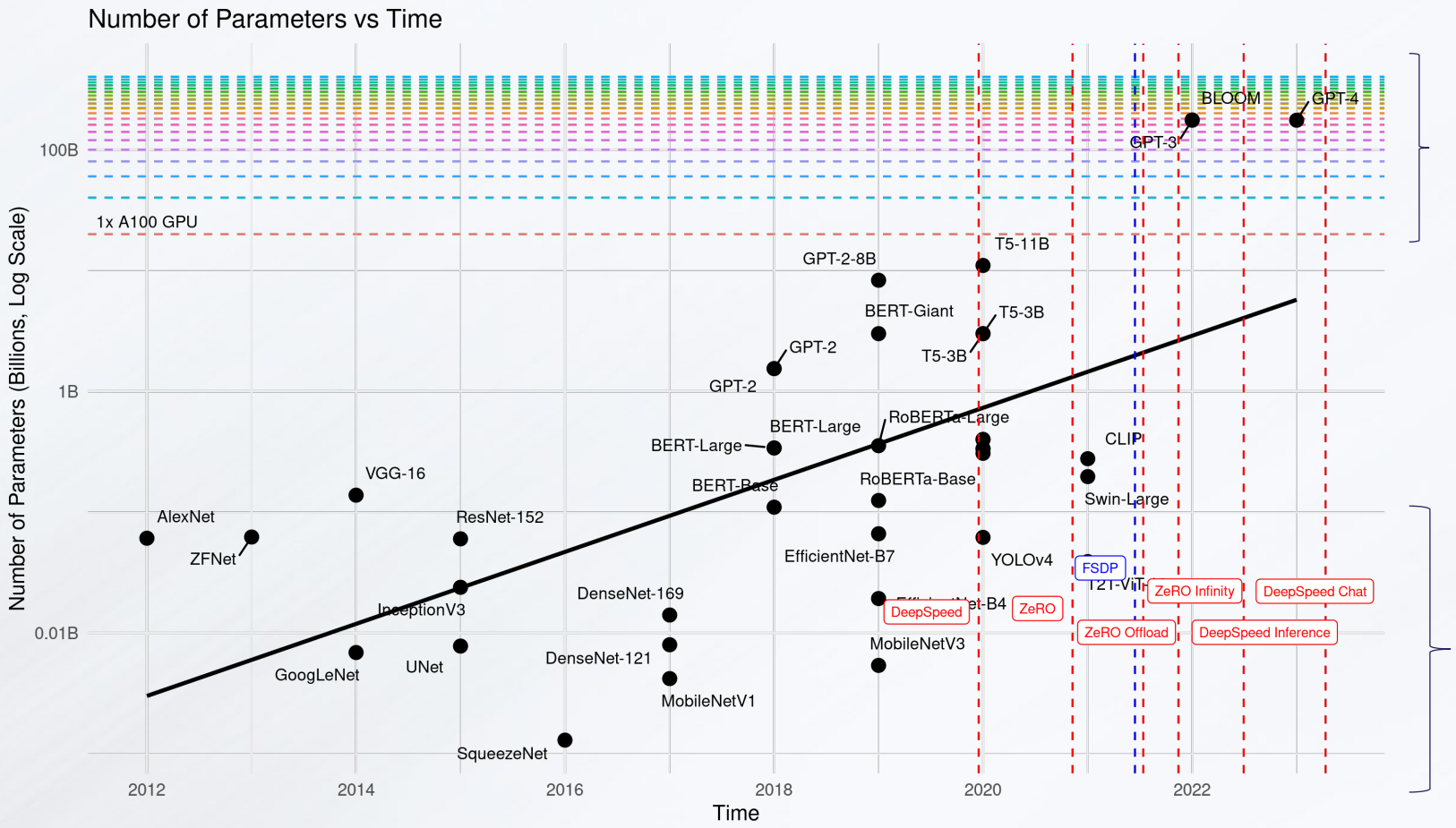
Transformer architecture





[ASH17] Ashish Vaswani, Noam Shazeer, Niki Parmar et al. *Attention Is All You Need*

Large Language Models and the GPU memory wall



In 3 yrs:
Model size Increased x1000
GPU memory increased x5

Software stacks (e.g: deepspeed) have been developed to handle the issue

- Better memory management (e.g: ZeRO)
- **Offloading (CPU/NVMe/Filesystem)**

Hyper Efficient Data Storage for NVIDIA BasePOD and SuperPOD

DDN Virtualization Technology **Eliminates cables, switches & servers**

Large Capacity Flash Drives **Doubles Capacity per Watt**

End-to-End Parallel Datapath **Doubles Performance per Watt**

Workload Focused Performance Optimization **Speeds up Applications**

100% linear scale out Software **Removes costly silos**

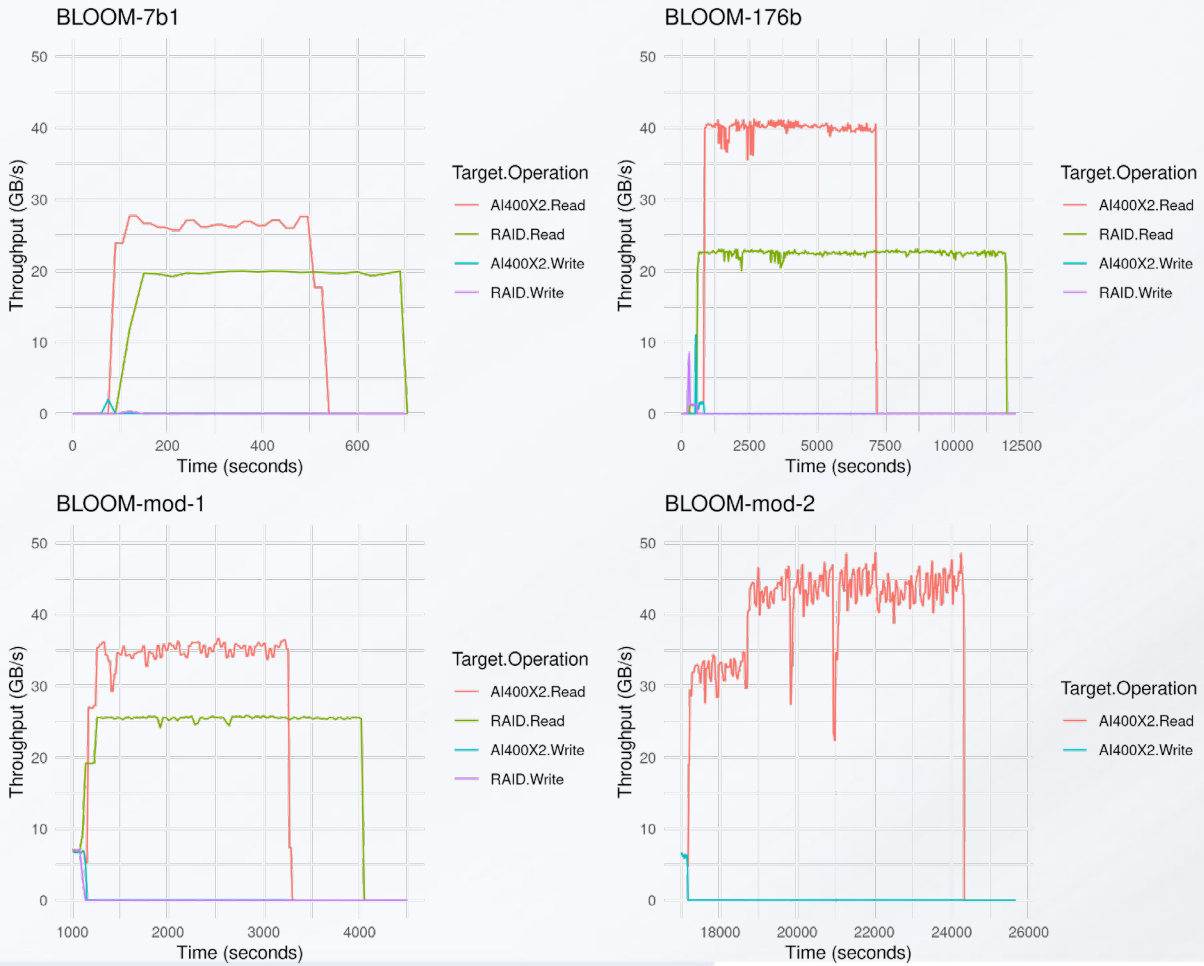
1.4PB Raw

LLM Offloading Experiment – IO throughput

The total amount of data transfered is the same between the local RAID and the AI400X2

The IO throughput determines the performance

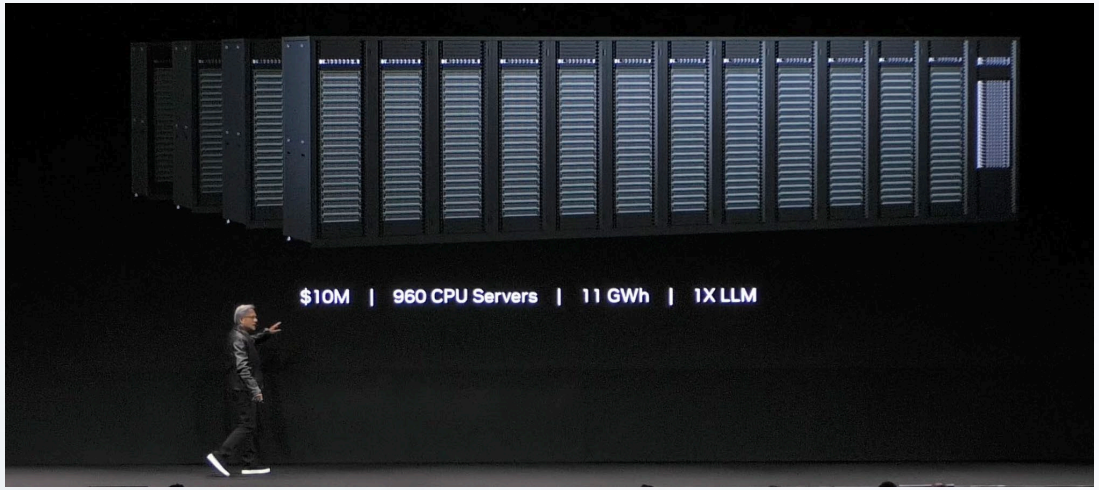
Transfer is overlapped with computation - It is a throughput problem



Opportunities of Accelerated Computing

$10m Computer.

BEFORE



AFTER – Accelerated Computing

1) Your Datacenter is power limited. (almost every DC is power limited)
2) You get 150x more performance with 3x more cost.

DDN gives you Up to 30x More Performance Per XXX

Traditional Systems

DDN Appliances

30x

MORE IOPs per Rack

10x

More Writes per Rack

6x

More capacity per watt

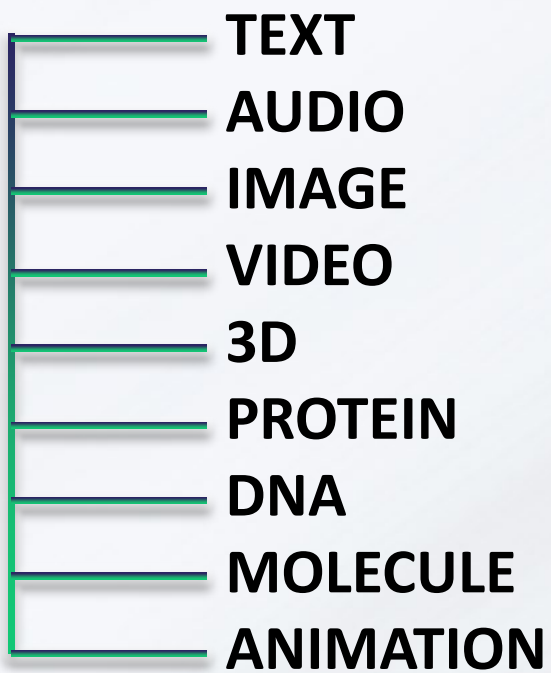
And These Systems ARE Tightly Coupled

[[AI Framework + GPU + Compute + Network + Filesystem + Storage]]

What’s Next?

What's Next?

- Now we can learn the language of any structured data – NDA, text, animation, protein.
- And now we can transform across languages. Text to chemicals. Images to text.
- A SW technology able to understand the representation of information of many modalities. We can now apply the instrument of our industry to areas that were impossible before.
- In order for AI to have a digital twin (remember so far only used for light industry (words, media, etc). For $50T heavy industry, manufacturing, pharma.. More has to be digitized.
- **We need to create the ability for their world to be created digitally.**



ddn

TABLE RONDE

Forum TERATEC 23
Unlock the future

Quelle stratégie l'Europe doit-elle adopter face au développement de l'IA générative ?

Animée par Julien Bergounhoux
14h30 – 15h30

Julien Bergounhoux
Rédacteur en chef,
L'Usine Digitale

Eric Bezille
Senior Presales Manager, Systems Engineering - CTO Ambassador,
Dell Technologies

Antoine Bordes
VP AI,**Helsing**

Laurent Daudet
CEO, **LightOn**

Pierre Puigdomenech
CEO, **Do It Now**

Stéphane Tanguy
CIO & CTO, **EDF**

KEYNOTE DE CLOTURE

Forum TERATEC 23
Unlock the future

De la recherche aux start-up, quelle place pour la France dans la course au quantique

Animée par Alain Aspect
15h30 – 15h45

Alain Aspect
Directeur de recherches, Emeritus Professor,
Prix Nobel de Physique 2022
CNRS

DISCOURS DE CLOTURE

Forum TERATEC 23
Unlock the future

La construction de l'Europe Numérique

Animée par Thomas Skordas
15h45 – 16h00

Thomas Skordas
Directeur Général Adjoint
Commission Européenne

European Commission

Forum TERATEC

Paris, 01 June 2023

Building the Digital Europe

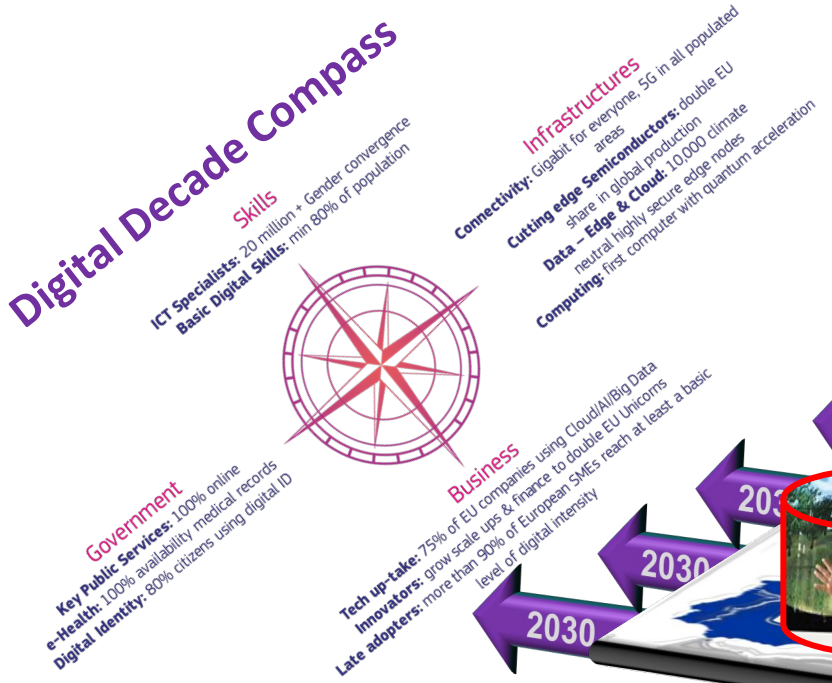
Thomas Skordas
Deputy Director General
DG CONNECT, European Commission

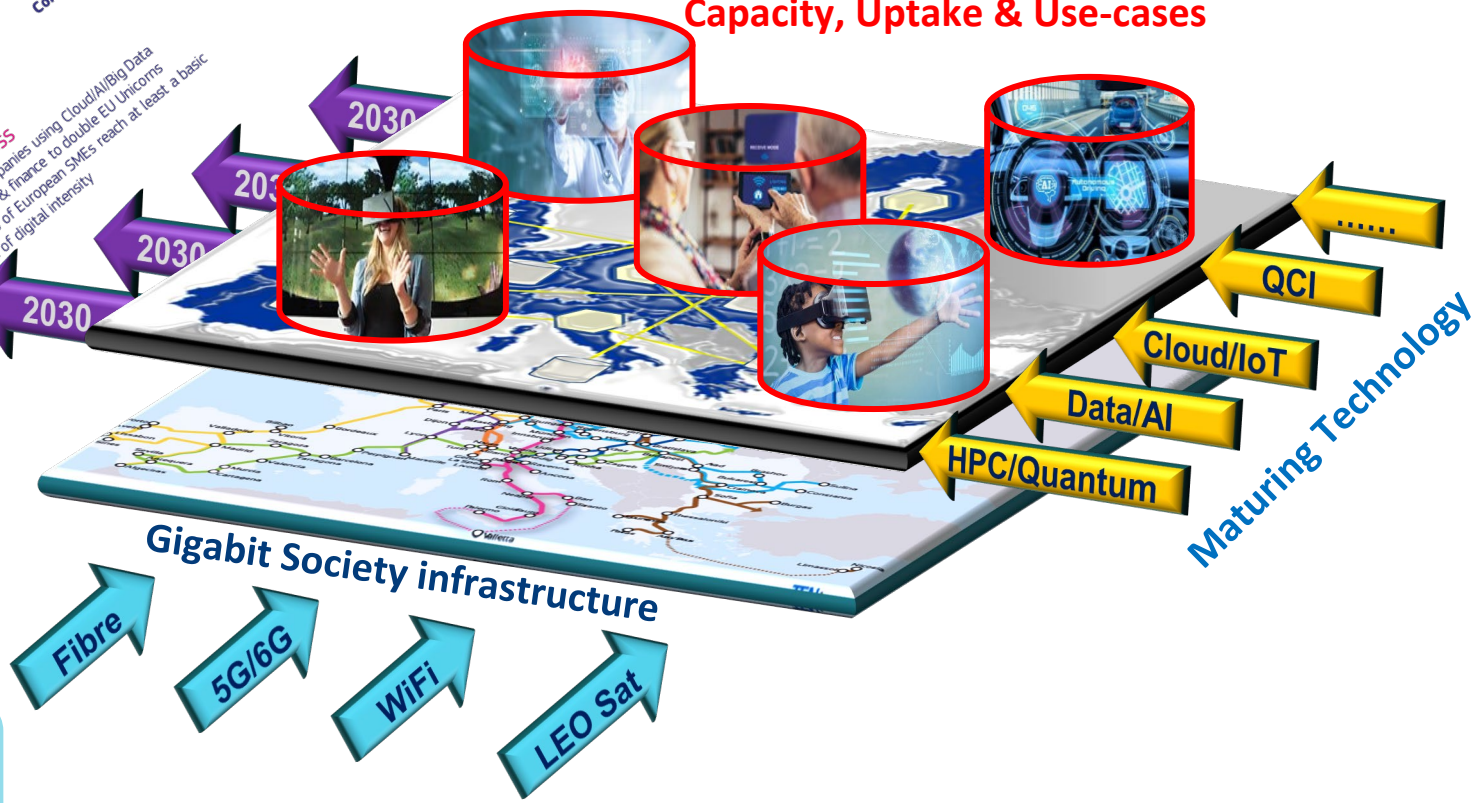
Digital in Europe (2021-27): ~20 B€

DIGITAL EUROPE (7.6 B€)

- EU-Wide deployment
- Strategic capacities
- Advanced digital skills
- Testing Facility

Capacity, Uptake & Use-cases





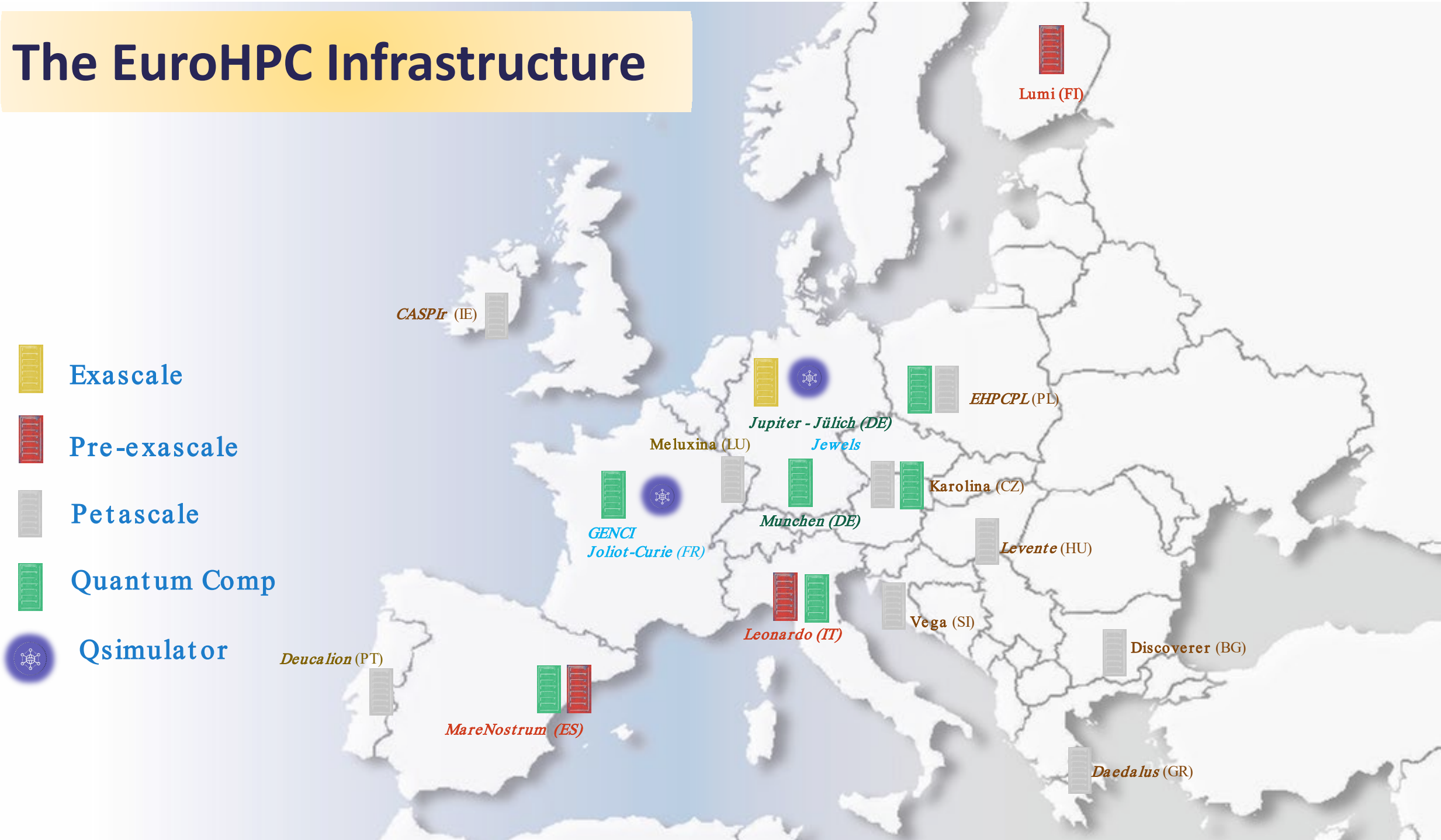
HORIZON EUROPE

- Preparing/maturing technology & progress (AI/Quantum…)
- By-design regulatory compliance (e.g. privacy friendly, unbiased AI)
- Leading & best in-class (e.g. strategic open autonomy, quantum)
- Prosperity, people & planet lens (green tech, societal challenges, industrial lead...)

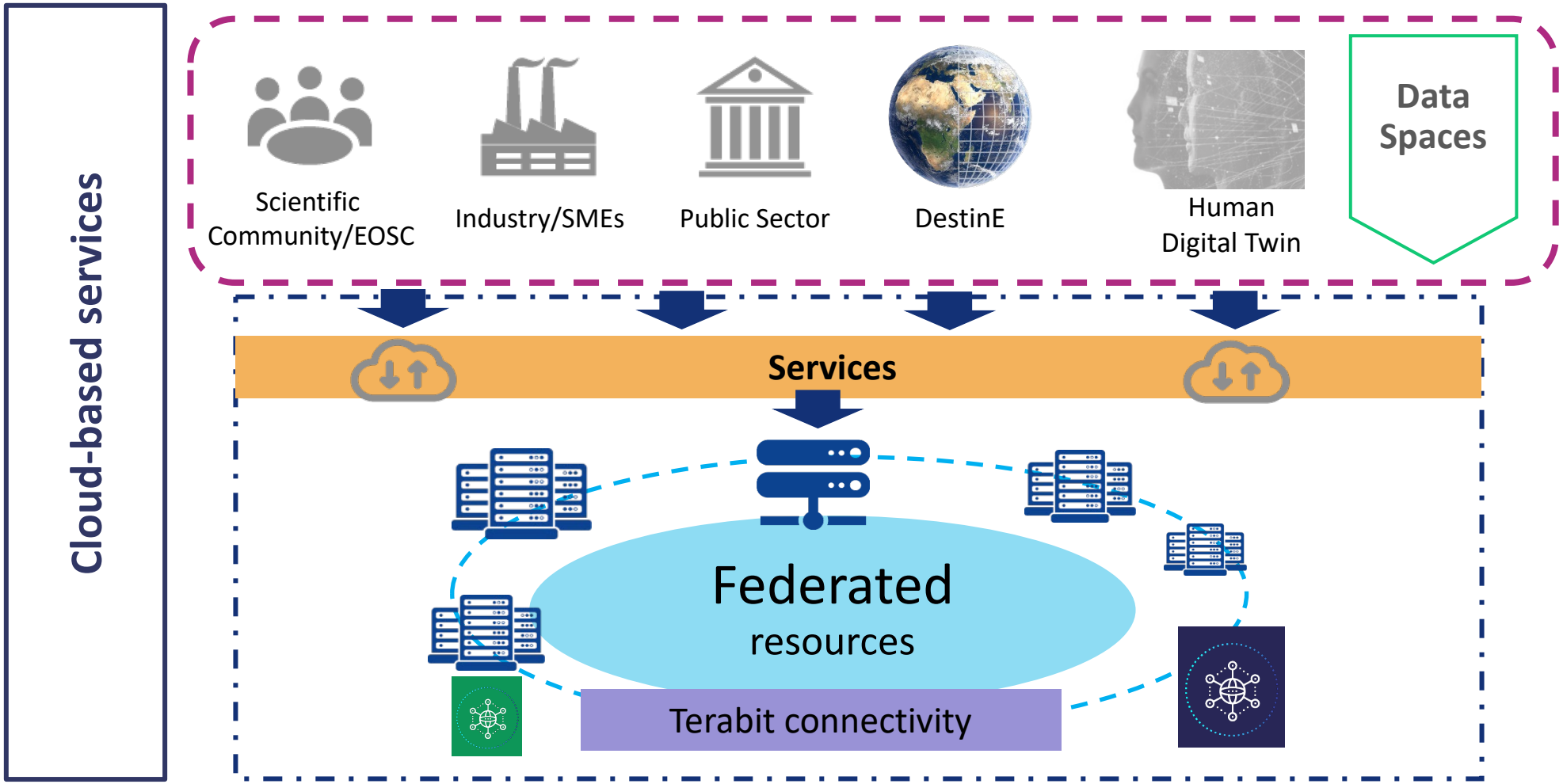
CONNECTING EUROPE FACILITY (2 B€)

- Strategic Backbones
- Fixed + Wireless Connectivity

The EuroHPC Infrastructure



EuroHPC: Hyper-connectivity & Federation



The EU's Chips Act: the 3 Pillars of Activity

EU + MS Investments: ~43 B€

European Chips Act

European Semiconductor Board (Governance)

Pillar 1

Chips for Europe Initiative

- Initiative on infrastructure building in synergy with the EU's research programmes
- Support to start-ups and SMEs

Pillar 2

Security of Supply

- First-of-a-kind semiconductor production facilities

Pillar 3

Monitoring and Crisis Response

- Monitoring and alerting
- Crisis coordination mechanism with MS
- Strong Commission powers in times of crisis

EU Investments in Quantum

~9 B€ Public funding in 2021-2027 (3 B€ EU + 6 B€ MS)

Equity Investments & Support to Start-ups

300 M€

IRIS2 & EuroQCI

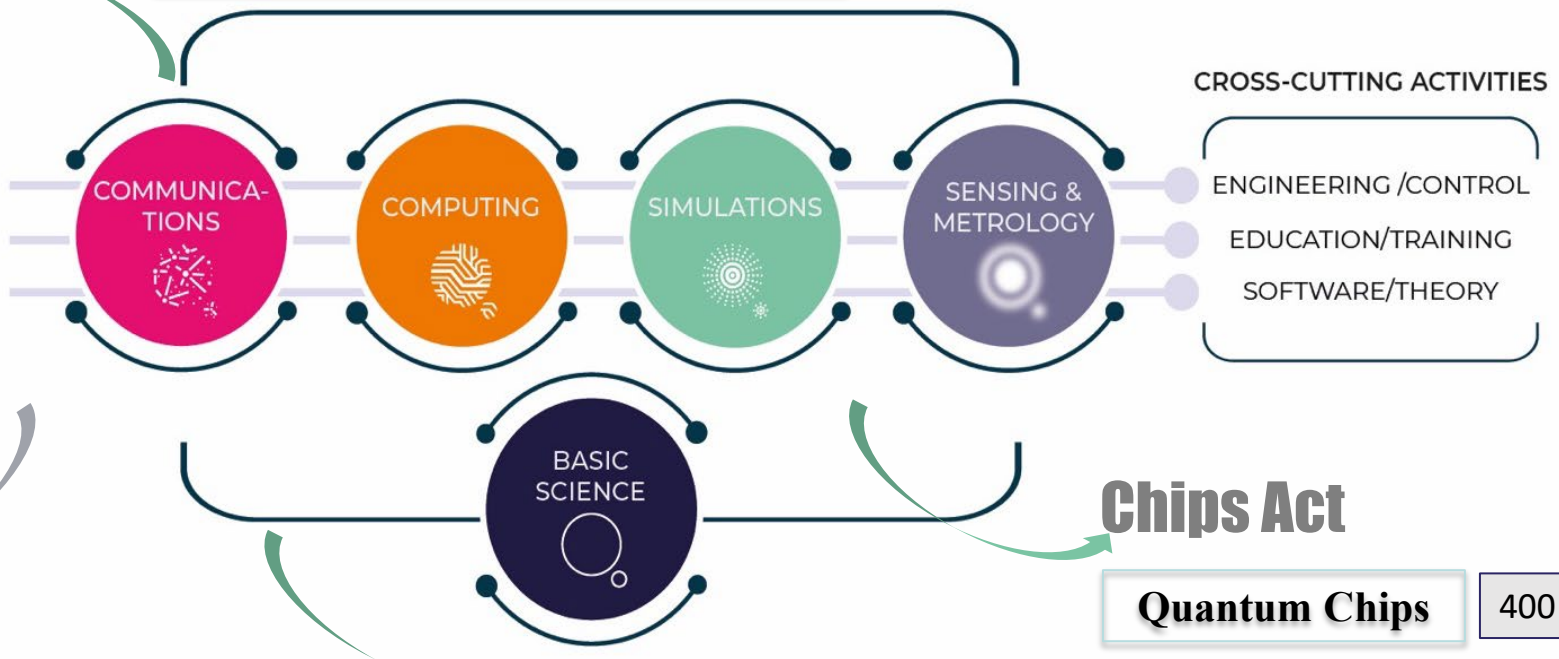
700 M€

300 - 400 M€

EuroHPC Joint Undertaking

1 000 M€

Quantum Flagship



Space Gravimetry

150 M€

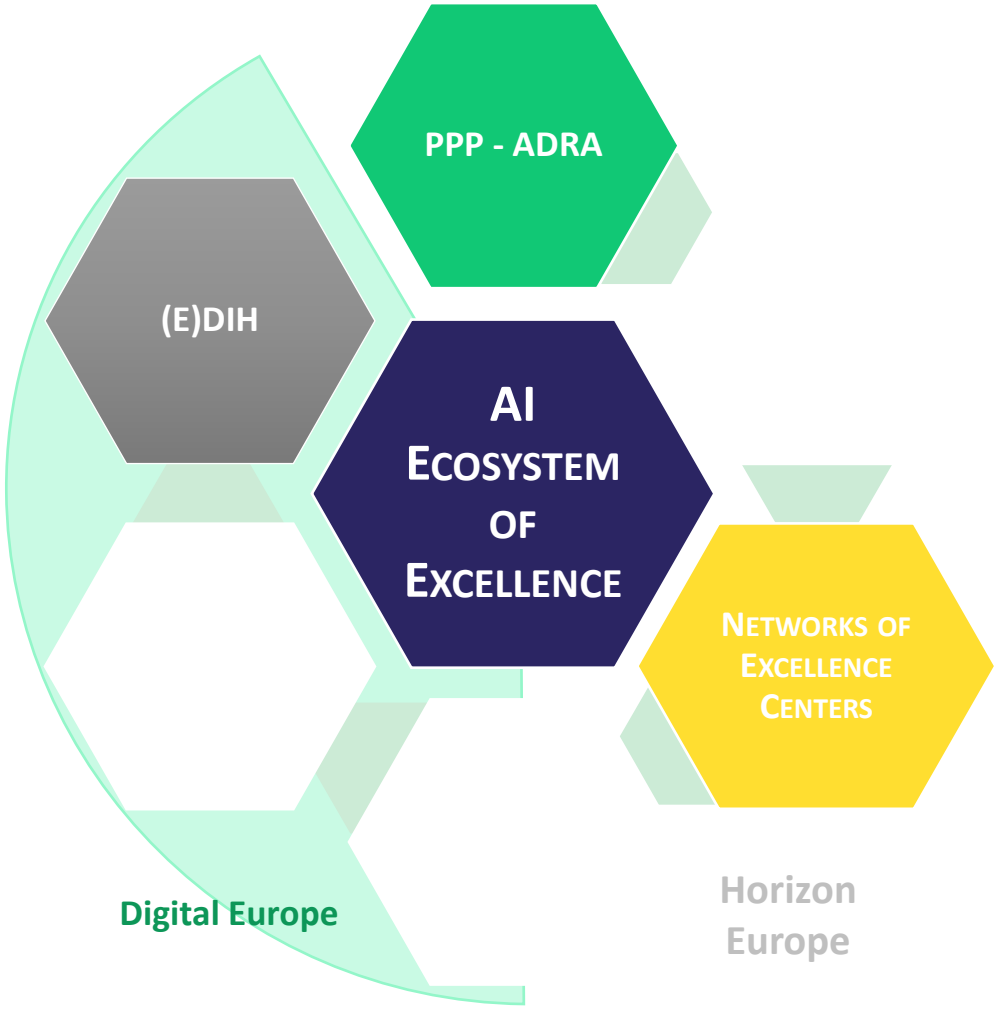
Chips Act

Quantum Chips

400 M€

Skills & Education

EU Excellence in AI: From the Lab to the Market



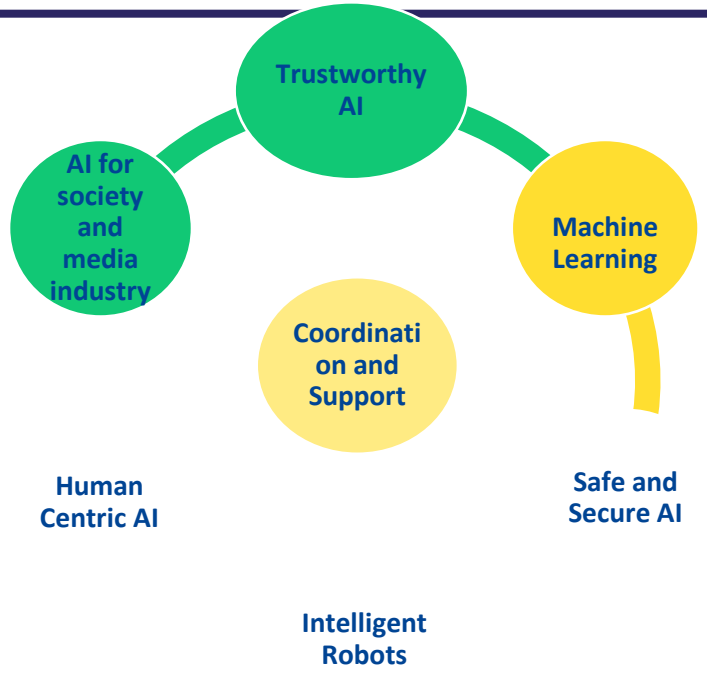
INVESTMENT COMMITMENTS (2021-2027)

➔ 1.3 B€ EU FUNDING

➔ 20 B€ investments by EU and MS

DIGITAL DECADE TARGETS:

➔ by 2030, 75% of European enterprises have taken up AI



Examples of EU initiatives in Digital Twins
Destination Earth and Destination Human

DestinE: Study the past, understand the present and predict the future

Building Virtual Human Twins

Powered by digital Research Infrastructures

- BRAIN DISORDERS
- CARDIOVASCULAR
- INFECTIOUS DISEASES

Cyber Security and the EU Cyber Solidarity Act

Cyber Solidarity Act

Strengthen solidarity at Union level in order to better detect, prepare and respond to cybersecurity threats and incidents

Pillar 1

- European Cyber Shield
- National SOCs
- Cross-border SOCs
- Information sharing

Pillar 2

- Cyber Emergency Mechanism
- Preparedness
- Response (EU Cybersecurity Reserve)
- Mutual Assistance

Pillar 3

- Cybersecurity Incident Review Mechanism
- Review and assessment of incidents
- Lessons learned and recommendations

Thank you!