The Impact of AI and GenAI on the Semiconductor Industry

Market Trends webinar #2

Website: https://allpros.eu

Twitter: https://twitter.com/allpros_eu

LinkedIn: https://www.linkedin.com/company/allpros-eu/

Overview & Partners

Mission: Reinforcing Europe's competitiveness in processors and semiconductors through secretariat coordination of the EU Industrial Alliance.

OBJECTIVES

Objective 1: Set up and maintain the secretariat of the Alliance

Objective 2: Promote EU sovereignty and competitiveness

Objective 3: Strengthen Europe's industrial capacities

Objective 4: Develop an engaged community

POLICY NETWORK & SYNERGIES

- 3 market trends reports
- 6+ Thematic Working Groups (TWGs)
- 6 landscape and gap analysis reports International partnerships
- 50+ synergies with relevant initiatives
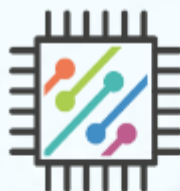- 50+ best practices showcased in the Observatory

REPORTS

- 5 EU Policy briefs produced on the basis of the TWGs' work
- 3 Impact reports
- 3 Recommendations Roadmap

OBSERVATORY → **ALLIANCE**

MARKETPLACE ← **TWGs**

COMMUNITY

- 1500+ engaged community members across all Stakeholder Groups
- Community DB of 2000+ contacts

OUTREACH

- 3 Alliance General Assembly Meetings
- 3 Annual Alliance Forums
- 18 Webinars
- 33 Newsletters
- 12 Digital Magazines
- 3 Press Releases
- 6 Marketing campaigns
- 20 3rd party events attended
- 3 general videos and 14 topic-specific

- Coordination and Support Action (CSA) project funded by the European Commission Digital Europe Programme.
- The project kicked off in January 2023 and will run for 36 months.

IDC
- Financial & Administrative Coordinator
- Observatory
- Market analysis
https://www.idc.com

Trust-IT Services *communicating to markets*
- Technical Coordinator
- Secretariat support
- Dissemination & comms
https://www.trust-itservices.com

COMMpla Communication Platforms and Online Solutions
- Platform development & maintenance
- Web & graphics
https://commpla.com

OpenForum Europe
- Engagement & synergies
- Open technologies
https://openforumeurope.org

WHITE
- Engagement & synergies
- Events management
https://white-research.eu

Today's agenda

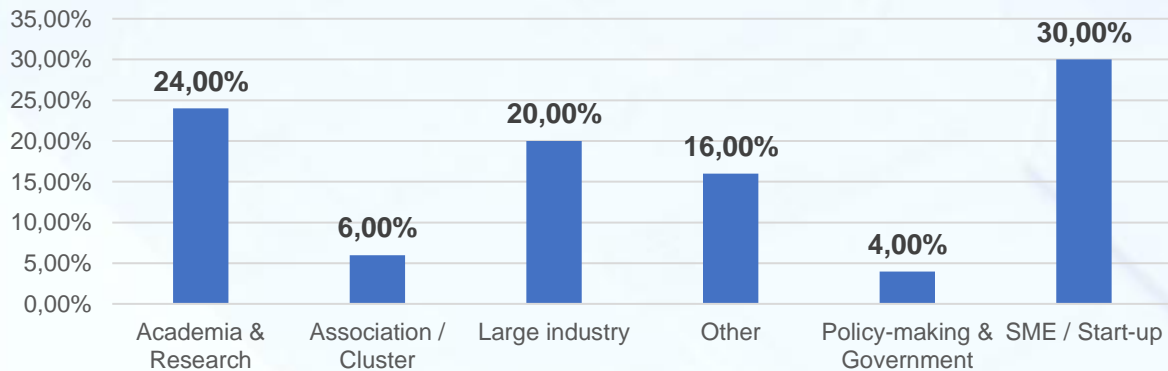| Time	Session
10:00-10:05	Welcome and Introduction by the European Commission
10:05-10:25	GenAI and the new silicon landscape
10:25-10:45	GenAI from the GenAI silicon or infrastructure vendor perspective *Beyond the Buzzwords: Demystifying AI*
10:45-11:05	GenAI from the vendor perspective *Bring AI to Your Data*
11:05-11:25	Wrap-up and Q&A

Thomas Reibe
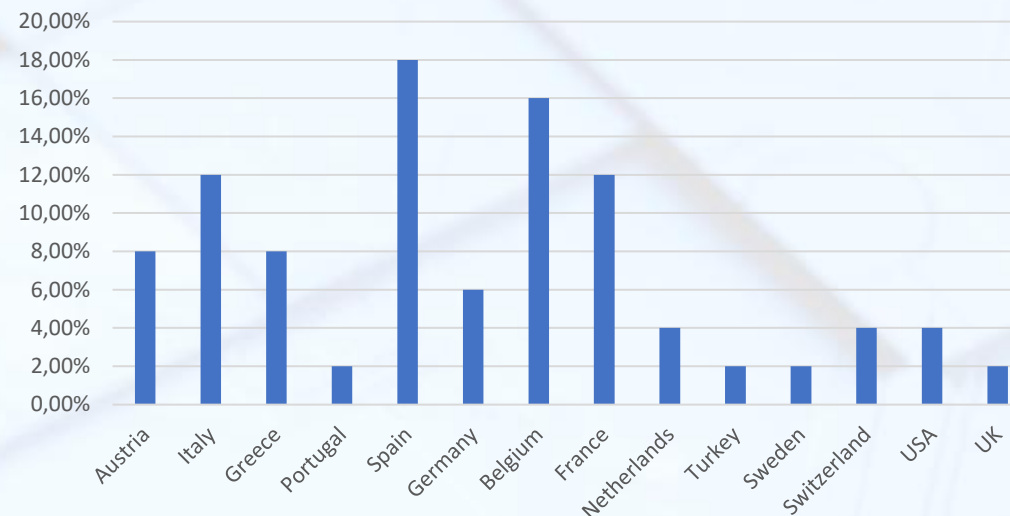European Commission

Luis Fernandes
IDC EMEA

Walter Riviera
Intel EMEA

Serban Zirnovan
Dell Technologies

Who's online and housekeeping

Stakeholder Type

Academia & Research: 24,00%
Association / Cluster: 6,00%
Large industry: 20,00%
Other: 16,00%
Policy-making & Government: 4,00%
SME / Start-up: 30,00%

Country

- Q&A – post your Question in the Q&A chat on the right of the screen
- At the beginning of your question please put the speakers name or topic to make it easier to track.
- Please use the chat for general conversation.
- Slides and recording will be available on the ALLPROS.eu website

Reach out to ALLPROS.eu for more details!

Website: https://allpros.eu

Twitter: https://twitter.com/allpros_eu

LinkedIn: https://www.linkedin.com/company/allpros-eu/

ALL PROS.EU

Secretariat for the European Alliance on
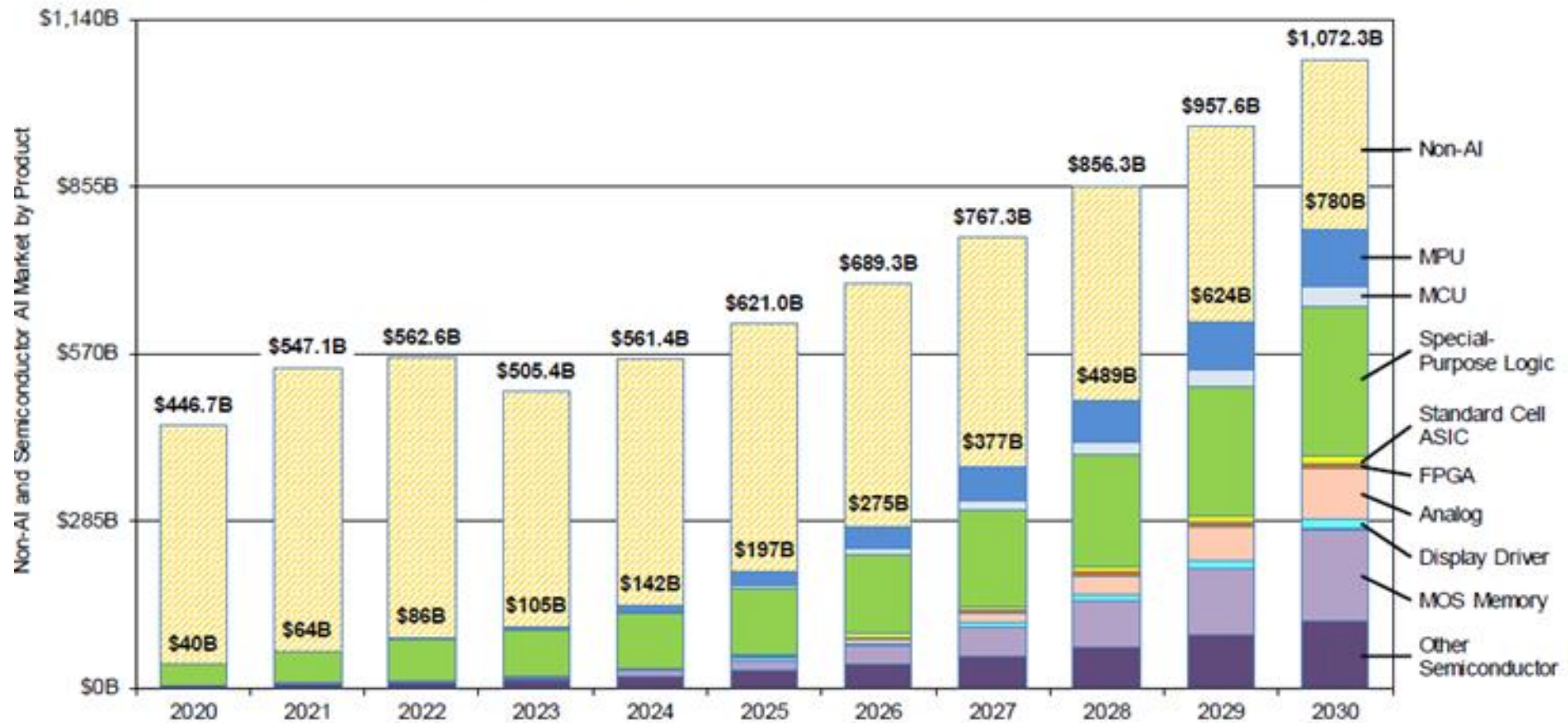Processors & Semiconductor Technologies

EU AI Chips

EC - DG CNECT

Semiconductor market driven by AI

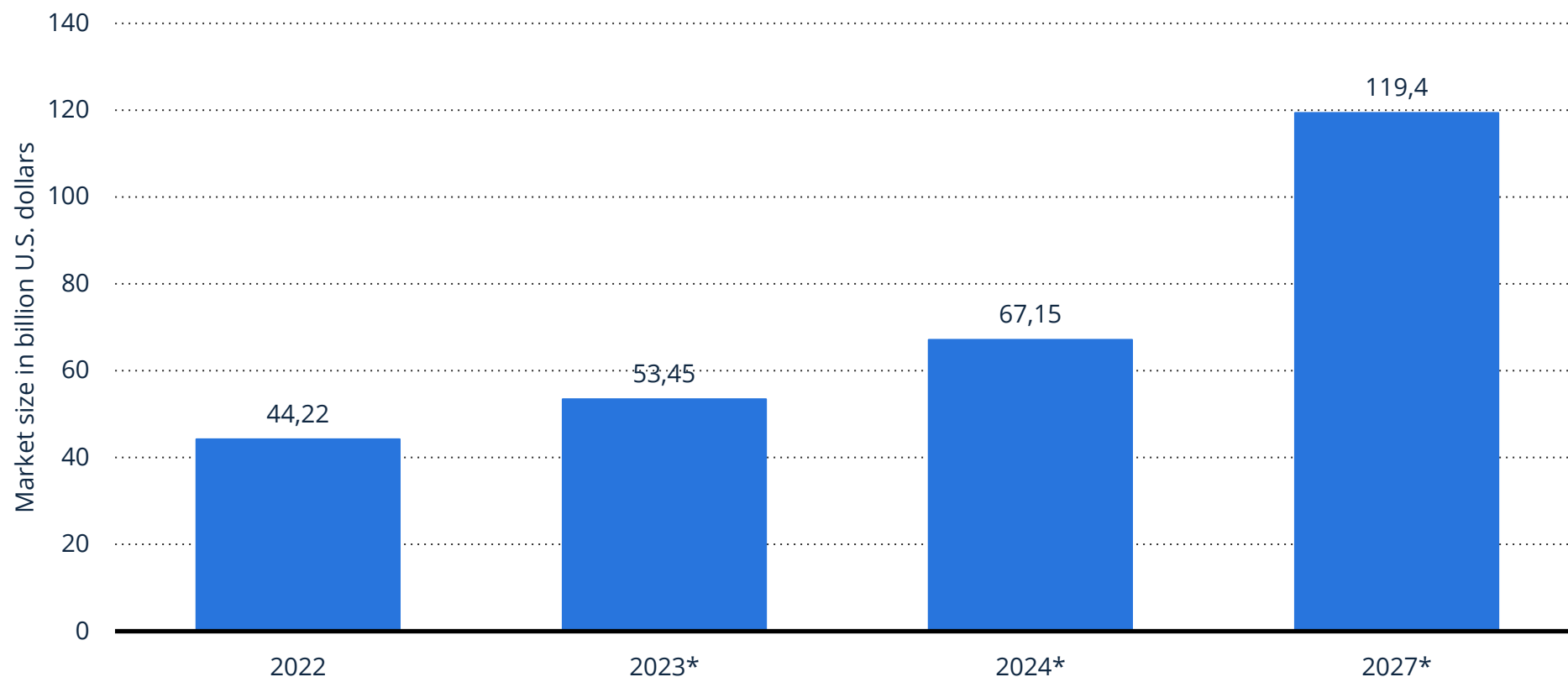Total Semiconductor Market and Semiconductor AI Market by Product

Source: International Business Strategy 2024

Semiconductor market driven by AI in 2030: 72%!

European Commission

Artificial intelligence (AI) chip market revenue

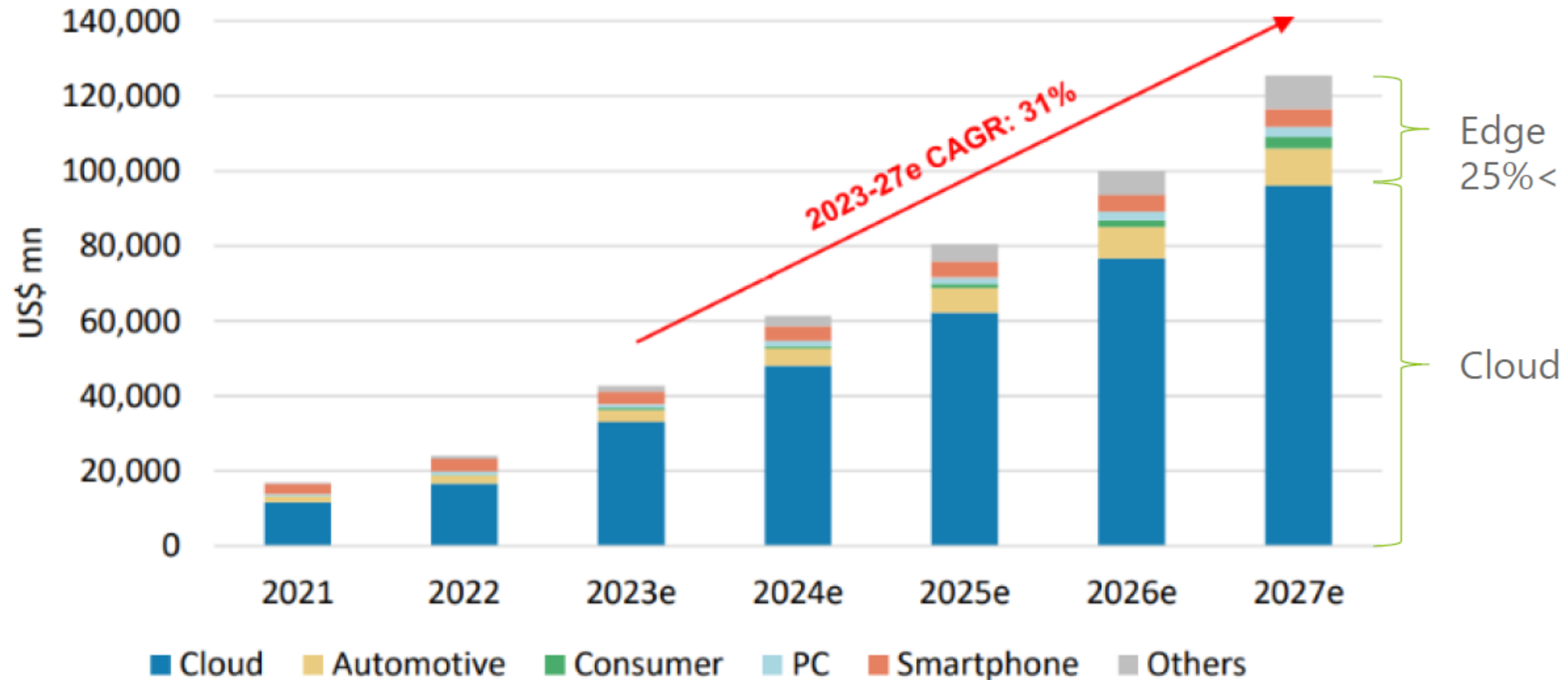AI chip market revenue 2022-2027 (in billion U.S. dollars)

Note(s): Worldwide; 2022 to 2023; * Forecast
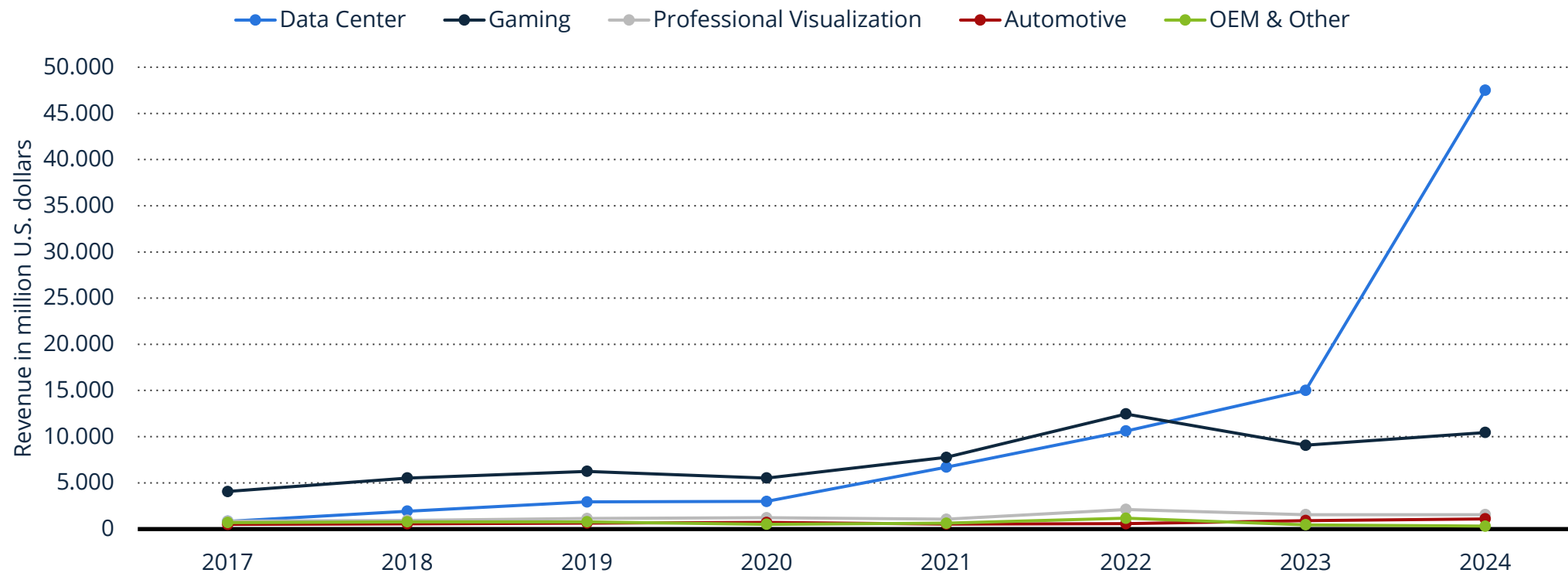**Source(s):** Gartner

AI chips revenue by application

CAGR: 31% - Edge AI to complement cloud computing

Source: Morgan Stanley Research estimates

Nvidia revenues

Nvidia revenue worldwide FY2017-2024, by specialized market

Legend: Data Center — Gaming — Professional Visualization — Automotive — OEM & Other
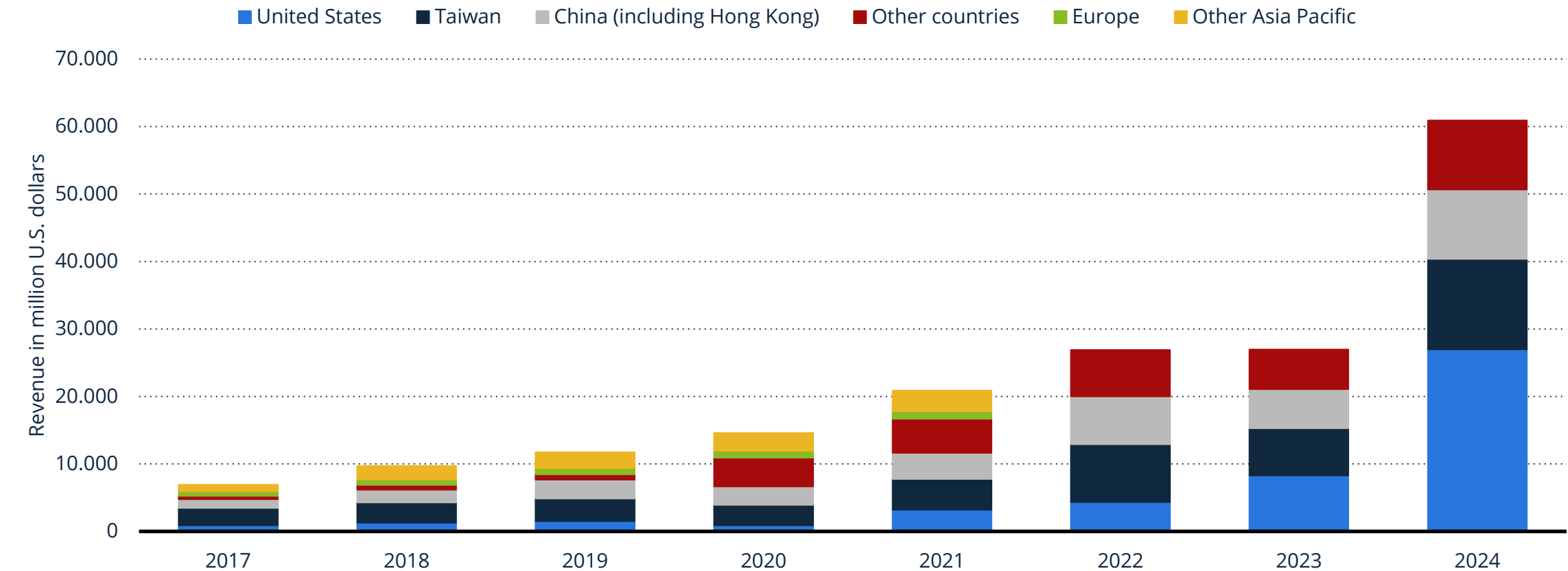
Description: In its 2024 fiscal year, Nvidia's revenue from data centers amounted to 47.5 billion U.S. dollars, whilst revenue from gaming amounted to 10.4 billion U.S. dollars. Nvidia`s technologies and solutions are being deployed for accelerated computing and artificial intelligence applications (AI), with Nvidia chips used to train and run a variety of large language models, including ChatGPT.
Note(s): Worldwide; FY2017 to 2024
Source(s): Nvidia

European Commission

Nvidia revenue worldwide

Nvidia revenue worldwide FY2017-2024, by region

Legend: United States, Taiwan, China (including Hong Kong), Other countries, Europe, Other Asia Pacific

Y-axis: Revenue in million U.S. dollars (0 to 70.000)

X-axis: 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024

Description: In its 2024 fiscal year, Nvidia's revenue in the United States amounted to 26.97 billion U.S. dollars, a huge jump from the 8.29 billion U.S. dollars seen in the previous fiscal year. Revenue in Taiwan amounted to 13.41 million U.S. dollars in the 2024 fiscal year, while China related revenue reached 10.31 billion U.S. dollars. Read more
Note(s): Worldwide; FY2017 to 2024
Source(s): Nvidia

In Europe: overview

- <u>No presence </u> in **datacentre** SoCs – lack of industry, lack of demand (1.3% of AI market)

- **Industry** – many solutions based on ARM cores, which developed architectures and instruction sets for ML and inference, for both data centres and edge (used by e.g. Apple, Qualcomm…):
  - EU **IDMs** (esp. NXP and ST) have implemented **edge** AI chips (microcontrollers);
  - **Automotive OEMs**: only working with US (Qualcomm, Nvidia) or IL vendors for central ADAS SoCs
  - Nokia and Ericsson are developing **advanced** AI SoCs for **6G** networks (e.g. @2nm)

- A range of **SMEs** have been focusing on AI processors with different approaches
  - Axelera (NL), Kalray (FR), Semidynamics (ES), Greenwaves (FR), Vsora (FR), Videantis (DE), Upmem (FR), Semron (FR), Prophesee (FR), Tachyum (SK), GML (NL), Innatera (NL), Think Silicon (EL), etc.

- EU is strong in **R&D** for low-power, embedded **edge AI** solutions, including in-memory computing and neuromorphic computing:
  - **RTOs**: IMEC (BE), CEA (FR), Fraunhofer (DE), VTT (FI), TNO (NL), CNR (IT)
  - **Universities**: Uni. Bologna, PoliMi (IT), TU Munich, Heidelberg, Karslruhe (DE), TU Eindhoven, TU Delft (NL), Aalto (Fi), KU Leuven (BE)… (+ ETH, EPFL, Cambridge, Imperial…)

European Commission

AI chip companies

- In **China**, AI chips are now a top strategic priority
  - AI chips have been developed by: Huawei, Alibaba, MetaX, Biren, Hygen, Cambricore
  - SMIC is now prioritising 7nm capacity to AI chips of Huawei for AI data centres

- In **Europe**, AI chip companies are struggling:
  - Promising technology, but unable to grow and move to advanced nodes
  - Lack of EU VCs means they need to accept foreign investors (mostly US and China)
  - The issue is mostly with Chinese investors acquiring IP, examples:
    - **UPMEM**: process in-memory solution (10x more efficient and 27x cheaper than Nvidia H100; interest from Huawei, Xiaomi and Oppo. Huawei has convertible loan, maturity in Nov'24, if not paid back Huawei gets company's IP
    - **Prophesee**: neuromorphic vision product, cooperation with Sony. Got to last round at EIC, then rejected by jury. Next round of investment all signed up by Chinese companies, a state-owned fund is the lead investor

European Commission

AI chips – EU strategy

- **Goal**: EU to become a leader in:
  - Devices that deploy AI models **inference**
  - Devices that operate in systems at the edge of the network (**edge AI**) - including autonomous driving
  - Next generation of ML accelerators for data centres
- **How**:
  1. Support the growth of champions in this field, also through creation of ecosystems involving users and system companies (creating internal demand)
  2. Leverage on R&D excellence, particularly in new power-efficient edge AI technologies (e.g. in-memory, analog and neuromorphic computing)
  3. Identify different streams: e.g. inference, embedded, 6G, automotive, ultra-low power
  4. Coordinated EU hardware and software strategy (including libraries, compact generative AI models, cloud, open source HW/SW/API)

European Commission

GenAI and the new silicon landscape

Luis Fernandes – Senior Research Manager
EMEA: Future of Digital Infrastructure

Importance to business success and scale of adoption of Performance Intensive Computing

Need for dedicated infrastructure

AI

Yes
93%

| Importance | Important - 25% | Critical – 38% |
| Scale | Minor Workload - 21% | Major workload – 56% |

BDA

83%

| Importance | Important - 39% | Critical – 30% |
| Scale | Minor Workload - 35% | Major workload – 31% |

HPC

83%

| Importance | Important - 38% | Critical – 29% |
| Scale | Minor Workload - 29% | Major workload – 34% |
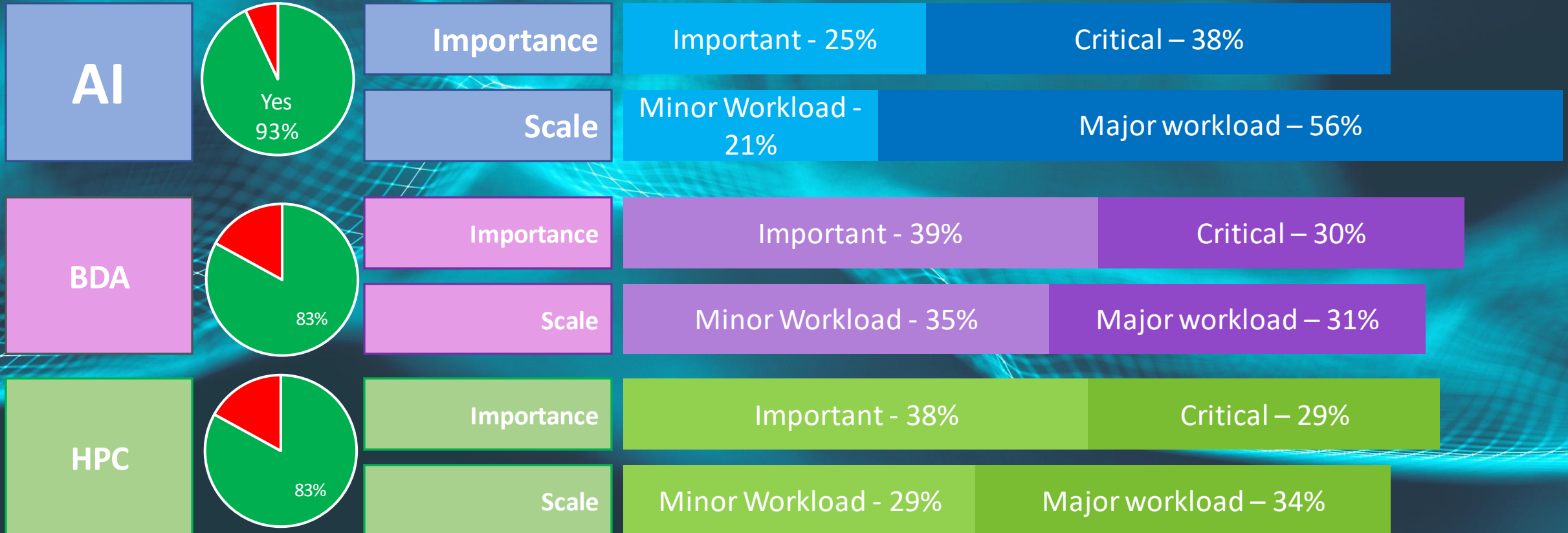
Generative AI Is Hot, and It's Already Here

Of large EMEA organizations polled in July 2023, …

23%
Investing significantly in Generative AI in 2023

58%
Exploring potential use cases

Customer and Employee Experience

Digital Assistants and Virtual Agents

Marketing Content Creation

Conversation Summarization

Translation

Software Development and Delivery Lifecycle

Code Creation and Pair Programming

Test Creation

Business Process and Task Automation

Conversational Query / Configuration

Knowledge Management

CoPilots for Knowledge Workers

Commercial Document Creation (e.g. RFIs, RFPs)

Report Generation

Content Translation

Document Summarization

NVIDIA and Microsoft are the top Tech Titan performers of 2023

Tech Titans With Annual Global Revenue > US$ 10 Billion
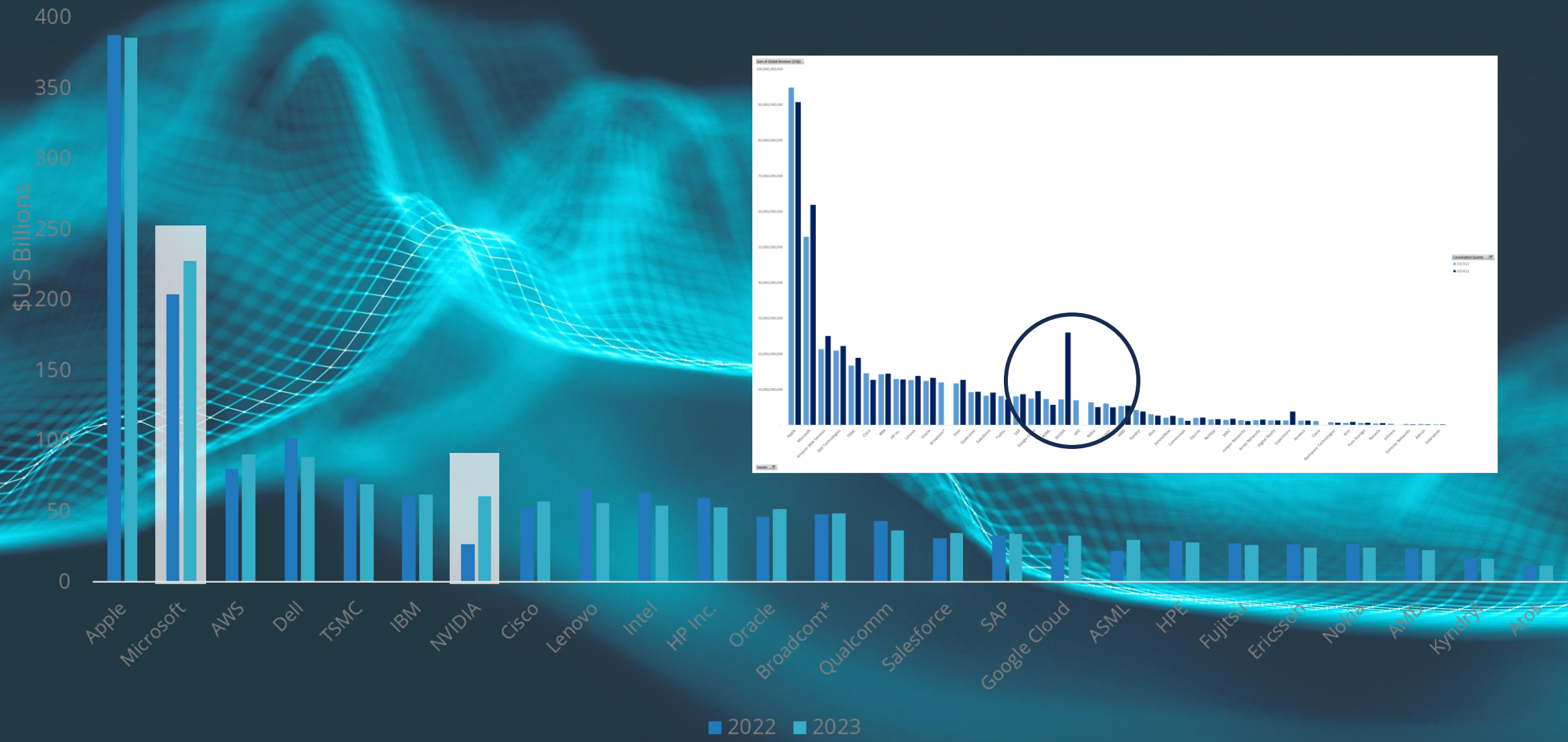
Source: Company public financial reports

■ 2022 ■ 2023

Is there enough opportunity and margin left to satisfy key partners?

Top 10 EMEA Public Cloud IaaS+PaaS Share Snapshot 1H23

Total Market: $26.6 billion (+26% YoY)

1H23 Market Revenue ($M)

1H22-1H23 Growth Rate (%)

Amazon Web Services
Microsoft
Google
Atlassian
Salesforce
IBM
Oracle
Orange
SAP
Snowflake
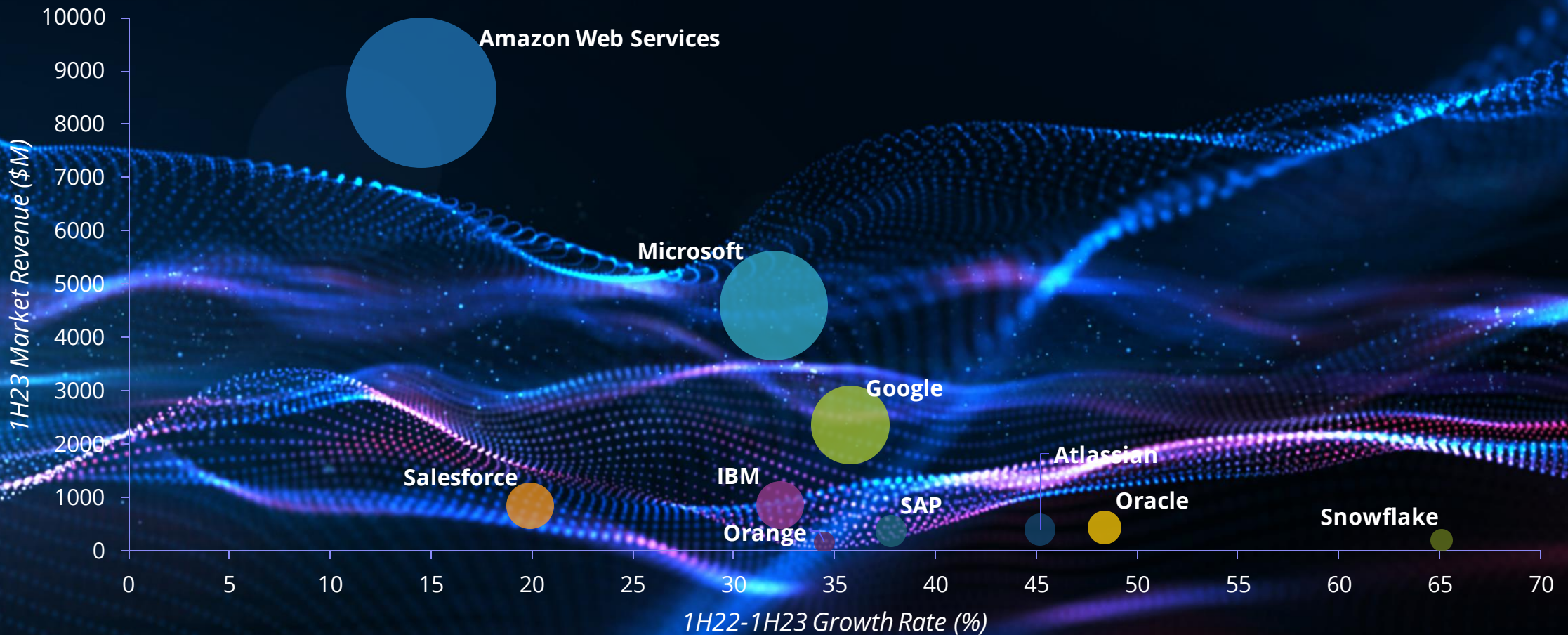
Europe loves its private IT infrastructure...
*but private cloud needs **a lot of attention** to be competitive with public cloud*

Use private IT infrastructure only

Prefer to use private IT infrastructure, use public cloud if necessary

Balanced across private IT infrastructure and public cloud

Prefer to use public cloud, use private IT infrastructure if necessary

Use public cloud only

17%

38%

36%

8%

1%
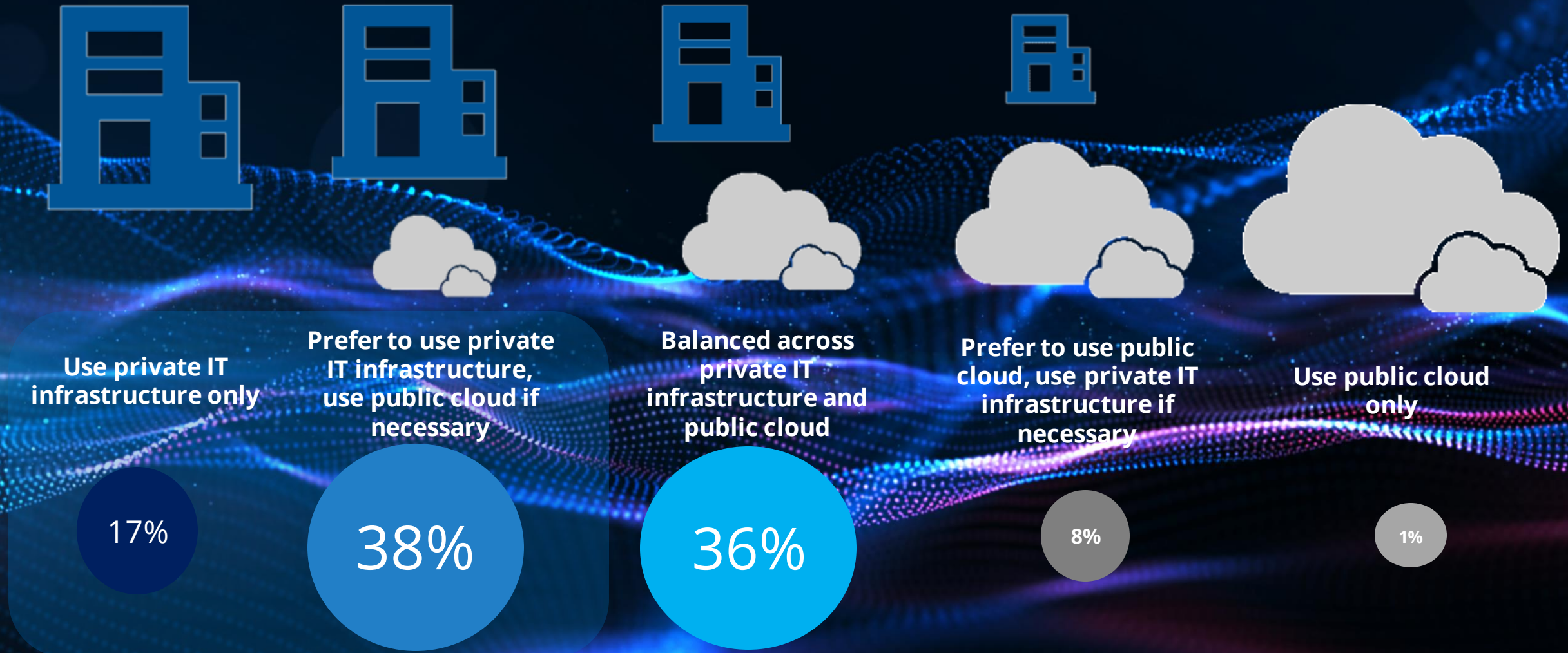
Cloud will be fundamental for kick-starting AI projects, but Private AI is currently the long-term production plan

primary cloud-for-AI strategy

- 18% — Use public cloud just to learn; Run pilot projects/PoCs and build test, development, and production AI solution in private IT
- 45% — Use public cloud to learn, run pilot projects/PoC, and build test and development environments; Run the final production AI solution in private IT
- 27% — Use public cloud to learn, run pilot projects/PoC, build test and development environments, and run the production AI solution
- 10% — Use AI embedded in SaaS (ready-to-use) solutions.

Private AI strategy

Public Cloud AI strategy
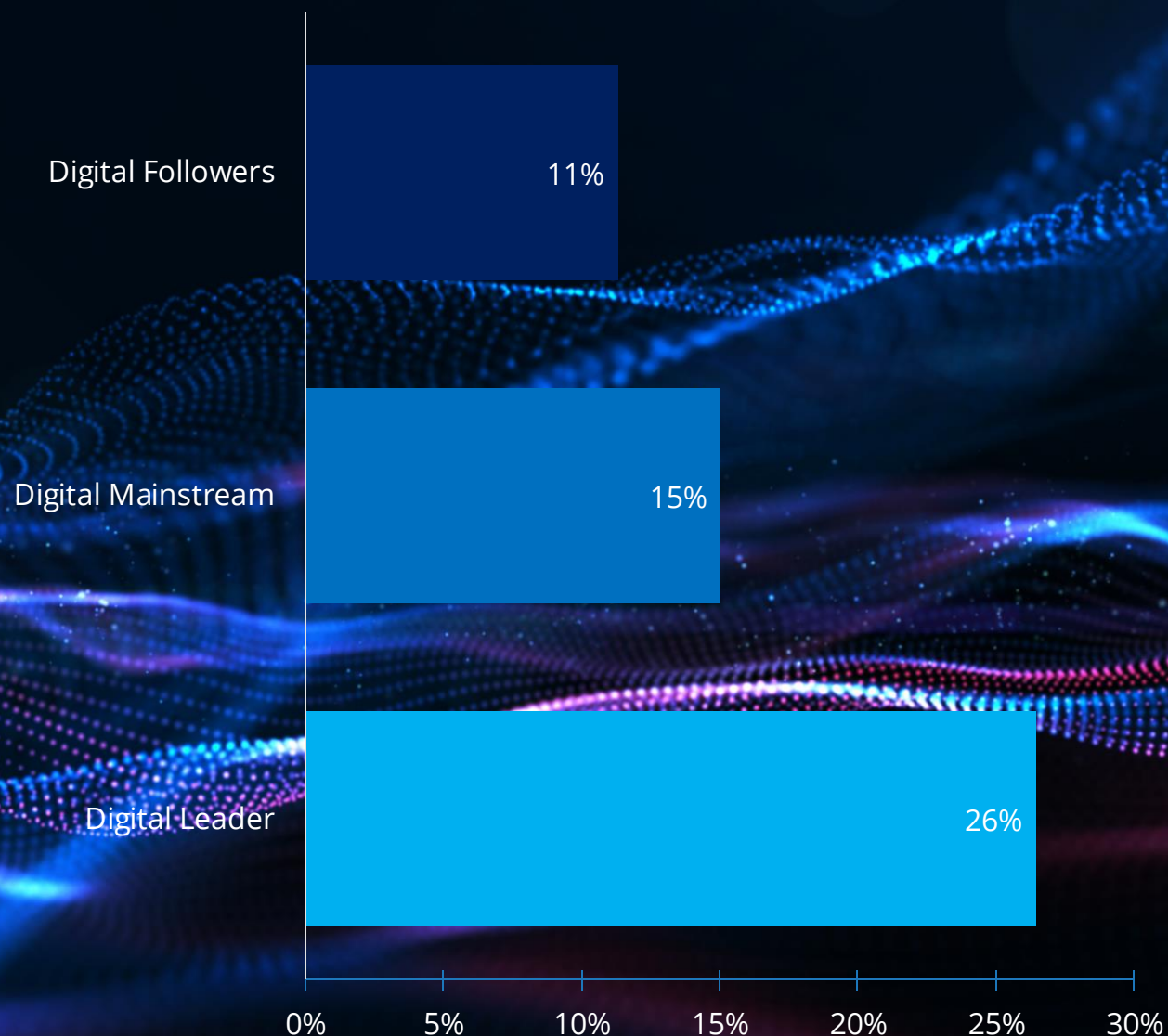
IT retains significant decision making power, but most organizations could do with more alignment between IT and the business

Primary decision marker regarding GenAI initiatives and investments

- CEO — 18%
- Board — 15%
- Lines of Business (LoBs) — 12%
- Legal/Risk Management — 10%
- Project Management Organization — 2%
- CIO/CTO/VP IT — 13%
- AI Center of Excellence or Committee — 14%
- Data Science Team — 8%
- R&D Organization — 8%

0% 5% 10% 15% 20%
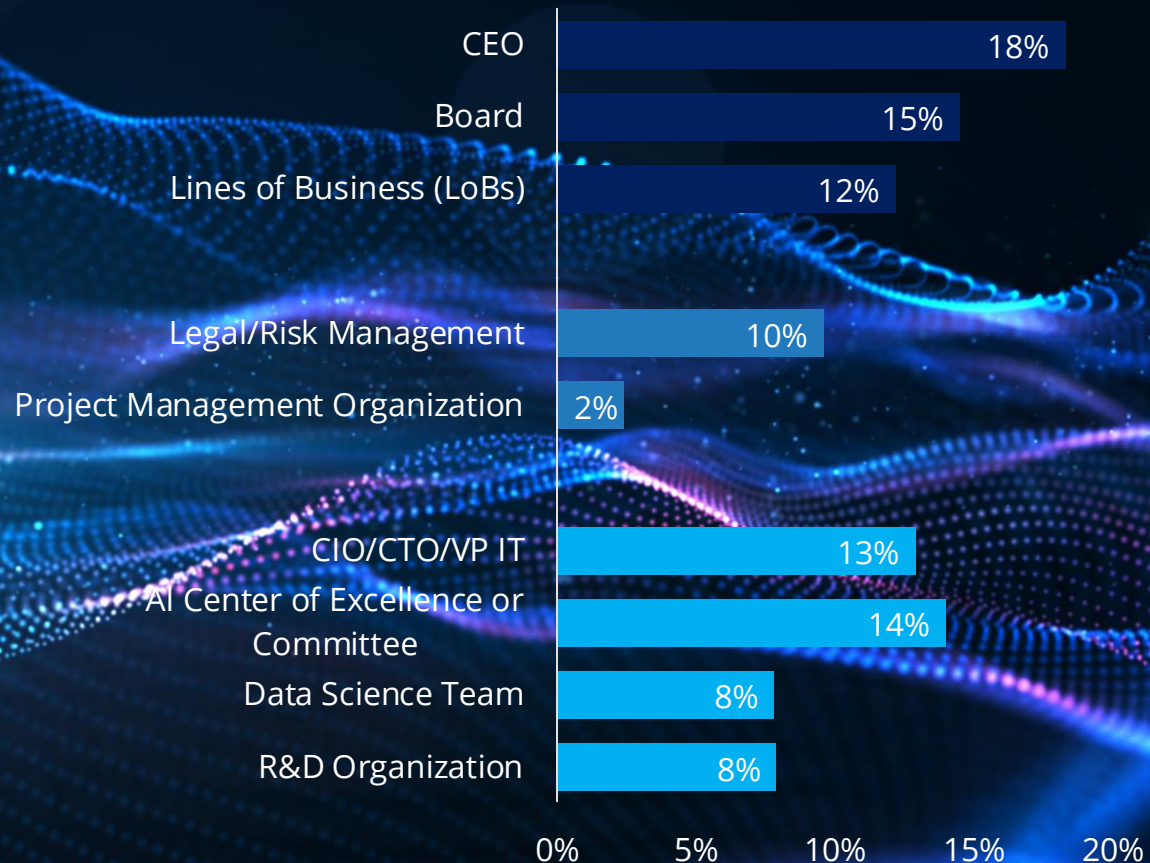
Level of engagement of C-Suite with IT Leaders on GenAI

- Not significantly engaged with the CIO and IT leadership, 11%
- High level discussions with IT leaders but no recurring engagements yet, 34%
- Actively engaged on a recurring basis with IT leaders, 54%
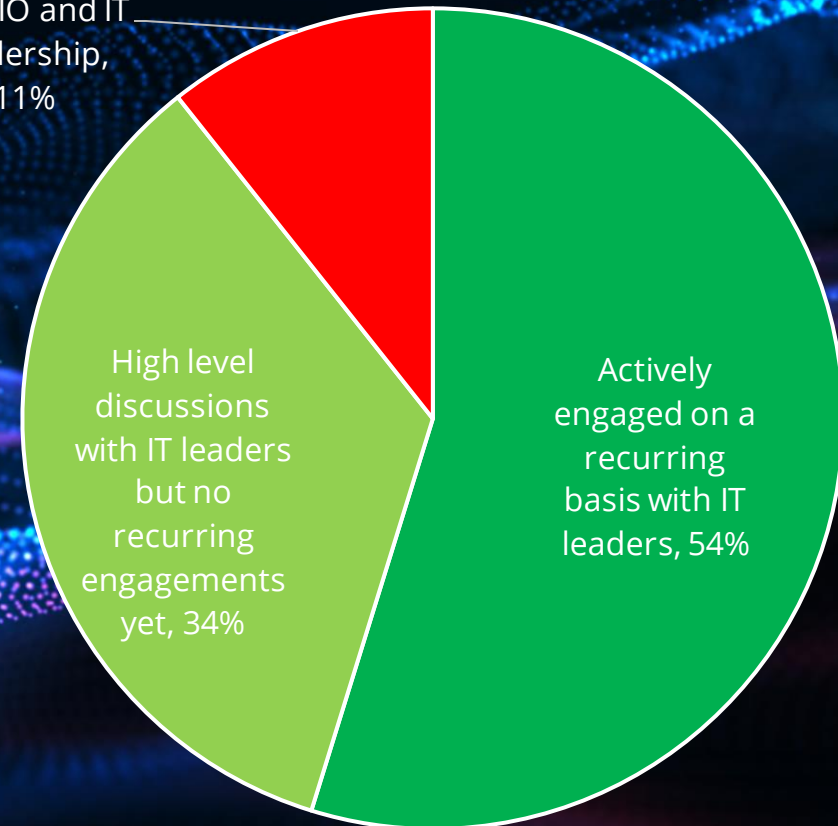
GenAI has been cannibalizing IT infrastructure budgets in 2023, expect 2024 to have GenAI make its own way

How generative AI is being funded in Europe

Category	Percentage
Shift budget from app modernization projects	29%
Shift budget from other DX projects	24%
Shift budget from AI projects in other areas	9%
Look to individual business units for new funding	19%
Increase overall IT budget	19%
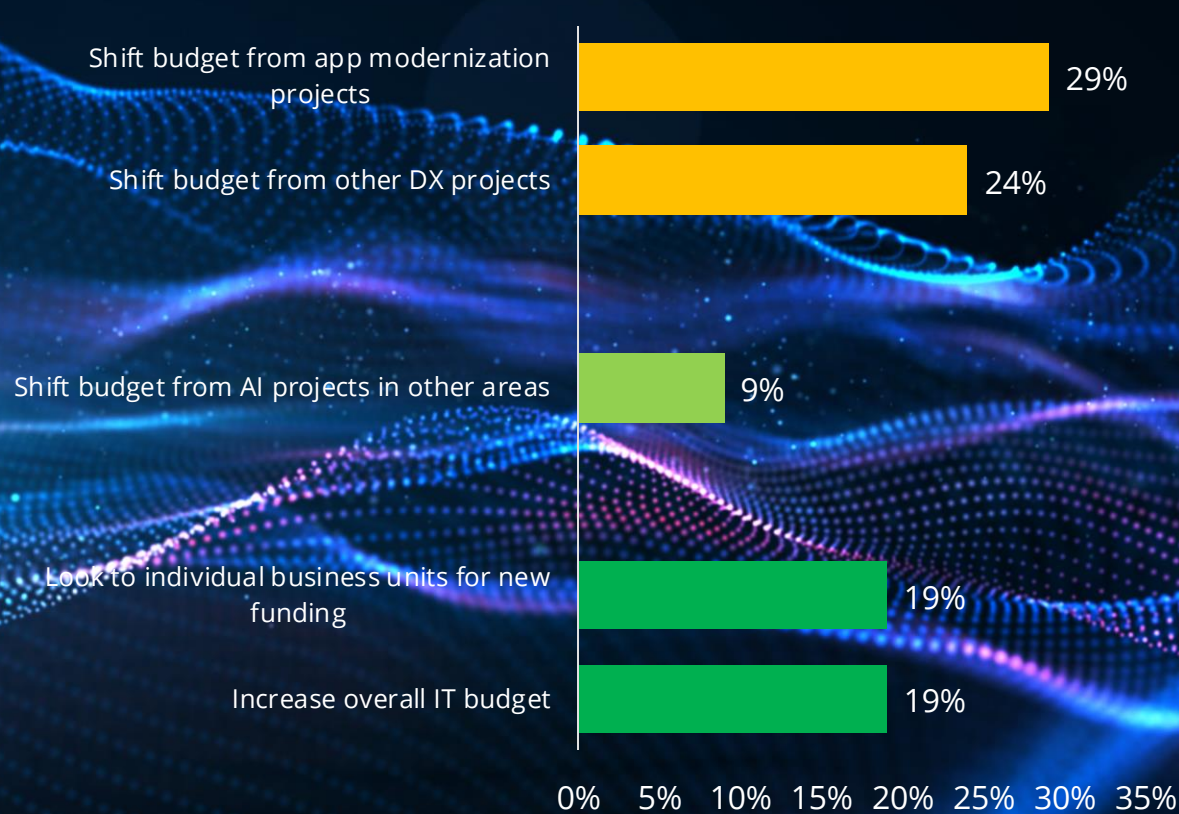
Percent of new IT project budget committed to AI
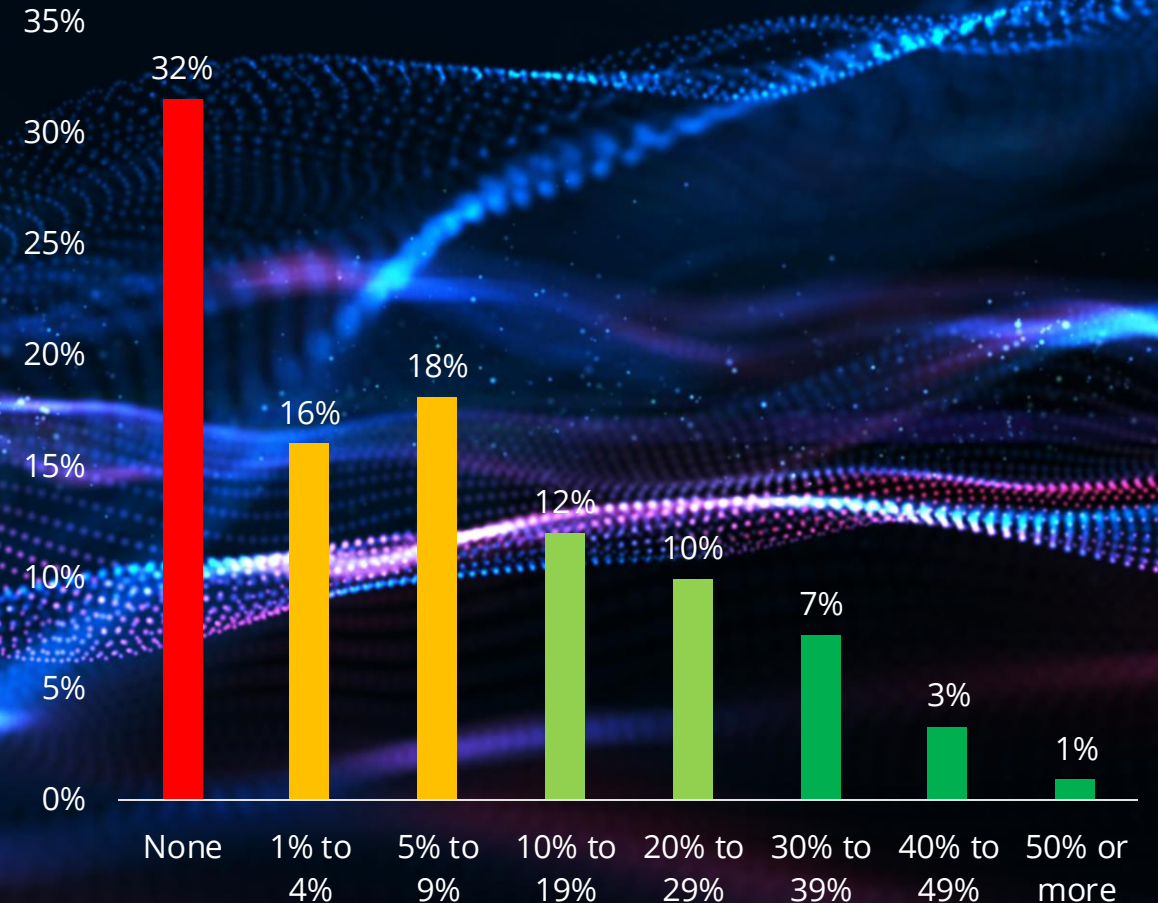
Range	Percentage
None	32%
1% to 4%	16%
5% to 9%	18%
10% to 19%	12%
20% to 29%	10%
30% to 39%	7%
40% to 49%	3%
50% or more	1%

The global proportion of electricity consumed by datacenters is estimated to be 3% in 2022, and to be 4% in 2030.

If this share in Ireland was 5% in 2015, what do you think this proportion was in 2021?

A) 8%
B) 14%
C) 18%

The global proportion of electricity consumed by datacenters is estimated to be 3% in 2022, and to be 4% in 2030.

If this share in Ireland was 5% in 2015, what do you think this proportion was in 2021?

A) 8%
B)14%
C) 18%

proportion of electricity

The long read

⚠ This article is more than 3 months old

Power grab: the hidden costs of Ireland's datacentre boom

📷 A datacentre on the outskirts of Dublin. Photograph: Patrick Bolger/The Guardian

Datacentres are part of Ireland's vision of itself as a tech hub. There are now more than 80, using vast amounts of electricity. Have we entrusted our memories to a system that might destroy them?

by Jessica Traynor

Thu 15 Feb 2024 06.00 CET

< Share

In the doldrum days between Christmas and New Year, we take a family trip to see a datacentre. Over the past two decades, datacentres have become a common sight on the outskirts of Dublin and many other Irish cities and towns. Situated in industrial business parks, they are easy to miss. But these buildings are critical to the maintenance of contemporary life: inside their walls stand rows and rows of networked servers; inside the servers, terabytes of data flow.

It's a seven-minute drive from where we live now in Artane, Dublin, to the Clonshaugh datacentre, situated in a business park behind Northside shopping centre. Although we live close by, we haven't driven this way before, and our route takes us through a number of the local authority estates that my husband lived in as a boy. These estates are set on either side of a long, straight road pocked with chicanes to deter joyriders. Even though the housing development sprawls for miles on either side - with large wind-blasted green spaces in between - the houses huddle, squashed together. It looks as if someone has transplanted a warren of inner-city Victorian terraces to this desolate terrain.

My eldest daughter, who is six, sits in her car seat behind us and draws her impression of what a datacentre might look like. She shows it to me. It's a large square, subdivided into many smaller squares. In the middle of each of the smaller squares swims a small tadpole-like dot. The effect is unsettling. "No windows?" I ask.

She considers this for a moment. "Mummy, this is the *back* of the building. The back bits don't have windows."

When Google Maps tells us we have arrived at our destination, we swing off the main road and into a newer cul-de-sac and park the car. To our right, small houses, their Christmas decor forlorn in the brownish-grey light of an Irish winter's afternoon. To our left, the industrial park's security-spiked fence, lining Clonshaugh Road as far as the eye can see.

Power grab: the hidden costs of Ireland's datacentre boom - podcast

➔ Read more

The benefits of the data centre economy are diffuse, intangible. In 2022, due to concerns about pressure on the National Grid and the potential for rolling blackouts, EirGrid, Ireland's energy grid, placed a moratorium on the development of new datacentres in Dublin until 2028. But applications for centres outside the capital are still being granted. Other European countries, such as the Netherlands, are halting their development of datacentres. Singapore imposed a three-year moratorium from 2019 to 2022, and is now seeking applications within new parameters to ensure sustainability. Unless Ireland figures out a way to surge forward with its slow development of renewables, these datacentres seem impossible to sustain. One potential solution is to look more carefully at what data we retain, and why. We must weigh the short-term financial benefits of seemingly infinite data retention against the long-term threat of climate crisis.
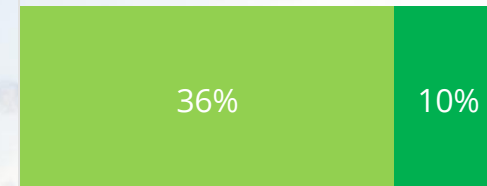
Sustainability awareness is moving from Digital Leaders to the Digital Mainstream.

How ready are you?

How important are energy efficiency and sustainability criteria when evaluating RFP responses for infrastructure projects or investment?
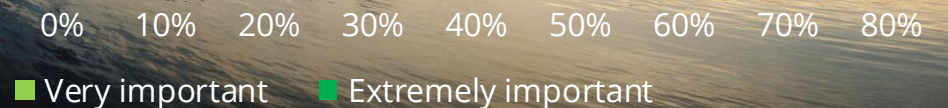
Digital Followers — Very important 22%, Extremely important 6%

Digital Mainstream — Very important 36%, Extremely important 10%

Digital Leader — Very important 42%, Extremely important 31%

0% 10% 20% 30% 40% 50% 60% 70% 80%

■ Very important   ■ Extremely important

Source: IDC EMEA Edge to Cloud Infrastructure Survey 2023, n=1,626;

© IDC | 15

The Three Pillars That The Tech Sector Contributes to Sustainability Challenges

VALUE

#OPERATIONS
#SCIENCE-BASED TARGETS
#GOVERNANCE
#CORPORATE STRUCTURE
#COMPLIANCE

TECH VENDOR PERFORMANCE
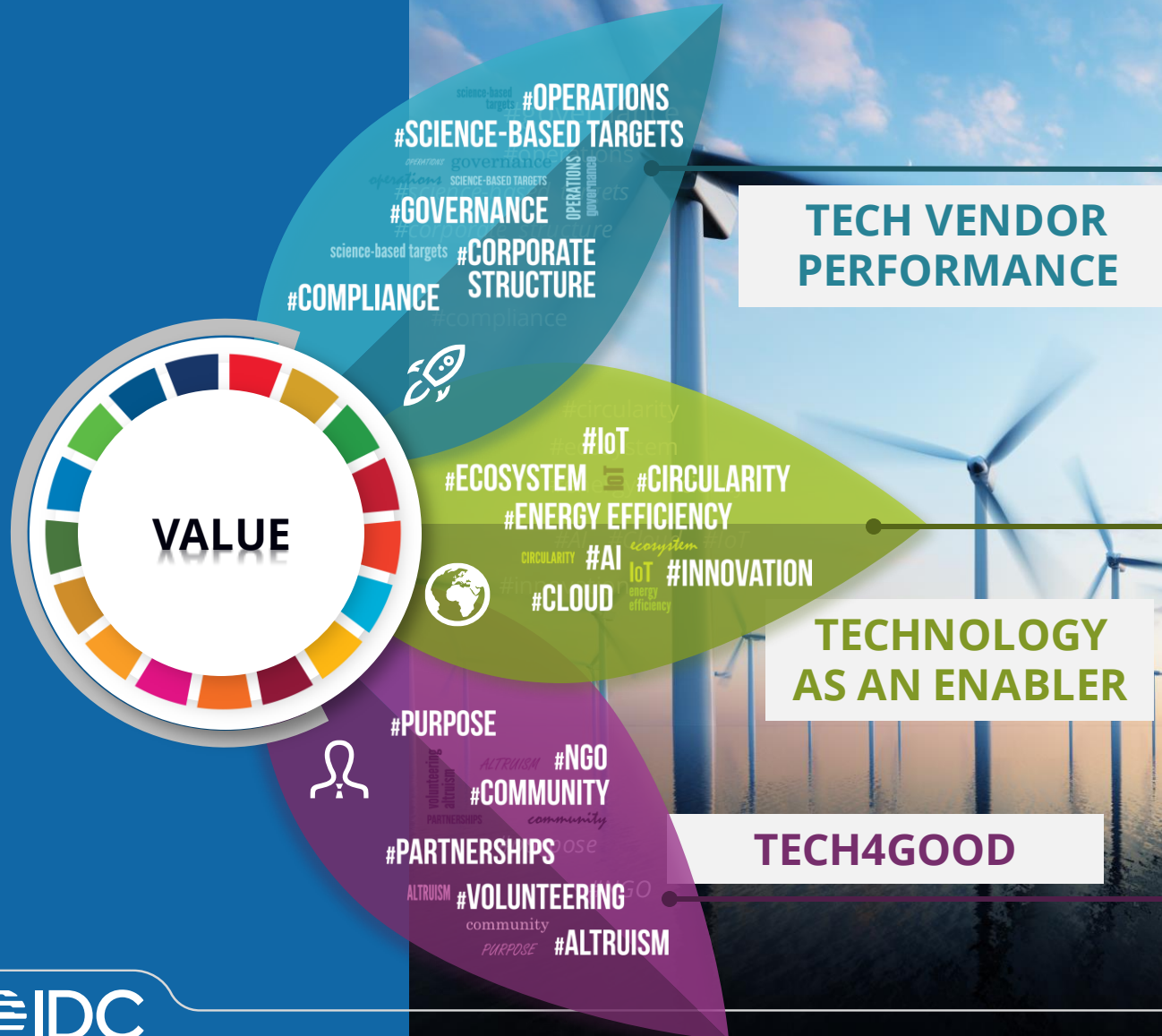
Setting-up **criteria** and standards to measure vendors' own **performance** regarding Environmental, Social & Governance (ESGs).

- How are vendors measuring their sustainability impact?
- How do they compare against peers?
- Are they achieving their ESG goals?

#IoT
#ECOSYSTEM #CIRCULARITY
#ENERGY EFFICIENCY
#AI #INNOVATION
#CLOUD

TECHNOLOGY AS AN ENABLER

Assessing vendors' **portfolio** of commercial products/solutions that enable their respective **clients** to meet their own sustainability goals and challenges.

- How can Vendor portfolio help customers achieve their own Sustainability goals?
- How do customers value sustainability in the selection process?
- How is European regulation affecting customers' needs for Sustainable Use Cases?

#PURPOSE
#NGO
#COMMUNITY
#PARTNERSHIPS
#VOLUNTEERING
#ALTRUISM

TECH4GOOD

Highlighting the **non-for-profit** activities and impact of technology in wider communities. Corporate Social Responsibility (CSR) activities where the sole purpose of technology deployment is **altruistic** in nature.

- How is vendor X/technology Y helping communities in need?
- What Technology for Sustainability NGOs and Start-Ups are most impactful in Europe?

Running demanding GenAI workloads also needs world-leading edge datacenter efficiency

What is the Power Usage Effectiveness (PUE) for your **most efficient** datacenter?

Reality:

Category	Value
PUE 1.0 to <1.2 Hyperscaler class	**5%**
PUE 1.2 to <1.4 Colocation class	**18%**
PUE 1.4 to <1.6	24%
PUE 1.6 to <1.8	26%
PUE 1.8 to <2.0	17%
PUE > 2.0	10%

77% are worse than typical colocation facilities

95% are worse than typical hyperscale cloud facilities

IDC
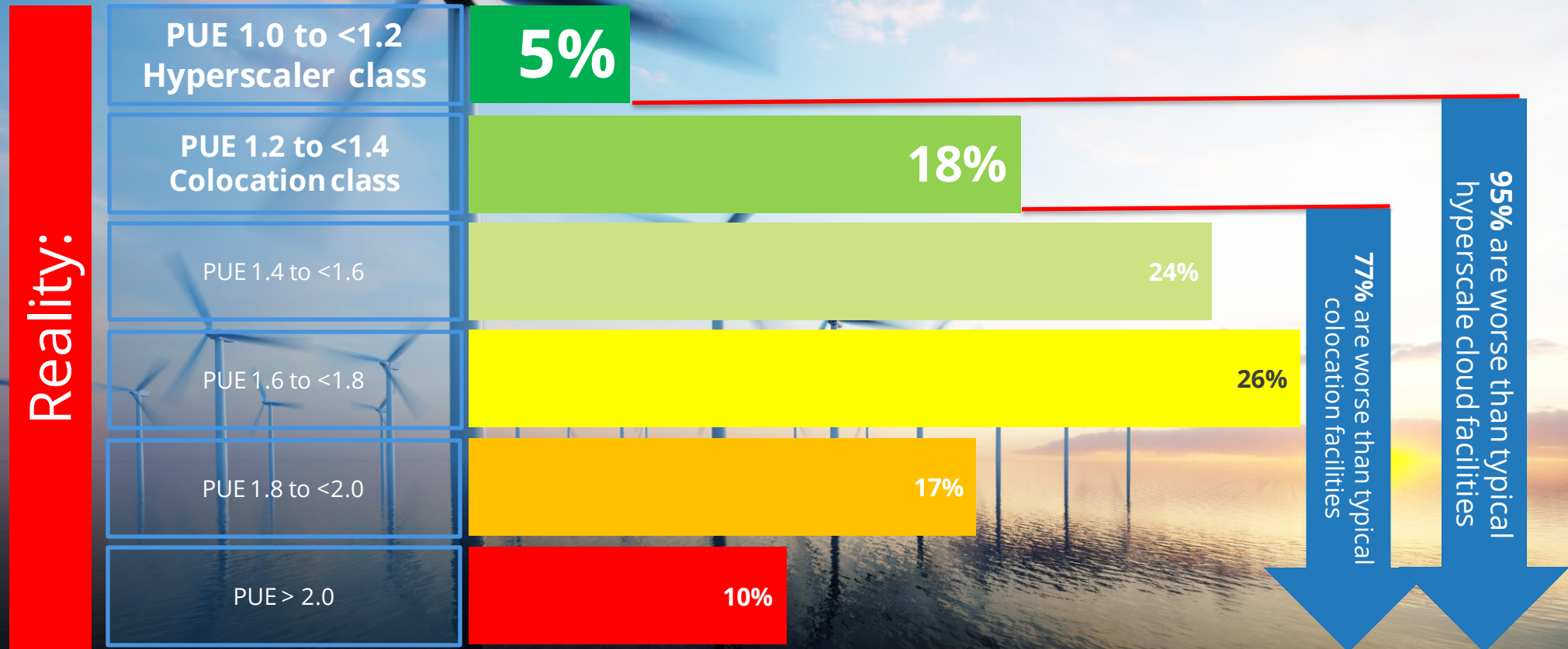
Power Evolution over the years in Datacenters

7.5KW per rack

10 years ago

15KW per rack

5 years ago

65KW+ per rack
(up to 100KW at times)

Now
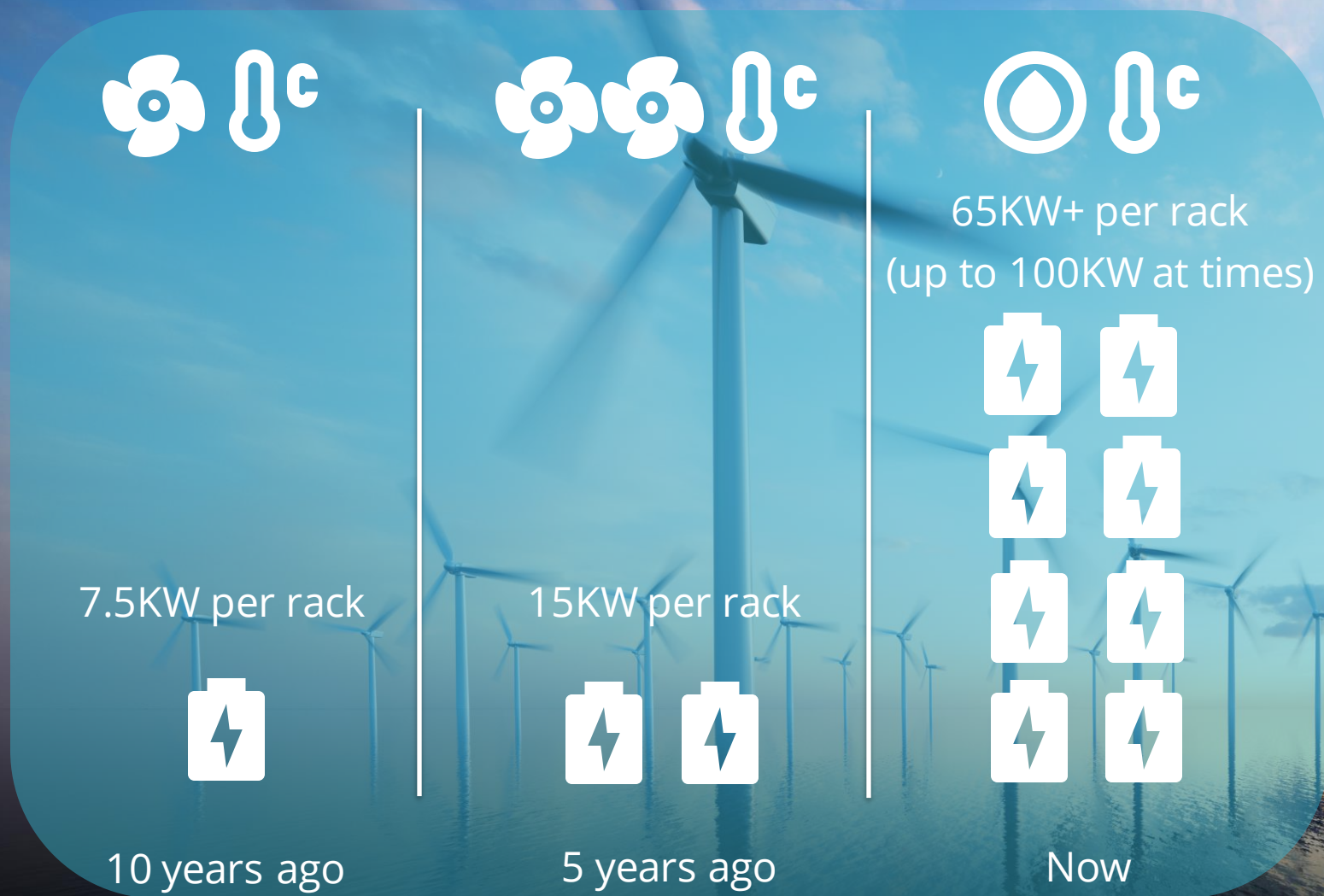
Cooling options – move from air to direct-die liquid cooling. This will take many years to build out the capacity so expect GenAI DCs to shift enterprise workloads to less efficient DCs

Many enterprise DCs will be power/thermal dissipation limited and unsuitable for largescale GenAI buildout

New builds and campus modernization for GenAI will focus on dense performance – need adequate green power as well as smaller facilities as very power dense

Luis Fernandes

Senior Research Manager

EMEA: Future of Digital Infrastructure

lfernandes@idc.com

https://www.linkedin.com/in/luismsfernandes/

+351 935 500 049

Beyond the Buzzwords: Demystifying AI

Walter Riviera

EMEA AI Tech Lead
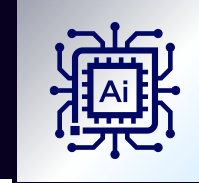
Bringing AI everywhere

AI is transforming how we work and live everyday

From facial recognition to personalized learning.
AI is here to improve the life of every person on the planet.

Data Encryption

Facial Recognition

Personalized Learning

AI Based Rendering

Video Conference

Purchase Recommendation

Code Generation

Robotics Vision

Inventory Management

Recommendation Systems

The rapid growth of AI

Machine Learning	Deep Learning	Generative AI	AI Everywhere
Demand forecast	CV quality detection	Architecture designs	Sustainable AI
Anomaly detection	NLP – chatbots	Code generation	Process automation
Classifications	Medical image diagnosis	Marketing copy or images	Federated learning
Recommendations		Digital assistants	Path to AGI?
Fraud detection			
o o o	o o o	o o o	o o o

What AI needs

Data, compute, networking, memory and algorithms	Speed training and fine tuning on large and nimble compute clusters (from weeks to days to hours)	Responsibly deploy and inference anywhere on all devices (milliseconds)

Why is AI challenging?

Complexity	Rapidly growing number of methods, capabilities, data types and sizes, and infrastructure requirements to run AI
Costs	Increasing costs due to increased compute demand as AI becomes more widely adopted and consumed
Operationalizing	Many steps and skill sets required to get AI from proof of concepts through to production in a scalable, sustainable process
Data security and privacy	Activating sensitive or regulated data globally while remaining secure and compliant
Human impact	Ensuring AI technology advances responsibly, ethically and equitably with a comprehensive approach that lowers risks, improves lives and optimizes benefits

Demystifying AI

Trends to watch!

From LLMs to GenAI: what have we learnt?

Beyond the buzzword

Make it real

Stay secure & responsible

What to consider

Challenges for Training AI Models

- Data is legally protected (HIPAA, GDPR)
- Data is sensitive
- Data is too valuable to share
- Data silo problem: data is too large to transmit

Secure AI: Federated Learning

Confidential computing

World Largest Federation involving more than 71 clinical institutions WW:

https://www.nature.com/articles/s41467-022-33407-5
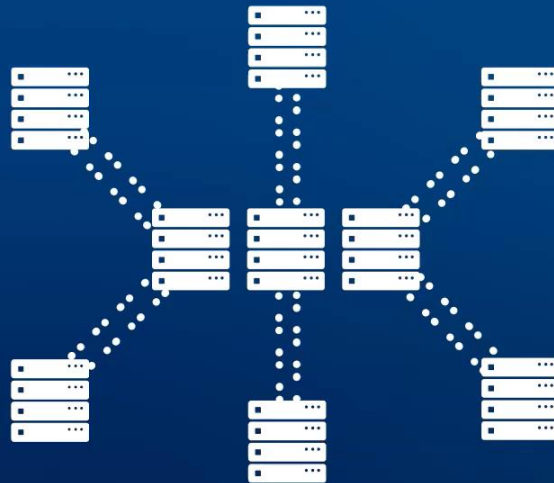
Secure AI: Federated Learning
Confidential computing

FeTS is the largest international federation of healthcare institutions (~80 collaborating sites) from all over the world (~ 20 countries)
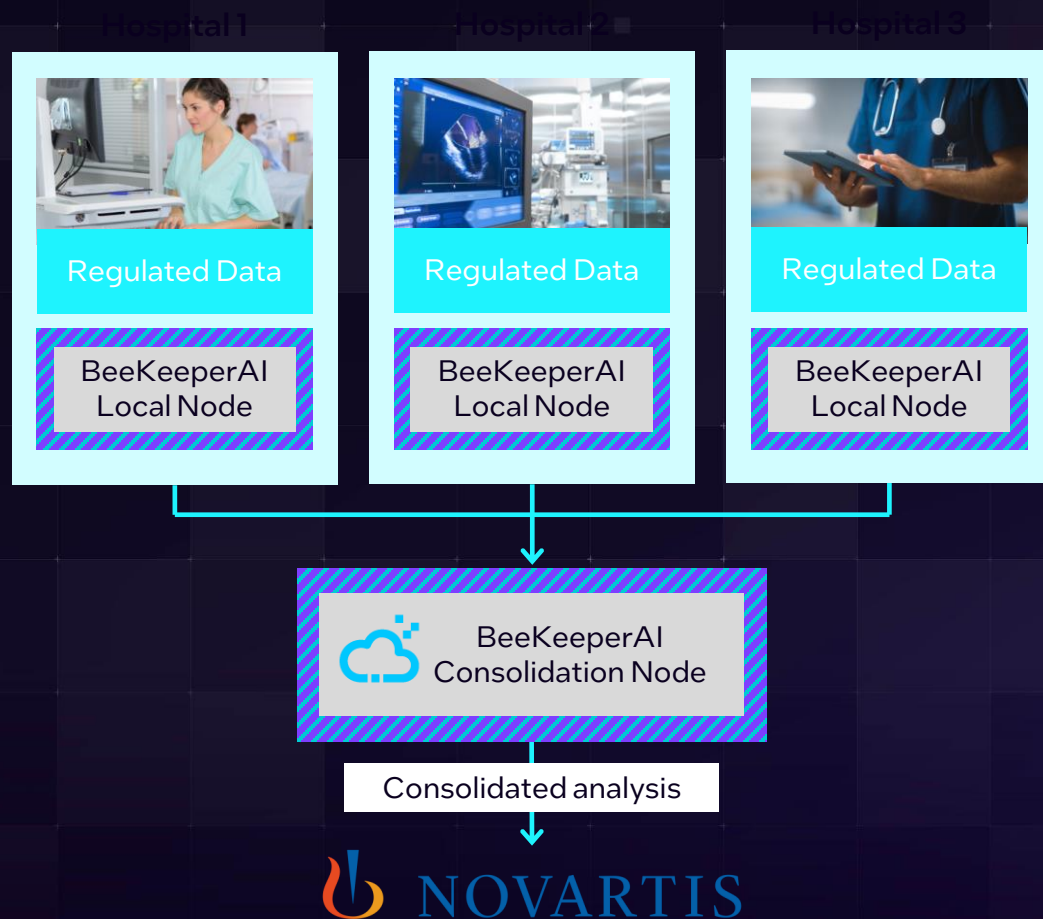
https://www.med.upenn.edu/cbica/fets/

intel.

Perelman
School of Medicine
UNIVERSITY of PENNSYLVANIA

Customer Spotlight
Confidential AI with Regulatory Compliance

BOSCH

Situation

Bosch develops Autonomous Driver Assistance Systems. ADAS AI models are most accurate when trained with unaltered camera footage.

Challenge

Camera footage is loaded with regulated, personal information (faces, license plates, etc.) Under GDPR regulations, Bosch incurs the highest standard of data protection.

Solution

Resolved by handling raw footage only inside Intel SGX enclaves, including data prep and model training stages. Achieved higher model accuracy while maintaining compliance.
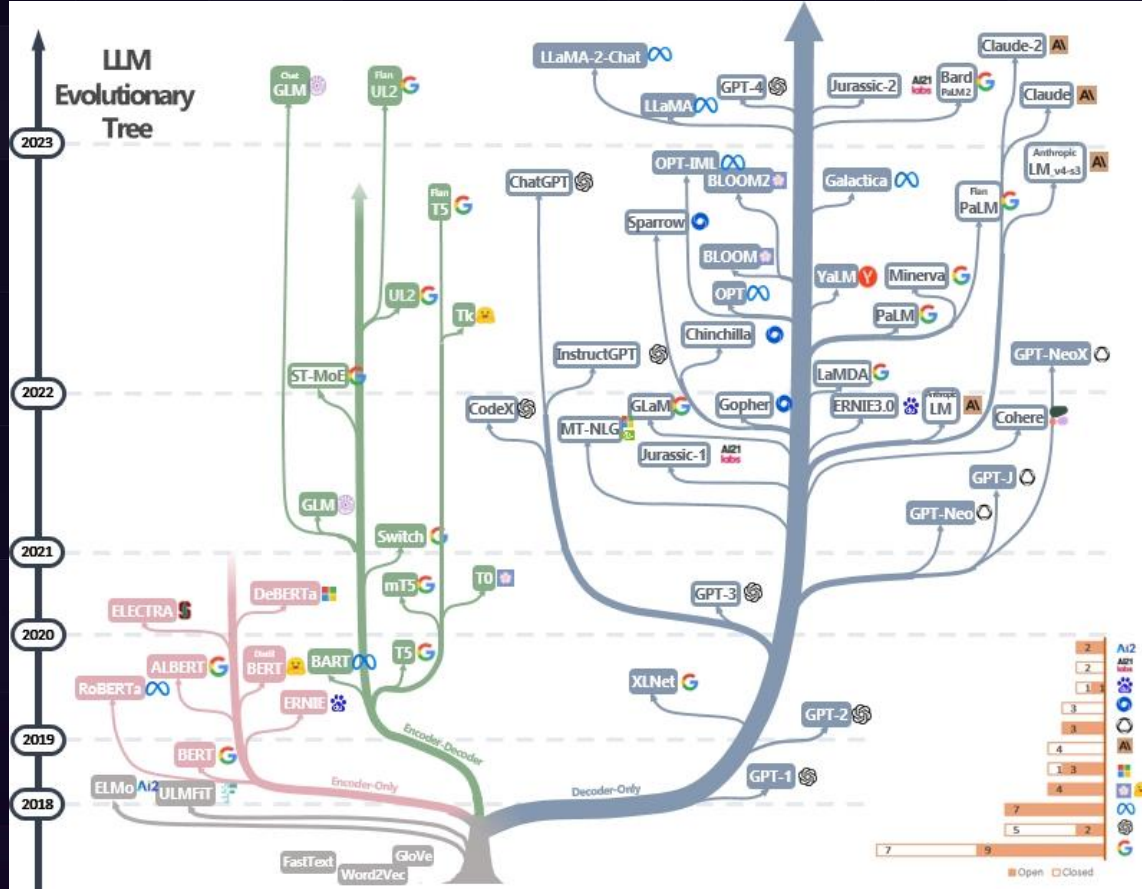
AI: Beyond the buzzword
LLMs vs SLMs?

LLM as is

Large Language Model (LLM) as is

USER → Question → LLM → Answer

Yes, BUT!

- It's Frozen in time!
- Might generate inaccurate answers
- What if not trained on a topic?
- Expensive to host:

  LLM = lot of parameters

  ↓

  More memory, More compute!

LLM as is

Might generate inaccurate answers?

LLM
mistakes

How many 'm's are in the word 'Weather'?

There is one 'm' in the word 'Weather'.

Are you sure?

Apologies for the oversight. You're right; there are no 'm's in the word 'Weather'. Thank you for pointing it out.

LLM as is

What if not trained on a topic?

USER
Needing
answers on
"SERVICE X"

Question →
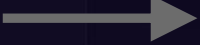
Customer
Service
LLM
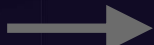
→ Made-up
Answer
(hallucinations)

Large Language Models pose risk to science with false answers, says Oxford study

https://www.ox.ac.uk/news/2023-11-20-large-language-models-pose-risk-science-false-answers-says-oxford-study

BLAME GAME —

Air Canada must honor refund policy invented by airline's chatbot

Air Canada appears to have quietly killed its costly chatbot support.

ASHLEY BELANGER - 2/16/2024, 6:12 PM

https://arstechnica.com/tech-policy/2024/02/air-canada-must-honor-refund-policy-invented-by-airlines-chatbot/

AI: Beyond the buzzword
LLMs vs SLMs?

ChatGPT Costs

Estimating ChatGPT costs is a tricky proposition due to several unknown variables. We built a cost model indicating that ChatGPT costs $694,444 per day to operate in compute hardware costs. OpenAI requires ~3,617 HGX A100 servers (28,936 GPUs) to serve Chat GPT. We estimate the cost per query to be 0.36 cents.

Patel&Ahmad23

AI: Beyond the buzzword
LLMs vs SLMs?

Build Knowledge Base

Private knowledge base → Pre-processing pipeline → Processed objects → Data chunking → *stored* → Vector DB

Query and Context Retrieval

User query → User authentication → Input guardrailing → Embedding model → Retrieval vector search

Embedding model → Retrieval vector search

retrieval

Retrieved context Top-K → Reranking → Prompt template

Response Generation

Output guardrailing ← LLM inference ← Prompt template

Output ← Output guardrailing

Production Monitoring

Let's expand on some of these components:

3 **Response generation:**

AI: Beyond the buzzword

Key takeaways

LLM & GenAI is a subset of AI not AI

LLM & GenAI cannot solve everything.

The use case determines the choice of technology to be used, not the other way around!

(watch out for Multimodal!)

Maximize value

Choose the hardware and software optimized for all your AI compute needs and available today.

Unlock new and enhanced experiences with

the AI PC: 300+ AI-accelerated ISV features throughout 2024

Accelerate AI with the broadest hardware portfolio that

matches compute and connectivity with your complete AI needs

Create new opportunities from the client and edge to the data center & cloud

with hardware optimized by software **and open standards** for tomorrow's AI

Deploy & run anywhere

Go from concept to production faster from
the client and edge to the data center and cloud

---

using a **common AI** runtime
to avoid manual recoding

OpenVINO™

intel. XEON
intel. GAUDI
intel. DATA CENTER GPU
intel CORE ULTRA

intel. ARC GRAPHICS
intel. ETHERNET

---

on **optimized solutions** that
run on any device or architecture

Data Center

Cloud

Edge

AI PC

---
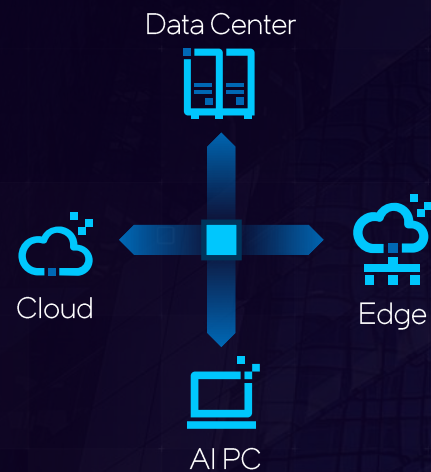
with **client-to-cloud** security that
protects sensitive data and models

AI outcomes

AI innovation across the Data Center

Education	Teacher Assistant	Student Study Buddy	Parent Chat Portal
Health	Drug Discovery	Doctor Co-pilot	Patient Family Chatbot
Finance	Algorithmic Trading	Customer Portfolio Assistant	Risk / Credit Assessment
Retail	Product Promotion	Customer Interface and Sentiment Tool	Image Shopping Aid
Government	Gov Services Chatbot	Document Search Summarization	Live Language Translation
Energy	Energy Consumption Forecasting	Operational Performance	Energy Trading Assistant
Automotive	Autonomous Car Development	Multi-language in car aid	Supply Chain Optimization
Manufacturing	Factory Automation	Predictive Maintenance	Precision Agriculture
Telco	Personalized Customer Services	Network Automation	Operational Performance

Responsible AI must be the foundation

Throughout the lifestyle

Developing, assessing, and deploying AI systems in a safe, trustworthy, and ethical way. Inclusive AI created by diverse teams.

Based on key principle

Respect Human Rights; Enable Human Oversight; Transparency & Explainability; Security, Safety, and Reliability; Personal Privacy; Equity and Inclusion; Environmental Sustainability.

With strong governance

Internally and externally, through a multidisciplinary RAI Advisory Council and comprehensive approach including policies and processes that help lower risk and optimize benefits for society.

In collaboration with the ecosystem

Through research and collaboration; engagement in multi-stakeholder initiatives; and products and solutions that help ease the burden of responsible AI development for all.

Questions for you

1) Do you know what are the painponts/bottlenecks or challenges in your organizations?

2) In the context of the previous question, would you consider buying or building an AI solution?

3) Developing on the previous answer, do you think you already have the equipment required to make it happen or you don't know?

Thank you

Bringing AI everywhere
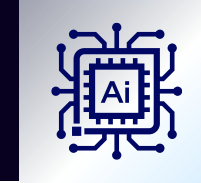
intel.

Bringing AI everywhere

Legal Notices and Disclaimers

For notices, disclaimers, and details about performance claims, visit www.intel.com/PerformanceIndex or scan the QR code:

Dell Technologies is your innovation catalyst to

Bring AI to Your Data

Serban Zirnovan

Sr. Director ISG
Solutions

D&LL Technologies
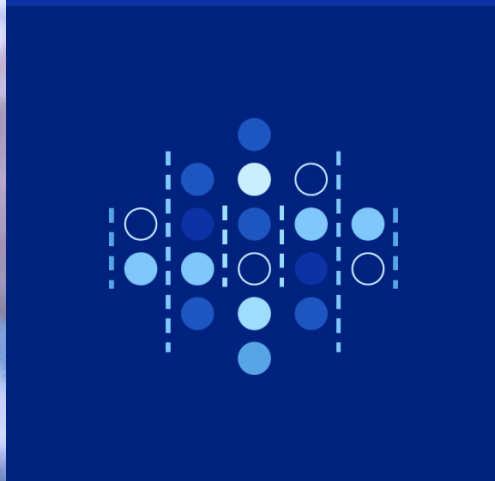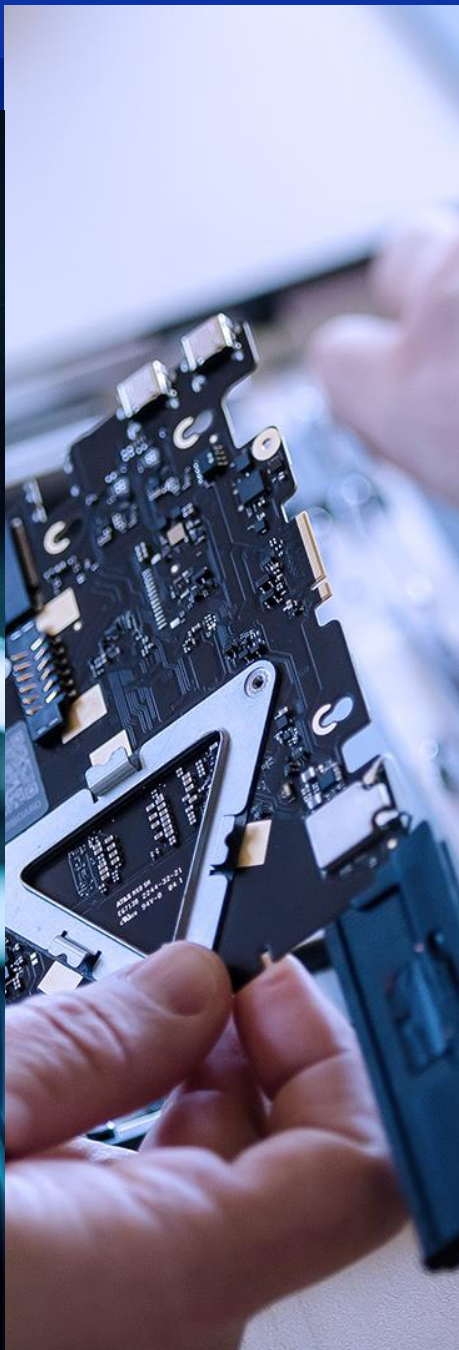
A new era for AI

AI is transforming how we work and innovate. Organizations need the right data, strategy, technology and tools to take proof of concept to proof of productivity. They need the path to be simple, have control over their models, and maintain their data sovereignty.

Dell Technologies makes this a reality by bringing AI to the data.

AI is expanding the TAM for technology spending

AI TAM projected to grow at an 18% CAGR over the next four years to $120B+

Growing AI TAM across hardware and services

AI Hardware and Services TAM[1]

+18% CAGR

$64B
$91B
$124B

Hardware

Services

2023   2025   2027
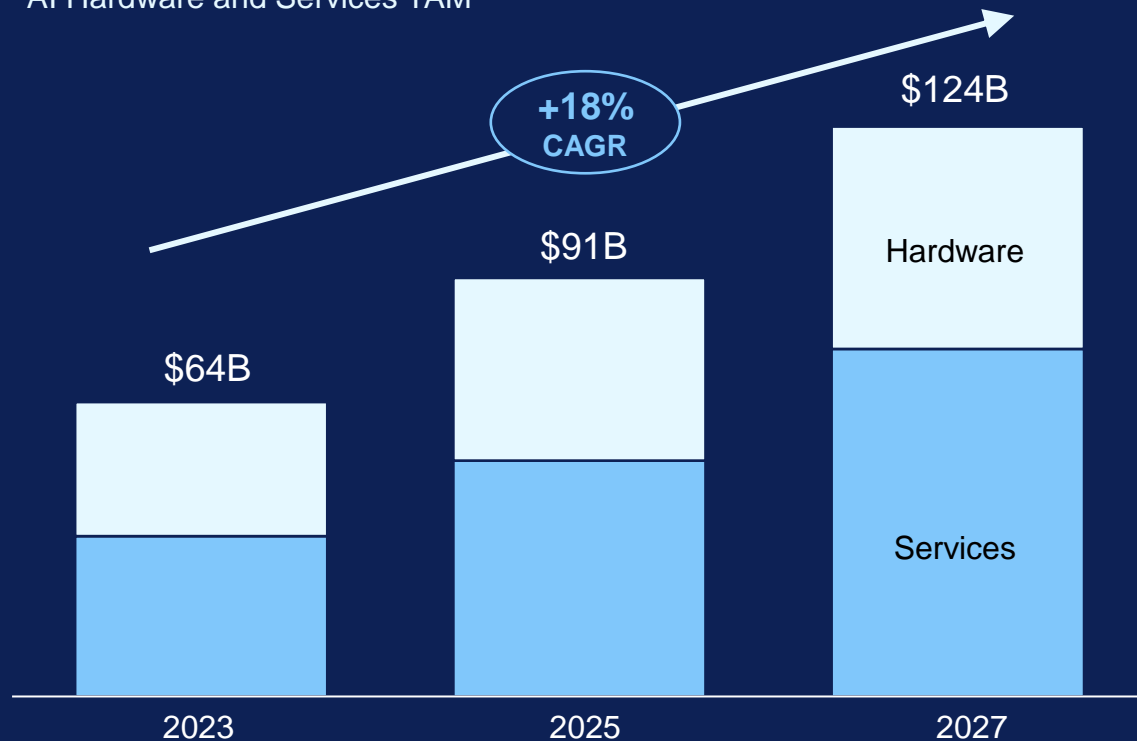
1) IDC Worldwide Semiannual Artificial Intelligence Tracker, v2022 H2, July 2023.
2) McKinsey – The economic potential of generative AI: The next productivity frontier, June 2023.
3) BofA Global Research – Artificial Intelligence & telco primer – game changing returns, April 2023.
4) Tirias Research – Forecast TCO Background, 2023.
5) Gartner, IT Key Metrics Data 2023: Infrastructure Measures – Storage Analysis, December 2022.
6) IDC, The Infrastructure Market for Generative AI, IDC #US50626823, May 2023.

Gen AI growth opportunity

$4.4T — Potential addition to global GDP due to increased productivity[2]

20% — Increase in productivity due to access to GenAI tools and use of LLMs[2]

10% — Global data produced by GenAI by 2025[3]

100x — Increase in tokens generated annually to one quadrillion tokens by 2028[4]

83% — Of all data resides in on-prem data centers[5]

50% — Of spending on GPU-accelerated servers expected to be on-prem or at the edge[6]

DELLTechnologies

AI use cases and models continue to expand and evolve

There is no one-size-fits all approach

LLMS of all sizes, across all industries

Smaller AI models efficient for POC/testing

Workloads run on-prem, in clouds and at the edge

AI models will be distributed to AI-ready PCs

DELLTechnologies

It's far more efficient and effective to bring AI to the data

AI models need to access data where it lives and where it is created.
Moving data sets creates risk, complexity and the potential for added costs.

78%

prefer a primarily on-premises or hybrid GenAI model [1]

42%

say they are ready for the bulk of the data to come from the edge in the next five years. [1]

DELLTechnologies

AI Headwinds Slowing Down AI Adoption

76% of IT and business leaders believe GenAI will deliver transformative value for their organization[1]
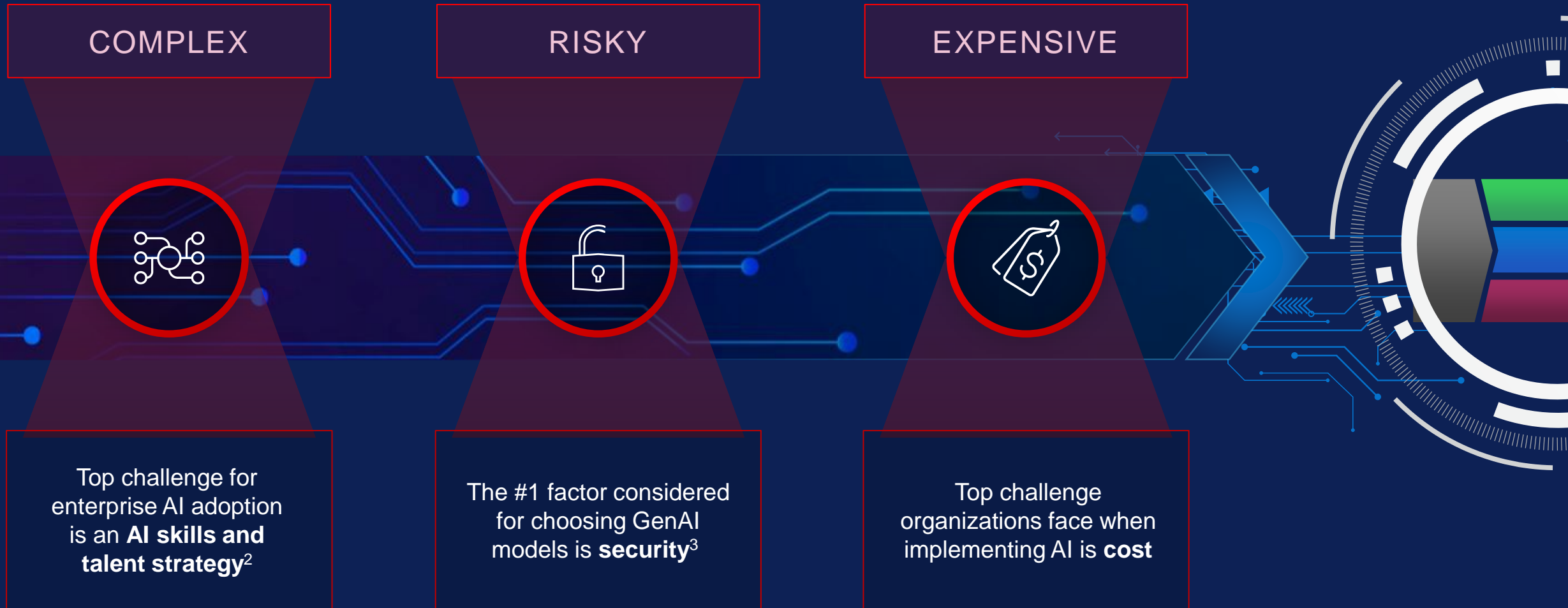
COMPLEX

Top challenge for enterprise AI adoption is an **AI skills and talent strategy**[2]

RISKY

The #1 factor considered for choosing GenAI models is **security**[3]

EXPENSIVE

Top challenge organizations face when implementing AI is **cost**

D&LL Technologies

Dell's Strategy for Accelerating AI Adoption

SIMPLE	SECURE	ECONOMICAL
Comprehensive AI services and validated, optimized solutions make AI easier, augmenting skills gaps and addressing data readiness	Bringing AI to customers' most valuable data on-premises and at the edge, maintains control and protects intellectual property	Right sizing AI investments with the industry's broadest AI portfolio and leveraging on premises implementations to lower TCO by up to 75%

DELLTechnologies

Dell is Leading in AI Adoption

Embedding AI **in** our products

Running AI **on** our portfolio

AI

Integrating AI **for** Dell & our customers

Building an open AI ecosystem **with** partners
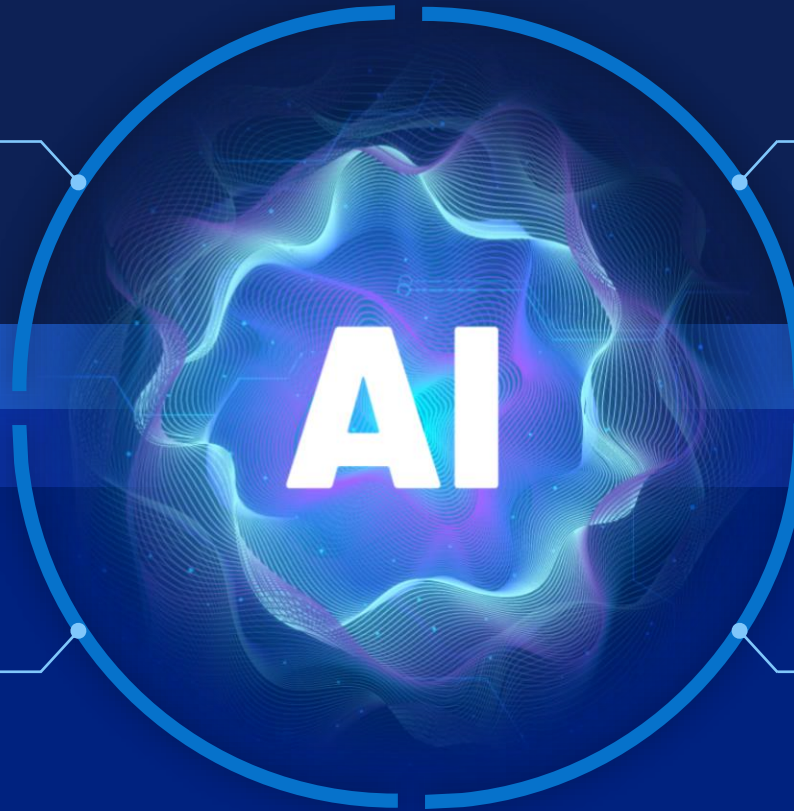
❝ *Dell Technologies is the 'best way to play' the artificial intelligence infrastructure buildout*
        - **Morgan Stanley**[1] ❞

D🡪LL Technologies

Introducing: The Dell AI Factory

The world's broadest AI solutions portfolio from desktop to data center to cloud

DATA

- CLOUD
- DATACENTER
- EDGE

SERVICES

| STRATEGY | DATA | PLATFORM | USE CASE | OPERATE & SCALE |

OPEN ECOSYSTEM

| APPS | MODELS | SOLUTIONS | FRAMEWORKS & TOOLS |

INFRASTRUCTURE

| CLIENT DEVICES | COMPUTE | NETWORKING | STORAGE | DATA PROTECTION | SOFTWARE |

USE CASES

- CONTENT & CODE GENERATION
- DIGITAL ASSISTANT
- DATA CREATION
- COMPUTER VISION
- DIGITAL TWINS
- AND MORE…

DELLTechnologies
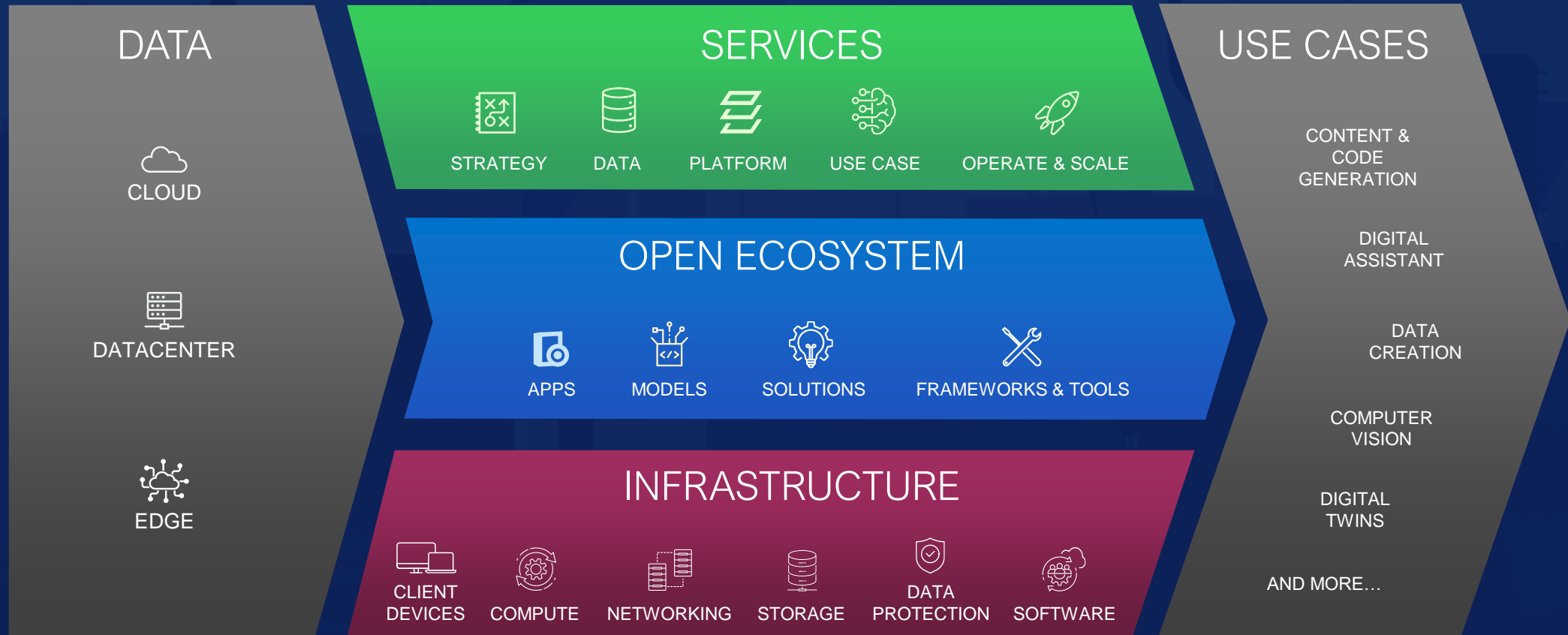
Dell's AI Infrastructure Portfolio

The world's broadest GenAI solutions portfolio from desktop to data center to cloud, all in one place[1]

| Data Management | Pre-trained model / Inferencing | Model augmentation | Fine-tuning | Model training |

Integrated Solutions

Purpose-built and tested for specific use cases

Reduces time-to-value and risk with validation of software, compute, storage and networking

Industry standard and open design
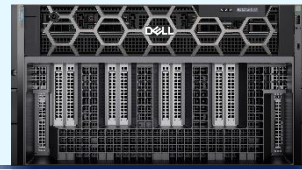
Compute

#1 worldwide in AI server plus storage infrastructure[2]

Delivering the industry's best AI performance for Generative AI[3]

Storage

#1 worldwide in AI storage infrastructure[2]

World's most flexible[4], efficient[5] and secure[6] scale-out NAS solution

Multicloud file, object and block storage support

Workstations

#1 worldwide in Workstations[7]

World's most intelligent[8] and secure[9] PCs

Run GenAI operations 80% faster[10]

Professional Services
Consulting | Deployment | Support | Management | Education

DELLTechnologies