McKinsey
& Company

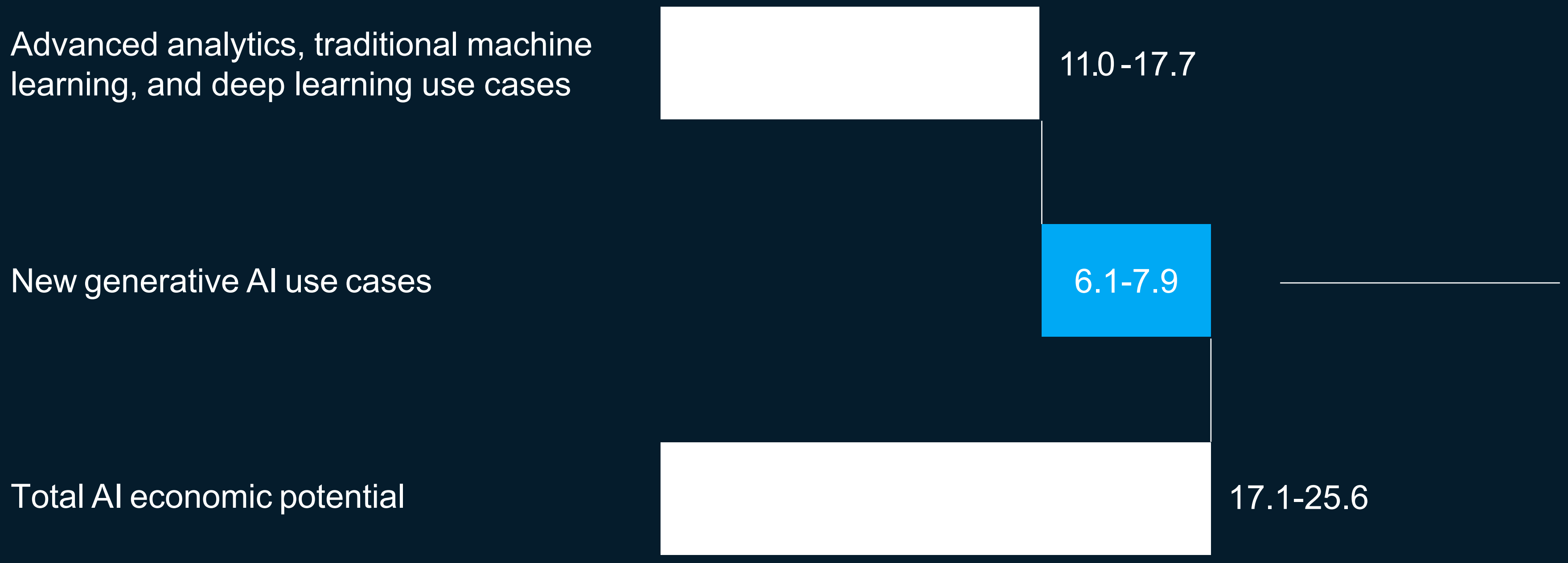
GenAI at the Edge: Overview & Outlook

October 30, 2024

GSA
Where Leaders Meet

AI’s annual economic impact based on business use cases estimation

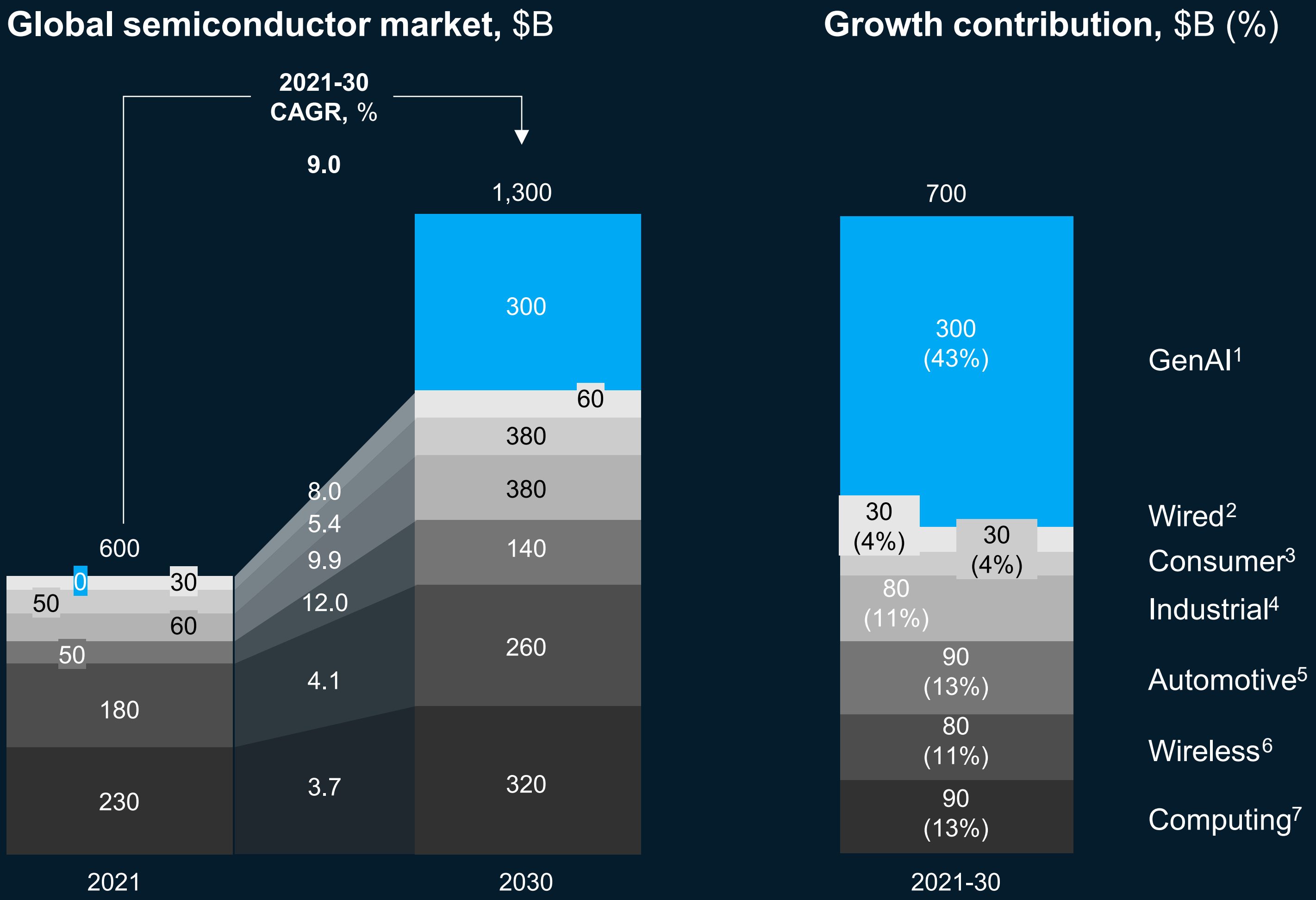
$ trillions



GenAI could create an additional ~ 35-70% of value above what other AI and analytics can unlock

Source: McKinsey Global Institute analysis

~50% of overall semiconductor market growth will be driven by GenAI until 2030

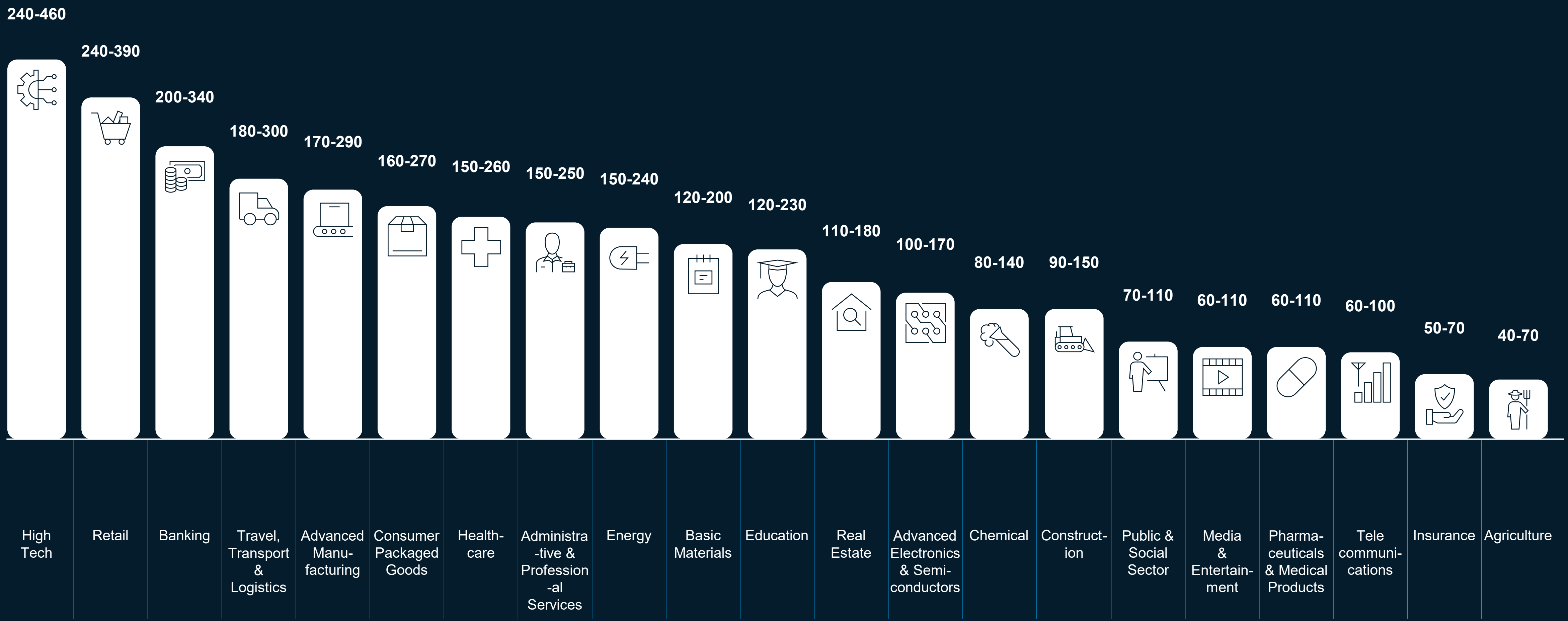


1. GenAI market based on leading edge & memory and base case scenario; 2. Switches & routers, aggregate equipment, CPEs; 3. TVs, Consoles, Smart watches, Home appliances, etc.; 4. Automation, Medical, Test & Measurement, Security, Buildings, Lighting, Power & Energy, Military, Other; 5. Connectivity, Telematics, Infotainment, Drivetrains, Powertrains, ADAS, Chassis, Body & Convenience, Other; 6. Mobile phones, smartphones, tablets, communications infrastructure

Source: McKinsey Analysis, OMDIA

GenAI will have a significant impact across all industry sectors

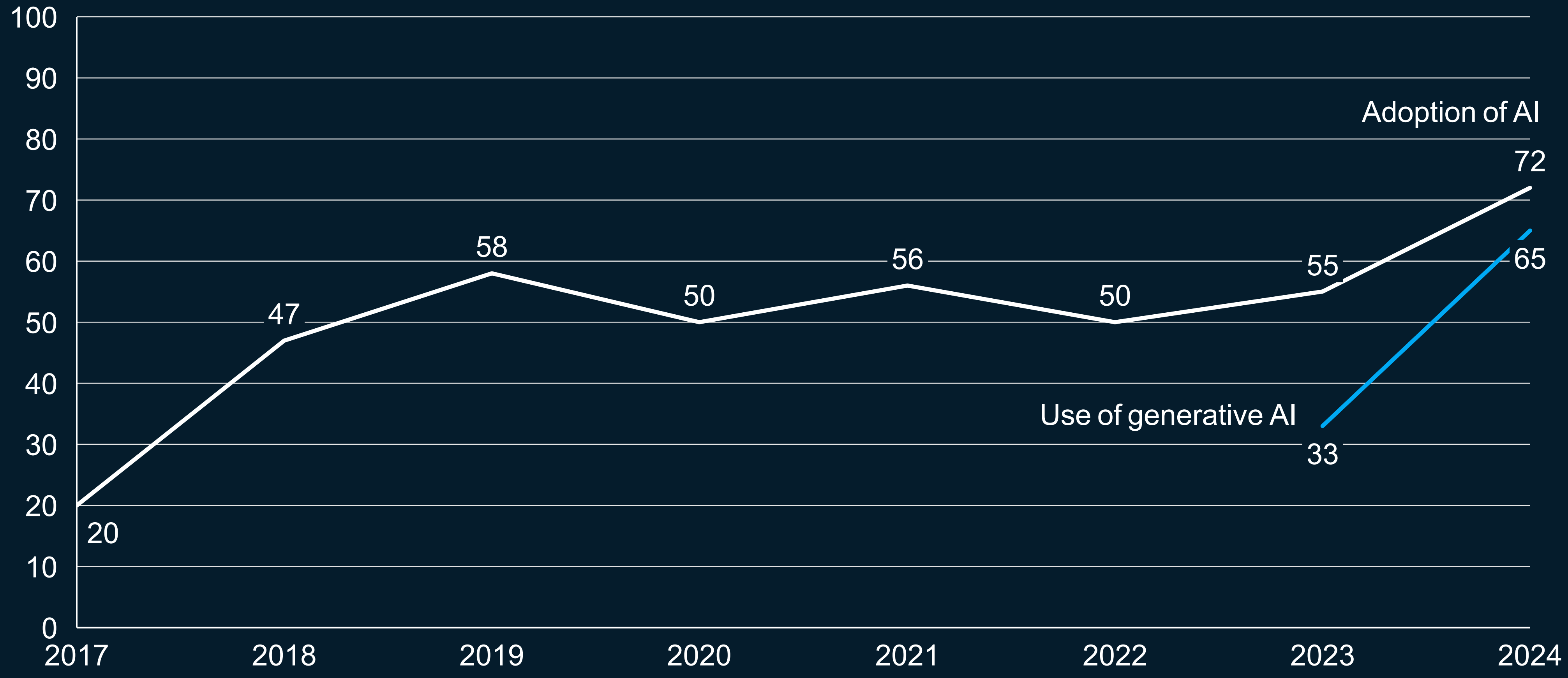
GenAI annual productivity impact by sector (Total, $ billion)



Source: McKinsey report: The Economic Potential of Generative AI: The Next Productivity Frontier

AI adoption worldwide has increased dramatically in the past year, after years of little meaningful change

Organizations that have adopted AI in at least 1 business function,[1] % of respondents



[1] In 2017, the definition for AI adoption was using AI in a core part of the organization's business or at scale. In 2018 and 2019, the definition was embedding at least 1 AI capability in business processes or products. Since 2020, the definition has been that the organization has adopted AI in at least 1 function.

Source: McKinsey Global Survey, *The state of AI in Early 2024: Gen AI adoption spikes and starts to generate value*; 1,363 participants at all levels of the organization, Feb 22–Mar 5, 2024

The GenAI solution architecture AI extends beyond the foundation models

Applications and models are required but not sufficient

UI/UX and applications to get GenAI into production at scale with the right UI/UX interface is critical

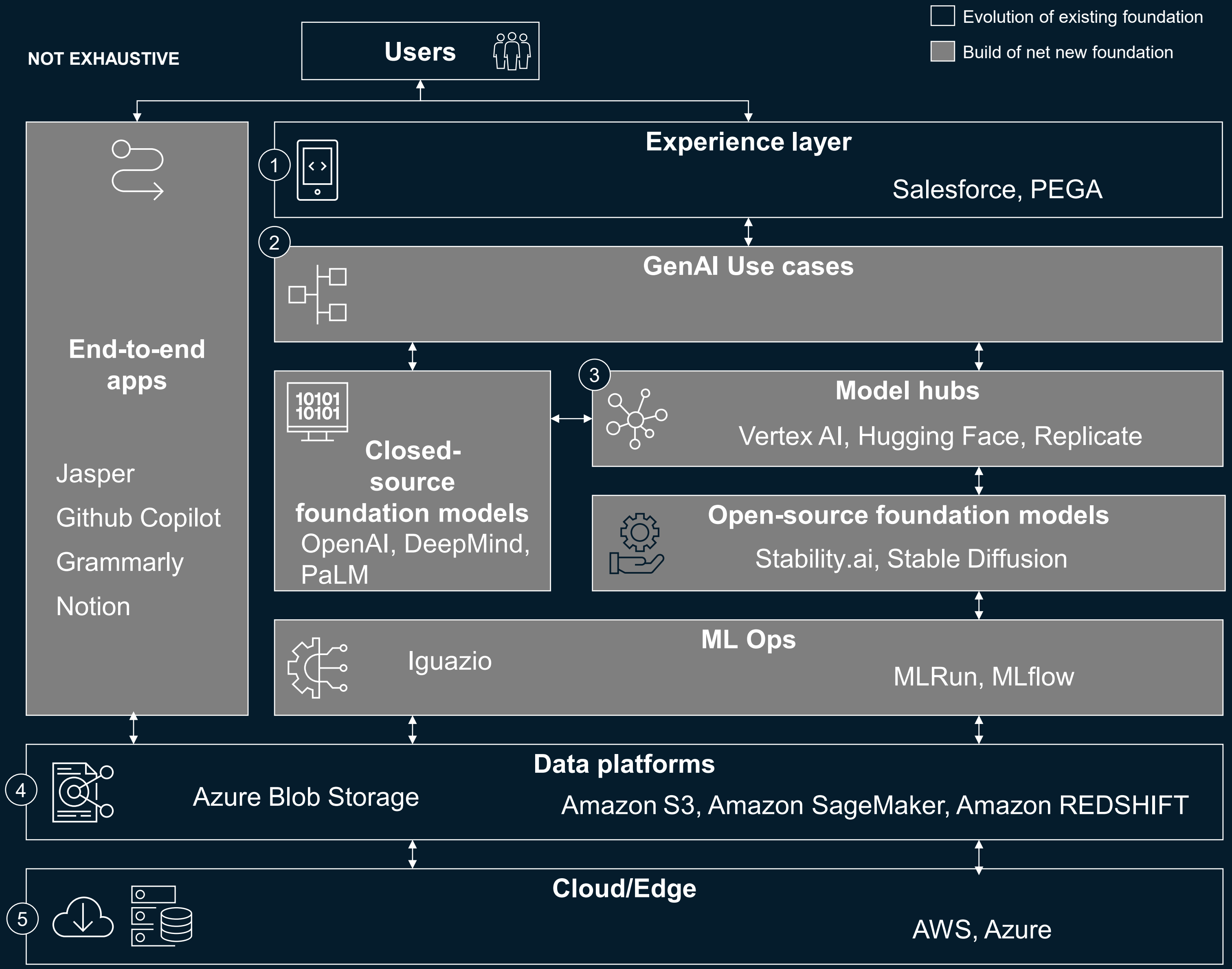
Data architecture will be a must, including access to large bodies of unstructured data

Infrastructure like Cloud and Edge will be in more demand than ever before

Processes and people implications will be critical to address for GenAI to unlock its full potential (“human in the loop”)

A GenAI tech stack will eventually need to be built on existing foundations

NOT EXHAUSTIVE



□ Evolution of existing foundation

▨ Build of net new foundation

Users

End-to-end apps

Jasper

Github Copilot

Grammarly

Notion

1 Experience layer

Salesforce, PEGA

2 GenAI Use cases

Closed-source foundation models

OpenAI, DeepMind, PaLM

3 Model hubs

Vertex AI, Hugging Face, Replicate

Open-source foundation models

Stability.ai, Stable Diffusion

ML Ops

Iguazio

MLRun, MLflow

4 Data platforms

Azure Blob Storage

Amazon S3, Amazon SageMaker, Amazon REDSHIFT

5 Cloud/Edge

AWS, Azure

Key evolutions of tech stack to support GenAI

Building and scaling GenAI will require evolving all layers of the client technology stack, including:

1. **Experience layer:** Front end channels (e.g., Salesforce, PEGA) will need to be integrated into GenAI workflows
2. **GenAI use cases:** E2E methodology and accelerators to implement GenAI use cases (prompt / context engineering, QA & risk controls, human feedback)
3. **Models and ML Ops:** access to different LLMs will need to be established, and automated ML Ops pipelines will be needed to deploy and adapt LLMs
4. **Data platforms:** Existing data stores (e.g., Amazon Redshift) will need to be vectorized and indexed to prepare data that will be ingested by LLMs
5. **Cloud/Edge:** foundations will need to be evolved to create modular isolation zones, along with implementing security

Key Benefits of GenAI at the Edge

Low Latency, Real-Time

GenAI at the edge processes data locally, providing real-time responses and interactions

Scalability and Cost Efficiency

Distributing AI processing across edge devices helps scale applications more efficiently and reduce implementation cost

Reliability and Low Bandwidth

Processing data locally improve reliability at low connectivity and reduces amounts data sent to centralized cloud servers

Security and Regulation Compliance

Local data processing minimizes the risk of data breaches and ensures sensitive information remains on the device

There are 7 trends driving need for Edge, 4 are driven by GenAI

1. Driven by GenAI, data creation continues to explode, increasing demand for datacenters

64ZB

Accumulated digital universe of data in 2021

19% Increase in installed base until 2025

2. GenAI at end devices requires operations at low latencies and high bandwidth

$13.2T

global economic value from 5G use cases will be made possible by 2035

3. GenAI attracts more stringent data regulations on local retention

Data regulations are mandating enterprises to retain sensitive data within regional boundaries; enterprises often react by running workloads outside of public cloud

>60

Countries had data protection and localization requirements in 2021

4. Moving data across environment is costly and prevents full utilization of data

<20%

of the data generated by enterprises is used due to challenges including latency and costs associated with moving data

5. For speed and security running GenAI, enterprises continue to take a hybrid of on-prem and cloud

90%

enterprises will continue to have significant amount of their IT hosting spend on on-prem and private infrastructure

6. The world's computing infrastructure is getting more distributed – Edge DCs

26%

of all servers shipped in 2024 (5.5 million) will be deployed at the edge – up from 20% in 2019

7. As attack grows, enterprises are looking to run isolated environments that minimize security risks

>33%

organizations worldwide have experienced a ransomware attack or breach that blocked access to systems or data – only 13% of them reported experiencing a ransomware attack/breach and not paying a ransom

Source: IDC, Omdia

Edge market is shaped by investment from suppliers, as well as innovation in GenAI

Not exhaustive

New megatrends in the Edge space…	…are accelerating the ecosystem and customer uptake	Examples	
Edge **use cases becoming more "intelligent"** with more complex algorithms being run	**65%** of Edge use cases will be AI driven by 2027	NVIDIA	Nvidia is pushing use cases in Edge AI such as energy forecasting, predictive maintenance in manufacturing, virtual assistants in retail, etc.
Large investments with significant focus on foundational use cases such as "content delivery networks" and "virtual network functions"	**$274bn** estimated worldwide spend on Edge in 2025, growing with a CAGR of ~16% from 2022	AWS Azure Google Cloud	Cloud hyperscalers heavily investing, e.g., Microsoft invested $5+ bn in the past 4 years. Edge share of cloud capex is expected to reach 60% of spend in 2025, up from only ~3% in 2020
Ecosystems and partnerships are evolving between Telcos and Hyperscalers	**35+** Telco & Hyperscaler partnerships announced in the last two years for mobile Edge	Amazon Google Cloud Anthos Azure Sphere	Instead of competing, hyperscalers and telcos are partnering to obtain different capabilities in the Edge stack e.g.; AT&T & GCP, KDDI & AWS
Edge services are gaining a higher share of wallet moving beyond mostly infrastructure spend today	**50%** of Edge spending will be on Edge services by 2025	Mutable.ai MobiledgeX EdgeConneX	Edge services will be a key segment in the next years with niche players offering new business models

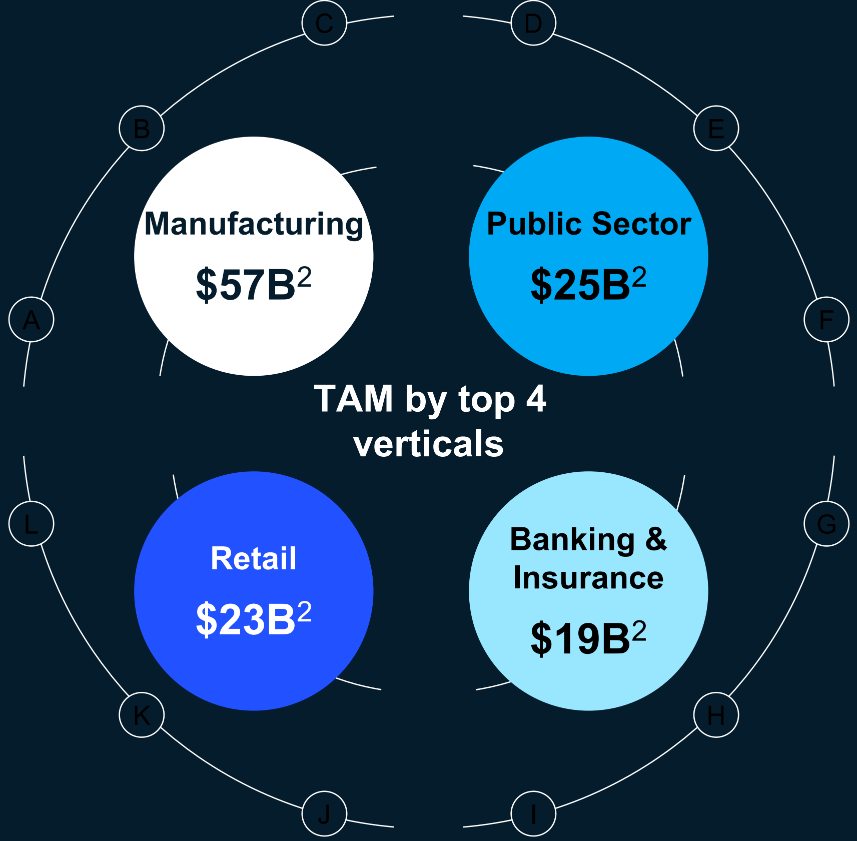
Source: IDC; Gartner; Expert interview, Press Search

Global Edge TAM will exceed $320B by 2026, driven by use cases demanding computing power or GenAI deployment

By 2026, global edge market is expected to be as significant as the public cloud market today[1]

$324B

Total global Edge TAM by 2026[2]



Key use cases[3]

Manufacturing

- Ⓐ AI-enabled visual quality inspections
- Ⓑ Digital twin and related use cases e.g., process optimization, automation
- Ⓒ Predictive maintenance

Public Sector

- Ⓓ Drones and other battlefield device integration
- Ⓔ Public Safety and Emergency Response
- Ⓕ Smart cities and traffic management

Retail

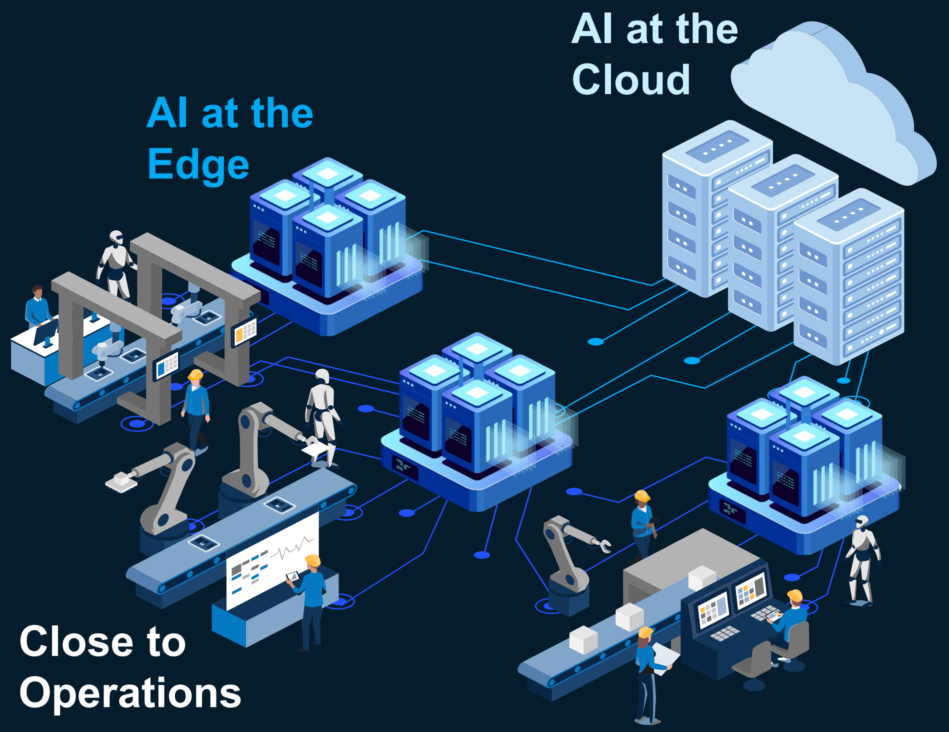
- Ⓖ Inventory monitoring / optimization
- Ⓗ Real-time personalized promotions
- Ⓘ Mobile scanning & self-checkout

Banking & Insurance

- Ⓙ Fraud analysis and investigation
- Ⓚ Real-time personalized promotions
- Ⓛ Automated threat intelligence and prevention

1. Worldwide Public Cloud Services Revenue was at $409B in 2021, according to IDC Worldwide Public Cloud Services Spending Guide | June (V1 2022)
2. Source: IDC Worldwide Edge Spending Guide - Forecast 2022 | Aug (V1 2022); Manufacturing industry combines Discrete and Process Manufacturing; Public Sector combines State/Local Government and Federal/State Government
3. McKinsey research

Manufacturing: AI at the Edge leads to immediate actionable insights with enhanced security and cost efficiency



Low Latency	**Need for insights** closer to data sources for timely decision making at manufacturing operations
Cost Effective	**Prohibit cost of transmitting large volumes of data** to and from the Cloud to local operation
Security and Compliance	**Compliance with data residency regulations,** especially for high-tech industries
Reliability	**Need for increased reliability** to ensure business continuity, especially in areas of low or no connectivity

Example use cases

- Real-time anomaly detection
- Digital manufacturing
- Safety supervision
- Remote location data analytics
- Autonomous navigation
- AR/VR applications

GenAI is accelerating the edge

Data regulation is taking center stage around the world

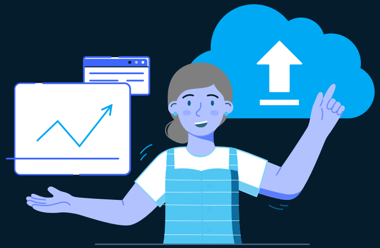
75%

Of all countries have some level of data localization rules[1], requiring rethinking of IT infrastructure which can be fulfilled by adoption of edge storage and computing

GenAI use cases help to drive growth of enterprise edge computing spending

~$324 billion

Projected global addressable market on edge computing in 2026[2], growing at a CAGR of ~14% between 2021-2026



Data volume and velocity is growing at an unprecedented pace

<20%

Share of data generated by enterprises that is ultimately used, due to challenges with latency and costs of moving data across environments

Distributed computing is getting more popular, unlocking real-time insights

26%

Forecast **share of servers** shipped in 2024 that will be **deployed at the edge**—up from 20% in 2019

Edge computing provides **flexibility for organizations** to **process data closer** to where it **originates** with **ultra-low latency,** achieve **data sovereignty**, greater **data privacy** (as compared to cloud) while unlocking a variety of use cases that rely on **real-time data processing**

1. Localization of data privacy regulations creates competitive opportunities, McKinsey & Company, 30th June 2022; 2. Worldwide Edge Spending Guide, IDC, February 2023

McKinsey
& Company