INTELLIGENCE TO SHAPE YOUR TOMORROW

Generative AI: Memory Market Impacts

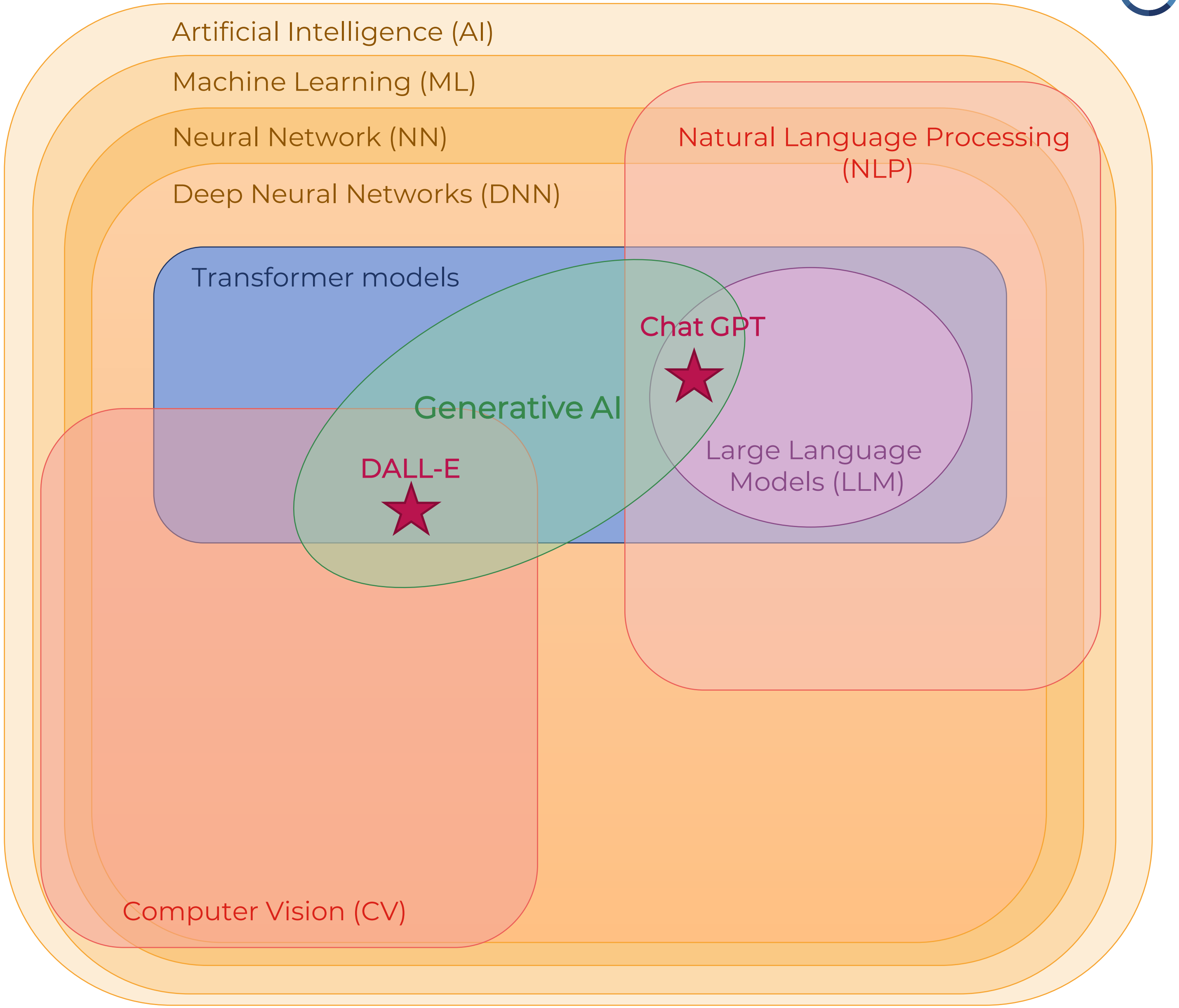
John Lorenz
Principal Analyst, Memory

FMS
the Future of Memory and Storage

YOLE Intelligence

AI DEFINITIONS

- The main goal of AI is for an artificial system to be autonomous. For that, understanding the environment and navigating through it make up a large part of the research.
- Machine learning (ML) is one of the most important fields of AI. The key idea behind ML is teaching an algorithm how to solve a problem by showing it examples of the solution without formally explaining how to reach it.
- An (artificial) neural network is a network of simple elements called neurons and is one of the ML methods. Neurons are organized in layers, and when a network has several layers, it is called a deep neural network (DNN). We are talking about deep learning (DL) when machine learning principles are applied to DNN.
- Transformer neural networks have a DNN architecture using a self-attention mechanism. These are designed to process sequential input. It comprises two main elements: an encoder network and a decoder network.
- Large language models (LLMs) are networks with a transformer architecture and a large number of parameters, trained on large quantities of unlabeled text using self-supervised and semi-supervised learning. They enable many NLP applications such as generating, summarizing, and translating texts.
- Generative AI is a type of AI capable of generating text, images, video, audio, code, 3D models, etc., in response to prompts. Generative AI can create new and original content on demand.



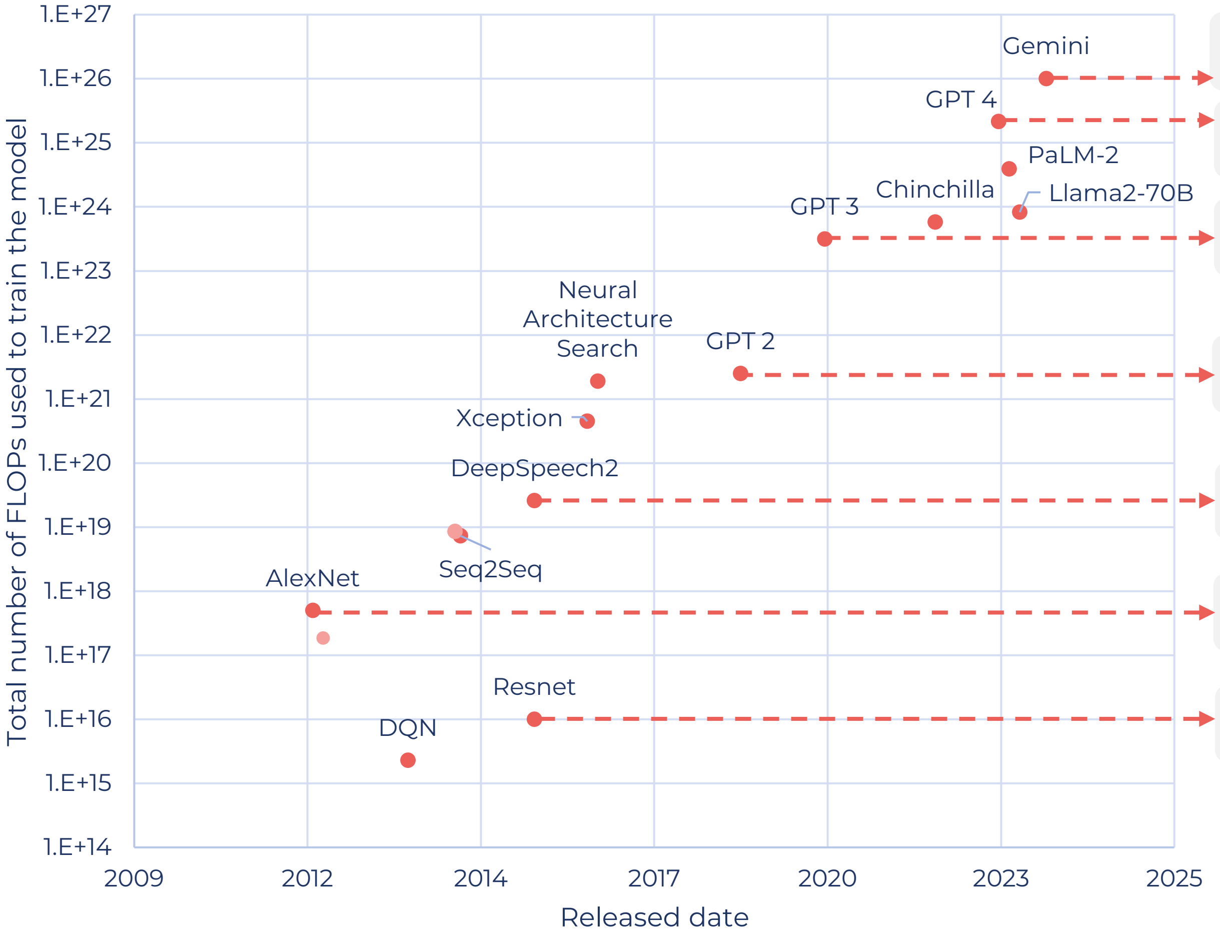
Artificial Intelligence (AI)
Machine Learning (ML)
Neural Network (NN)
Deep Neural Networks (DNN)
Natural Language Processing (NLP)
Transformer models
Chat GPT
Generative AI
DALL-E
Large Language Models (LLM)
Computer Vision (CV)

PROCESSORS FOR ARTIFICIAL INTELLIGENCE – TRAINING FOCUS

Processor requirements* by AI model

* With Yole estimation

Evolution of the number of FLOPs for training AI models



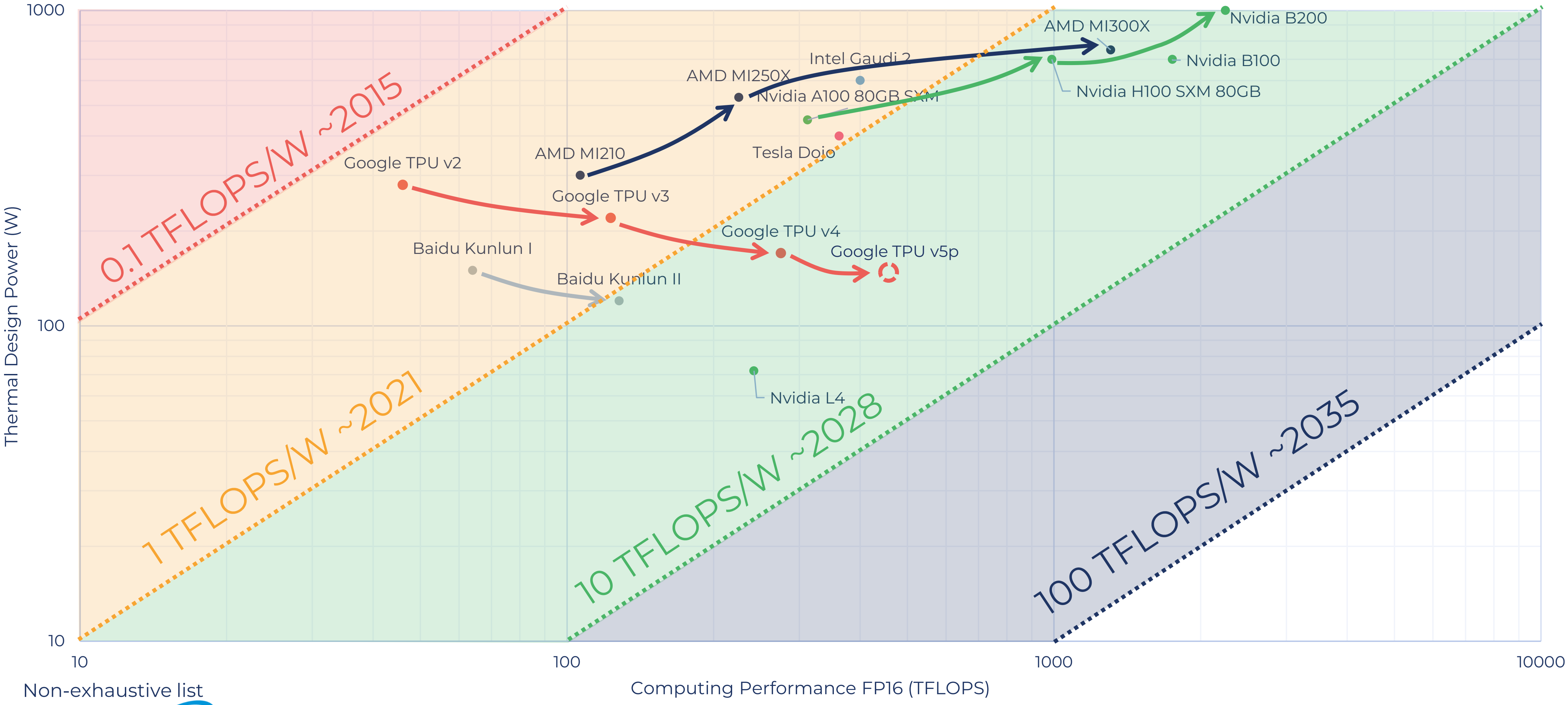
H100 or	TPU v5p or	AWS trn1	Time[1]	Cost[2]	Power[3]
x21,800	x47,000	x113,800	5 Months	$327M	41MW
x8,000	x17,200	x41,600	3 Months	$70M	15MW
x1,500	x3,240	x7,820	1 Week	$1M	3MW
x84	x180	x435	1 Day	$8k	157kW
x21	x45	x109	1h	$85	39kW
x5	x10	x25	5min	$1.63	9kW
x3	x6	x15	10s	$0.03	5kW

[1] AI accelerators utilization is assumed to be ~70% and in TF32 precision.
[2] Google TPU v5p rental = $1.89/hour for a 3-year commitment
[3] 4 Nvidia H100 for 1 CPU Intel Xeon Platinum, at 90% of MAX thermal dissipation power (TDP) and ~60% for other electronic components and cooling.

DATACENTER AI PROCESSORS

Computing performance for artificial intelligence in datacenter

Performance per watt (TFLOPS/W)



Non-exhaustive list

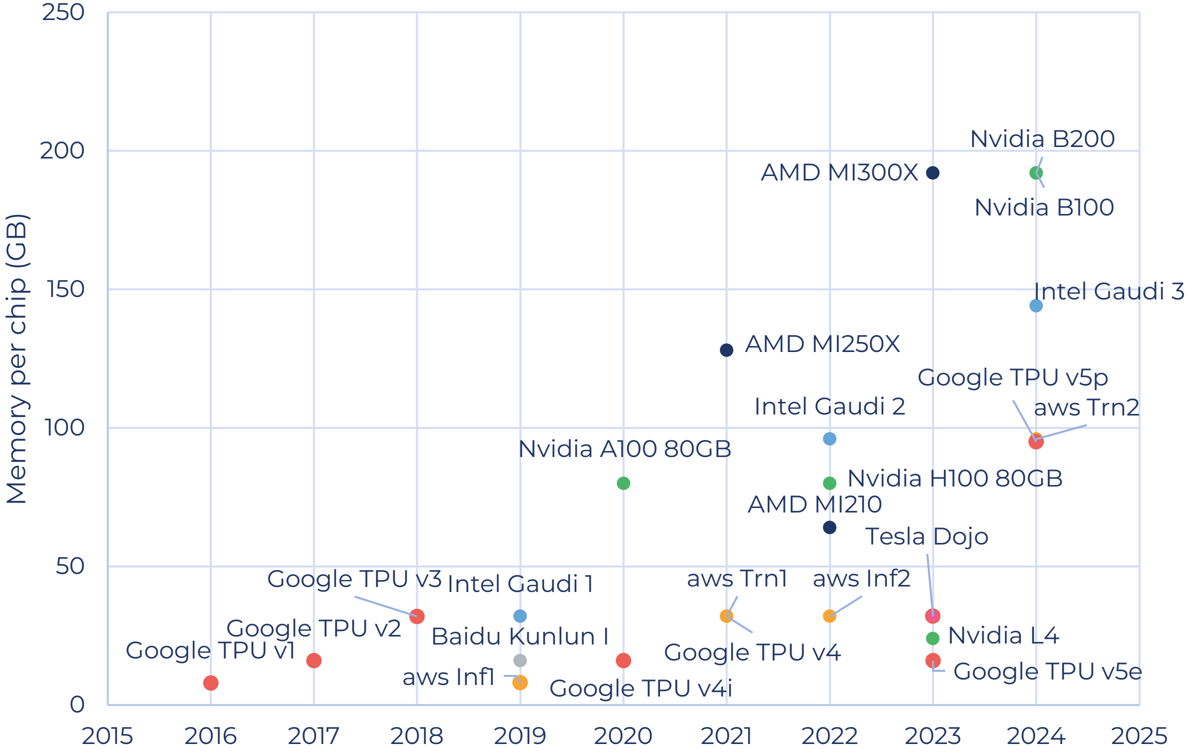
DATACENTER AI PROCESSORS

Memory for artificial intelligence

A GPU's memory capacity is essential for artificial intelligence. Memory has two main purposes:

1. Store the AI model (~parameters)
2. Store the KV cache, i.e., the matrices K and V used for calculating attention in generative ai.

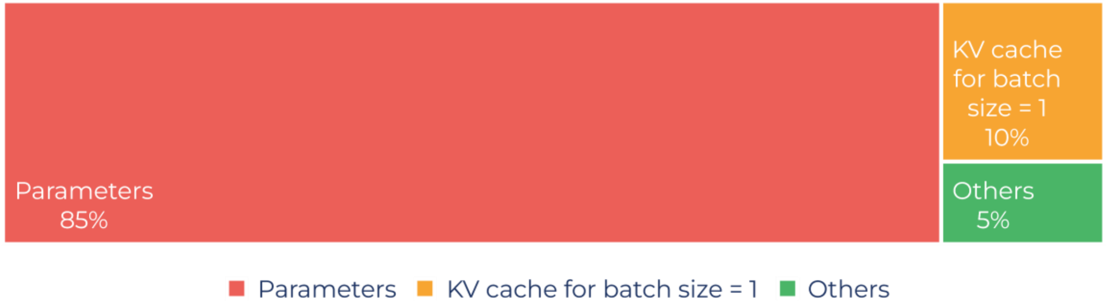
Memory per chip (GB)



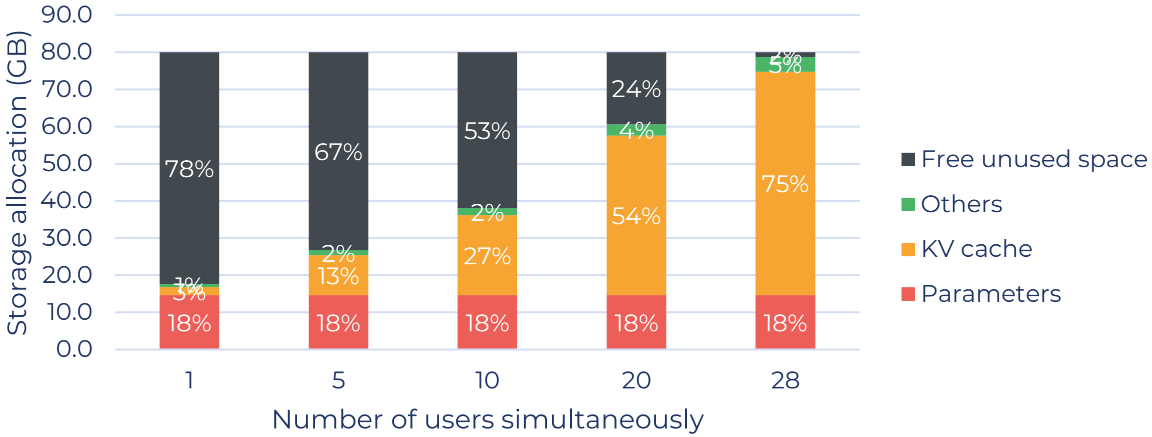
Non-exhaustive list

In the inference phase (the model has been trained before) it is possible to have several users simultaneously. However, this will depend on several parameters, such as the capacity of the HBM memory.

Average (x)PU memory usage for generative AI for one user



Storage allocation for Llama2 7GB (FP16) in inference, by number of users on a single Nvidia H100 80GB



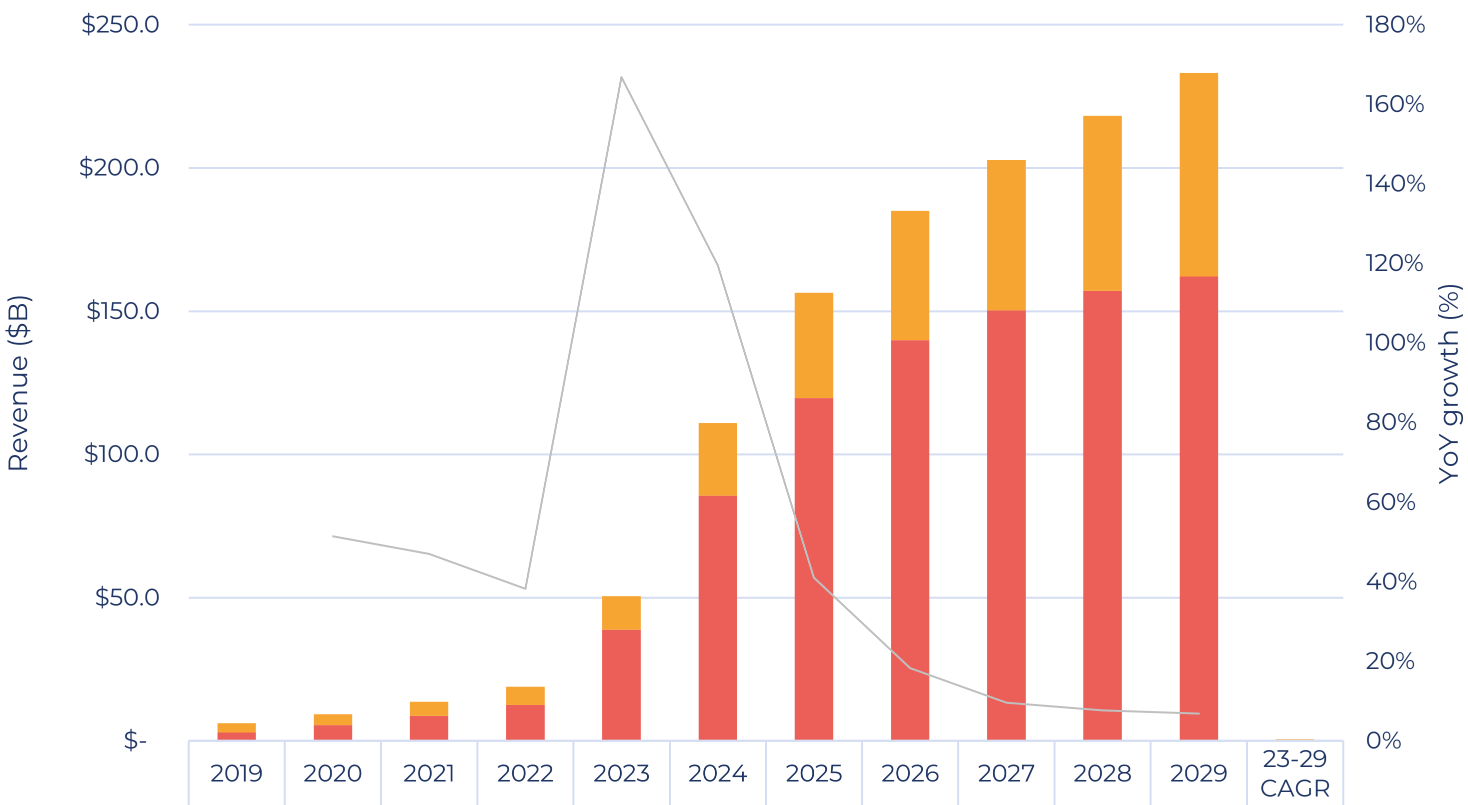
Note: these are theoretical results obtained using the Llama2 model. These values may differ in real-life conditions.

DATACENTER – GPU AND AI ASIC FORECAST

Revenue forecast by type of processor, in $B

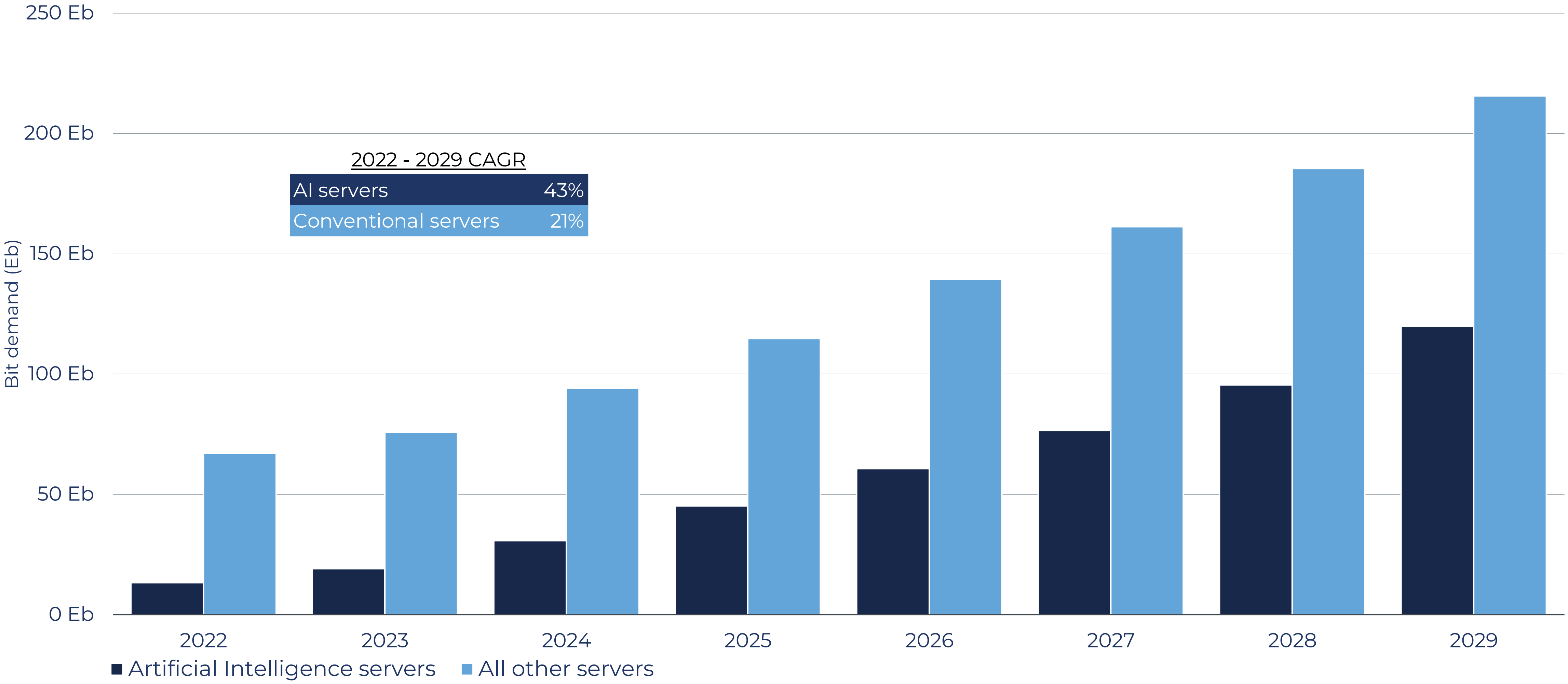
- The massive growth that the datacenter GPU and AI ASIC market experienced in 2023 (167% YoY) is expected to continue in 2024 before stabilizing in the year following. We expect this stabilization since the number of companies able to massively buy GPUs and AI ASIC is limited, and because the lifecycle of these components is also growing on average. However, we don't expect a revenue decrease after this big growth, since AI progress is very fast, the model size is still expanding, and the corresponding applications are far from all being discovered. We expect that the ratio of GPU and AI ASIC used for AI inferences will grow in the coming years.
- The total market is expected to reach more than $150B in 2025 and more than $230B in 2029. It represents a CAGR$_{23\text{-}29}$ of 29%.

Datacenter GPU and AI ASIC revenue forecast, in $B



	2019	2020	2021	2022	2023	2024	2025	2026	2027	2028	2029	23-29 CAGR
🟧 AI ASIC	$3.2	$3.8	$5.0	$6.3	$11.8	$25.3	$36.8	$45.1	$52.5	$61.1	$71.1	35.0%
🟥 GPU	$3.0	$5.5	$8.7	$12.6	$38.8	$85.6	$119.	$139.	$150.	$157.	$162.	26.9%
Total	$6.2	$9.3	$13.7	$18.9	$50.5	$110.	$156.	$185.	$202.	$218.	$233.	29.0%
— YoY (%)		51%	47%	38%	167%	120%	41%	18%	10%	8%	7%	

AI DEMAND IS OUTPACING THE OVERALL SERVER MARKET

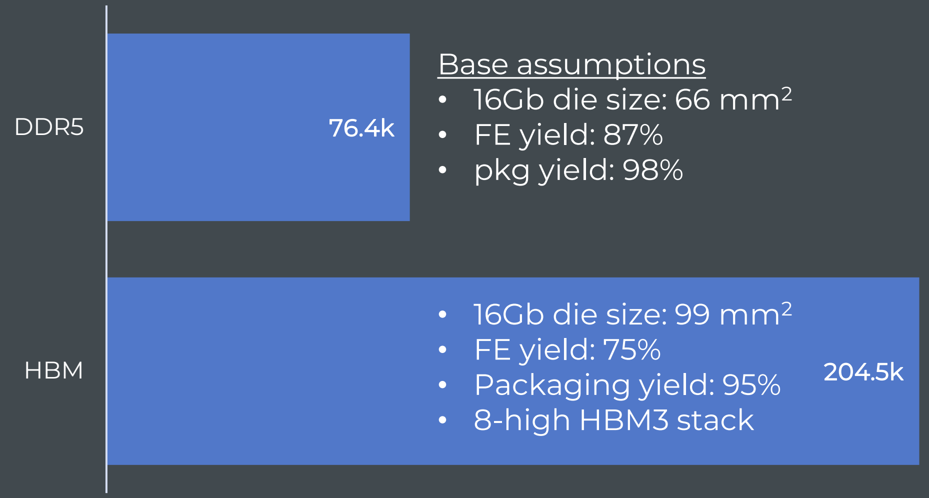


2022 - 2029 CAGR

AI servers	43%
Conventional servers	21%

HBM: MORE COMPLICATED PRODUCTION EFFORT

Wafer Starts needed to ship 1 billion Gb

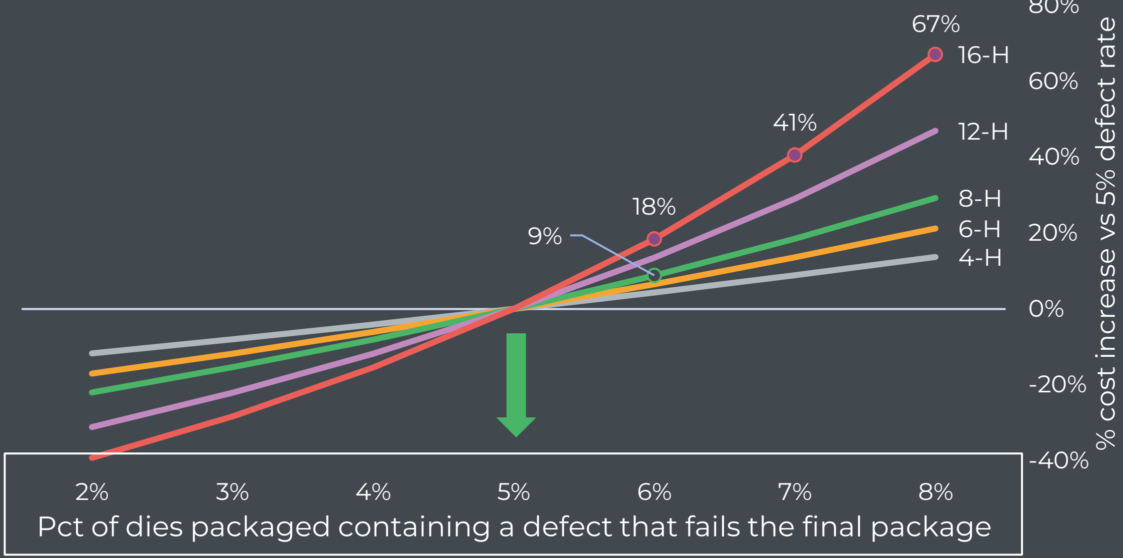


Base assumptions

- 16Gb die size: 66 mm^2
- FE yield: 87%
- pkg yield: 98%

- 16Gb die size: 99 mm^2
- FE yield: 75%
- Packaging yield: 95%
- 8-high HBM3 stack

Compounding yield impact on cost per Gb



HBM requires almost 3X as many wafer starts for same bit output as DDR5

- Die size, TSV area, TSV process yield
- Packaging yield and compounding yield effect

Small variations in yield can greatly distort the product cost

- In an 8-high package, 1 ppt of worse yield results in 9% higher cost/bit
- At 16-high, 1 ppt of worse yield results in 18% higher cost/bit

FROM BASIC AI FEATURE TO GEN AI

Softwares with AI accelerated features

Smartphones with basic AI features

Use of a specific AI* for targeted applications

Models with les than 150M parameters

OR

- Face recognition
- Picture touch up / tagging (apple)
- Assistant

New devices with embedded Gen AI integrated in OS, softwares and games.

New devices with embedded LLM AI

Use of LLM multimodal AI* for assisting the user in all his tasks.

Hybrid

AND

OS

- Personal assistant
- Predictive UI
- …

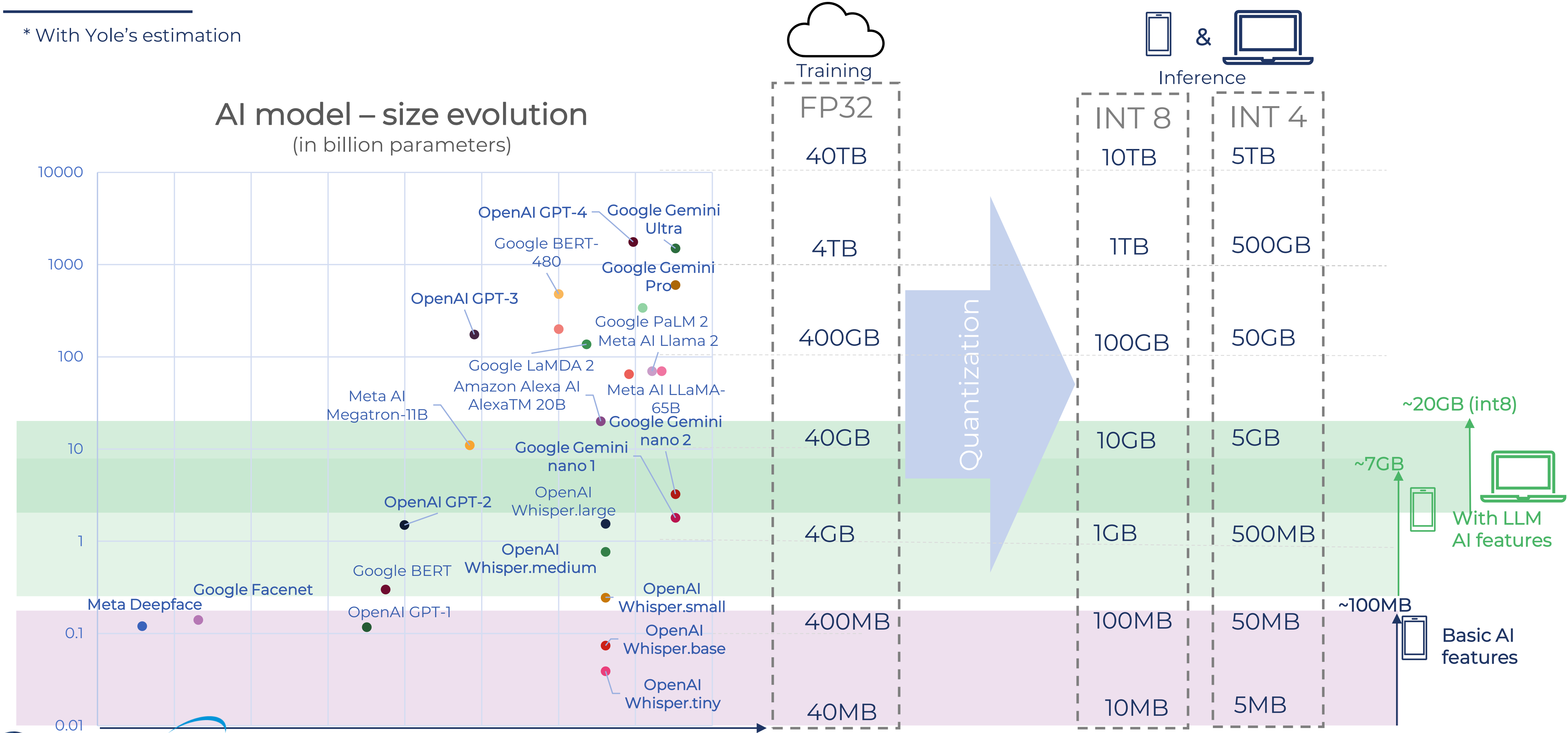
Softwares / App

- Creativity
- Gaming
- Productivity

* AI inference

MARKET TRENDS – BIT-DEMAND GROWTH DRIVEN BY AI

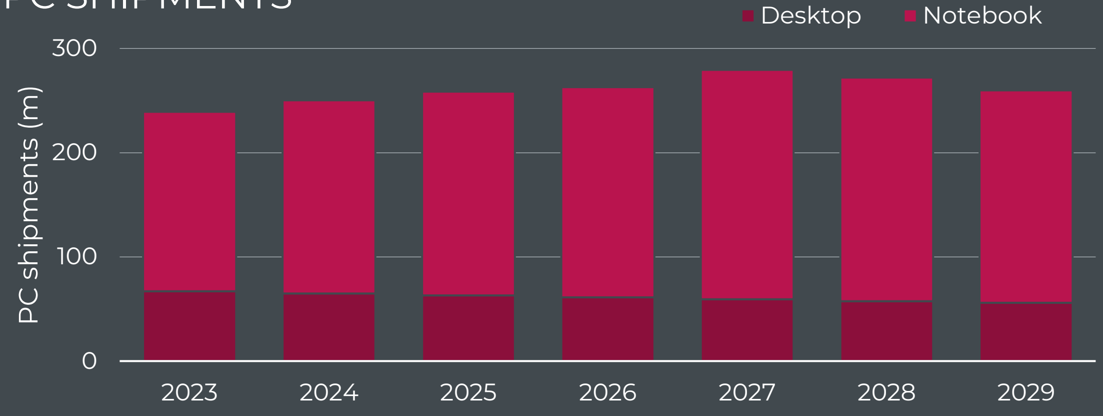
DRAM requirements for AI-model training and inference



* With Yole's estimation

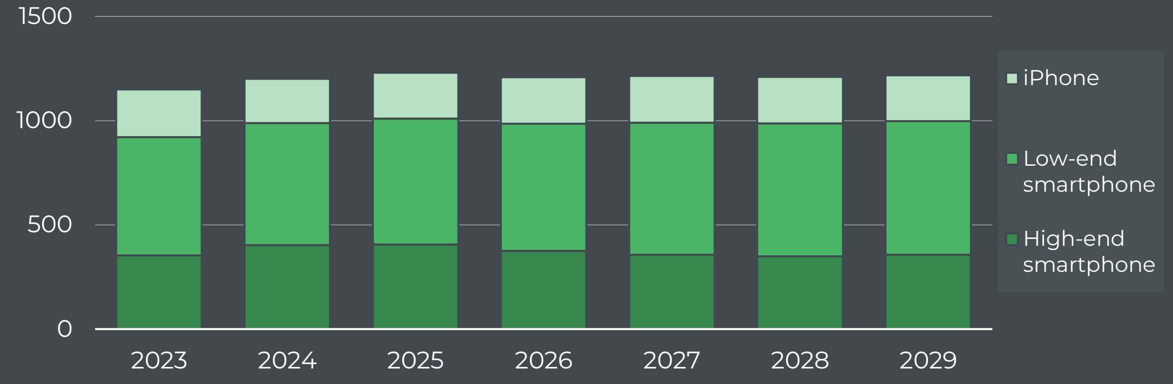
PC AND SMARTPHONE DEMAND SUMMARY

PC SHIPMENTS



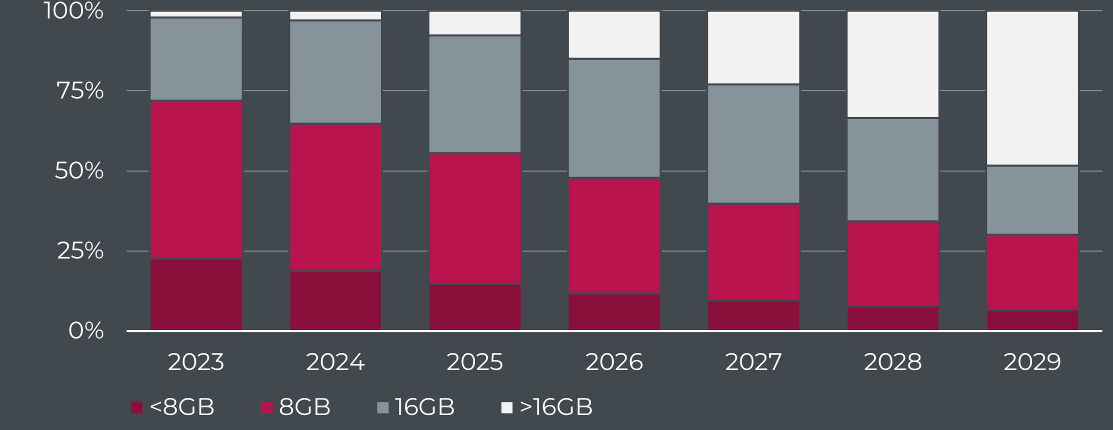
Unit growth	-14%	5%	3%	2%	6%	-3%	-4%

SMARTPHONE SHIPMENTS (m)

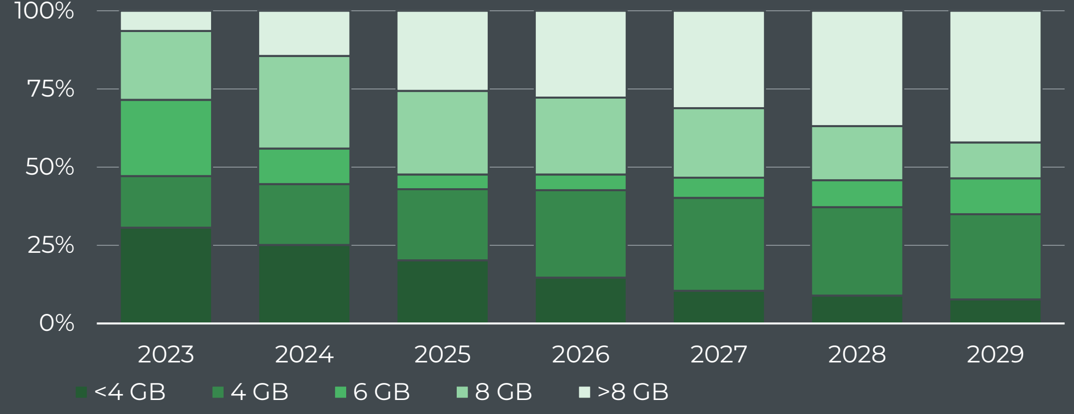


Unit growth	-4%	4%	3%	-2%	1%	0%	1%
5G attach rate	69%	76%	83%	87%	91%	95%	100%

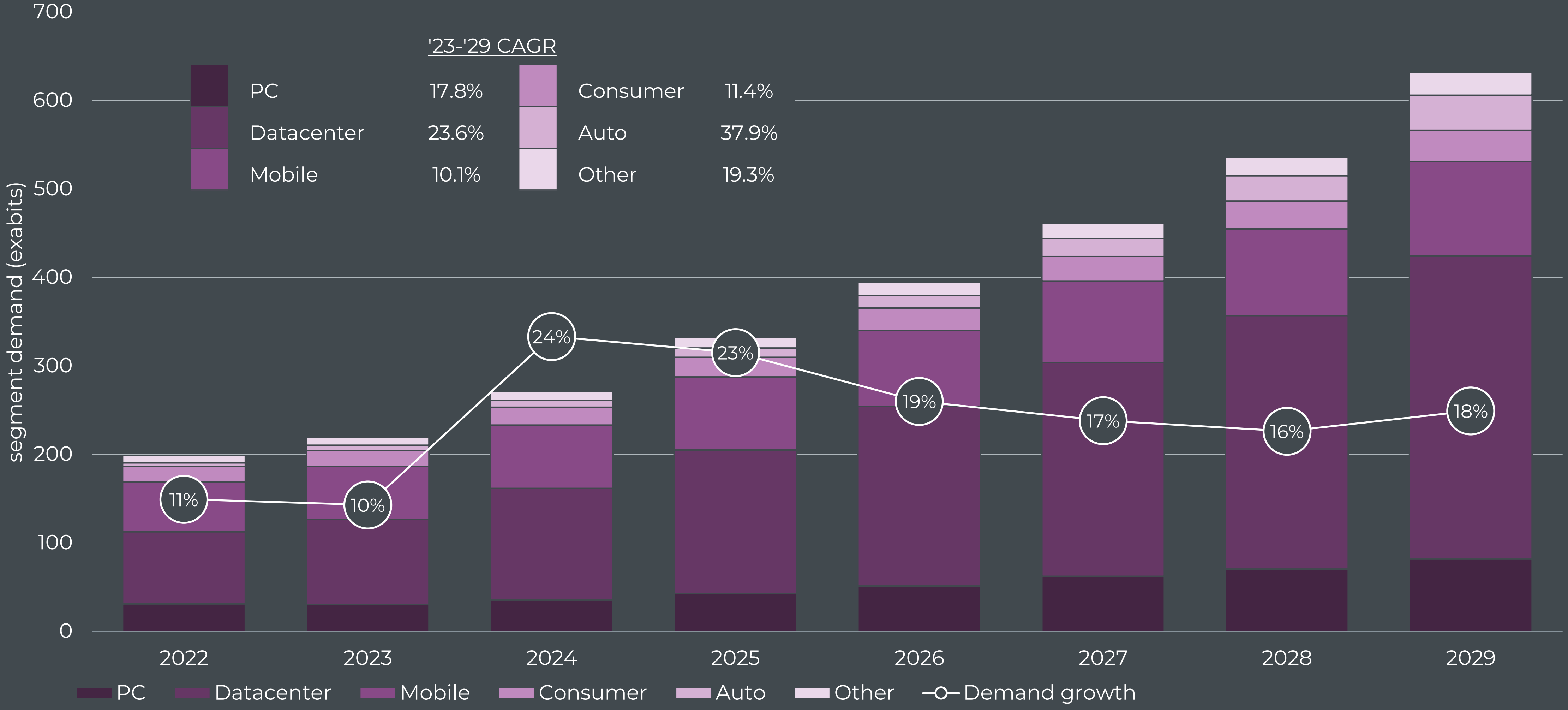
PC DRAM DENSITY MIX (% of units)



SMARTPHONE DENSITY MIX (% of units)



DRAM SEGMENT BIT DEMAND



the Future of Memory and Storage