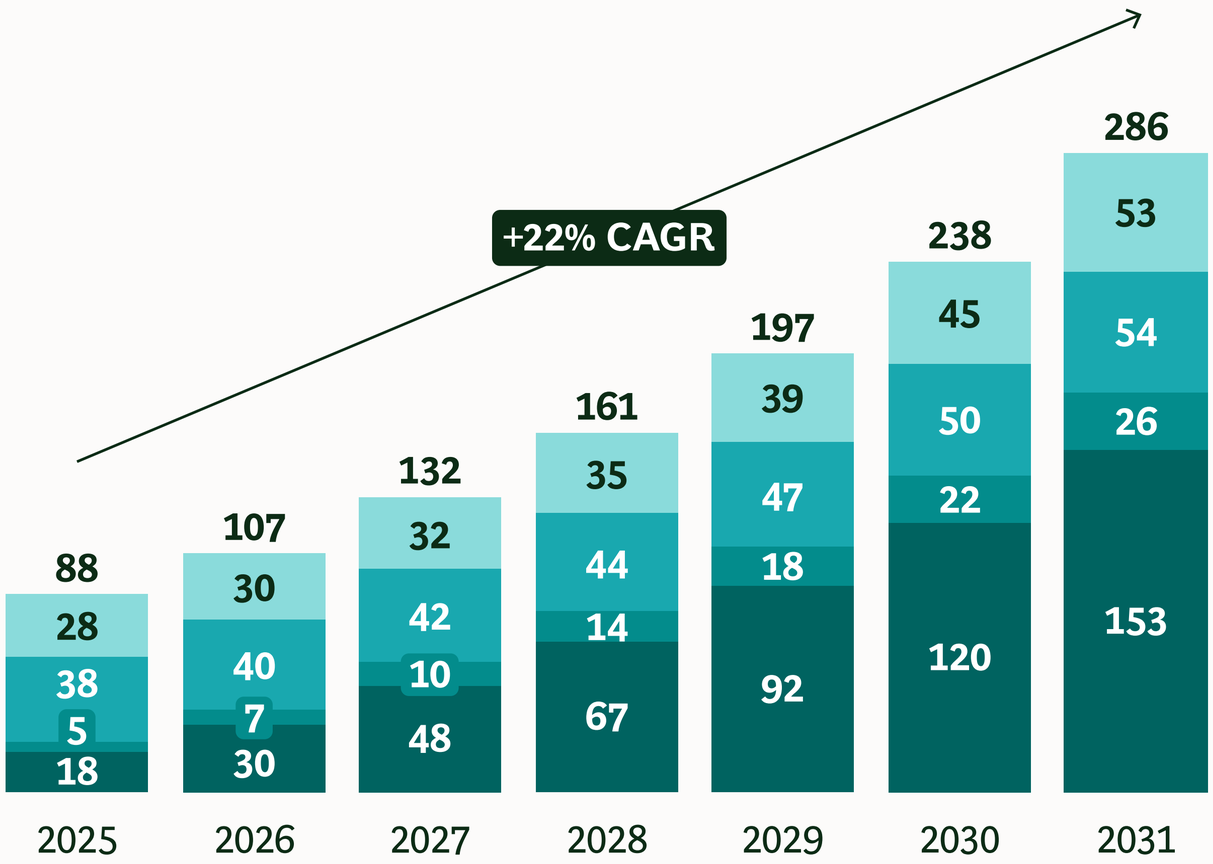
How the AI Compute Shortage Will Create Winners and Losers

November 2026

Global DC supply is expected to reach about 286 GW by 2031, growing at a CAGR of more than 20% over the next five years

Base case for global DC supply (GW)[1]



Drivers of DC demand

CAGR 2025–2031

CPUs[2]
Driven by a need for CPUs for agentic orchestration +11%

Other traditional DC components
(e.g., storage) +6%

High-performance computing +33%

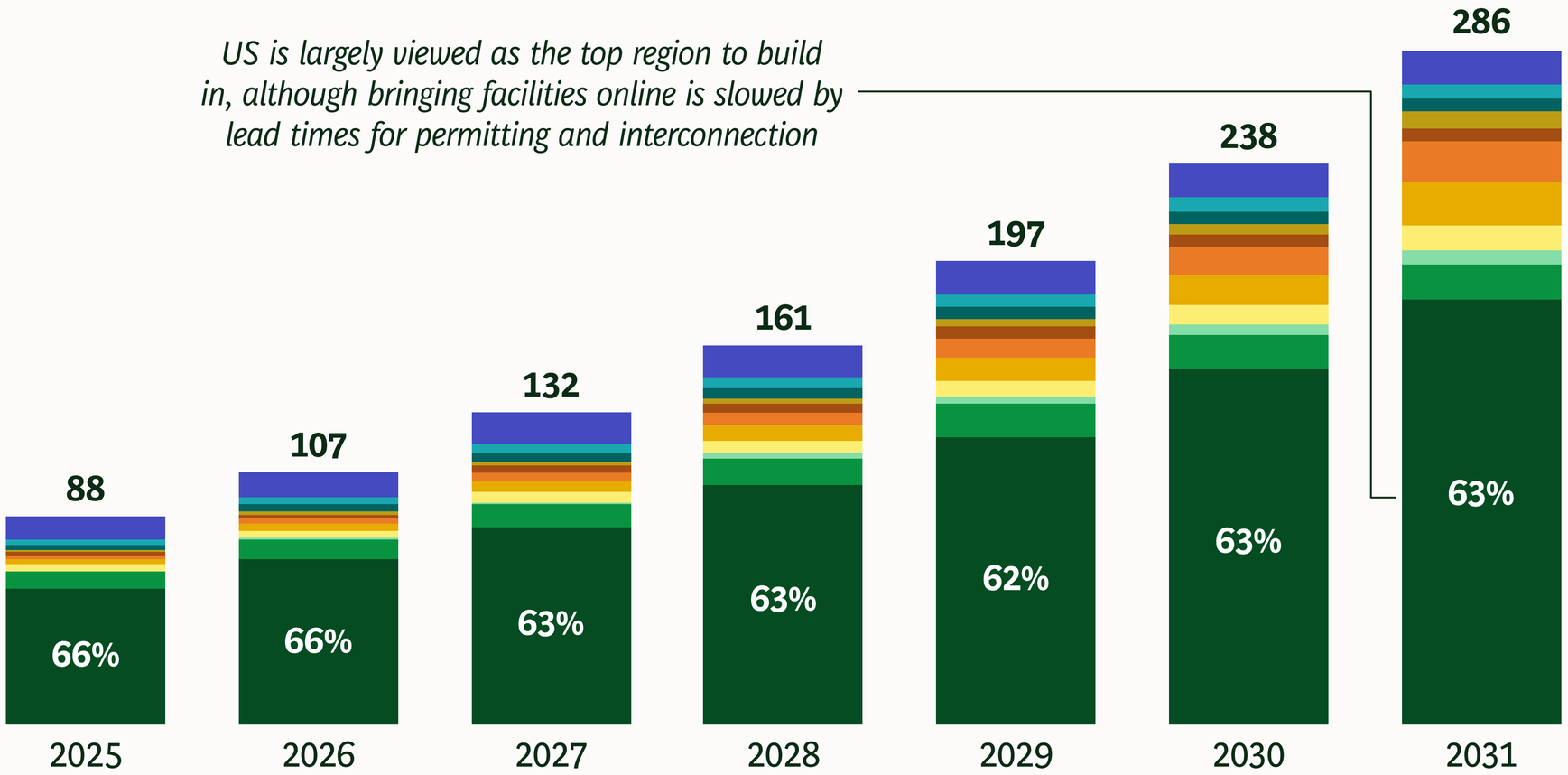
AI chips[2]
Primary DC market driver (training and inference) +43%

Source: BCG analysis.
Note: DC = data center; gigawatt; CPU = central processing unit. The numbers may not tally due to rounding.
[1]Excluding DC supply needed by China and cryptocurrency systems (expressed in critical IT load). [2]CPUs included as part of AI accelerator modules (e.g., Nvidia's Vera CPUs in Vera Rubin racks) are accounted for as part of the AI chips segment.

The US is expected to continue representing about 60% or more of installed DC supply, with the rest distributed across a few global hotspots

Base case for global DC supply by region (GW)[1]

US is largely viewed as the top region to build in, although bringing facilities online is slowed by lead times for permitting and interconnection



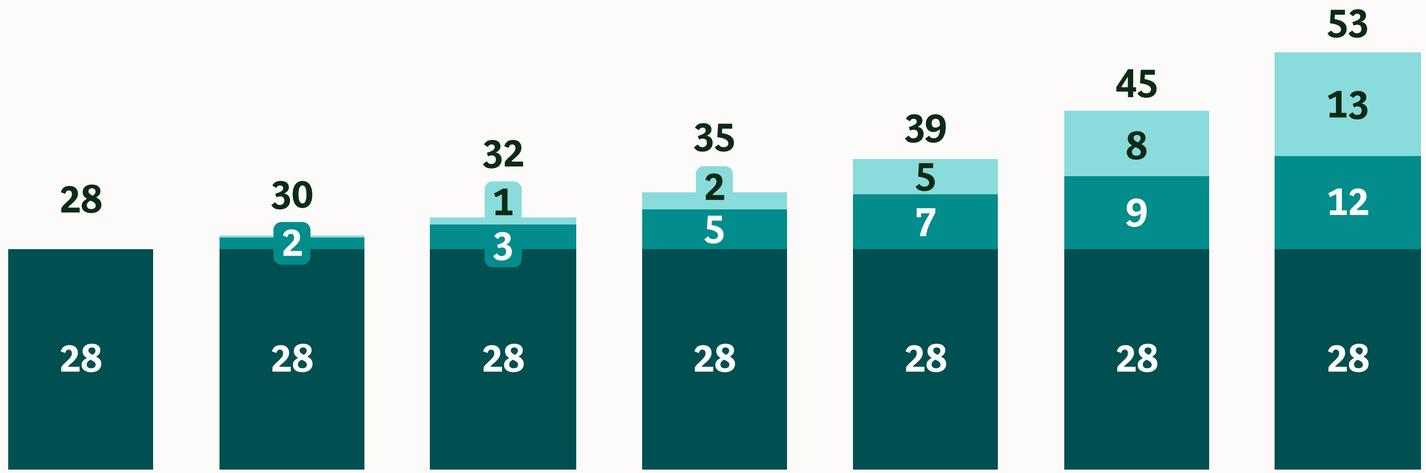
Expected supply, 2025–2031

Others	+6 GW
India	+4 GW
Australia	+3 GW
Canada	+6 GW
Malaysia	+4 GW
Middle East	+16 GW
Rest of Europe	+16 GW
Nordics[2]	+8 GW
Iberia[2]	+5 GW
FLAP-D[2]	+8 GW
US	+122 GW

Source: BCG analysis.
Note: DC = data center; GW = gigawatt; FLAP-D = Frankfurt, London, Amsterdam, Paris, and Dublin. The numbers may not tally due to rounding. Given the limited visibility to date, the US forecast is not yet adjusted for the potential impact of the US midterms.
[1]Excluding DC supply needed by China and cryptocurrency systems (expressed in critical IT load). [2]Within Europe, FLAP-D's power is constrained, while the Nordics and Northern Spain face other challenges with scaling, including access to labor and components that are needed for a buildout that is many times larger than what the region has previously undertaken.

Agentic AI is expected to increase demand for CPUs in the medium term

Base case for CPU installed base growth (GW)[1, 2]



	2025	2026	2027	2028	2029	2030	2031
Share of cumulative growth from AI uplift (%)		6	22	31	39	47	53
Share of cumulative growth from traditional CPUs (%)		94	78	69	61	53	47

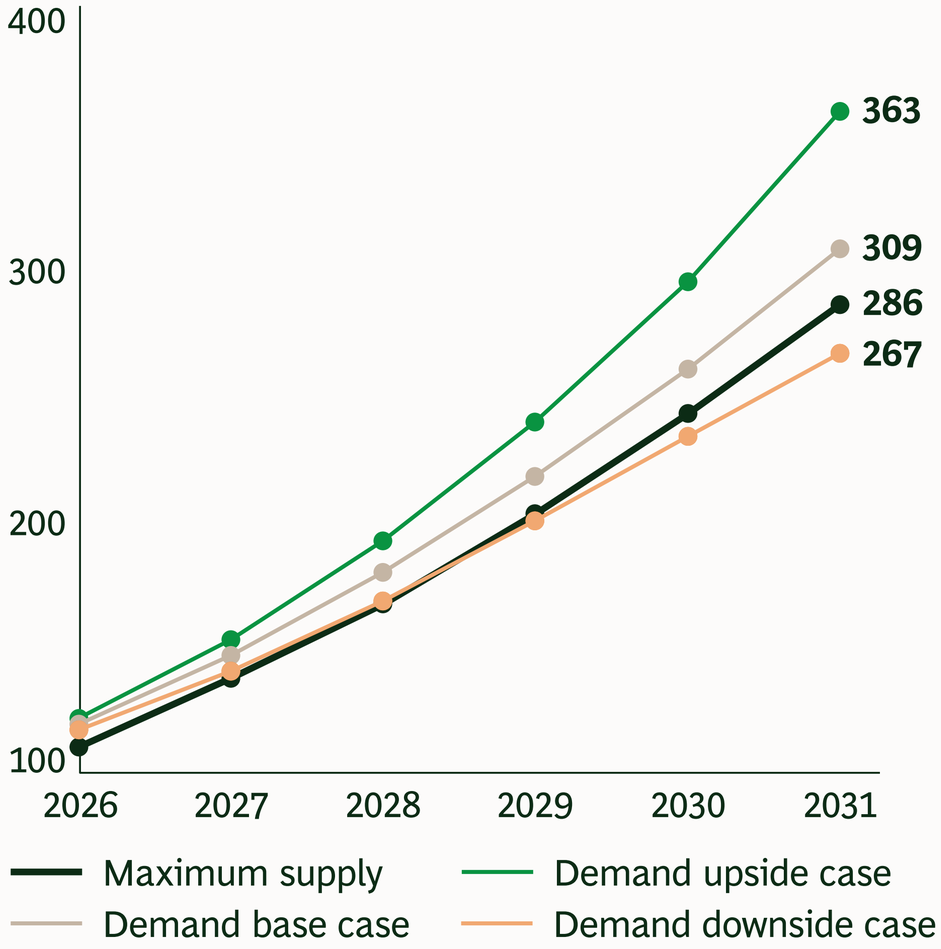
AI uplift Traditional CPUs Current installed base

- CPU capacity labeled as “AI uplift” represents capacity required to support agentic AI workloads (e.g., for orchestration across agents)
- CPU capacity labeled as “traditional CPUs” represents capacity required for all other workloads, regardless of agentic AI needs
- About 50% of CPU growth by 2031 is likely to be attributable to AI uplift and about 50% to traditional CPU growth

Source: BCG analysis.
Note: CPU = central processing unit; GW = gigawatt. The numbers may not tally due to rounding.
[1]Excluding DC supply needed by China and cryptocurrency systems (expressed in critical IT load). [2]Excludes CPUs included as part of AI accelerator modules (e.g., Nvidia’s Vera CPUs in Vera Rubin racks).

Both the base and upside cases expect DC demand to continue outpacing DC supply into the next decade

Estimated DC supply and demand (GW)[1]



DC demand scenarios

Upside case

AI increasingly substitutes for human work, with autonomous agents operating end-to-end workflows

- Faster capability gains expand task scope and duration
- Enterprise adoption shifts from copilots to autonomous workflows
- Higher usage intensity and demand for training accelerate growth in DC demand

Base case

AI remains primarily a tool used by humans but becomes embedded across everyday enterprise and consumer workflows

- Capabilities improve steadily and adoption broadens across use cases
- Enterprises scale copilots, AI-enabled software, and APIs
- Usage and training grow consistently but remain mediated by humans

Downside case

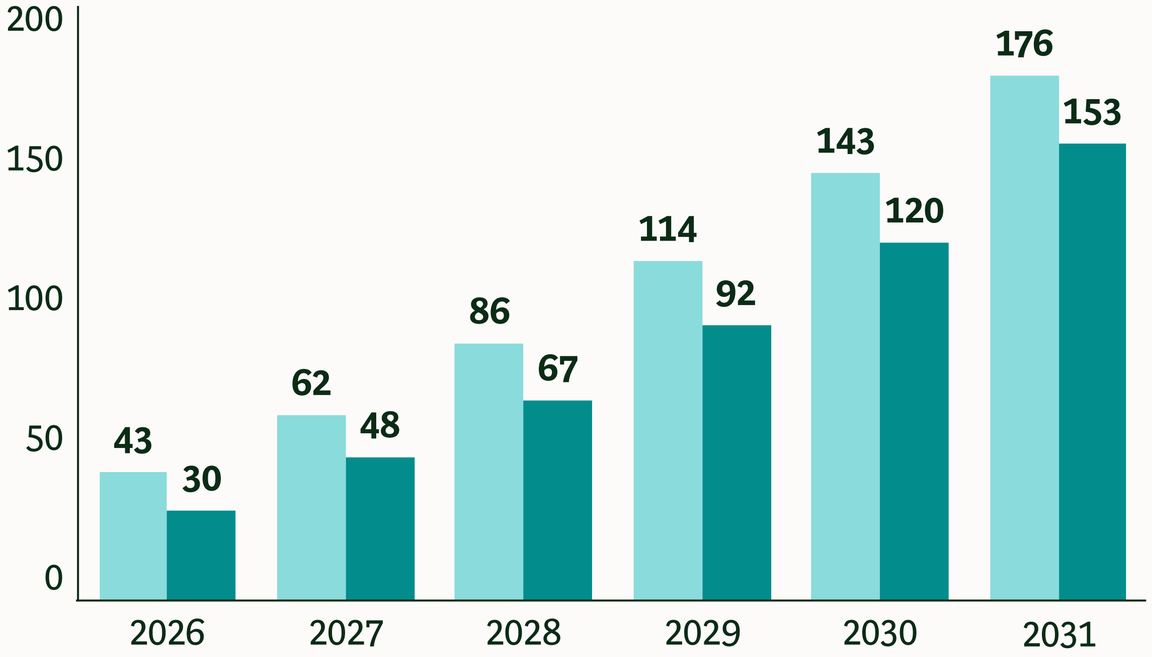
Gains in AI capabilities slow and adoption remains constrained, limiting both usage intensity and training demand

- Improvements become more incremental, especially for complex tasks
- Enterprise deployment remains constrained by data, legal, and ROI barriers
- Lighter usage and slower growth in training reduce DC demand

Source: BCG analysis.
Note: DC = data center; API = application programming interface.
[1]Excluding DC supply needed by China and cryptocurrency systems (expressed in critical IT load). Assumes non-AI supply and demand stay in equilibrium.

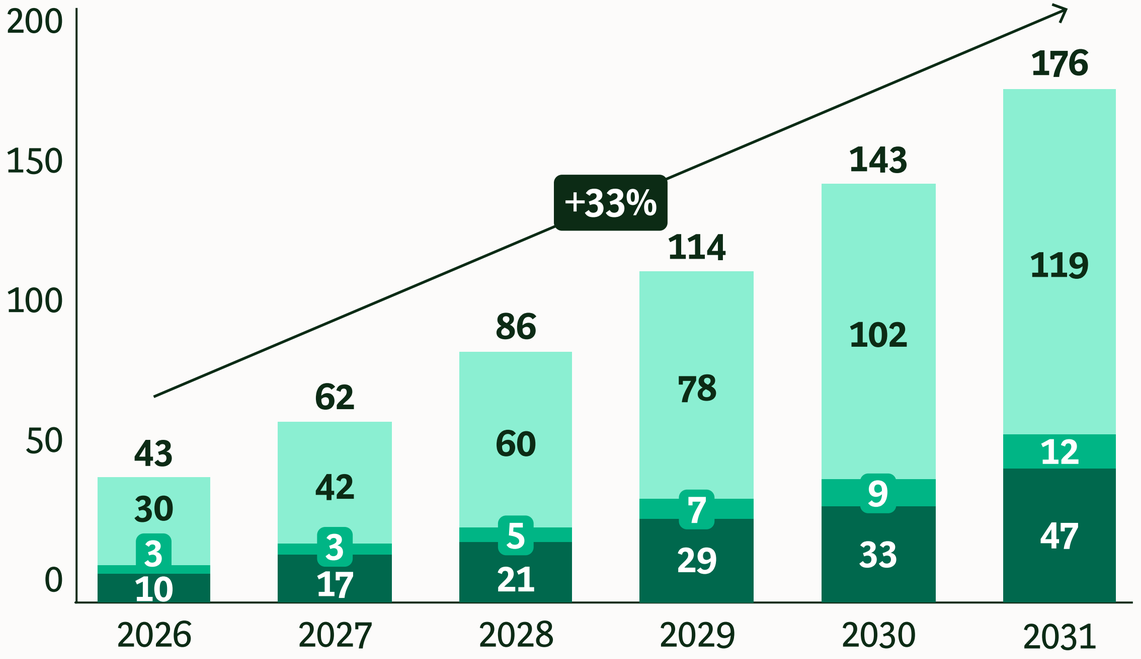
The base case estimates AI demand to be about 15% higher than AI supply by 2031

Base case for AI supply versus demand (GW)[1]



	2026	2027	2028	2029	2030	2031
Gap (GW)	13	14	19	21	23	23
Gap (%)	42	30	28	23	19	15

Base case for AI demand by workload (GW)[1]

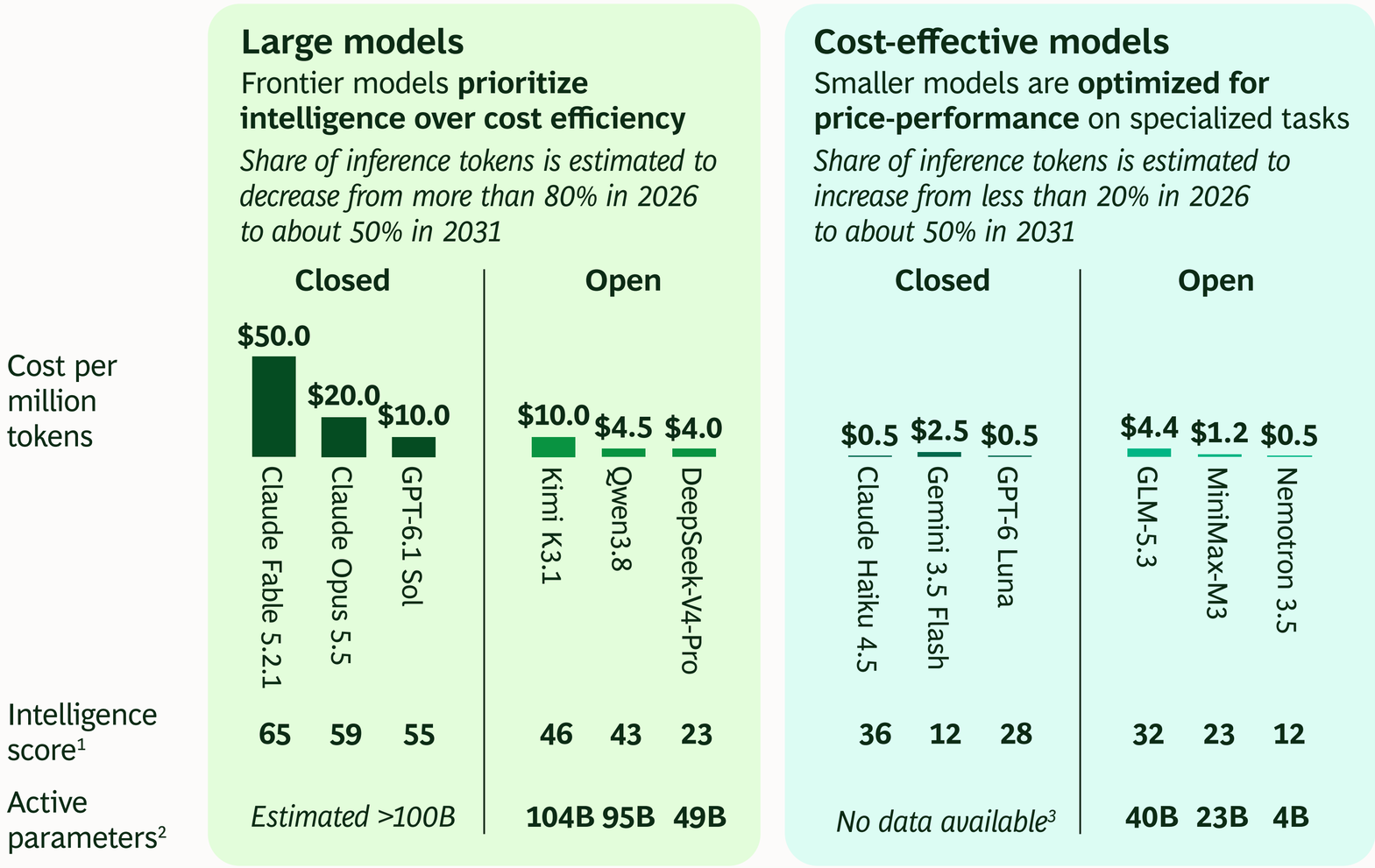


Enterprise inference	Other inference[2]	AI training
+32% CAGR	+31% CAGR	+36% CAGR

Source: BCG analysis.
Note: The numbers may not tally due to rounding.
[1] Expressed in critical IT load or cumulative server power.
[2] Includes consumer, government, and physical AI inference.

Enterprises are gradually shifting toward smaller, more cost-effective models across proprietary and open-source software

Base case for model price and intelligence score by model class



Key elements

AI model architectures are becoming more efficient
MoE and other technical efficiencies are reducing the active compute required per token

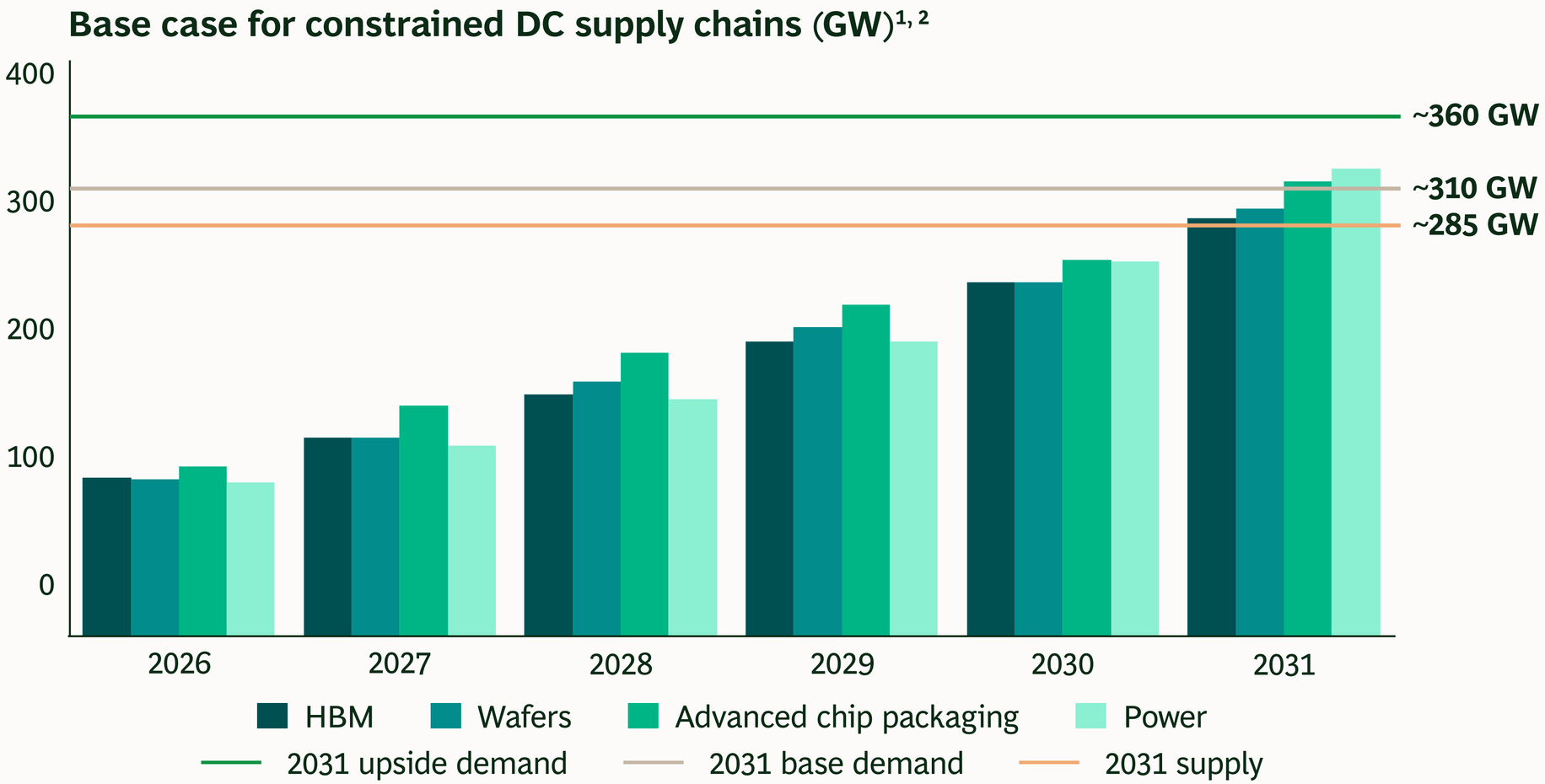
Enterprises are increasingly focused on reducing token spending
Rising AI usage, particularly from agentic workflows, is increasing token consumption per task, pushing buyers to consider cost-effective models when frontier intelligence is not required

Open-weight and smaller models are becoming viable alternatives
Smaller models are offering comparable intelligence for specific tasks; intermediaries such as inference providers are reducing deployment requirements for open-weight models

The use of model-routing software is rising
Model routing is accelerating multimodel AI adoption, with platforms such as OpenRouter and Cursor directing workloads to the best-fit model on the basis of cost, capability, and latency

Sources: Artificial Analysis; Epoch AI; company price sheets; BCG analysis.
Note: MoE = mixture of experts (which is a machine learning technique).
[1]Based on Artificial Analysis Intelligence Index v5.0.1. [2]Active parameters are an indicator of model size. [3]Active parameters are not disclosed by closed-source labs.

Major constraints continue to affect supply chains simultaneously, with power constraint alleviating in the outer years

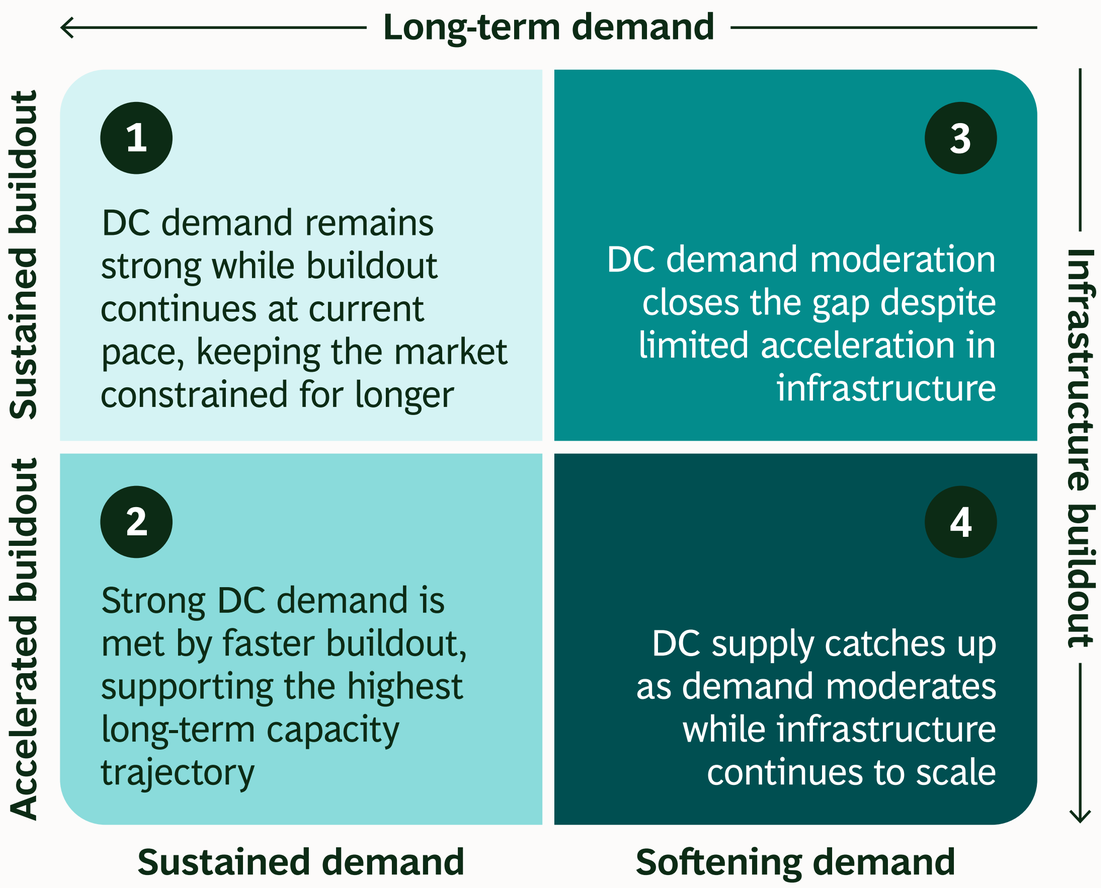


- **HBM** is not ramping up fast enough
- **Wafer** outlook remains a potential concern
- **Advanced chip packaging** is tight, but it's not the limiting constraint
- **Power** is a near-term bottleneck

Source: BCG analysis.
Note: DC = data center; GW = gigawatt, HBM = high-bandwidth memory.
[1]Component capacities are modeled based on the assumption that the input in question is the constraint throughout (that is, what does the output look like if all other inputs are not a constraint). For simplicity in components calculations, conversion to gigawatts was based on extending model's base shipments to GW ratio.[2]Excluding DC supply needed by China and cryptocurrency systems (expressed in critical IT load).

After 2031, five factors are critical to determining how fast DC supply can catch up with the base case for DC demand

DC supply is likely to depend on DC demand and infrastructure buildout scenarios[1]



Five key factors

Demand signals

- Model capabilities: End-to-end task performance on complex tasks
- Enterprise diffusion: Scaling of AI across enterprise use cases and functions
- Token economics: Trajectory of inference costs and model routing toward the most economical and efficient models (large-model versus small-model mix)

Hardware efficiency and utilization

- Chip utilization (MFU): Improvements in realized DC supply versus chip capacity
- Power efficiency: Pace of improvement in effective FLOPs per watt

Manufacturing capacity

- Logic wafers: Trajectory of leading-edge foundry capacity
- HBM: Capacity expansion and chipmaker ability to reduce HBM dependency
- Advanced packaging: Pace of advanced-packaging capacity expansion

Power and physical infrastructure absorption

- Power generation: Pace of new power availability
- Grid interconnection and scaling of behind-the-meter power: Trajectory of time-to-power timelines
- Sites and construction: Pace of permitting, buildout, and fiber expansion

Nonterrestrial DCs

- Economics: Trajectory of launch and operating costs
- Scalability: Pace of technical maturation and commercial deployment

Source: BCG analysis.
Note: DC = data center; MFU = model FLOPS utilization; FLOPS = floating-point operations per second; HBM = high-bandwidth memory.
[1]Assuming base case through 2031.