

Agentic AI – The Surge Begins

May 11, 2026 03:15 PM GMT

Agentic AI is no longer emerging but already reshaping industries – a new class of digital workforce is driving results. This means a rewrite of the AI infra equation – one with far more CPU and memory LTA capacity than earlier AI assumptions suggested, and hence much higher growth opportunity.

The 'agentic economy' is scaling faster than initially anticipated. This shift is already showing up in valuations – since mid-April (see our recent report [Rise of the AI Agent – Global Implications](#)), CPU and memory stocks have reached new highs. Recent 1Q26 commentary from AMD, Arm and Intel points to a recent surge in high-core-count CPUs and orchestration-heavy workloads, as hyperscaler data center deployments increasingly require denser CPU infrastructure around GPU clusters. At the same time, reported 3- to 5-year memory LTAs are emerging as a critical structural shift for the industry, extending pricing visibility and reinforcing durability in the cycle.

What's changed? 1Q26 has been an agent AI breakout quarter with a surge in 'agent-first' strategies across many companies, moving from experimental pilot programs to full-scale enterprise deployment. With accelerating enterprise adoption, CPU capacity requirements rise sharply and more measurable as agentic workloads scale. Now, in 2Q26, the debate is no longer about the potential of agentic AI, but about the speed of its integration. We also revise our memory framework to reflect higher orchestration CPU demand and rising DRAM content per CPU, driving incremental ~74EB in our base case, and ~228EB in a bull case by 2030.

Upgrading our CPU/memory TAM: We revise our framework to reflect faster agentic AI adoption, higher concurrency and more instructions per token. Our base-case total server CPU TAM rises 25% to US\$125bn by 2030, driven by a US\$79bn orchestration CPU layer on top of US\$45bn host/cloud CPU TAM; our top-down bull case reaches US\$283bn as dedicated agentic CPU racks scale. Applying the same CPU-led framework to memory, this implies ~74EB of incremental DRAM demand in our base case and ~221EB in our bull case (or 1.7x to 4.9x larger than the entire 2026 DRAM market).

Stay full-stack: The AI data center not only needs more GPUs but a stronger host CPU layer that can keep accelerators, memory, networking and software agents synchronized. Our preferred exposure is unchanged across CPUs, DRAM, NAND/ eSSD, HDD, ABF substrates, foundry, memory interface, BMC, sockets/connectors and semi-cap, where content growth and supply constraints capture the "second leg" of the AI infrastructure cycle ([Exhibit 1](#)).

MORGAN STANLEY & CO. INTERNATIONAL PLC

Shawn Kim

Equity Analyst

+84 90 7677-1018

Lee Simpson

Equity Analyst

+84 90 7677-8338

Nigel van Putten

Equity Analyst

+84 90 7675-2803

Amelia M Scicluna

Research Associate

+84 90 7675-6040

Cindy Huang

Equity Analyst

+84 90 7675-2918

MORGAN STANLEY ASIA LIMITED

Duan Liu

Equity Analyst

+852 3939-7357

Exhibit 1 : Likely beneficiaries of agentic AI take-up

Global Exposure Across the Stack 7 Names by exposure					
CPU	DRAM	NAND	HDD	FOUNDRY	IC-design
• NVDA • Intel • AMD • Arm	• Samsung • Hynix • Micron	• Kioxia • SanDisk	• Seagate • WDC • TDK	• TSMC	• GUC • Egis
PCB/Substrate/CCL & Materials			BMC, CPU & Memory interface		
• SEMCO • Unimicron • NYPCB • Ibiden • Nittobo • MEC			• Aspeed • Renesas • Montage • WPG • AP Memory		
MLCC & CPU socket		ODM	SPE		
• Murata • TDK • Yageo • FIT Hon Teng • Lotes		• Wuxin • Hon Hai	• ASML • ASMi • AMAT • Beje • KLAC • Tokyo Electron • Ulvac • Wonik		

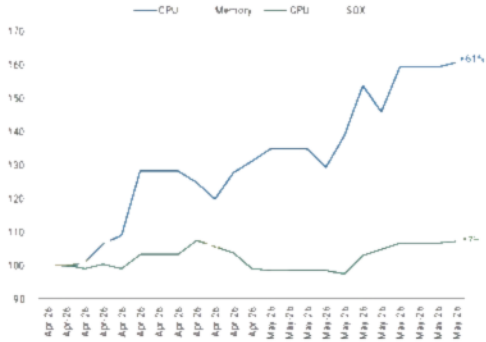
Source: Morgan Stanley Research

Morgan Stanley does and seeks to do business with companies covered in Morgan Stanley Research. As a result, investors should be aware that the firm may have a conflict of interest that could affect the objectivity of Morgan Stanley Research. Investors should consider Morgan Stanley Research as only a single factor in making their investment decision.

For analyst certification and other important disclosures, refer to the Disclosure Section, located at the end of this report.

+ = Analysts employed by non-U.S. affiliates are not registered with FINRA, may not be associated persons of the member and may not be subject to FINRA restrictions on communications with a subject company, public appearances and trading securities held by a research analyst account.

Exhibit 2: CPU and memory have significantly outperformed the broader AI ecosystem since our last report



Source: Morgan Stanley Research. Note: CPU is an average performance of ARM, INTEL, AMD. Memory: Samsung, SK Hynix, Micron. GPU: NVIDIA

Agentic AI – Key Developments

We expect a surge in agentic AI competition to manifest between hyperscalers, frontier model companies, incumbent software vendors and new startups as companies race to build out money-making AI tools. But we are still early in the price discovery stage of the Agentic AI economy and still don't know yet what kind of unlock it is going to ultimately bring to drive the AI TAM, early in the structural shift in AI infrastructure from pure compute toward orchestration, memory and system-level coordination.

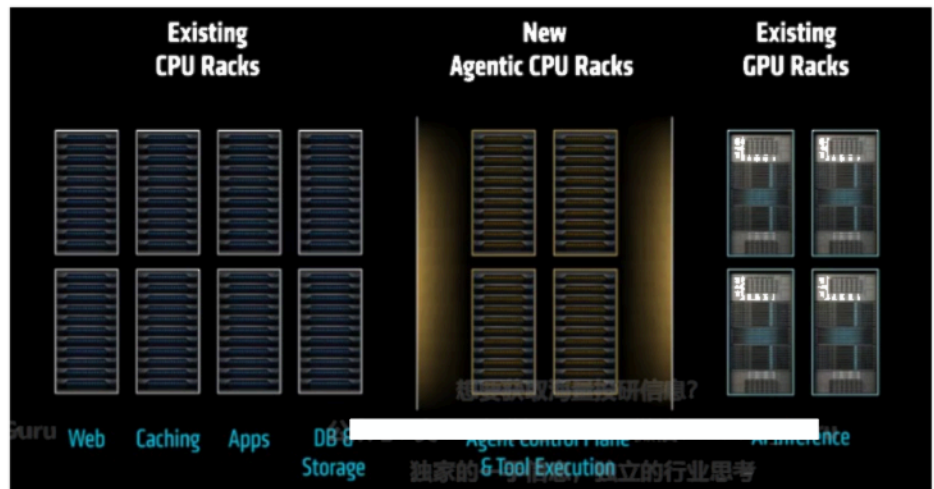
A pivot from search to action. We believe AI platforms are now shifting from cost centres to revenue infrastructure with the deployment of AI agents at scale in commerce, advertising and enterprise productivity. Agents drive more engagement, utility and customers stickiness on their platforms, due to their continuing learning ability and user context gained over time. Agents that conduct transactions becomes a significant value driver for hyperscale providers. Both Meta and Google have been reported to be working on AI agents, with Meta building a highly personalized AI assistant to carry out everyday tasks for its users (*Financial Times*, 6 May), and Google is developing a 24/7 personal agent for work, school and daily life, powered by Gemini (*Business Insider*, 6 May).

Agentic AI needs a rewrite of the AI infrastructure equation. Building an Agentic AI system does not mean putting a few more CPUs next to the same GPU-heavy rack design. It is a much more complex structural shift in data center architecture by adding a newly engineered CPU compute layer. Agentic AI requires entirely new racks of CPU servers that sit alongside GPU infrastructure and run to power the work of all these agents. The AI system in the future will look like a distributed system consisting of:

1. **GPU racks for dense model compute**, fast networking and a software stack that can keep it all observable, secure and efficient;
2. **Agentic CPU racks for orchestration**, processing data and tool execution.

In the agentic era, performance will not come from one processor doing everything but the right architecture – with CPUs and GPUs working together to move AI from answers to action. What will matter more for enterprises deploying AI agents is a balanced architecture, where: 1) the CPU tier is large enough to minimize GPUs wait, 2) efficient networking supporting agents, and 3) a data path allowing minimal latency. The orchestration layer is designed for concurrency to minimize cost and complexity.

Exhibit 3: Agentic AI infrastructure architecture – the CPU/GPU shift is far complex than adding more CPUs



Source: AMD, Morgan Stanley Research

A bigger TAM

We now expect the total addressable market for server CPUs and memory to grow meaningfully faster than our prior framework implied.

- We lift our base case total server CPU TAM to US\$125bn by 2030, from c.US \$100bn+ previously, as we account for agentic sessions, concurrent usage and instructions per token.
- In our bull case, using a top-down AI data center capacity framework, we estimate total server CPU TAM could reach US\$283bn by 2030.
- For memory, the same CPU-led framework implies a material uplift in DRAM demand, with agentic orchestration workloads driving ~74EB of incremental DRAM demand in our base case (or 1.7x 2026 total 45EB shipment) and ~221EB (4.9x this year's DRAM market) in our bull case by 2030.

Recent developments

AMD raised the server CPU TAM bar materially

AMD's 1Q26 call is the strongest CPU validation. Management now expects the server CPU TAM to grow >35% annually to >\$120bn by 2030, up from its prior ~18% CAGR framing, explicitly citing agentic AI as a driver of CPU compute for orchestration data movement and parallel execution. AMD also reported Data Center revenue of \$5.8bn

+57% YoY, with server CPU revenue up >50% YoY and Q2 server CPU revenue expected to grow >70% YoY.

Arm's AGI CPU opportunity is scaling faster than expected

Arm reported record Q4 FYE26 revenue of \$1.49bn and full-year revenue of \$4.92bn, while data center royalties more than doubled YoY. More importantly, customer demand for the new Arm AGI CPU is now >\$2bn across FYE27-28, more than double what was indicated at Arm Everywhere. Arm also says its CPU compute share is around 50% among

top hyperscalers, with the AGI CPU developed with Meta and positioned for agentic AI infrastructure.

Intel validated the CPU / foundry / packaging angle

Intel's 1Q26 release explicitly said the shift from foundational models to inference to agentic AI is increasing demand for Intel's CPUs, wafer capacity and advanced packaging. Intel also reported DCAI revenue of \$5.1bn, +22% YoY, highlighted Xeon 6 as the host CPU for NVIDIA DGX Rubin NVL8, and noted Google's continued Xeon deployment plus a custom ASIC infrastructure processing unit collaboration.

Meta-AWS Graviton deal is the clearest real-world agentic CPU deployment

Meta signed an agreement with AWS to deploy tens of millions of Graviton cores for agentic AI workloads. AWS framed this as CPU-intensive demand from real-time reasoning, code generation, search and multi-step task orchestration, while Meta said agentic AI is evolving to require more CPU and that no single chip architecture can serve every workload.

Microsoft explicitly framed the cycle as the "agentic computing era"

Microsoft's FY3Q26 results used the phrase "agentic computing era", with revenue of \$82.9bn, +18% YoY. Management said it is focused on delivering cloud and AI infrastructure and solutions to help customers "eval-max" outcomes in the agentic computing era.

Alphabet / Google Cloud reinforced full-stack AI infrastructure demand

Google Cloud revenue grew 63% YoY to US\$20.0bn in 1Q26, supported by enterprise AI solutions and AI infrastructure, while the company positioned its new Gemini Enterprise Agent Platform as the connective layer for building, deploying and managing agents at scale. Google also launched an Agentic Data Cloud to help agents organize, understand and act on enterprise data in real time, committed US\$750mn to accelerate partner-led agentic AI development across its 120,000-member ecosystem, and introduced new TPU hardware (TPU 8t / 8i) designed for agentic workloads.

Memory LTAs are the most important structural change

Samsung and SK hynix have indicated they are moving toward 3-5 year LTAs with Big Tech customers. Memory and advanced packaging specifications for custom AI chips are increasingly locked in at the design stage, and that LTAs may include mechanisms such as price floors, supply volume commitments and binding agreements with very large up-front payments.

Base Case CPU TAM to US\$125bn

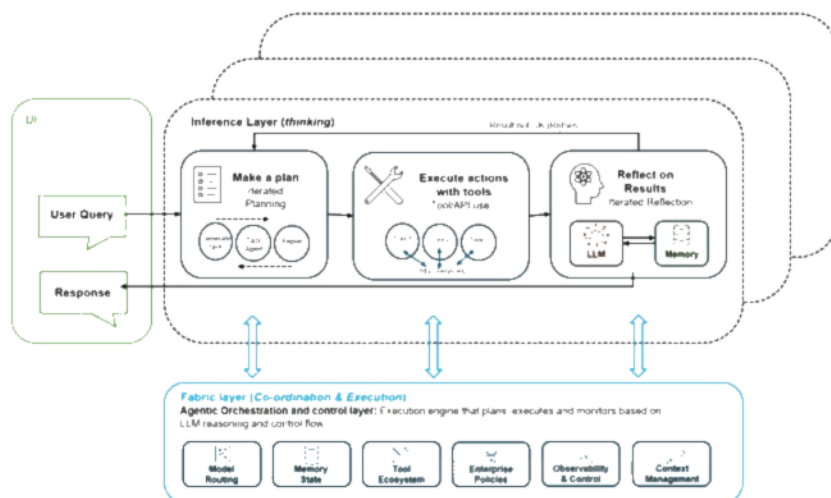
We revisit our CPU TAM model. As per our initial note, [Rise of the AI Agent - Global Implications](#), we maintain our bear case TAM of US\$77bn. However we now lift our base case to US\$125bn by 2030 as we account for agentic sessions, concurrent use and instructions per token. We also see a higher bull case which using a Top-Down model we estimate could reach US\$283bn.

A reminder of our thesis

We believe agentic AI will increase the CPU-to-GPU mix in AI systems by adding more orchestration, memory, and tool-use work around each model call. This should not reduce GPU demand, but it does increase overall system complexity and shifts incremental infrastructure spend toward CPUs, networking, and memory. In this environment, the advantage is less about owning the accelerator and more about owning the system architecture.

The initial GenAI wave was dominated by GPU-centric model-serving, with relatively light control-plane overhead. Agentic inference introduces multi-step workflows (plan, retrieve, call tools, execute, iterate), greater reliance on persistent memory and external tools/APIs, and higher orchestration complexity.

Exhibit 4: The inference fabric framework



Source: Morgan Stanley Research

From US\$100bn CPU TAM to US\$125bn

In our initial note, [Rise of the AI Agent - Global Implications](#), we parsed the Host CPUs from Orchestration CPUs. We used Mercury data to estimate the Host CPU market (Grace, Vera, Graviton etc) at US\$45bn by 2030. We continue to maintain this assumption in our new model. However, our previous estimates for the orchestration CPU TAM were conservative in particular our assumption for an AI semi TAM of US\$1.2Trn, note Nvidia are talking to this being as large as US\$3/4Trn. Our previous orchestration TAM of US\$30-

60bn implied a CPU TAM of US\$105bn, in-line with Arm's commentary at Arm Everywhere ([link](#)). However, AMD are now talking to a level closer to US\$120bn.

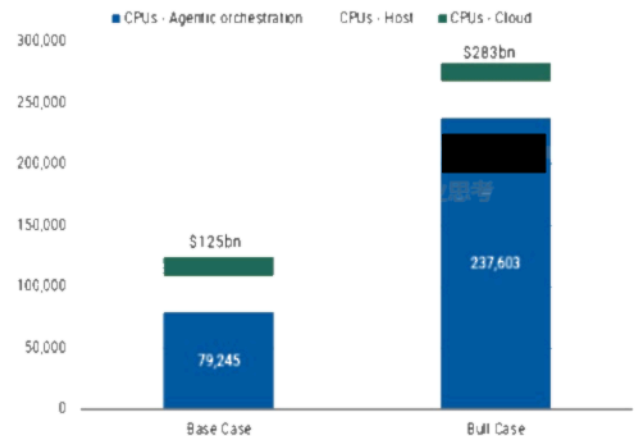
In deriving the Orchestration CPU TAM we take two different approaches: (i) a **bottom-up** model identifying agents per session, concurrent usage of compute and instructions per token; and (ii) a **top-down** model identifying GW installed per annum, implied racks and the ratio of CPU:GPU.

Exhibit 5: We lift our base case Orchestration CPU TAM to \$79bn (prior \$60bn).

TAM 2030 (\$bn)	Orch.	Host + cloud	Total
Bear Case	32	45	77
Base (current) model	60	45	105
Base (new) Bottom-Up model	79	45	125
Bull Top-Down model	238	45	283

Source: Morgan Stanley Research estimates

Exhibit 6: Our bull case implies a \$238bn CPU orchestration TAM.

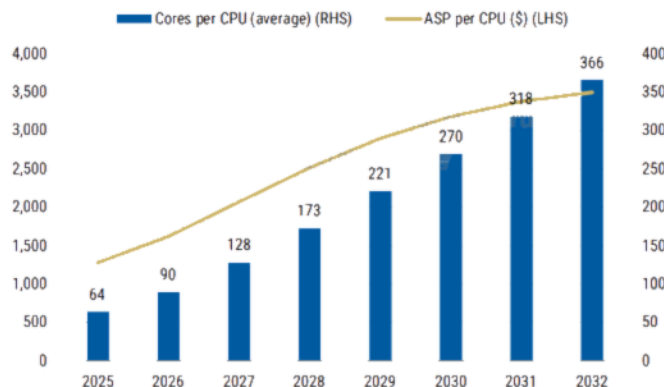


Source: Morgan Stanley Research estimates (e)

Assumptions underlying our model

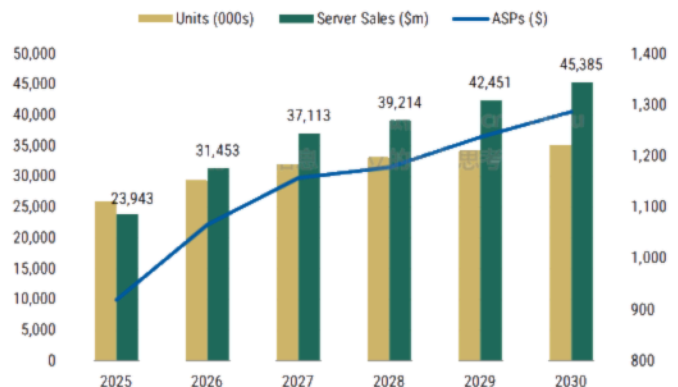
We assume in both our bottom-up and top-down model that a new generation of CPUs per annum implies that pricing per core declines 10% yy. However, in-line with current core trends we expect core counts per CPU to grow from 64 today to c.130 by FY27 and to 200-500+ by 2030. All in, this implies the FY27 ASP per CPU to come in at US \$2,000+ growing to at least US\$3,000 by 2030. For the head node market we model a combination of head node CPUs and regular CPU cloud servers. We expect CPU server sales to grow from US\$31bn FY26 to US\$45bn FY30.

Exhibit 7: We model the ASP per CPU to grow in-line with cores per CPU.



Source: Morgan Stanley Research estimates

Exhibit 8: We continue to expect CPU server sales to grow to \$45bn by 2030.

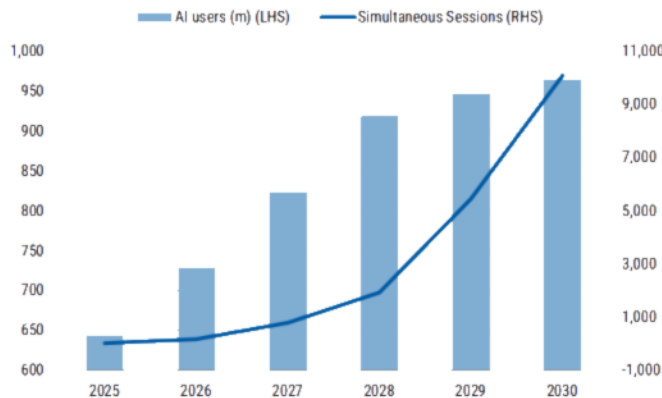


Source: Morgan Stanley Research estimates

A US\$79bn agentic CPU TAM base case...

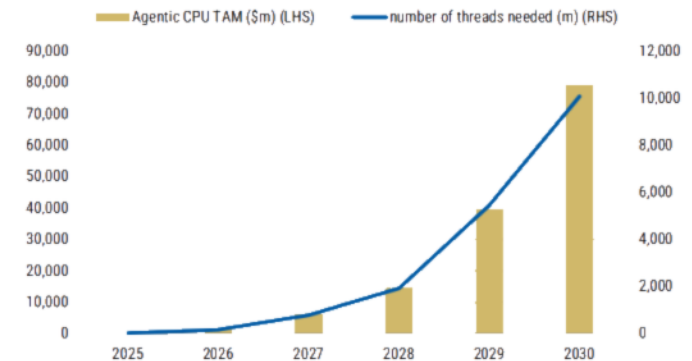
In deriving our base case we lean on the heavy use of agentic workflows by knowledge workers while assuming that single thread CPUs dominate. In this bottom-up model we estimate c.1bn knowledge workers globally by 2032 with the geographic mix as 50% Asia, c.25% Europe and 10 to 15% US. We estimate an AI take-up amongst knowledge workers of 78% for FY26 growing to 99% by 2030, while we model the occurrence of concurrent sessions to grow from 4% in FY26 to 19% in 2030. We assume from therein concurrency flattens at 15-20%. While 2 threads per core is the standard today, we see that reducing on the growing incidence of single threaded CPU cores amidst the need for power efficiency. We model the hyperthreading multiplier to decrease from 2 in 2026 to c.1.5 in 2030 implying the true availability of cores to grow at a 33% CAGR FY25 to FY29. On agentic use today we think today 6 agents per sessions is an accurate representation. However we expect this could almost reach 100 agents per session by 2032. We note this may still be conservative as exponential growth would imply trillions of agents per session. Assuming 1 session per thread, this implies 25mn CPUs by 2030. Applying our pricing assumptions this sets our base case orchestration TAM at US\$79bn FY30. When including our estimate for the Host and Cloud CPU we estimate a US\$125bn CPU TAM.

Exhibit 9: We model simultaneous sessions to grow at a 176% CAGR from FY26 to FY30.



Source: Morgan Stanley Research estimates

Exhibit 10: Our base case implies a \$79bn Agentic CPU TAM by 2030.

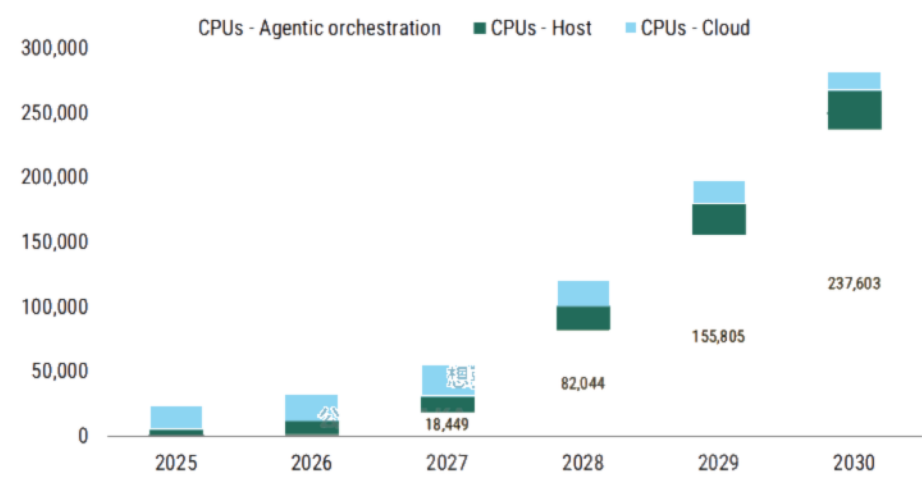


Source: Morgan Stanley Research estimates

... with our bull case implying a US\$238bn agentic CPU TAM.

To derive our bull case we look top-down at what AI data centre installed GW capacity may imply for the CPU:GPU ratio. Our Morgan Stanley global estimates imply AI data centre installed capacity grows from 24GW today before flattening to 35GW in the outer years. Using Nvidia racks as a market proxy we take the power budget per rack to calculate the GPU and CPU power draw per rack and what that would imply for the number of CPUs and GPUs per rack. Assuming pricing for Host CPUs are largely the same as orchestration CPUs this implies a \$25bn Host CPU TAM. We assume the CPU:GPU ratio ramps from 1:2 to 2:1 by 2030 implying orchestration CPUs grow at a 196% CAGR FY26 to FY30. This sets our orchestration CPU TAM at US\$238bn FY30. Accounting for host and cloud CPUs, our Bull Case CPU TAM is US\$283bn.

Exhibit 11: In our top-down model, we estimate the Agentic CPU TAM to grow at a 251% CAGR from FY26 to FY30.



Source: Morgan Stanley Research estimates

Raising Memory TAM

The revised CPU model materially raises the memory opportunity. Our prior framework estimated 15-45EB of incremental DRAM demand by 2030. We now estimate ~74EB in our base case and ~221EB in our bull case.

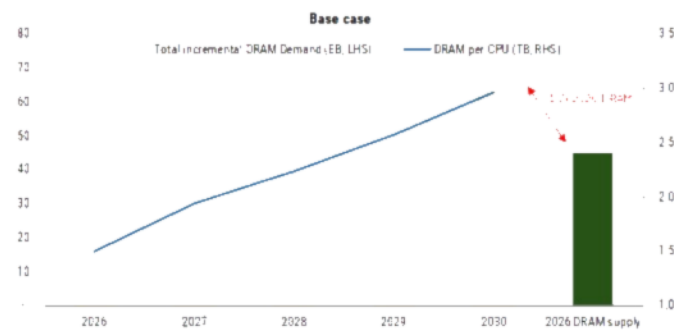
Agentic AI changes memory TAM in two ways: It raises the effective context per request and increases the amount of CPU-side memory provisioned per AI rack. As a result, the marginal memory demand shifts away from hot HBM alone and increasingly into host DRAM and rack SSD, where retained context, intermediate state and warm KV live closest to the CPU.

We keep the model intentionally simple:

$$\text{Incremental DRAM demand} = \text{orchestration CPU units} \times \text{DRAM per CPU}$$

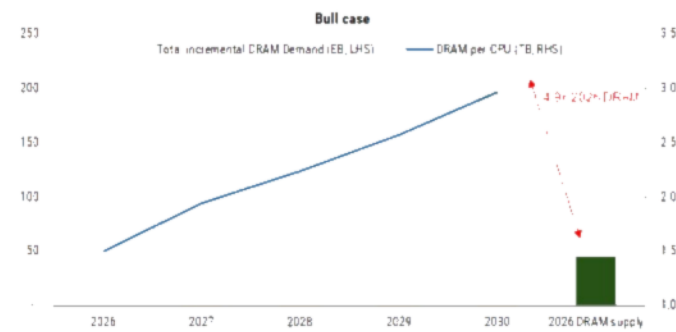
Our previous framework assumed 10-15mn incremental orchestration CPUs and 1.5-3.0TB DRAM per CPU, implying 15-45EB of incremental DRAM demand by 2030. The updated CPU TAM model points to a much larger orchestration CPU base. In our new base case, the US\$79bn orchestration CPU TAM implies roughly 25mn orchestration CPUs by 2030. Assuming DRAM per orchestration CPU rises toward 3TB as agent concurrency, memory services, RAG, KV-cache offload and persistent context scale, this implies roughly 74EB of incremental DRAM demand. In the bull case, the US\$244bn orchestration CPU TAM implies a much larger CPU pool and roughly 221EB of incremental DRAM demand by 2030.

Exhibit 12: Base case: agentic orchestration drives ~74EB incremental DRAM by 2030



Source: Morgan Stanley Research

Exhibit 13: Bull case: DRAM demand reaches ~221EB as orchestration CPU racks scale



Source: Morgan Stanley Research



Agentic AI – Global Beneficiaries

Exhibit 14: Agentic AI Beneficiaries

[illegible]

Source: Factset, Morgan Stanley Research