Peak
Insights

HBM Opportunity Within Memory

After two years of cyclical declines, the global compute and memory market is at an inflection point, creating a unique opportunity for our thematic investing framework. We believe we are at one of these points of structural change within the DRAM (dynamic random-access memory) market driven by the emergence of High Bandwidth Memory (HBM) and its applicability to AI-related compute. AI and HBM in particular are going to alter the memory landscape for many years, potentially driving higher growth and profitability for the companies involved , creating a secular structural growth opportunity within the cyclical memory market.

Christopher Smith
Portfolio Manager

20 Years Investment Experience

Memory Background

DRAM is the primary memory in modern computers and graphic cards. DRAM can be thought of as the short-term memory that allows a processor to quickly access and process information. As compute speed grows thanks to faster logic chips, input/output speeds from memory have become a critical constraint for overall performance. This dynamic has led to the creation of new DRAM microarchitectures. HBM, a 3D version of stacked DRAM, is very fast and energy efficient and will increasingly be paired alongside AI graphics processing unit (GPU) chips.

DRAM has historically been a hyper-cyclical market marked by booms and busts where the latter periods often resulted in negative gross margins and large operating losses for industry participants. Over the last several decades, a large number of US and Japanese players entered this fiercely competitive market, but failed to survive as returns were driven down by new market entrants. Today, the industry is consolidated to three large players: Korea-based Samsung Electronics and SK Hynix and US-based Micron Technology. Greater consolidation has led to better supply discipline during economic downturns and made the industry more attractive from a long-term investing standpoint.

DRAM, as an industry, has seen annual binary digit (bit) growth (a measure of the quantity of information produced) in the mid-teens with total through-cycle revenue growth in the single digits. Price, while cyclical, has tended to decline over time. Each successive generation of product and process node advancement leads to faster data transfer rates, higher memory densities and improved energy efficiency. Technological advancements in DRAM production have been highly deflationary as each successive generation lowered the cost per bit.

This marks a shift in this deflationary paradigm as the cost to produce HBM is substantially higher due to its larger die size, 3D stacking and additional expensive components.

ARTISAN PARTNERS

Peak Insights

AI and HBM's Role

While the growth of AI applications and AI infrastructure is unlikely to be linear, it is still in very early stages with a long runway for growth. Lisa Su, the CEO of Advanced Micro Devices, made headlines in December 2023 when she forecasted an addressable market for datacenter AI chips of $400B in 2027 from $45B in 2023, suggesting a 70% compound annual growth rate (CAGR). While this number is eye-popping, the mathematical rationale and basis for this estimate is not unreasonable. Estimates can justify 10 million AI servers in place by 2027 using what we believe are reasonable assumptions. Exhibit 1 does a great job of walking through, piece by piece, how one might arrive at this figure.

Exhibit 1: Usage and Complexity: Reasonable Assumptions Can Justify 10 Million AI Servers By 2027

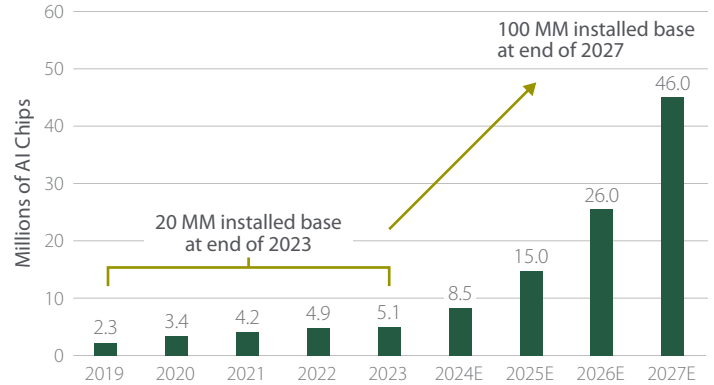|	2023 - 2027 INCREASE	RATIONALE
Model complexity	~100X	10X every 2 years. Deceleration vs today (10X every year).
Chip efficiency	~120X	6X every 18MM (every new generation). In line with today.
Chip time per unit of usage	**0.8X**	**Model sized divided by chip efficiency**
Users	~5X	200MM to 1B. Internet users grew from 200MM to 800MM from 1999 to 2003.
Usage per user	~5X	From 1min of full inference server usage per user per day in 2023 to 5min in 2027, as use cases are still limited. Imagine, for instance, GenAI getting into gaming or office productivity.
Usage	**25X**	**Users x usage per user**

INCREASE IN THE SERVER INSTALLED BASE		# OF SERVERS		
		2023		2027
Training $\sqrt{\begin{array}{l}\text{Usage}\\\text{Increase}\end{array}}$ x $\begin{array}{l}\text{Chip time / unit}\\\text{of usage}\end{array}$		1.7MM	4X	6.8MM

$\sqrt{25 \times 0.8} = 4$

Assumes # of users of the same training is growing as much as the volume of training

Inference Usage increase x Chip time / unit of usage		0.18MM	20X	3.6MM

$25 \times 0.8 = 20$

	Total AI servers	**1.9MM**	**5X**	**10MM**

A rapid but reasonable adoption and development of AI services can easily require 10MM AI servers in 2027

Source: New Street Research.

Assuming 8-10 chips per server, 10MM AI servers implies approximately 80MM to 100MM AI chips. In order to get to an installed base of 10MM AI servers in 2027, AI chip deployments would have to grow rapidly.
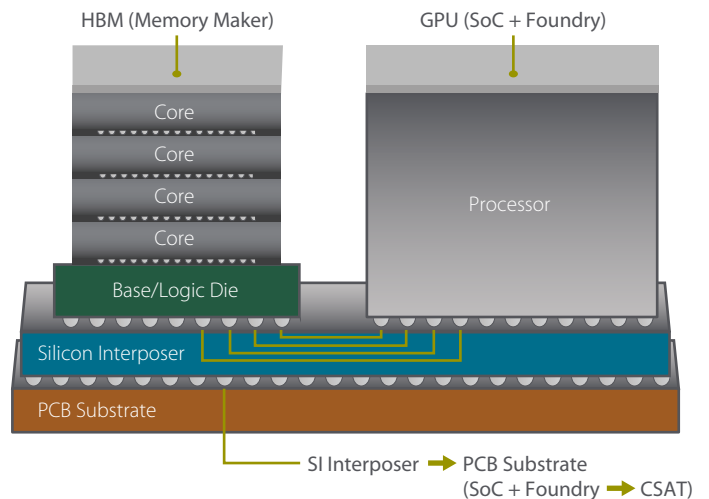
Exhibit 2: Global Datacenter AI Chip Installed Base

100 MM installed base at end of 2027

20 MM installed base at end of 2023

2019	2020	2021	2022	2023	2024E	2025E	2026E	2027E
2.3	3.4	4.2	4.9	5.1	8.5	15.0	26.0	46.0

Millions of AI Chips

Source: New Street Research, published 6 Jan 2024.

Given the inordinately large amount of data that needs to be processed by large clusters of AI accelerators in both training and inferencing, memory bandwidth has become a key constraint and enabler of performance. HBM's 3D vertical stacking of DRAM dies enable much wider memory interfaces, allowing for the "high bandwidth" in "high bandwidth memory." In addition, HBM requires less power than alternative technologies. All of this makes HBM particularly well suited and necessary for the training and inferencing associated with AI Large Language Models.
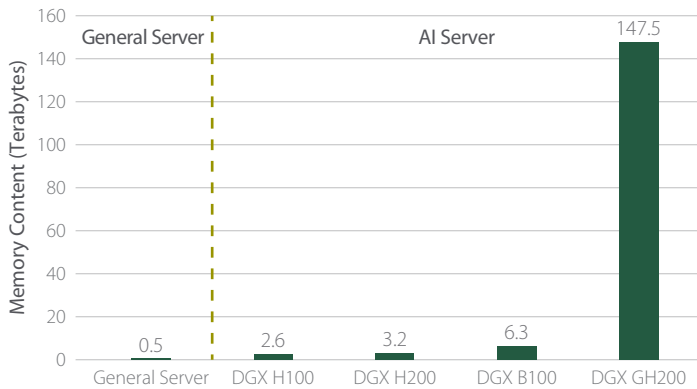
Exhibit 3: How HBM Attaches To A GPU

HBM (Memory Maker)

GPU (SoC + Foundry)

Core
Core
Core
Core

Base/Logic Die

Processor

Silicon Interposer

PCB Substrate

SI Interposer → PCB Substrate (SoC + Foundry → CSAT)

Source: Citi Research.

Peak Insights

Exhibit 4: DRAM Content Per System: General Server Versus AI Servers

General Server | AI Server

	Memory Content (Terabytes)
General Server	0.5
DGX H100	2.6
DGX H200	3.2
DGX B100	6.3
DGX GH200	147.5

Beyond the rapid growth of AI accelerators themselves, memory content of AI accelerators is 5X-294X larger than that of a traditional server (Exhibit 4). This is key as memory consumption per accelerator is set to grow enormously to yield better accelerator performance. As an example, Nvidia's X100 GPU is expected to be released sometime in 2025 and will have 5X the HBM content as its current H100 GPU. The H100 is currently the most coveted flagship GPU and has been hailed as the world's most advanced chip, but it is priced at a massive $30,000 USD.

Exhibit 5: HBM Density Per Nvidia AI GPU

	HBM Density per GPU (Gigabytes)
L40	48
A100	80
H100	80
H200	141
B100	192
X100	400
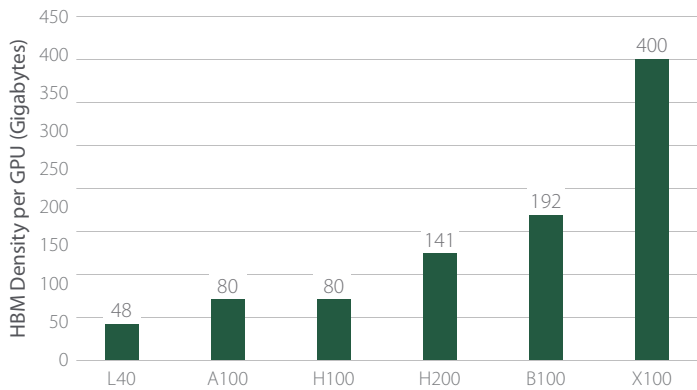
All of these factors could drive a 7X increase in HBM between 2023 and 2025 and an increase in HBM as a percent of the total DRAM market from the single digits to ~25% (Exhibit 6), driving an acceleration in the overall DRAM bit growth. Exhibit 7 shows the projected growth inclusive of AI demand versus previous numbers.

Exhibit 6: HBM Share Of DRAM Market
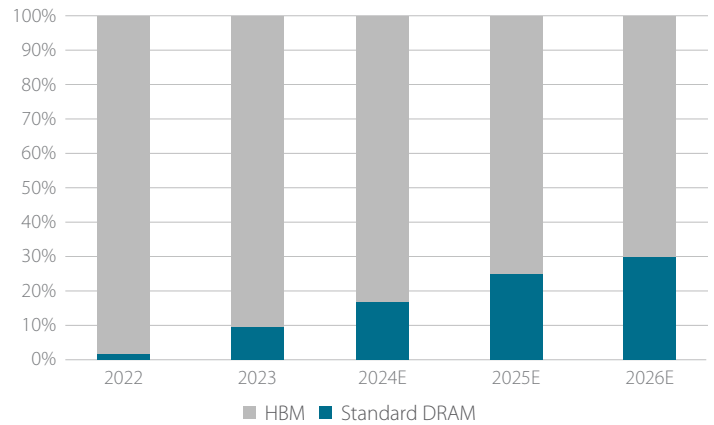
■ HBM ■ Standard DRAM

Exhibit 7: Global DRAM Demand Growth—Revised For New AI Demand

21% | 27% | 22% | 16% | 22% | 1% | 3% | 21% | 17% | 23% | 16%

—— Revised With AI Demand   —— Previous Estimates Without AI
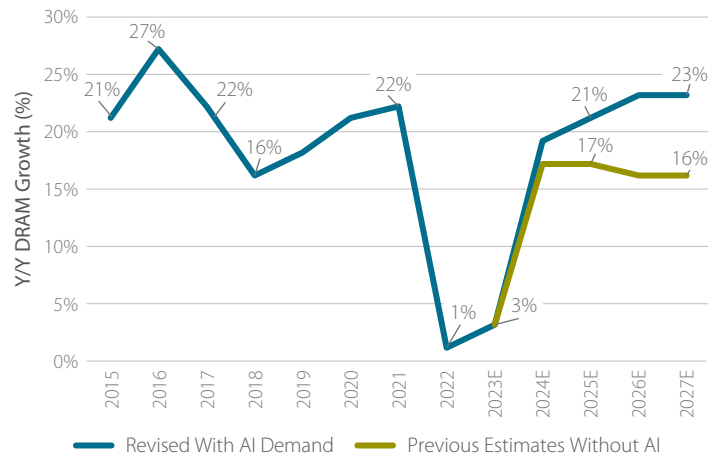
Memory has historically been a commodity that was *standardized* to fit into Intel CPUs, but in the future, it will be increasingly optimized and *customized* to meet the requirements of different AI chips. The increased product diversification and complexity of memory products driven by AI will lead to memory looking more like the logic foundry market—an attractive one to us at the Antero Peak Group. Customization, in our view, is synonymous with more differentiation, higher price and higher margin. This should benefit the HBM companies who have historically been faced with commodity pricing curves for their legacy memory products.

Peak Insights

Where Are We In The Cycle?

We are at the beginning of an upturn in the memory cycle after a prolonged and severe downturn which began in 2022. Demand for smartphones, PCs and general-purpose servers—the key end-markets for DRAM—has shifted to positive year-over-year growth in Q4 2023 after two years of declines. The COVID-driven demand surge in 2020 and 2021 led to overconsumption of goods, overproduction of components (including DRAM) and the build-up of excess inventory. The DRAM industry reduced production in 2023 by about 30% from highs to combat weak demand and excess inventory. The industry is still working down that excess inventory, but we believe there is now line of sight to normalization within the next couple of quarters. The supply cuts that we have seen combined with the bottoming out of end-markets and ongoing inventory normalization has led to pricing beginning to rise sequentially. In the last few cycles, trough to peak pricing increased between 77%-225% (Exhibit 8). In the current cycle, pricing has only increased 16% so far, suggesting we have quite a bit more room to go.

Exhibit 8: Trough to Peak DRAM Spot Price Change By Cycle

	2013 - 2014	2016 - 2018	2020 - 2021	2023 - CURRENT
Trough	$2.35	$1.54	$2.15	$1.19
Peak	$4.19	$5.00	$3.81	$1.38
Trough to Peak Increase	78%	225%	77%	16%

Source: Antero Peak Group. As of 31 Jan 2024.

Furthermore, unlike prior cycles, HBM presents an additional catalyst to the DRAM market today. While we have already discussed how HBM has attractive growth characteristics fueled by AI, HBM also acts as a functional supply cut for the DRAM market. Because the die size is 2X as big as double data rate (DDR) and yields are lower, for every bit of capacity allocated to HBM, the overall DRAM market loses 2 bits. Thus, if 7% of bit capacity were to move to HBM by 2025 (a reasonable estimate), overall DRAM capacity inclusive of HBM would decline by 14%, accelerating price growth. We believe existing supply discipline, the effective capacity reduction from the shift to HBM and a demand recovery is likely to drive a meaningful amount of additional pricing growth from here.

Summary

DRAM sits at an attractive inflection point—the beginning of an AI-driven memory upcycle with cyclical and secular tailwinds. We expect memory products to become semi-customized in the future due to the emergence of AI. This should result in increased product diversification and complexity. As products become more customized to meet specific needs, the ability of semiconductor companies to differentiate their products and charge a higher price will become apparent. Customers will be less able to switch between competing memory providers, leading to higher margins, stickier relationships and higher returns on capital over time. We see acceleration in the growth of leading-edge DRAM and are focusing our research on the companies in the space with exposure to HBM memory, including Samsung, SK Hynix, Micron and ASML—the only company in the world that makes the equipment necessary for the chip companies to make leading edge memory. In short, our strong belief in the coming importance of AI implies strong growth in HBM and an upcoming paradigm shift in the DRAM market.

For more information: Visit www.artisanpartners.com

A R T I S A N **A** P A R T N E R S