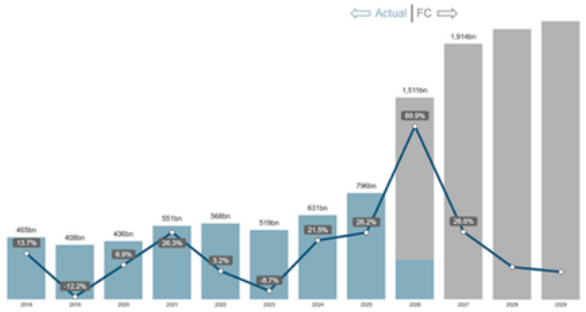
Memorie e AI: la fine della ciclicità?

Quando si parla di Intelligenza Artificiale, l'immaginario collettivo corre quasi sempre ai grandi progettisti di chip logici e acceleratori come NVIDIA o AMD. È un'associazione corretta, ma racconta solo metà della storia. Per funzionare, un acceleratore AI non ha bisogno soltanto di "cervello", ovvero di pura capacità di calcolo (compute), ma necessita di un sistema altrettanto sofisticato che gli fornisca dati alla velocità richiesta dagli algoritmi. Quel sistema è la memoria a semiconduttore, diventata uno dei segmenti più contesi, critici e redditizi dell'intera industria dei semiconduttori.

Fino a pochi anni fa, la memoria era considerata una commodity: un prodotto standardizzato, a basso margine, soggetto a cicli economici prevedibili legati alla domanda di PC e smartphone. I produttori facevano notizia principalmente per le oscillazioni dei prezzi o per i tagli alla produzione durante le fasi di surplus del mercato. L'avvento dei Large Language Models (LLM) e dell'AI generativa ha ribaltato questo scenario in meno di tre anni, trasformando la memoria nel vero ago della bilancia delle prestazioni computazionali.



Global semiconductor market. Actual and forecast in billion US dollars, YoY growth in percent. Credit: WSTS

Fondamenti delle memorie

La memoria a semiconduttore rappresenta l'hardware elettronico digitale dedicato alla registrazione e alla conservazione dei dati all'interno dei moderni sistemi informatici. Costruiti su circuiti integrati in silicio tramite la tecnologia dei transistor, questi dispositivi sostituiscono i vecchi sistemi storici come i nuclei magnetici in ferrite, offrendo dimensioni infinitamente più compatte e risposte quasi istantanee calcolate in nanosecondi. Dal punto di vista strutturale, l'informazione digitale viene memorizzata sotto forma di carica elettrica o di stato logico all'interno di una fitta griglia di celle elementari. Ogni cella gestisce un singolo bit di informazione, che corrisponde al linguaggio binario dei computer composto da zeri e uni, garantendo una densità di archiviazione che permette di racchiudere miliardi di dati in pochissimi millimetri di spazio.

Il funzionamento di questo sistema si basa sul lavoro coordinato di circuiti che individuano la posizione esatta del dato all'interno della griglia di silicio, proprio come se dovessero trovare una cella in un foglio di calcolo attraverso le coordinate di riga e colonna. Questi componenti, storicamente legati ai segnali di selezione delle righe e delle colonne, si collegano a speciali registri temporanei che funzionano da zone di sosta per i dati in transito. Da qui, le informazioni vengono inviate al processore o modificate tramite segnali elettrici dedicati alla lettura o alla scrittura. Le prestazioni generali di questo meccanismo vengono misurate attraverso indicatori come il tempo di accesso, cioè la velocità di risposta alla richiesta del processore, e il consumo di energia. Nei grandi centri di

calcolo, l'affidabilità viene inoltre protetta da tecnologie hardware di correzione d'errore che rilevano e riparano automaticamente i piccoli sbalzi elettrici nei dati.

L'universo di queste memorie si divide in due macro-famiglie basate sulla persistenza dei dati quando l'apparecchio viene spento. Le memorie volatili, note come RAM, richiedono un flusso costante di energia elettrica e perdono ogni informazione allo spegnimento. Questa famiglia si divide in memorie statiche, estremamente veloci e costose usate per le cache dei processori, memorie dinamiche sincrone, usate per la memoria centrale dei computer che richiedono continui cicli di rinfresco elettrico, e memorie ad ampia larghezza di banda, che collegano più chip in verticale per i calcoli legati alla intelligenza artificiale. Le memorie non volatili mantengono invece i dati stabili nel tempo anche senza alimentazione. Tra queste troviamo le classiche memorie di sola lettura utilizzate per l'avvio dei sistemi, le memorie flash di tipo NOR ad accesso rapido per i firmware, e le diffuse memorie flash di tipo NAND, che organizzano i dati in blocchi e rappresentano il cuore tecnologico delle chiavette USB e dei moderni dischi a stato solido per l'archiviazione di massa.

Memorie nella filiera dell'AI

L'infrastruttura che rende possibile l'intelligenza artificiale può essere immaginata come una filiera a cinque livelli posizionati uno sopra l'altro: alla base c'è l'energia che alimenta fisicamente i data center, sopra si trovano i semiconduttori e il calcolo, poi l'infrastruttura, più in alto i modelli, cioè gli algoritmi veri e propri e infine le applicazioni che tutti utilizziamo ogni giorno. Il livello dei semiconduttori, quello su cui si concentra questo report, non è un blocco unico, ma si divide in due ecosistemi complementari ed economicamente molto diversi. Da un lato c'è la logica di calcolo, che comprende le GPU, gli ASIC personalizzati e le CPU, mentre dall'altro c'è la memoria, ovvero i chip che immagazzinano e trasportano i dati. Il primo ecosistema è quello che fa notizia, mentre il secondo è quello che ha registrato la dinamica di prezzo e di scarsità più marcata degli ultimi due anni. All'interno di questo livello, l'interconnessione tra l'unità di calcolo logico e il sistema di memoria determina l'efficienza reale di un server per intelligenza artificiale.

Nella fase di addestramento, in cui il modello analizza immensi dataset per apprendere i legami statistici tra le informazioni, la memoria lavora per alimentare continuamente i core di calcolo con nuovi blocchi di dati, evitando che i costosi transistor di calcolo rimangano inattivi. La fase di inferenza, ovvero il momento in cui un utente pone una domanda a un modello già addestrato come ChatGPT, mette invece a dura prova la velocità pura della memoria di prossimità. Quando l'utente preme invio, la richiesta viene convertita in vettori matematici e inviata all'acceleratore. Per generare la prima parola di risposta, la GPU deve leggere l'intera struttura dei parametri del modello, che rappresentano l'enciclopedia delle conoscenze apprese e che risiedono fisicamente nei chip di memoria adiacenti al processore. Questo processo di lettura si ripete in modo ciclico e sequenziale per ogni singolo elemento o parola successiva che il modello decide di generare.

Durante questo ciclo continuo, la memoria deve gestire contemporaneamente tre flussi informatici distinti:

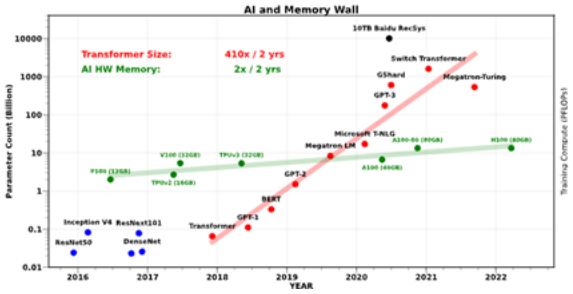
- Pesi del modello: una sorta di enciclopedia numerica che racchiude tutto ciò che il sistema ha imparato. Nei modelli più avanzati occupa da centinaia di gigabyte a oltre un terabyte, e deve essere riletta per intero dal chip di memoria a ogni frazione di secondo per generare ogni singola parola.

- Memoria di contesto: una sosta temporanea che conserva i dati della conversazione in corso, agendo come una memoria a breve termine affinché il sistema si ricordi dei messaggi precedenti inviati dall'utente.
- Attivazioni temporanee: dati di passaggio e risultati matematici intermedi generati dai calcoli dei vari strati dell'algoritmo, che vengono creati, modificati e distrutti continuamente.

Poiché la logica può calcolare i dati a una velocità infinitamente superiore rispetto al tempo che la memoria impiega a trasferirli attraverso i canali elettrici, l'intera operazione di inferenza non è limitata dalla potenza del processore, ma dalla velocità di trasmissione dei dati. Senza un'architettura di memoria capace di sostenere questo flusso colossale, l'architettura logica subisce un rallentamento forzato, rendendo l'erogazione dei servizi di intelligenza artificiale lenta e insostenibile dal punto di vista energetico.

Memory Wall

Il concetto di blocco prestazionale introdotto in precedenza si traduce nel fenomeno tecnico noto come Memory Wall, ovvero il divario prestazionale e il limite fisico per cui la velocità di calcolo dei processori cresce a un ritmo nettamente superiore rispetto all'ampiezza di banda con cui la memoria riesce a consegnare i dati. Il problema non risiede nella capacità aritmetica degli acceleratori, in grado di eseguire trilioni di operazioni al secondo, ma nell'infrastruttura di trasporto delle informazioni. Questa asimmetria deriva da un cambiamento storico: se nel ventesimo secolo la densità e la velocità di logica e memoria crescevano in parallelo, negli ultimi venti anni la miniaturizzazione dei condensatori della memoria ha incontrato severi limiti fisici, con incrementi di densità inferiori a due volte per decennio a fronte dei costanti progressi dei processori. Oltre a rallentare il calcolo, il trasporto fisico degli elettroni tra i chip di memoria e il processore attraverso le interfacce tradizionali è diventato la voce di costo più dispendiosa dal punto di vista energetico, richiedendo molta più energia per lo spostamento di un singolo bit rispetto a quella necessaria per eseguire l'operazione matematica stessa.
Fino all'avvento dell'intelligenza artificiale generativa, l'hardware standard doveva bilanciare carichi di lavoro generici per computer o elettrodomestici, alternando costantemente operazioni di lettura e scrittura. L'introduzione dei modelli basati su architettura Transformer ha reso il Memory Wall un fattore critico proprio a causa di uno squilibrio strutturale nella fase di inferenza, dove il rapporto tra letture e scritture supera la proporzione di mille a uno. Nella pratica, l'interrogazione di un modello linguistico è un processo quasi interamente di sola consultazione: per generare anche una singola parola o token, il processore non deve compiere calcoli complessi, ma ha l'obbligo di rileggere da capo centinaia di gigabyte di dati che compongono i pesi del modello e la memoria di contesto accumulata. L'unica operazione di scrittura si riduce al singolo token appena calcolato. Costringere le memorie tradizionali a eseguire oltre mille letture massive per ogni singola scrittura satura costantemente i canali di comunicazione, lasciando i core di calcolo delle GPU inattivi in attesa dei dati e facendo impennare la latenza e i consumi energetici dei data center.



Come evidenziato dal grafico, mentre la dimensione e i parametri dei modelli Transformer crescono a un ritmo vertiginoso (circa 410x ogni due anni), la capacità di memoria dell'hardware AI avanza a un passo drasticamente inferiore (soltanto 2x ogni due anni).

Tecnologie a confronto (DDR5 e HBM)

Per superare i limiti imposti dal Memory Wall e soddisfare la richiesta di dati dei processori per intelligenza artificiale, l'industria dei semiconduttori ha dovuto abbandonare le architetture generiche in favore di soluzioni specializzate. Il confronto tecnologico ed economico tra la memoria dinamica tradizionale, nota come DRAM, e la memoria ad ampia larghezza di banda, chiamata HBM, mette in luce due filosofie ingegneristiche opposte. Le memorie di tipo DDR5 o LPDDR5 si basano su una struttura piana a due dimensioni, in cui i chip finiti e racchiusi nel loro involucro protettivo vengono saldati a distanza dal processore sulla scheda madre. Al contrario, la tecnologia HBM sfrutta un'architettura tridimensionale, dove i singoli strati di memoria vengono assottigliati e impilati verticalmente sopra un chip logico di controllo alla base, proprio come i piani di un grattacielo. Questo blocco verticale viene poi posizionato nello stesso involucro fisico della GPU, riducendo al minimo la distanza che i dati devono coprire.
La differenza fondamentale tra i due sistemi risiede nella struttura del canale di comunicazione, chiamato bus, che trasporta le informazioni:

- DRAM tradizionale: adotta una logica a canali stretti, solitamente larghi trentadue o sessantaquattro linee parallele per ciclo. Per compensare questo limite e trasferire volumi elevati di dati, i segnali vengono fatti viaggiare a frequenze elettriche altissime, un meccanismo che genera forte rumore elettrico, dispersione termica e consumi energetici elevati.

- Memoria HBM: segue una logica basata su canali larghissimi, con un bus dati che tocca i milleventiquattro fili paralleli nelle versioni attuali e raddoppia a duemilaquarantotto linee nello standard di nuova generazione HBM4. Questa struttura permette il passaggio di terabyte di dati al secondo mantenendo frequenze e tensioni operative inferiori, abbattendo il consumo energetico per singolo bit trasferito.

Mentre le memorie di tipo DDR5 e LPDDR5 restano lo standard economico e modulare per l'informatica di massa, i computer aziendali e gli smartphone, la tecnologia HBM è diventata il pilastro degli acceleratori di calcolo avanzati. Dal punto di vista industriale, tuttavia, la fabbricazione dei moduli HBM comporta una complessità e costi senza precedenti. Un gigabyte di memoria HBM costa oltre tre volte rispetto alla DRAM standard e il suo peso economico arriva a rappresentare dal quarantacinque al sessanta per cento dell'intero costo di produzione di una GPU per intelligenza artificiale. Questa impennata dei costi è guidata da tre fattori produttivi:

- Consumo di silicio: in quanto i processi di perforazione verticale richiedono da tre a quattro volte la quantità di wafer di silicio rispetto alle memorie piane
- Rese di fabbricazione: registrano livelli di scarto superiori del venti o trenta per cento a causa del rischio che un singolo difetto in uno strato rovini l'intero blocco verticale
- Packaging: necessità di ricorrere a complessi processi di packaging avanzato per saldare con precisione microscopica la memoria ai core di calcolo.

Caratteristica	DRAM tradizionale (DDR5 / LPDDR5)	HBM (HBM3e / HBM4)
Architettura	Chip planari, montati a distanza dal processore	Stack 3D verticale (8-16 die impilati) collegati da TSV
Ampiezza del bus dati	32-64 bit	1.024 bit (HBM3e) — 2.048 bit (HBM4)
Banda passante	Decine di GB/s per chip	Fino a 1,5-2 TB/s per stack (HBM4)
Posizionamento fisico	Su scheda madre, slot dedicati	A ridosso del die logico, su interposer di silicio
Processo di assemblaggio	Saldatura/montaggio standard	Packaging avanzato (es. CoWoS di TSMC)
Impiego tipico	PC, smartphone, server generici	Acceleratori AI, GPU per data center, HPC

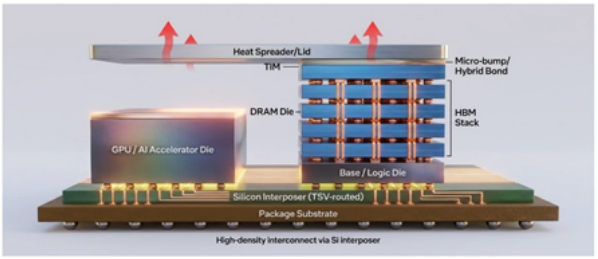
Advanced packaging

La necessità di avvicinare il più possibile i dati ai core di calcolo per superare il Memory Wall ha spinto l'industria dei semiconduttori verso una rivoluzione strutturale, segnando il passaggio dal packaging tradizionale a quello avanzato. Per decenni, l'approccio classico ha previsto la fabbricazione separata di processori e memorie, inseriti poi in involucri protettivi di resina e saldati l'uno accanto all'altro sulle schede madri. Le connessioni elettriche dovevano così percorrere tracce di rame lunghe diversi centimetri sul circuito stampato, introducendo ritardi temporali e una massiccia dispersione di energia per lo spostamento dei segnali. L'advance packaging supera questo limite inserendo più chip non confezionati, detti chiplet, all'interno dello stesso modulo integrato, riducendo le distanze di collegamento a pochi millimetri o micrometri e trasformando i componenti in un unico superchip.
Nello sviluppo degli acceleratori per intelligenza artificiale, questa tecnologia si articola su due architetture geometriche complementari, note come integrazione 2.5D e integrazione 3D.
L'integrazione 2.5D utilizza uno strato di silicio ultrasottile intermedio, chiamato interposer, posizionato alla base del modulo. Poiché le normali schede madri non hanno una densità di piste sufficiente per gestire i bus larghissimi delle memorie moderne, l'interposer funge da ponte ad alta velocità sopra il quale il processore logico e i moduli di memoria vengono montati fisicamente l'uno accanto all'altro. Lo standard industriale di riferimento in questo campo è il processo chiamato CoWoS (Chip-on-Wafer-on-Substrate), sviluppato dalla fonderia taiwanese TSMC per l'assemblaggio di quasi tutti i chip avanzati del mercato, come i sistemi Nvidia Blackwell. All'interno di questo ecosistema, le memorie si integrano verticalmente tramite il packaging 3D, che consiste nell'impilare fino a sedici strati di silicio per moltiplicare la capacità di memoria senza occupare spazio orizzontale. Il collegamento elettrico tra questi piani sovrapposti è garantito da migliaia di minuscoli canali verticali scavati direttamente nel silicio e riempiti di rame, chiamati Through-Silicon Vias o TSV, coordinati da un base die logico alla base che si occupa di gestire il traffico dei dati verso il processore.
Questa straordinaria sofisticazione ingegneristica ha però trasformato l'Advanced Packaging nel vero collo di bottiglia per la produzione globale di hardware per intelligenza artificiale, a causa di severi vincoli industriali e fisici. Il primo ostacolo riguarda la drammatica caduta delle rese di fabbricazione

(yields) dovuta all'accumulo dei difetti nell'assemblaggio verticale: se un solo strato presenta un'imperfezione o se un singolo canale microscopico fallisce il contatto, l'intero stack di memoria diventa inutilizzabile, portando a tassi di scarto superiori del venti o trenta per cento rispetto alle memorie classiche. Il secondo limite è di natura termica, poiché l'impilamento verticale ostacola la dissipazione del calore, trattenendo l'energia termica negli strati inferiori e generando pericolosi hotspot interni che possono danneggiare il silicio o causare thermal throttling. A ciò si aggiunge lo stress meccanico indotto dall'obbligo di assottigliare i wafer di silicio a livelli micrometrici (thinning), esponendo le piastre a rischi di deformazione fisica (warpage). L'insieme di questi fattori, unito alla limitata capacità produttiva delle linee di assemblaggio mondiali, fa sì che il packaging e la memoria HBM arrivino a rappresentare oltre il sessanta per cento del costo totale di fabbricazione di una GPU avanzata, condizionando i volumi di consegna di tutta l'industria tecnologica.



Figure 4: Integrated multiple chips into one single package using a sub-micron scale silicon interface (interposer).



HBM 3D Packaging

Supply chain globale (SK Hynix, Samsung e Micron)

Il mercato globale della memoria avanzata a semiconduttore ha superato la sua storica natura di commodity ciclica e volatile per trasformarsi in un pilastro della sovranità tecnologica e geopolitica. Il settore ha vissuto un profondo consolidamento industriale che ha portato alla nascita di un triopolio dominante, composto dalle sudcoreane Samsung Electronics e SK Hynix e dallo statunitense Micron Technology. Questa stretta concentrazione industriale ha permesso ai tre grandi produttori di gestire l'offerta con rigore, preparando il terreno per rispondere al deficit strutturale di chip innescato dagli investimenti monumentali dei grandi fornitori di servizi cloud. La produzione di memorie ad ampia larghezza di banda come la tecnologia HBM richiede infatti una densità di risorse e uno spazio in camera bianca quattro volte superiore rispetto alle memorie di tipo DRAM standard, riducendo di conseguenza la disponibilità di silicio per i settori tradizionali come l'automotive e l'industria.
Nel segmento strategico delle memorie HBM, la dinamica competitiva ha registrato una forte evoluzione. SK Hynix difende la sua leadership di mercato con una quota vicina al cinquanta per cento, beneficiando del vantaggio di essere stata la prima a muoversi sul mercato e di una storica partnership con Nvidia, supportata dall'uso della tecnologia di assemblaggio termico MR-MUF. Samsung Electronics ha registrato un recupero molto aggressivo, portando la sua quota al trentatré per cento grazie all'avvio della produzione in serie dell'architettura di nuova generazione HBM4 sul nodo produttivo DRAM 1c. Micron Technology, l'unico produttore occidentale rimasto nel segmento di fascia alta, mantiene una quota stabile intorno al diciotto per cento, focalizzandosi sull'efficienza energetica dei suoi moduli e impegnando l'intera capacità produttiva fino al 2027. Fuori

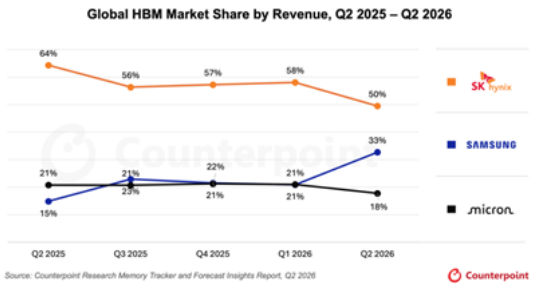
dall'oligopolio dei tre grandi produttori si muove la cinese CXMT, la quale tuttavia opera principalmente su mercati tradizionali ed è esclusa dalle forniture occidentali più avanzate a causa delle restrizioni all'esportazione di tecnologie.

Se si allarga l'analisi ai mercati convenzionali della memoria volatile DRAM e dello storage non volatile NAND Flash, le strategie e i rapporti di forza si frammentano:

- Mercato DRAM complessivo: Samsung consolida la prima posizione globale con il trentanove virgola quattro per cento del mercato grazie ai grandi volumi per i server aziendali. SK Hynix segue al venticinque per cento, dopo aver operato un forte passaggio di risorse dalle memorie tradizionali verso i più redditizi moduli HBM. Micron detiene il ventitré virgola tre per cento con un focus sui server di fascia alta, mentre la cinese CXMT insidia i leader storici raggiungendo il dieci per cento del mercato e una resa di fabbricazione del novanta per cento sui moduli di tipo DDR5.

- Mercato NAND Flash ed Enterprise SSD: la memoria di archiviazione per data center è diventata fondamentale per gestire i dataset dell'intelligenza artificiale. Samsung guida il settore con il ventotto per cento del mercato attraverso le sue architetture eSSD. SK Hynix, insieme alla divisione Solidigm, detiene il diciannove per cento grazie alla leadership nelle memorie ad altissima densità. Micron si attesta al quindici per cento in seguito all'uscita dal mercato al dettaglio, tallonata dalla cinese YMTC al quattordici per cento e dalla giapponese Kioxia al quattordici per cento.

I tre grandi produttori rispondono a questa pressione competitiva attraverso tre modelli di business radicalmente diversi. Samsung Electronics punta sulla totale integrazione verticale, combinando internamente lo sviluppo della memoria, le fabbriche di silicio e i processi di packaging avanzato per offrire una soluzione completa e indipendente. Al contrario, SK Hynix persegue una strategia di specializzazione d'eccellenza, preferendo rinunciare all'integrazione interna per stringere un'alleanza strategica con la fonderia taiwanese TSMC, unendo la propria memoria al packaging CoWoS leader del settore. Micron Technology ha invece adottato un modello focalizzato sulla sicurezza geopolitica, abbandonando i prodotti di consumo per diventare un fornitore puramente aziendale. Sfruttando i sussidi governativi americani del CHIPS Act, Micron sta costruendo nuovi poli produttivi negli Stati Uniti e a Singapore, offrendo ai clienti occidentali una catena di fornitura protetta e immune dalle tensioni internazionali nello Stretto di Taiwan.



Traiettorie future

L’infrastruttura hardware contemporanea si trova ad affrontare un limite strutturale in cui la separazione fisica tra l'unità che elabora i dati e quella che li conserva blocca le prestazioni dei

sistemi. Per superare questa barriera nei carichi di lavoro dell'intelligenza artificiale, l'industria ha smesso di puntare sulla sola potenza dei processori, concentrandosi sulla riprogettazione dei chip di memoria. Secondo le analisi di Yole Group e TrendForce, la memoria rappresenta oggi l'asse portante attorno a cui si riorganizza lo sviluppo dell'hardware globale. Questa trasformazione si articola su tre catalizzatori tecnologici fondamentali: lo sviluppo dello standard tridimensionale HBM4 in stretta sinergia con le fonderie logiche, la disaggregazione delle risorse nei grandi centri di calcolo tramite il protocollo CXL e l'introduzione del calcolo direttamente a bordo chip supportato dalla fotonica del silicio.

- **HBM4 e fonderie logiche**

Lo standard HBM4 rappresenta una discontinuità fondamentale rispetto alle precedenti generazioni di memorie 3D. Fino alla tecnologia HBM3E, lo strato logico posizionato alla base della torre di memoria, noto come base die, veniva realizzato dagli stessi produttori di DRAM utilizzando i medesimi processi produttivi della memoria. Con l'avvento di HBM4, la necessità di orchestrare flussi informativi immensi richiede l'adozione di nodi logici avanzati a 3 e 4 nanometri, tipici dei processori di calcolo più sofisticati. Questo cambiamento costringe i produttori di memorie a integrarsi direttamente con fonderie logiche esterne come TSMC per la fabbricazione del base die. Dal punto di vista tecnico, lo standard HBM4 raddoppia l'ampiezza del bus dati portandolo a 2048 bit, permettendo di raggiungere una banda passante compresa tra 1,5 e 2 terabyte al secondo per singolo stack e capacità complessive fino a 64 gigabyte. Questa complessità strutturale comporta un forte impatto economico. Secondo le rilevazioni degli analisti di TrendForce, la produzione di HBM4 introduce un premio di prezzo superiore al trenta per cento rispetto alla generazione precedente. Fonti di settore indicano che SK Hynix, nelle prime negoziazioni con Nvidia fornitrice della futura piattaforma Rubin, ha strutturato accordi commerciali con prezzi superiori fino al settanta per cento rispetto ai moduli HBM3E, espandendo ulteriormente il peso finanziario della memoria sul costo totale degli acceleratori.

- **CXL e memory pooling**

Se l'architettura HBM4 risponde al bisogno di massima velocità di prossimità per le GPU, il protocollo CXL (Compute Express Link) risolve l'efficienza complessiva dei data center attraverso la disaggregazione delle risorse. Nelle architetture server tradizionali, la memoria ad accesso casuale è rigidamente vincolata alla scheda madre del singolo processore. Questo isolamento fisico genera il fenomeno della memoria inutilizzata o bloccata, definita tecnicamente stranded memory, ovvero banchi di DRAM che rimangono inerti su un server mentre una macchina adiacente sperimenta un deficit di risorse. Sfruttando le linee di interconnessione ad alta velocità PCIe, il protocollo CXL abilita l'architettura del memory pooling. Questa tecnologia permette di estrarre la memoria dai singoli nodi fisici e di convogliarla in grandi pool condivisi, a cui i diversi processori e acceleratori grafici del data center possono attingere in modo flessibile e dinamico a seconda del carico di lavoro. L'evoluzione di questo mercato è confermata dai dati di Fortune Business Insights, che stimano una crescita del segmento dei sistemi di pooling CXL da 213,6 milioni di dollari a oltre 7,2 miliardi di dollari entro la metà del prossimo decennio. Il CXL non nasce quindi come uno strumento per aumentare la banda passante di picco del singolo chip, ma come una soluzione strutturale per ottimizzare il costo totale di possesso dei grandi centri di calcolo.

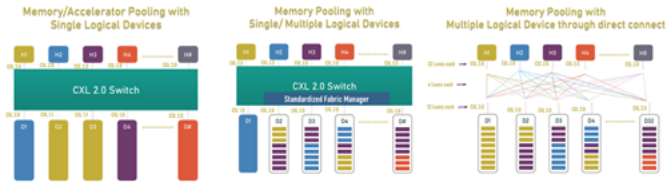


Figure 3: Pooling of CXL devices with or without a CXL 2.0 switch

- **Calcolo in memoria e fotonica**

La traiettoria più dirompente sul lungo termine è rappresentata dal paradigma del Compute-In-Memory, o CIM, unito all'introduzione delle interconnessioni ottiche. Il calcolo in memoria punta a eliminare alla radice il costo energetico del trasferimento dei dati eseguendo le operazioni logiche e matematiche direttamente all'interno della griglia di silicio della memoria, sia essa una comune DRAM o una memoria emergente ad alta ritenzione. Elaborando le informazioni sul posto, si azzera la necessità di far viaggiare i bit lungo i bus elettrici esterni, eliminando la principale causa di surriscaldamento del sistema. I recenti simposi industriali come il Flash Memory Summit hanno confermato che i moduli dotati di elaborazione integrata stanno uscendo dai laboratori per diventare una realtà commerciale. Tuttavia, questa tecnologia deve affrontare severi limiti fisici: integrare la logica di calcolo e i condensatori di memoria nello stesso spazio verticale aumenta la densità di calore in una struttura già soggetta a forte isolamento termico. Per mitigare questa criticità e superare i colli di bottiglia elettrici, l'industria sta sviluppando l'integrazione della fotonica del silicio direttamente all'interno dei moduli di memoria. Sostituendo i tradizionali conduttori in rame con microscopici canali in grado di guidare impulsi luminosi, i chip possono scambiare dati a frequenze altissime tramite la luce. L'adozione dei collegamenti ottici chip-to-memory permette di abbattere drasticamente la latenza e i consumi legati al trasporto degli elettroni, tracciando la strada per i supercomputer del prossimo decennio.

Le memorie in Borsa: dal ciclo al superciclo

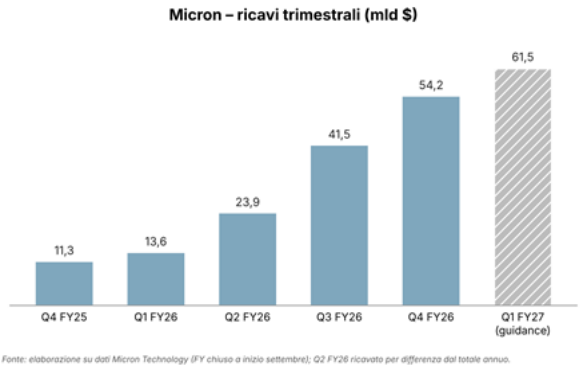
Se le sezioni precedenti hanno spiegato perché la memoria è diventata il collo di bottiglia dell'intelligenza artificiale, i mercati finanziari hanno tradotto questa scarsità in numeri che, fino a due anni fa, sarebbero sembrati impensabili per un settore considerato una commodity. Storicamente i titoli delle memorie hanno seguito un copione noto: prezzi in salita, investimenti in nuova capacità, eccesso di offerta, crollo dei prezzi e delle quotazioni. Gli investitori li hanno sempre trattati come titoli ciclici puri, da comprare quando gli utili sono ai minimi e da vendere quando toccano i massimi.

Il 2026 ha messo alla prova questo schema come mai prima d'ora. La domanda legata all'AI ha assorbito tutta la capacità produttiva disponibile, i prezzi contrattuali di DRAM e NAND hanno registrato i rialzi più forti della loro storia e i bilanci dei produttori hanno raggiunto livelli di redditività più tipici del software che dell'hardware. Secondo le stime di TrendForce, il mercato globale delle memorie dovrebbe valere circa 552 miliardi di dollari nel 2026 e superare gli 840 miliardi nel 2027.

Trimestrali da record

Il caso più emblematico è Micron. Nel quarto trimestre dell'esercizio fiscale 2026, chiuso il 3 settembre, la società ha riportato ricavi per 54,2 miliardi di dollari, quasi cinque volte gli 11,3 miliardi

dello stesso trimestre dell'anno precedente, con un margine lordo rettificato dell'87% contro il 45,7% di un anno prima. L'intero esercizio si è chiuso con 133,2 miliardi di ricavi, contro i 37,4 miliardi del 2025, e un utile netto di circa 85 miliardi. Il solo segmento Core Data Center ha generato 18 miliardi di ricavi con un margine lordo del 90%, quando un anno prima ne valeva 1,6. La guidance per il trimestre in corso indica ricavi per circa 61,5 miliardi, sopra i 57,9 attesi dagli analisti, e l'amministratore delegato Sanjay Mehrotra ha dichiarato di non avere visibilità su quando domanda e offerta torneranno in equilibrio, prevedendo una domanda superiore all'offerta sia nel 2027 sia nel 2028.



Micron – ricavi trimestrali in miliardi di dollari

Numeri analoghi arrivano dalla Corea del Sud. SK Hynix ha chiuso il secondo trimestre 2026 con ricavi per 79,3 trilioni di won (circa 55 miliardi di dollari, +257% su base annua) e un utile operativo di 60,5 trilioni di won (+557%), pari a un margine operativo di circa il 76%: il quinto trimestre consecutivo di profitti record. Nello stesso periodo la società ha avviato le spedizioni in volume di HBM4 e ha firmato accordi di fornitura di lungo periodo con circa dieci clienti.

Samsung Electronics, che a differenza dei concorrenti comprende anche smartphone, televisori ed elettrodomestici, ha comunicato l'8 ottobre i risultati preliminari del terzo trimestre: utile operativo di 107,4 trilioni di won (circa 80 miliardi di dollari), in crescita del 782% rispetto a un anno prima. È la prima società coreana a superare i 100 trilioni di won di utile in un solo trimestre. Secondo le stime degli analisti, la divisione memorie avrebbe generato da sola circa 110 trilioni di utile operativo, compensando le perdite della divisione consumer, anche grazie alle consegne di HBM4 per la piattaforma Vera Rubin di Nvidia. I risultati definitivi di Samsung e quelli di SK Hynix sono attesi entro la fine di ottobre.

Il boom non riguarda solo la DRAM. Sandisk, separata da Western Digital all'inizio del 2025 e specializzata in memorie NAND, ha chiuso l'esercizio fiscale con un ultimo trimestre da 8,97 miliardi di dollari di ricavi (+372% su base annua) e un margine lordo dell'84,6%, trainata dagli SSD per data center, i cui ricavi sono passati in un anno da 213 milioni a quasi 3 miliardi di dollari.

La performance in Borsa

I mercati hanno anticipato e amplificato questi risultati. Micron ha chiuso il 2025 a 285 dollari per azione, a fine maggio ha superato per la prima volta i mille miliardi di dollari di capitalizzazione e al 12 ottobre quotava circa 984 dollari, con un rialzo da inizio anno del 245%. SK Hynix è passata da circa 650.000 won di fine 2025 a circa 1,78 milioni di won a inizio ottobre, quasi triplicando il proprio

valore, mentre Samsung Electronics nello stesso periodo è più che raddoppiata, superando i 270.000 won.

Il peso dei due colossi dei chip ha trascinato l'intero listino coreano: già a inizio maggio il KOSPI aveva guadagnato quasi il 65% da inizio anno, sfiorando quota 7.000 punti. Ancora più estremo il caso di Sandisk, il cui titolo tra giugno 2025 e giugno 2026 è passato da circa 60 a quasi 2.000 dollari, moltiplicando il proprio valore di circa trenta volte. Anche Western Digital, oggi focalizzata sugli hard disk per data center, segna un +126% da inizio anno.

Il paradosso delle valutazioni

Rialzi di queste dimensioni farebbero pensare a valutazioni estreme, eppure i multipli raccontano una storia diversa. Annualizzando l'utile per azione indicato da Micron nella guidance per il trimestre in corso (38,15 dollari), il titolo tratta a circa 6-7 volte gli utili, mentre a luglio SK Hynix quotava circa 5,8 volte gli utili attesi: valori ben al di sotto di quelli medi dell'indice S&P 500. È il classico paradosso dei titoli ciclici: quando gli utili sono ai massimi il mercato applica multipli bassi, perché non ritiene quei profitti sostenibili nel tempo. In altre parole, gli investitori non stanno pagando gli utili di oggi, ma scontano la probabilità che il ciclo si inverta.

Il cuore del dibattito tra analisti è proprio questo: la memoria è diventata un'infrastruttura strategica con utili strutturalmente più alti, e quindi meritevole di multipli più elevati, oppure stiamo osservando il picco di un ciclo eccezionalmente forte, ma comunque destinato a chiudersi? A favore della prima tesi giocano i contratti di lungo periodo, che spesso prevedono prezzi minimi e pagamenti anticipati da parte dei clienti: le passività contrattuali non correnti di Micron, che riflettono questi anticipi, sono passate in un anno da 0,57 a 12,9 miliardi di dollari. A favore della seconda c'è la storia del settore, in cui ogni periodo di scarsità è stato seguito da un'ondata di investimenti e, prima o poi, da un eccesso di offerta.

Ottobre 2026: la correzione e il fattore CXMT

Le ultime settimane hanno mostrato quanto il sentiment resti fragile. Tra il 12 e il 13 ottobre i titoli delle memorie hanno subito una delle correzioni più violente dell'anno: Micron ha perso circa il 4% in una sola seduta, portandosi intorno al 20% sotto i massimi di giugno, mentre a Seul SK Hynix ha ceduto circa il 10% e Samsung oltre il 7%, trascinando il KOSPI alla peggiore seduta da luglio, con un calo superiore al 5%.

Il catalizzatore è arrivato dalla Cina. CXMT, il principale produttore cinese di DRAM, ha annunciato progressi su un'architettura di cella più densa, chiamata 4F², che a parità di nodo produttivo consente in teoria circa il 50% di densità in più rispetto all'attuale standard 6F², e ha dichiarato di voler lanciare entro fine anno moduli DDR5 per server, cioè proprio nel segmento più redditizio del mercato. Sulle quotazioni coreane hanno pesato anche, secondo quanto riportato dalla stampa finanziaria, la conclusione anticipata dei programmi di riacquisto di azioni proprie di Samsung e SK Hynix, per un valore complessivo di circa 40 miliardi di dollari, e cinque sedute consecutive di vendite nette da parte degli investitori esteri.

Più in generale, il mercato ha iniziato a chiedersi se gli utili abbiano già toccato il picco: negli ultimi 90 giorni le stime di consenso sugli utili del terzo trimestre di Samsung e SK Hynix sono state riviste al ribasso di quasi il 4%. Va però sottolineato che il prodotto server di CXMT è per ora solo un

obiettivo dichiarato: dovrà superare i lunghi processi di qualificazione presso i clienti e fare i conti con le restrizioni statunitensi sull'export di macchinari per la produzione di semiconduttori.

Rischi e segnali da monitorare

Per chi guarda al settore nei prossimi mesi, le variabili da tenere d'occhio sono principalmente quattro:

- **Nuova capacità produttiva:** Micron avvierà la produzione nel nuovo impianto dell'Idaho a metà 2027, mentre SK Hynix prevede l'apertura della prima camera bianca del polo di Yongin all'inizio dello stesso anno. Storicamente, l'aumento della spesa in conto capitale è stato il principale segnale anticipatore di un eccesso di offerta.
- **Dinamica dei prezzi:** dopo rialzi del 15-20% nel terzo trimestre, TrendForce stimava per il quarto trimestre aumenti dei prezzi contrattuali della DRAM per PC compresi tra il 3% e l'8%. Il rallentamento è fisiologico, ma un appiattimento dei prezzi è spesso il primo segnale di inversione del ciclo.
- **Margini dell'HBM:** secondo Bernstein, l'HBM potrebbe diventare meno redditizia della DRAM tradizionale già nel 2027, con una normalizzazione dei prezzi nel 2028 e margini lordi in calo da circa il 90% a valori sotto l'80%.
- **Concorrenza cinese e domanda AI:** l'avanzata di CXMT e YMTC, l'evoluzione delle restrizioni all'export e la sostenibilità dei piani di investimento degli hyperscaler restano le principali variabili esterne in grado di cambiare lo scenario.

Conclusioni

Il settore delle memorie sta vivendo la fase più redditizia della sua storia, e la parte tecnica di questo report spiega perché: senza memoria ad alta banda, il calcolo per l'intelligenza artificiale semplicemente non scala. Per gli investitori, però, la domanda non è se la memoria sia importante, ma quanto a lungo i produttori riusciranno a mantenere margini vicini al 90% prima che nuova capacità, nuovi concorrenti o un rallentamento della spesa in AI riportino il mercato in equilibrio. Le valutazioni contenute dicono che il mercato crede ancora nel ciclo; i risultati record e i contratti di lungo periodo dicono che, almeno fino al 2027, questo ciclo potrebbe essere diverso dai precedenti. La correzione di ottobre ricorda che, in un settore con questa storia, la volatilità resta parte integrante dell'investimento.

Autori: Salvatore Gulisano (Asset Management Senior Associate); Leonardo Valinotto (Asset Management Associate)

Revisione: Francesco Manzi (Head of Asset Management Division)