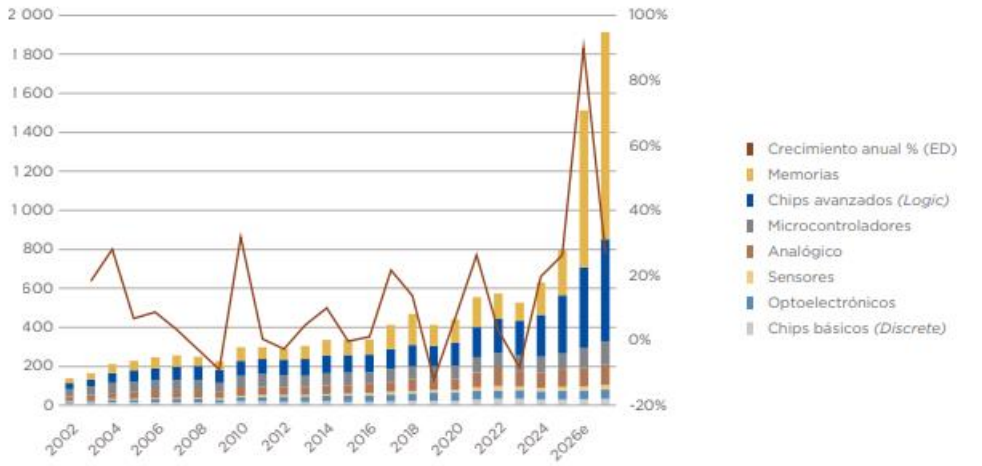
Semiconductores e inteligencia artificial: ¿cuál es la situación del ciclo de inversión?

Un ciclo de inversión sin precedentes, pero con una dispersión creciente

Tres años después del lanzamiento de ChatGPT, la oleada de inversión en infraestructuras de inteligencia artificial (IA) no muestra ningún indicio de agotamiento, sino todo lo contrario. Los cuatro grandes hiperescaladores1 estadounidenses (Microsoft, Alphabet, Amazon y Meta) prevén dedicar juntos cerca de 725.000 millones de dólares en gastos de inversión en 2026, lo que supone un aumento de más del 75% respecto al nivel ya récord de unos 410.000 millones alcanzado en 2025. Esta tendencia no es fruto de una simple apuesta: responde a una demanda que, trimestre tras trimestre, supera las capacidades operativas. El cambio en el uso de los modelos de IA, de la fase de entrenamiento a la de explotación a gran escala (la inferencia), transforma una demanda puntual de cálculo en un consumo permanente.

En este contexto, el mercado de semiconductores globales, del que no se esperaba superar el nivel simbólico de un billón de dólares de ventas anuales antes de 2028 / 2029 (gráfico 1), debería registrar a partir de este año ventas de cerca de 1,5 billones, frente a los 800.000 millones de 2025. El consejero delegado de TSMC, Che-Chia Wei, que a finales de 2025 hablaba de una demanda «aún mayor de lo que preveíamos tres meses antes», confirmó a principios de junio que la demanda seguía sin mostrar indicios de debilitamiento, y que su grupo se esforzaba ahora por no convertirse en el cuello de botella de la cadena. Por otra parte, los primeros indicios de monetización son tangibles: los ingresos de las divisiones de nube de los hiperescaladores se están acelerando y sus carteras de pedidos contractuales prácticamente se han duplicado en un año. Todos ellos son elementos que justifican, a los ojos de sus direcciones, el continuo esfuerzo de equipamiento en una carrera en la que ninguno puede permitirse ceder terreno.

Gráfico 1 — El mercado global de semiconductores debería superar el billón de dólares a partir de este año
Evolución del mercado global de semiconductores en miles de millones de dólares, por segmento y año desde 2002, y evolución anual en % (escala derecha)



Edmond de Rothschild, WSTS

Los semiconductores, primeros beneficiarios de esta bonanza, se impusieron como el principal catalizador de la rentabilidad de los índices mundiales, ya que el índice de referencia del sector (SOX) registró una subida del 102% desde principios de año, y su peso en el S&P 500 se ha cuadruplicado con creces desde 2020 (gráfico 2). Sin embargo, detrás de esta dinámica de conjunto se esconde una creciente dispersión entre los subsegmentos de la cadena de valor. Entender dónde se forma el valor, y a qué precio se paga, sigue siendo, en nuestra opinión, un ejercicio esencial para los inversores en 2026.

Gráfico 2 — La valoración de los semiconductores en el S&P 500 alcanza un récord, mientras que el sector del software se deteriora
Evolución de la ponderación de la capitalización bursátil de los sectores de los semiconductores y el *software* en el índice S&P 500, desde 2004, en %

Bloomberg, Edmond de Rothschild

Aceleradores: los GPU siguen dominando, ASIC en pleno auge Los procesadores gráficos (GPU o Graphics Processing Unit) siguen siendo el núcleo principal de la infraestructura de IA. Concebidos en un principio para la reproducción de imágenes, estos chips, también denominados aceleradores, son excelentes en el

procesamiento paralelo masivo2 que exigen el entrenamiento y la inferencia de los grandes modelos. Nvidia reina en este segmento sin una gran competencia, con más del 80% del mercado y los ingresos de centros de datos que han aumentado un 92% interanual en el último trimestre publicado. Su fuerza no se debe únicamente a sus chips: su ecosistema de software CUDA, adoptado por millones de desarrolladores desde hace casi veinte años, constituye una barrera de entrada que sus competidores, con AMD a la cabeza, tienen dificultades para superar a pesar de las ganancias progresivas de cuota de mercado en la inferencia. Sin embargo, dado que los costes de computación ascienden a decenas de miles de millones de dólares al año, los gigantes de la nube desarrollan en paralelo sus propios chips a medida, los ASIC3, optimizados para una tarea precisa y que ofrecen un coste por cálculo y un consumo energético en gran medida inferiores cuando la carga de trabajo lo permite. Google es pionero con sus TPU o Tensor Processing Units4, ya disponibles para clientes externos, Amazon acelera el despliegue de sus chips Trainium, mientras que Meta, Microsoft y OpenAI continúan con sus propios programas.

La lógica es doble: abaratar el coste unitario del cálculo y disminuir la dependencia de un proveedor que tiene una posición dominante en precios y márgenes. Los grandes ganadores de este enfrentamiento son los desarrolladores especializados que apoyan a los hiperescaladores, encabezados por Broadcom y Marvell, y sobre todo el fundidor taiwanés TSMC, que fabrica la práctica totalidad de estos componentes avanzados, sea cual sea la bandera con la que estén diseñados.

La ampliación de la cadena de valor: CPU, óptica y equipos

De forma más inesperada, la oleada de IA beneficia ahora también a los procesadores generalistas (CPU o Central Processing Unit), que durante mucho tiempo se han limitado al papel de directores de orquesta de las GPU. En este sentido, el aumento de la inferencia reequilibra la composición de los servidores especializados: mientras que un centro de datos de IA contaba históricamente con una CPU para ocho GPU, esta ratio ya se ha reducido a alrededor de uno por cuatro, e Intel y AMD prevén una convergencia hacia la paridad. Esta demanda imprevista ha cogido a la industria por sorpresa, ya que los precios de las CPU para servidores han aumentado en los últimos meses, hasta el punto de que Intel ha reorientado su capacidad de producción en detrimento de los chips de consumo. Los factores de este reequilibrio se deben en gran medida a la naturaleza de las nuevas cargas de trabajo de IA, como los agentes autónomos, que consumen más capacidad de cálculo. Más allá en la cadena, a medida que los centros de datos concentran cientos de miles de chips, el factor limitante se desplaza hacia la red que los conecta: el transporte de datos entre chips, servidores y centros de datos sitúa las interconexiones ópticas en el centro del juego, hasta la co-packaged optics5, que integra la conversión óptica lo más cerca posible del procesador, un nicho en el que se posicionan tanto Nvidia y Broadcom como especialistas de la fotónica como Coherent o Lumentum.

El ensamblaje avanzado (incluida la tecnología CoWoS6 de TSMC) sigue siendo una de las capacidades más restrictivas de la cadena. Por último, los fabricantes de máquinas de producción de chips, con ASML a la cabeza con su monopolio en la litografía EUV7, ofrecen una exposición indirecta pero ineludible: cualquiera que sea el chip que gane, se fabricará en sus máquinas. Esta lógica de «pico y pala» sigue impulsando las expectativas de crecimiento por encima de los aceleradores.

Las memorias: de un componente banalizado a un cuello de botella estratégico

Las memorias, consideradas durante mucho tiempo como un segmento cíclico y poco diferenciado, han pasado a ser un cuello de botella estratégico del cálculo de IA. En efecto, el rendimiento de un acelerador depende tanto de la velocidad a la que llegan los datos como de su potencia de cálculo bruta. Es precisamente el papel de las memorias de alto ancho de banda (HBM8), apiladas verticalmente lo más cerca posible del procesador, ya que cada nueva generación de GPU incorpora cantidades cada vez mayores. Este mercado está bloqueado por un oligopolio de tres actores: el coreano SK Hynix, que es el líder del mismo y cuyas capacidades ya están totalmente reservadas para el año en curso, su compatriota Samsung y la estadounidense Micron. La demanda de IA absorbe una parte tan grande de las capacidades de producción que la tensión se ha propagado al conjunto del mercado de la DRAM9 convencional, cuyos precios se han más que duplicado en un año, hasta el punto de que Microsoft ha asignado unos 25.000 millones de dólares del aumento de su presupuesto de inversión de 2026, únicamente al encarecimiento de los costes de suministro de componentes de memoria. Es importante señalar que la subida de precios no ha mermado la demanda, lo que supone un indicio de la voluntad de los hiperescaladores de invertir «cueste lo que cueste» para ganar la batalla de la IA. Esta escasez ha seguido impulsando las cotizaciones de compañías, como Seagate, SK Hynix, Micron o Western Digital, niveles de progresión que requieren una vigilancia especial sobre la que volveremos mas adelante.

Un análisis de inversión constructivo, pero que sigue siendo selectivo Identificamos varios elementos que favorecen una continuación del ciclo. Las carteras de pedidos están llenas, las capacidades de producción de los eslabones críticos están reservadas con varios trimestres de antelación, y la visibilidad en 2026 y 2027 es inusualmente elevada para un sector históricamente cíclico. Sin embargo, esta confianza no puede ignorar algunas valoraciones alcanzadas. La subida de las cotizaciones desde principios de año, especialmente marcada en las memorias, ya integra una parte sustancial de la ejecución de las carteras de pedidos previstas. Por ello, consideramos que el análisis debe centrarse en el crecimiento de los beneficios esperados en relación con los múltiplos pagados (ratios tipo: valor de empresa / beneficio de explotación o precio sobre beneficio ajustado a la tasa de crecimiento), así como en la dinámica de las revisiones, en la calidad y diversificación de las carteras de pedidos y en el poder de fijación de precios de cada eslabón.

Por otra parte, los riesgos clásicos del sector no han desaparecido. Una concentración extrema de la demanda de procesadores procedentes de unos pocos hiperescaladores, cuyo menor cambio en las previsiones de inversión ejercería un impacto violento en toda la cadena, el carácter cíclico histórico de las memorias, cuya escasez siempre ha terminado por reabsorberse, y un contexto geopolítico en el que las restricciones a la exportación a China y la dependencia de Taiwán siguen siendo factores de riesgo estructurales.

En este contexto, seguimos convencidos de que la pertinencia de la temática no prejuzga el precio al que es razonable exponerse. La selectividad, relegada a un segundo plano durante la fase de expansión indiscriminada del ciclo, vuelve a ser una herramienta de gestión del riesgo primordial para los inversores.

Están cambiando los factores que determinan la demanda

Queda la pregunta: ¿qué respaldará la próxima fase de crecimiento, una vez absorbida la actual oleada de equipos? En nuestra opinión, la respuesta se debe menos a la formación de modelos cada vez más grandes que a su uso masivo y permanente, la inferencia, cuya participación en la demanda de cálculo no deja de crecer. Es precisamente este cambio hacia agentes de inteligencia artificial autónomos (IA agéntica, véase cuadro comparativo), consumidores continuos de cálculo más que usuarios puntuales, lo que ahora conviene examinar.

El LLM sigue siendo central, pero su uso cambia: se pasa de un chatbot (o agente de conversación) que responde al agente autónomo que ejecuta. Un agente planifica, descompone una tarea, recurre a herramientas, comprueba sus resultados e itera. Esta cadena de acción, y no el tamaño de los modelos, impulsará la próxima oleada de demanda de cálculo.

Cuadro comparativo — IA generativa 1.0 (*chatbots*) frente a 2.0 (agéntica)
La IA agéntica supone una evolución significativa respecto a los *chatbots* estándar

	IA generativa 1.0	IA generativa 2.0
Paradigma	*Chatbots*	IA agéntica
Capacidades de IA	Responden a *prompts*, dan consejos	Planifica, actúa e itera de forma autónoma
Modelo de interacción	Intercambio único, respuesta a un *prompt*	En varias etapas, orientada hacia un objetivo
Acción	Salida únicamente en forma de texto	Llama a las API, escribe y ejecuta código, navega por la web
Dependencia del ser humano	Elevado, la intervención humana es necesaria	Baja, el agente funciona de forma autónoma
Relación humana/IA	1:0, 1	De 1:1 a varios
Consumo de tokens	1x	1000x

Edmond de Rothschild

Sobre todo, la demanda pasa del entrenamiento a la inferencia. El entrenamiento sigue siendo un evento «por oleadas» (lanzamiento de un nuevo modelo, ajustes de fine tuning), mientras que la inferencia corresponde a un consumo continuo y masivo, multiplicado por: el aumento de los agentes y los recursos a herramientas, los contextos más largos y, la necesidad de iterar/autocorregir. En otras palabras, incluso si el crecimiento de los presupuestos de entrenamiento se desacelera, la explosión del volumen de inferencia puede seguir impulsando la demanda de cálculo. A modo de comparación histórica, es un poco como el smartphone, solo se compra una vez cada dos o tres años, pero se utiliza todos los días, para todo (mensajes, vídeo, geolocalización, pagos, etc.). A largo plazo, las ventas puntuales de teléfonos no son las que más pesan, sino el consumo diario de datos y servicios móviles. Del mismo modo, en la IA, el uso continuo de los modelos (la inferencia) acaba revistiendo mucha más importancia que los costes iniciales de entrenamiento.

Por último, la inferencia se difunde más allá de Nvidia (ASIC de los hiperescaladores, aceleradores alternativos, etc.). En este modo de funcionamiento (workflow), la CPU recupera peso y dispara también los servidores CPU y la memoria no HBM (DRAM/capacidad). Para el inversor, el tema se convierte en una búsqueda de cuellos de botella: a medida que se diversifica la combinación de chips, el valor migra hacia los eslabones limitantes (capacidad de memoria, interconexiones, ensamblaje, equipamiento) en lugar de hacia un solo «campeón».

La economía de los tokens: un consumo multiplicado por mil y más

El punto clave es cuantitativo: una tarea agéntica consume aproximadamente 1.000 veces más fichas (o tokens11) que un intercambio con un chatbot. El razonamiento paso a paso, las repetidas llamadas de herramientas y la autocorrección multiplican el cálculo por solicitud. Sobre todo, el uso pasa del entrenamiento puntual a la inferencia permanente: un agente gira de forma continua, lo que transforma un gasto de cálculo puntual en consumo recurrente.

Más allá de la infraestructura, el reto se convierte también en el de la adopción: ¿quién adoptará la IA en primer lugar y a qué velocidad? A medida que los agentes empiezan a ejecutar procesos, procesan, generan y producen potencialmente más datos que los humanos. Los editores de software, históricamente centrados en interfaces y workflows «human–first», pasan a productos «agent–first»: APIs12, permisos, logs, gobernanza y observabilidad diseñados para actores no humanos. En concreto, se confía a las personas las tareas de gran valor (juicio, relación, creatividad, decisión), y se utiliza la tecnología para ayudarles, simplificar o automatizar. Son los agentes los que dirigen el flujo de trabajo diario, inician y encadenan las tareas, y solicitan la intervención humana únicamente cuando es necesario (control, arbitraje, gestión de casos complejos).

Las interacciones sociales y económicas llevarán cada vez más la huella de agentes, lo que da inicio a un importante proyecto: el desarrollo de marcos de seguridad agénticos para permitir interacciones y comercios dirigidos por la IA de manera fiable y a gran escala.

En materia de identidad y cumplimiento normativo, esto sugiere una extensión natural de los componentes existentes: el tradicional KYC (Know Your Customer o «conocer al cliente») deberá integrar un KYA (Know Your Agent o «conocer al agente»). En un mundo agéntico, no solo hay que controlar la identidad del cliente, sino también la del agente, a lo que tiene acceso, a lo que está autorizado a decidir y a pagar en su nombre. La administración de identidades también deberá gestionar identidades de agentes, y el análisis conductual como la autenticación multifactor deberá integrar la intención del agente. Se trata de nuevas oportunidades para quienes gestionan identidades, derechos de acceso, auditoría y cumplimiento, más allá del cómputo de tokens e inversiones en infraestructuras.

Codificar con la IA (vibe coder) no es hacer software

El código se está convirtiendo en una idea nativa generalizada para la inteligencia artificial: la IA ya produce una parte mayoritaria de los nuevos proyectos. La diferenciación se desplaza hacia los agentes y la orquestación, que industrializan workflows y hacen que las aplicaciones profesionales sean más sencillas de desplegar y operar. De ahí una distinción útil: generar código se ha vuelto casi gratuito; convertirlo en un producto fiable, integrado, seguro y monetizable sigue siendo el paso decisivo.

Y el efecto puede ser aún más marcado en los editores verticales, que ya disponen de datos propios, de workflows especializados y de una gran capacidad de distribución en un sector determinado: Veeva en el sector sanitario, CCC Intelligent Solutions en el sector de los seguros, Workiva para la documentación reglamentaria, o incluso IQVIA sobre los datos de venta de medicamentos y la I+D. En estos modelos, la IA no es solo una capa de asistencia: se convierte en un catalizador de productividad y de conformidad directamente integrado en el sistema de producción del cliente.

Es lo que separa a los proveedores de modelos (OpenAI, Anthropic) de las plataformas horizontales que industrializan a los agentes en los procesos empresariales (Microsoft con Copilot, Salesforce con Agentforce, ServiceNow, Palantir) y de las plataformas verticales que captan el valor allí donde el workflow y el dato son los más específicos. El valor duradero se desarrolla en este nivel de integración.

Monetización: de la facturación por usuario al resultado entregado, y del usuario humano al agente de IA

El modelo por usuario (per seat o por asiento) da paso a una tarificación con resultados obtenidos, es decir, con la tarea realizada, con el expediente tratado o con el ticket resuelto. Si este movimiento se confirma, el mercado objetivo cambia de naturaleza: ya no se limita al presupuesto de software, sino que apunta a una fracción de los costes de mano de obra que el agente automatiza, un depósito mucho más amplio. Salesforce ya factura a Agentforce a una conversación resuelta; este tipo de modelo, y la prueba de que genera un retorno cuantificable para el cliente, validará la sostenibilidad de los ingresos.

Este modelo es «nativo» de las plataformas cloud, pero puede convertirse en un puente entre software y servicios: la empresa ya no compra una herramienta (licencia) ni un equipo (días-persona), compra un resultado producido por una cadena de API. ¿Quién capta el valor? Intermediarios de infraestructuras como Cloudflare o Akamai. Estos actores ilustran bien este desplazamiento: una capa de la infraestructura situada lo más cerca posible tanto de los usuarios como de la seguridad, que puede monetizarse por cada solicitud y función ejecutada (enrutamiento, protección, observabilidad, pasarelas, etc.). En un mundo dominado por agentes, esta capa se torna estratégica porque se sitúa: (i) lo más cerca posible del usuario y los datos (latencia), (ii) en el punto de control de las identidades, permisos y políticas, y (iii) en la interfaz entre el tráfico «humano» y el tráfico «agente».

El retorno sobre el capital invertido (ROIC) no es, en este momento, la cuestión central

A largo plazo, el mercado exigirá la prueba mediante el retorno de la inversión o ROIC real, no la promesa ni la perspectiva a largo plazo. Sin embargo, una parte de los gastos de inversión en infraestructuras de IA todavía no se ha arbitrado sobre la base de rendimientos claramente identificados, sino sobre una hipótesis de demanda futura: una estimación del consumo de tokens y de la carga de inferencia en un horizonte de 2 a 5 años. Dicho de otro modo, hoy se financia una capacidad apostando por los volúmenes de uso del futuro, con una visibilidad aún imperfecta sobre (i) la tarificación del token, (ii) la parte captada por cada eslabón (nube, silicio, memoria, red), (iii) la velocidad de difusión de los casos de uso realmente monetizables, y (iv) la intensidad competitiva que podría comprimir los márgenes.

En este contexto, el riesgo es que el mercado bursátil espere más un relato de escasez —la escasez de GPU, HBM, CoWoS, de capacidad de centros de datos y de electricidad— que los fundamentales de creación de valor para el accionista (flujo de efectivo sostenible, disciplina de asignación, rendimiento sobre el capital invertido). Un relato designa a algunos ganadores, aumenta su peso en los índices, le siguen las compras pasivas y, después, las de los gestores activos, a veces con apalancamiento y opciones.

La fórmula «attention is all you need» o «solo es necesario prestar atención» describe la mecánica de los LLM; funciona casi como una metáfora de los mercados actuales: la atención prestada al relato (y a los condicionantes de oferta visibles) tiende a dominar la atención prestada a los rendimientos económicos. La fragilidad no se debe necesariamente a la solidez de las empresas, sino al diferencial entre expectativas y ROIC, y a la concentración de los titulares.

Anthony Toupin, analista sénior de inversión global del Grupo Edmond de Rothschild

Xiadong Bao, gestor de carteras de renta variable internacional de Edmond de Rothschild Asset Management