A oportunidade mais mal compreendida do boom de IA

Introdução e Resumo em 3 Minutos

Caro leitor,

Sabemos que você é ocupado. Por isso, antes de entrar na carta completa preparamos um resumo para quem tem apenas três minutos. Três minutos honestos, sem aquela leitura dinâmica de LinkedIn em que você finge que leu mas só viu o título.

Se depois desses três minutos você quiser parar, sem ressentimentos. Mas apostamos que vai querer dedicar alguns minutos a mais.

A carta em 3 minutos

Existe hoje, escondida sob décadas de reputação como negócio cíclico e comoditizado, uma das oportunidades mais mal compreendidas e precificadas do boom de inteligência artificial: os fabricantes de chips de memória.

A tese é simples na essência, mas profunda nas implicações. A inteligência artificial moderna - dos grandes modelos de linguagem aos agentes que já estão transformando o desenvolvimento de software - funciona, em última instância, como um exercício brutal de álgebra linear em escala colossal. Trilhões de parâmetros precisam ser multiplicados, movidos e atualizados bilhões de vezes. E para que isso aconteça, os processadores precisam ser alimentados constantemente com dados em velocidades absurdas. Se os dados não chegam rápido o suficiente, GPUs ficam parados. Em grandes data centers, processadores ociosos representam desperdício medido em milhões de dólares por hora.

É aí que entra a *High Bandwidth Memory* (HBM), uma tecnologia de memória que entrega velocidade ordens de magnitude superior à memória tradicional. Apenas três empresas no mundo - Samsung, SK Hynix e Micron - conseguem fabricá-la, e as barreiras de entrada não param de crescer. Até Elon Musk, que pretende internalizar praticamente toda a cadeia de hardware de IA, reconhece que a memória de alta performance é o componente mais difícil de replicar.

O que está acontecendo agora é uma mudança de identidade. A memória deixou de ser um componente genérico e periférico para se tornar o gargalo central do desempenho de IA - o fator que mais determina o custo por token, a eficiência energética e a velocidade dos sistemas. E os números confirmam: a Micron saiu de margens brutas entre 20-30% para próximas de 60%, com a HBM saltando de 5% da receita para algo entre um quarto e um terço projetado para 2026.

Apesar de tudo isso, o mercado continua precificando essas empresas como se nada tivesse mudado. Micron e SK Hynix negociam a 8-10x lucros projetados - múltiplos típicos de meio de ciclo para empresas de commodity. A narrativa dominante permanece a mesma de sempre: a oferta vai alcançar a demanda, os preços vão cair, e o velho ciclo de *boom e bust* vai se repetir. Pode ser. Mas se este ciclo for diferente - e há evidências crescentes de que é -, a assimetria de retorno é extraordinária.

No fundo Equitas High Convictions, esses investimentos já respondem por 36% do ganho total do fundo desde sua criação e pela totalidade da rentabilidade dos últimos 12 meses.

A carta completa que se segue explora essa tese em profundidade: a mecânica técnica que torna a memória tão crítica, a evolução do *profit pool* da IA, porque isso não é uma bolha como a da internet, e os riscos que poderiam invalidar a tese. Se você chegou até aqui e tem mais do que três minutos, continue. Acreditamos que essa transformação ainda está nos estágios iniciais - e que pode representar uma rara oportunidade de longo prazo. O restante da carta explica o porquê.

Imagine este cenário

Imagine o que poderia acontecer com a ação de uma empresa que, ao longo de cinco anos, passa de ser percebida pelo mercado como uma produtora cíclica de commodities - com margens brutas de 25–35% e normalmente negociada a 10–12x lucros - para se tornar a fornecedora de um dos insumos mais críticos da tecnologia mais transformadora já desenvolvida pela humanidade, com margens brutas de 65–75% e negociada a múltiplos compatíveis com sua posição única de mercado.

Agora imagine que apenas três empresas no mundo possuem a capacidade tecnológica, a escala de fabricação e o know-how acumulado necessários para produzir esse insumo. Imagine que as barreiras de entrada não estejam diminuindo, mas aumentando rapidamente; que a demanda esteja acelerando mais rápido do que a oferta consegue realisticamente acompanhar; e que o mercado continue avaliando essas empresas como se ainda estivessem presas ao mesmo ciclo de commodities que definiu sua história por décadas.

Esta é a oportunidade que acreditamos existir hoje em Micron e SK Hynix e Sandisk.

Identificamos essa tese pela primeira vez no segundo semestre de 2024 e iniciamos uma posição em Micron no fundo Equitas High Convictions por volta de US$100 por ação. Em abril de 2025, durante a volatilidade associada ao que ficou conhecido como “Liberation Day”, a ação chegou a negociar brevemente na casa dos US$60. Em vez de reduzir, mantivemos nossa posição. Ao longo do segundo semestre de 2025, à medida que as primeiras evidências que sustentam nossa tese começaram a se materializar, ampliamos o investimento para incluir SK Hynix e, mais recentemente, Sandisk – um case na indústria de memória exposto as mesmas dinâmicas que Micron e SK Hynix mas com um portfólio de produtos diferente. Desde o seu início, em julho de 2022, o Equitas High Convictions acumula valorização de 176% — o equivalente

a 33% ao ano — superando com ampla margem tanto o seu benchmark quanto o Ibovespa, que avançaram 42% e 84% no período, respectivamente. Embora os investimentos em fabricantes de chips de memória sejam recentes, eles já respondem por 36% do ganho total do fundo desde a sua criação e foram responsáveis pela totalidade da rentabilidade nos últimos 12 meses.

O que começou como uma hipótese está se tornando, cada vez mais, uma realidade observável.

Para entender o porquê, precisamos começar pela tecnologia que está no centro dessa transformação: a High Bandwidth Memory - HBM.

Por que a velocidade da memória determina o futuro da inteligência artificial

Os sistemas modernos de inteligência artificial são construídos sobre redes neurais inspiradas na estrutura do cérebro humano. Esses sistemas consistem em camadas de operações matemáticas organizadas como grandes matrizes de parâmetros que transformam dados de entrada em representações cada vez mais abstratas. Modelos de ponta atualmente contêm centenas de bilhões a trilhões de parâmetros - modelos da classe GPT-4 possuam mais de 1 trilhão de parâmetros - distribuídos em dezenas a centenas de camadas.

Treinar esses modelos consiste em multiplicar repetidamente matrizes enormes usando conjuntos massivos de dados. Cada etapa de treinamento envolve ajustar parâmetros por meio de *gradient descent* - essencialmente atualizando bilhões de pesos por meio de multiplicação de matrizes.

A inferência - a geração de um token - também é multiplicação de matrizes, na qual vetores de entrada (*embeddings*) se propagam pela rede neural. Durante a inferência, o modelo treinado usa esses parâmetros para gerar previsões, respostas ou ações.

Tanto no treinamento quanto na inferência, a operação computacional dominante é a multiplicação de matrizes e vetores em escala gigantesca. É o uso anabolizado de álgebra linear para simular raciocínio. Um método que, apesar de extremamente ineficiente quando comparado ao funcionamento de um cérebro biológico, funciona aplicando força bruta extrema na forma de dados e energia para resolver o problema.

O avanço que levou a inteligência artificial ao *mainstream* ocorreu em 2017, quando pesquisadores do Google publicaram o artigo ***Attention Is All You Need***[1]. Esse trabalho introduziu a arquitetura *transformer*[2], que substituiu o processamento sequencial das redes neurais anteriores por uma estrutura radicalmente mais paralelizável baseada em mecanismos de atenção.

O *transformer* tornou possível processar sequências inteiras de dados simultaneamente, em vez de passo a passo. Essa mudança permitiu um aumento dramático na eficiência do treinamento e na escala dos modelos. Nos anos seguintes, essa arquitetura tornou-se a base dos grandes modelos de linguagem, culminando no lançamento público do ChatGPT no final de 2022 - o momento em que a inteligência artificial deixou de ser um campo de pesquisa e entrou no *mainstream* global.

A arquitetura *transformer* permitiu desempenho sem precedentes, mas o fez aumentando drasticamente a necessidade de computação paralela. Foi nesse momento que as GPUs se tornaram tão importantes. GPUs, sigla para *Graphics Processing Units*, foram originalmente projetadas para executar tarefas de processamento em paralelo, em vez de

[1] https://arxiv.org/pdf/1706.03762

[2] O link apresenta uma explicação detalhada do funcionamento da arquitetura Transformer: https://poloclub.github.io/transformer-explainer/

sequencialmente como as CPUs (*Central Processing Units*), o componente central dos nossos computadores pessoais. Sua aplicação original era processar pixels na tela dos computadores, uma tarefa que exige grande volume de processamento e é melhor executada em paralelo. Descobriu-se que essa capacidade de processamento paralelo é perfeitamente adequada para multiplicação de matrizes - o núcleo do processamento de sistemas de IA baseados em redes neurais. Treinar um modelo moderno envolve milhares de GPUs trabalhando simultaneamente, cada uma processando fragmentos de conjuntos de dados gigantescos e trocando continuamente resultados intermediários. Essa troca envolve armazenamento e movimentação constante de dados entre unidades de processamento e unidades de memória.

Essa arquitetura criou um novo e inesperado gargalo. O fator limitante do desempenho da IA deixou de ser a capacidade de processamento - passou a ser a capacidade de alimentar os processadores com dados rapidamente.

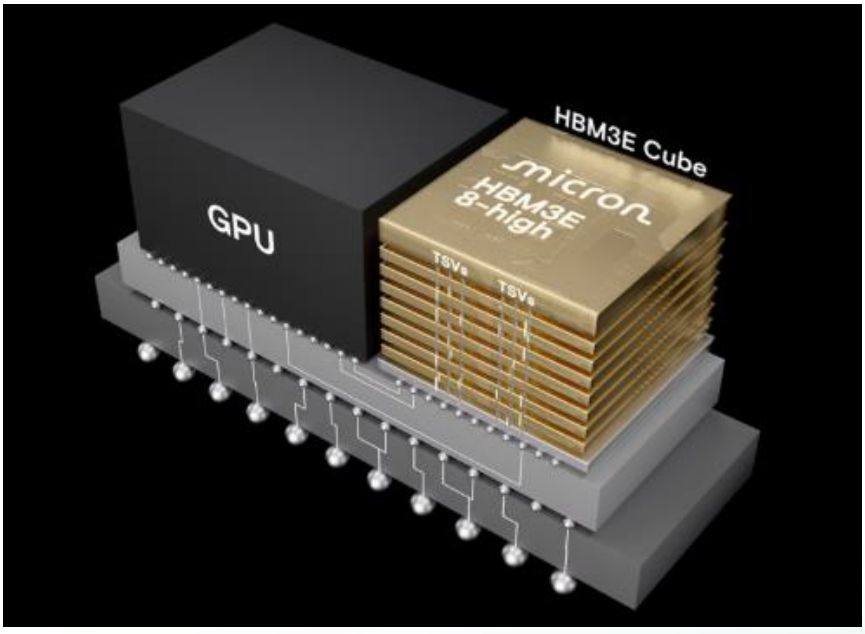
As GPUs são projetadas para realizar um número extraordinário de cálculos em paralelo. No entanto, elas só operam com eficiência máxima se os dados chegarem de forma contínua e em altíssima velocidade. Se o fluxo de dados desacelera, os processadores ficam ociosos. Em grandes clusters de IA, o custo de processadores ociosos é medido em milhões de dólares por hora.

É aqui que a *High Bandwidth Memory* (HBM) se torna essencial.

Por que a *High Bandwidth Memory* é fundamentalmente diferente da memória DRAM tradicional

Os chips de memória volátil DRAM tradicionais foram projetados para computação de propósito geral. Eles não foram concebidos para suportar milhares de processadores executando simultaneamente cargas de trabalho massivamente paralelas.

A *High Bandwidth Memory* representa uma ruptura radical com essa arquitetura. Em vez de posicionar chips de memória lado a lado em uma placa de circuito, o HBM empilha múltiplos *dies* de memória verticalmente e os conecta por meio de interconexões microscópicas verticais conhecidas como *Through-Silicon Vias* (TSVs). Essa estrutura tridimensional reduz drasticamente a distância física percorrida pelos sinais e, principalmente, permite um número muito maior de conexões paralelas entre memória e processamento, resultando em ordens de magnitude superiores de largura de banda - que determina a velocidade de transferência de dados entre a memória e a unidade de processamento.



	DRAM Tradicional	High Bandwidth Memory (HBM)
Arquitetura	Chips lado a lado na placa	Dies empilhados verticalmente com TSVs
Largura de Banda	~50 GB/s	~1.000+ GB/s (HBM3E)
Aplicação Principal	Computadores, smartphones, servidores	GPUs de IA, data centers de alta performance
Complexidade	Fabricação convencional	Manufatura avançada + empacotamento 3D
Barreiras de Entrada	Baixas (múltiplos fabricantes)	Muito altas (apenas 3 empresas)
Margem Bruta Típica	25–35%	65–75% (estimativa)
Custo no GPU NVIDIA	Componente secundário	~60% do custo total de produção

Construir essas pilhas é extraordinariamente difícil. A tecnologia exige manufatura avançada de semicondutores, empacotamento de alta precisão, sistemas complexos de fornecimento de energia e gestão térmica sofisticada para dissipar o enorme calor gerado por memórias densamente empilhadas operando em velocidades extremas.

O resultado é que a capacidade de armazenar dados e entregá-los cada vez mais rápido aos processadores tornou-se a principal restrição ao desempenho, à eficiência energética e ao custo dos sistemas de inteligência artificial. O custo de gerar um único token de IA depende cada vez menos apenas da computação e cada vez mais da velocidade e eficiência com que a memória consegue alimentar os processadores com dados.

Essa mudança é sutil, mas profunda. Durante décadas, a computação foi limitada pelo poder de processamento. Hoje, ela é limitada pelo movimento de dados e pelo consumo de energia.

A largura de banda de memória tornou-se o novo gargalo da computação. A importância do HBM fica clara quando se analisa a estrutura de custos de produção dos tão procurados GPUs da NVIDIA. Em seu relatório *The Memory Wall: Past, Present, and Future of DRAM*[3], a SemiAnalysis estima que o HBM representa atualmente aproximadamente **60% do custo de produção** dos GPUs da NVIDIA.

Como isso se traduz em uma vantagem competitiva duradoura

Ao longo das últimas décadas, a indústria global de memória acumulou um vasto e altamente especializado corpo de conhecimento. Os três principais fabricantes - Samsung, SK Hynix e Micron - detêm, em conjunto, dezenas de milhares de patentes que cobrem técnicas de empilhamento, processos de encapsulamento, arquiteturas de fornecimento de energia, ciência de materiais e gestão térmica. Replicar apenas essa propriedade intelectual provavelmente exigiria muitos anos e dezenas de bilhões de dólares.

Ainda assim, patentes, por si só, raramente criam vantagens competitivas verdadeiramente duradouras em tecnologia. O que, em última instância, distingue uma commodity de uma tecnologia difícil de replicar não é apenas a proteção legal, mas a existência de uma curva de desenvolvimento íngreme e continuamente evolutiva.

Os produtos se tornam "comoditizados" quando o desempenho atinge um ponto de utilidade marginal decrescente, como havia acontecido com a grande maioria dos chips de memória há décadas. Com o tempo, melhorias incrementais deixam de gerar valor adicional relevante para os clientes. Quando o desempenho se torna "bom o suficiente", é apenas uma questão de tempo para os players menos avançados fazerem o "*catch-up*" tecnológico e passarem a competir em preço com os líderes da tecnologia, o que leva a forte compressão de margens. Durante décadas, a memória se encaixou nessa

[3] https://newsletter.semianalysis.com/p/the-memory-wall

descrição. A capacidade de DRAM continuou crescendo, o custo por bit caiu de forma constante e as melhorias de desempenho - embora tecnicamente impressionantes - passaram a oferecer diferenciação econômica cada vez menor. O valor marginal de maior velocidade de memória se achatou, e a indústria passou a ser sinônimo de ciclos de *boom* e *bust*.

A inteligência artificial está mudando essa dinâmica de forma profunda.

A principal restrição dos sistemas modernos de IA não é mais apenas a capacidade de processamento, mas a velocidade e a eficiência energética com que os dados são movimentados entre memória e processadores. À medida que redes neurais escalam para trilhões de parâmetros e clusters de IA se expandem para dezenas de milhares de GPUs, o custo de mover dados tornou-se o componente dominante tanto do desempenho quanto do consumo de energia. Em grandes sistemas de IA, mover dados consome muito mais energia do que realizar os cálculos em si.

Pela primeira vez em décadas, memória mais rápida deixou de ser uma melhoria marginal - tornou-se um determinante de primeira ordem do desempenho do sistema, da eficiência energética e do custo por token. O avanço de sucessivas gerações de *High Bandwidth Memory* tornou-se, portanto, central para a evolução de todo o ecossistema de IA.

A transição atual para o HBM4 ilustra claramente essa dinâmica. A quarta geração de *High Bandwidth Memory* está sendo desenvolvida e qualificada por meio de um processo longo e complexo de *co-design* e certificação envolvendo a NVIDIA e os principais *hyperscalers*. Esse processo leva anos e exige profunda colaboração ao longo de toda a cadeia de suprimentos de semicondutores. A memória deixou de ser um componente genérico; tornou-se uma parte estreitamente integrada da arquitetura do sistema.

Mesmo que essa fosse a única mudança estrutural em curso, já implicaria um ciclo muito diferente das dinâmicas históricas de *boom* e *bust* da indústria de memória. O ciclo atual de demanda não é impulsionado por melhorias incrementais em computadores pessoais ou smartphones, mas pela construção global de infraestrutura de IA. Trata-se de um ciclo de implantação com características de longa duração, potencialmente se estendendo por muitos anos.

Se esse ciclo será eventualmente prolongado ou interrompido por novos paradigmas de computação ainda é incerto. No entanto, a direção da pesquisa, tanto em *hardware* quanto em *software*, sugere uma tendência mais profunda e duradoura.

Em um nível fundamental, os computadores modernos continuam extraordinariamente ineficientes quando comparados ao cérebro humano. O cérebro opera com cerca de 20 watts de potência enquanto realiza tarefas cognitivas que permanecem muito além das capacidades dos sistemas de IA mais avançados de hoje. Os maiores data centers de IA já exigem gigawatts de energia. A diferença de eficiência energética é impressionante.

Uma parte significativa dessa lacuna decorre da separação entre memória e processamento - característica central da arquitetura tradicional de von Neumann[4], ainda usada nos GPUs mais atuais. Nos sistemas atuais de IA, volumes massivos de energia são consumidos apenas no deslocamento de dados entre a memória e as unidades de processamento. Esse obstáculo, conhecido como "parede da memória" (*memory wall*), tornou-se o principal gargalo da computação.

Como resultado, a próxima fronteira da computação gira cada vez mais, em torno da redução da distância entre memória e processamento. Pesquisas em computação na memória (*in-memory computing*), arquiteturas de processamento em memória e sistemas neuromórficos refletem o esforço de aproximar a computação digital da eficiência dos sistemas biológicos.

[4] modelo clássico de computadores no qual processamento (CPU) e memória ficam separados e se comunicam continuamente por um barramento. https://www.geeksforgeeks.org/computer-organization-architecture/computer-organization-von-neumann-architecture/

Isso levanta uma questão estratégica fundamental: quem está mais bem posicionado para liderar essa convergência?

Serão as empresas dominantes de plataforma, como a NVIDIA, capazes de estender seu controle sobre toda a camada tecnológica (*stack*)? Ou as empresas que já dominam a tecnologia de memória avançada se aproximarão do centro da arquitetura de computação?

A resposta ainda é incerta. Mas está cada vez mais claro que a memória está migrando da periferia para o núcleo do design de sistemas. Até que surja um paradigma de computação radicalmente novo, a necessidade de encurtar a distância entre memória e processamento continuará sendo o fator determinante para a evolução do desempenho e da eficiência energética. Essa dinâmica, por si só, pode sustentar um ciclo muito mais longo e estruturalmente diferente dos observados no passado.

Se a convergência entre memória e computação for, de fato, o próximo grande salto de paradigma, os fabricantes de memória poderão ocupar uma posição ainda mais central na cadeia de valor do que ocupam hoje.

Primeiras evidências de que a tese está funcionando

Ao longo do último ano, o desempenho financeiro da indústria global de memória apresentou uma melhora expressiva, oferecendo as primeiras evidências concretas de que as forças estruturais descritas nesta carta já começaram a se materializar. Receitas, margens brutas, lucro operacional e geração de fluxo de caixa livre cresceram fortemente entre os principais nomes do setor, com a *High Bandwidth Memory* (HBM) emergindo como o motor mais visível dessa transformação.

A expansão da rentabilidade foi particularmente notável ao analisarmos os resultados trimestrais recentes e as projeções (*guidances*). A Micron reportou margens brutas próximas de 60% no último trimestre - uma recuperação impressionante frente às margens negativas registradas no fundo do ciclo em 2023, e um salto significativo em relação à sua média histórica, que gira em torno de 20%. A SK Hynix também apresentou resultados sólidos, com projeções que indicam a continuidade dessa expansão, impulsionada pela demanda sustentada por HBM e memórias voltadas para IA. Ambas as empresas registraram lucros operacionais recordes e forte geração de caixa, evidenciando a rapidez com que a indústria migrou de prejuízos severos para níveis de rentabilidade típicos de topo de ciclo.

A evolução do mix de produtos da Micron ilustra essa mudança com clareza. Historicamente, a HBM representava uma fatia pequena e altamente especializada da receita da companhia, estimada em cerca de 5% há poucos anos. Hoje, o consenso de mercado aponta que essa participação deve atingir entre um quarto e quase um terço do faturamento em 2026. A administração espera que a HBM se torne um dos principais pilares de crescimento e lucratividade, com a capacidade de produção para 2026 já totalmente comprometida. O que antes era um produto de nicho tornou-se o coração do negócio. Outra mudança importante no comportamento dos clientes reforça essa transformação estrutural: enquanto a memória era historicamente vendida no mercado *spot* ou em contratos de curto prazo, um número crescente de grandes clientes passou a buscar contratos de fornecimento de longo prazo.

É importante notar que a expansão das margens não se limitou ao aumento da fatia de HBM no mix. Os preços das memórias tradicionais também subiram consideravelmente. Ao longo de 2024 e no início de 2026, dados do setor indicam que os preços de DRAM subiram entre 30% e 50% desde as mínimas cíclicas, enquanto os preços de NAND de alta capacidade (*Enterprise SSDs*) subiram de forma ainda mais agressiva. Relatórios recentes continuam apontando para uma oferta restrita e novos reajustes em segmentos específicos.

Essa força generalizada nos preços reflete uma combinação poderosa entre oferta e demanda. Do lado da demanda, a construção da infraestrutura global de IA está impulsionando um aumento expressivo no consumo de memória em data

centers. Embora a HBM esteja no topo da hierarquia, as GPUs modernas dependem de múltiplas camadas auxiliares de memória e armazenamento. À medida que os *clusters* de IA escalam, a demanda cresce não apenas por HBM, mas também pelas tecnologias DRAM e NAND tradicionais utilizadas em toda a arquitetura do sistema.

Do lado da oferta, as restrições produtivas desempenham um papel crucial. A fabricação de memória exige instalações (*clean rooms*) altamente especializadas, extremamente caras e que demandam tempo para serem expandidas. O rápido aumento da produção de HBM consome a mesma capacidade fabril que antes era dedicada às memórias convencionais. Com a reconfiguração das fábricas e a realocação de capacidade, a oferta de DRAM tradicionais encolheu. Essa dinâmica contribuiu para a alta de preços em todo o setor e, em alguns casos, produtos mais comoditizados tiveram valorizações superiores à da própria HBM.

Esse componente da recuperação de lucros remete a ciclos passados, nos quais a alta de preços acaba incentivando o aumento da oferta, levando à normalização. Com o tempo, a oferta tende a se equilibrar nos segmentos de *commodities*. A diferença fundamental hoje é a migração acelerada para a HBM e produtos de alto valor agregado. Dependendo da velocidade da resposta da oferta, a indústria pode não voltar a sofrer a compressão abrupta de margens que marcou os ciclos anteriores por muitos anos.

Apesar da melhora robusta nos fundamentos, os múltiplos de avaliação permanecem curiosamente próximos aos níveis históricos. No final de 2024, Micron e SK Hynix negociavam entre 8x e 15x o lucro projetado (*forward P/E*). Hoje, mesmo após revisões para cima nas estimativas de lucro, continuam sendo negociadas em torno de 8x a 12x - patamares condizentes com múltiplos históricos de meio de ciclo, apesar da maior resiliência estrutural dos produtos atuais. Em suma, os lucros cresceram muito mais rápido do que os múltiplos, o que sugere que ainda não houve uma reavaliação de mercado (*re-rating*) relevante.

Essa ausência de *re-rating* indica que o mercado ainda enxerga o cenário atual sob a ótica dos ciclos passados. A narrativa predominante é a de que a oferta eventualmente alcançará a demanda, os preços cairão e o setor voltará ao velho padrão de *boom e bust*.

Acreditamos que essa visão subestima a profundidade das mudanças em curso.

É possível que o momento atual acabe se assemelhando aos ciclos anteriores; o veredito final ainda não foi dado. No entanto, a falta de uma reavaliação nos múltiplos melhora significativamente a relação risco-retorno do investimento. Se o ciclo se estender além do previsto, os analistas serão forçados a revisar suas projeções sucessivamente. Até lá, o mercado parece dar pouco crédito à possibilidade de que este ciclo seja estruturalmente diferente de tudo o que vimos antes.

Alguns comentários recentes de líderes da indústria de IA que sustentam nossa visão

Uma das validações mais contundentes da nossa tese vem justamente de quem está construindo a maior infraestrutura de IA do mundo.

Em entrevista recente a Dwarkesh Patel (5 de fevereiro de 2026)[5], Elon Musk ofereceu uma visão excepcionalmente direta sobre como os desenvolvedores de IA de fronteira enxergam o futuro da infraestrutura de computação. Musk descreve sua ambição de verticalizar toda a camada tecnológica (*stack*) de IA: do desenvolvimento de modelos ao design de chips, passando pela fabricação de semicondutores e até a produção de equipamentos de litografia. O objetivo estratégico é claro: reduzir a dependência de fornecedores críticos em toda a cadeia de valor.

[5] https://www.youtube.com/watch?v=BYXbuik3dgA&t=1s

No entanto, por volta do minuto 27 da entrevista, Musk destaca uma exceção crucial. Ao discutir quais partes do hardware seriam as mais difíceis de internalizar, ele aponta especificamente a memória como o componente mais difícil de replicar. Ele observa que, mesmo para uma organização disposta a investir dezenas de bilhões de dólares, os desafios para construir uma capacidade de fabricação competitiva seriam colossais. Ou seja, Musk está afirmando que pretende internalizar em suas empresas o que a NVIDIA, TSMC e até a ASML fazem, mas acredita que o difícil será internalizar o que a Micron, SK Hynix e Samsung fazem. Nos minutos finais, Musk cita a Micron nominalmente, explicando que a demanda da SpaceX e da xAI é tão forte que eles incentivaram a empresa a expandir sua capacidade e que poderiam, na prática, absorver toda a produção da Micron. Poucas declarações ilustram tão bem o quão central a memória avançada se tornou para a expansão da IA.

Essa perspectiva é corroborada por Dylan Patel, fundador e analista-chefe da *SemiAnalysis*. Em sua participação no *The MAD Podcast* com Matt Turck (fevereiro de 2026)[6], Patel explorou o que chama de "tokenomics" emergente da IA. Ele explica que, em modelos modernos baseados em agentes - como o Claude ou ferramentas de programação -, o custo dominante não está mais na fase de *decode* (a geração de tokens), mas no estágio de *pre-fill* do cache *key-value*. Como os agentes alternam constantemente entre contextos massivos, como grandes repositórios de código ou bases de conhecimento corporativas, volumes enormes de dados precisam ser carregados na memória antes mesmo que a geração comece. Nesse novo paradigma, o custo da inteligência passa a ser o custo de mover e armazenar contexto. A largura de banda de memória - e especificamente a HBM - torna-se, portanto, o principal determinante do custo por token.

Essa visão é reforçada de forma consistente pelos líderes das maiores *Big Techs*. Durante sua visita a Taiwan no início de 2026, Jensen Huang descreveu o mercado global de armazenamento e memória como "completamente subatendido hoje", prevendo que este se tornará o maior segmento do setor à medida que evolui para sustentar a "memória de trabalho das IAs do mundo". Ele enfatizou que a futura arquitetura Rubin e a ascensão dos agentes de IA exigirão níveis sem precedentes de memória de contexto, tornando-a tão estrategicamente vital quanto o próprio processamento.

Satya Nadella expressou uma visão semelhante em conferências recentes, descrevendo a evolução da infraestrutura do Azure em direção ao que chamou de "potência de dados síncronos distribuídos" (*distributed synchronous data power*). Para ele, a viabilidade econômica da IA em escala - como o Copilot - depende fundamentalmente da eficiência da HBM3E e da capacidade de entregar largura de banda massiva em larga escala. Nadella observou que cada camada da estrutura tecnológica está sendo moldada pela densidade de memória necessária para o raciocínio dos agentes (*agentic reasoning*).

Sundar Pichai fez comentários comparáveis ao discutir a infraestrutura dos modelos Gemini e a próxima geração de TPUs. Ele tem destacado o "TPU Ironwood" (7ª geração) e sua forte dependência de parcerias de HBM (com SK Hynix e Samsung) para garantir que as janelas de contexto multimodal do Gemini - que já ultrapassam 2 milhões de tokens - mantenham o desempenho.

Em conjunto, esses depoimentos formam uma narrativa extremamente coerente: para os construtores dos maiores sistemas de IA do planeta, a memória deixou de ser um componente de suporte para se tornar a restrição central do sistema. Para o investidor, essa mudança pode representar uma das transformações mais profundas - e ainda pouco compreendidas - do cenário tecnológico atual.

A evolução do *profit pool* da IA

Compreender a importância estratégica da *High Bandwidth Memory* (HBM) é essencial para entender como o valor econômico está distribuído hoje no ecossistema de IA - e como essa distribuição deve evoluir.

[6] https://www.youtube.com/watch?v=DqBMzuzxZog&t=1s

Atualmente, a NVIDIA captura uma parcela extraordinária do valor gerado pela infraestrutura de IA. Seu domínio baseia-se em uma rara convergência de vantagens: liderança em arquitetura de GPUs, controle do ecossistema de software CUDA e a capacidade de projetar sistemas completos que integram processamento, rede, interconexões e memória em uma única plataforma otimizada. Essa combinação permitiu que a NVIDIA deixasse de ser apenas uma fabricante de chips para se tornar uma empresa de sistemas que, na prática, define o padrão de construção dos *clusters* modernos de IA.

As consequências econômicas dessa posição são evidentes. A NVIDIA tornou-se a empresa mais valiosa do mundo; suas receitas cresceram cerca de oito vezes em apenas três anos e suas margens brutas atingiram aproximadamente 75% - patamar historicamente restrito às empresas de software mais lucrativas. Isso reflete seu papel como principal orquestradora de toda a camada de hardware de IA. Na arquitetura atual, a NVIDIA especifica os requisitos de memória, dita o design dos sistemas e captura a maior parte dos ganhos econômicos gerados pela expansão da infraestrutura.

Dentro desse paradigma, os fabricantes de memória ocuparam, historicamente, um papel de dependência. A HBM é desenvolvida em estreita colaboração com a NVIDIA e passa por um longo e rigoroso processo de qualificação antes de ser integrada às suas plataformas. O resultado é uma cadeia de valor na qual os fornecedores de memória entregam um componente crítico, mas permanecem economicamente subordinados ao dono da plataforma.

No entanto, essa estrutura começa a mudar.

Os maiores compradores de infraestrutura de IA - os *hyperscalers* e laboratórios de IA de fronteira - estão cada vez mais desconfortáveis com o nível de concentração e dependência no mercado de GPUs. Nos últimos dois anos, praticamente todos os grandes provedores de nuvem e laboratórios aceleraram investimentos em chips proprietários: o Google com as TPUs, a Amazon com o Trainium e o Inferentia, a Microsoft com o Maia, a Meta com o MTIA e novos entrantes, como a xAI, desenvolvendo suas próprias iniciativas de hardware. O objetivo estratégico não é necessariamente eliminar a NVIDIA, mas diversificar o fornecimento e recuperar poder de negociação.

Esse movimento ganhou um catalisador concreto em 2025, quando a AMD lançou o ROCm 7 - a nova geração de sua plataforma de software para GPUs, concorrente direta do CUDA. A atualização trouxe ganhos expressivos de desempenho e, crucialmente, redesenhou a camada de portabilidade para tornar mais simples migrar código da NVIDIA para GPUs AMD. À medida que a barreira de software que canalizava a demanda por HBM através da NVIDIA se reduz, abre-se um segundo grande canal de consumo de memória avançada - justamente num momento em que a oferta já é insuficiente.

Essa mudança tem implicações profundas para a indústria de memória. Na arquitetura atual, a NVIDIA atua como a principal porta de entrada (*gatekeeper*) para a demanda de HBM. À medida que surgem plataformas alternativas de computação, o número de compradores de memória avançada se expande significativamente. Em vez de atender a uma única plataforma dominante, os fabricantes de memória passam a fornecer simultaneamente para múltiplos ecossistemas concorrentes.

Em termos econômicos, a base de clientes de memória avançada está se ampliando, enquanto a oferta permanece concentrada em um pequeno número de produtores. Essa dinâmica tem o potencial de reequilibrar o poder de barganha ao longo da cadeia de valor.

Ao mesmo tempo, a crescente complexidade do design dos sistemas de IA está empurrando a memória para o centro das decisões de arquitetura. Como discutimos anteriormente, o desempenho e o custo dos sistemas de IA são cada vez mais limitados pela largura de banda de memória, pela eficiência energética e pela movimentação de dados. Essas restrições não são facilmente resolvidas por software; elas exigem um desenvolvimento conjunto (*co-design*) profundo entre processamento, encapsulamento e memória.

Isso cria uma mudança sutil, mas fundamental, no papel dos fabricantes de memória: de meros fornecedores de componentes, eles passam a ser parceiros estratégicos no design dos sistemas. Quanto mais a memória se aproxima do

núcleo da arquitetura, maior é a influência de seus fornecedores sobre o desempenho, a eficiência e, em última instância, o custo por token.

O *profit pool* da IA ainda está altamente concentrado, mas as forças estruturais sugerem que sua distribuição deve se ampliar gradualmente. A memória está deixando a periferia da camada tecnológica para ocupar o centro - saindo de uma posição de fornecimento para uma parceria estratégica vital no ecossistema de hardware de IA.

A inteligência artificial como a maior transformação tecnológica da história

A inteligência artificial tem o potencial de remodelar setores cujo peso econômico supera, de longe, o próprio setor de tecnologia. A indústria global de software, sozinha, aproxima-se da marca de US$ 1 trilhão, com gastos projetados para ultrapassar US$ 1,2 trilhão este ano. A economia global de serviços é ainda maior: serviços profissionais e comerciais já movimentam mais de US$ 6 trilhões anuais, enquanto o mercado global de serviços como um todo deve chegar perto dos US$ 24 trilhões até 2029. A inteligência artificial situa-se exatamente na intersecção desses dois mundos. Ela é, simultaneamente, um novo paradigma de software e uma tecnologia de automação de propósito geral, capaz de transformar tanto as ferramentas que as organizações utilizam quanto os serviços que elas entregam.

Diferente de revoluções tecnológicas anteriores, a curva de adoção da IA avança em uma velocidade sem precedentes. O ChatGPT atingiu 100 milhões de usuários ativos mensais apenas dois meses após o lançamento - a adoção mais rápida de um aplicativo de consumo na história. Para efeito de comparação, o Instagram levou mais de dois anos para alcançar esse marco, e o TikTok, nove meses. Em menos de dois anos, o ChatGPT já era utilizado semanalmente por cerca de 10% da população dos Estados Unidos - número que saltou para aproximadamente 30% poucos meses depois. Hoje, a plataforma atende centenas de milhões de usuários globais semanalmente. Nenhuma tecnologia de propósito geral anterior alcançou escala de massa com tal rapidez.

A internet exigiu décadas de investimento em infraestrutura antes que sua adoção plena fosse possível. Redes de fibra ótica precisaram ser instaladas, torres de telefonia construídas e bilhões de dispositivos fabricados. Já a inteligência artificial está sendo implantada sobre uma base já existente: plataformas globais de nuvem, data centers em hiperescala e bilhões de aparelhos conectados. Hoje, mais de 70% da população mundial usa celulares e há mais de sete bilhões de smartphones em circulação. A IA não precisa de uma nova infraestrutura para escalar - ela pode alavancar imediatamente os investimentos realizados nas últimas duas décadas.

Para entender por que o momento atual é crucial, é útil contrastar o software tradicional com os sistemas modernos de IA. O software convencional é determinístico: a mesma entrada produz sempre a mesma saída. Essa previsibilidade historicamente tornou o software confiável para aplicações críticas. Já os sistemas de IA são probabilísticos. Seus resultados baseiam-se em probabilidades estatísticas aprendidas a partir de dados, o que significa que as respostas podem variar e são inerentemente sujeitas a incertezas. As primeiras gerações de modelos de linguagem de larga escala (LLMs) eram, portanto, limitadas pelas "alucinações" - respostas confiantes, porém incorretas -, o que restringia o uso corporativo a experimentos e protótipos.

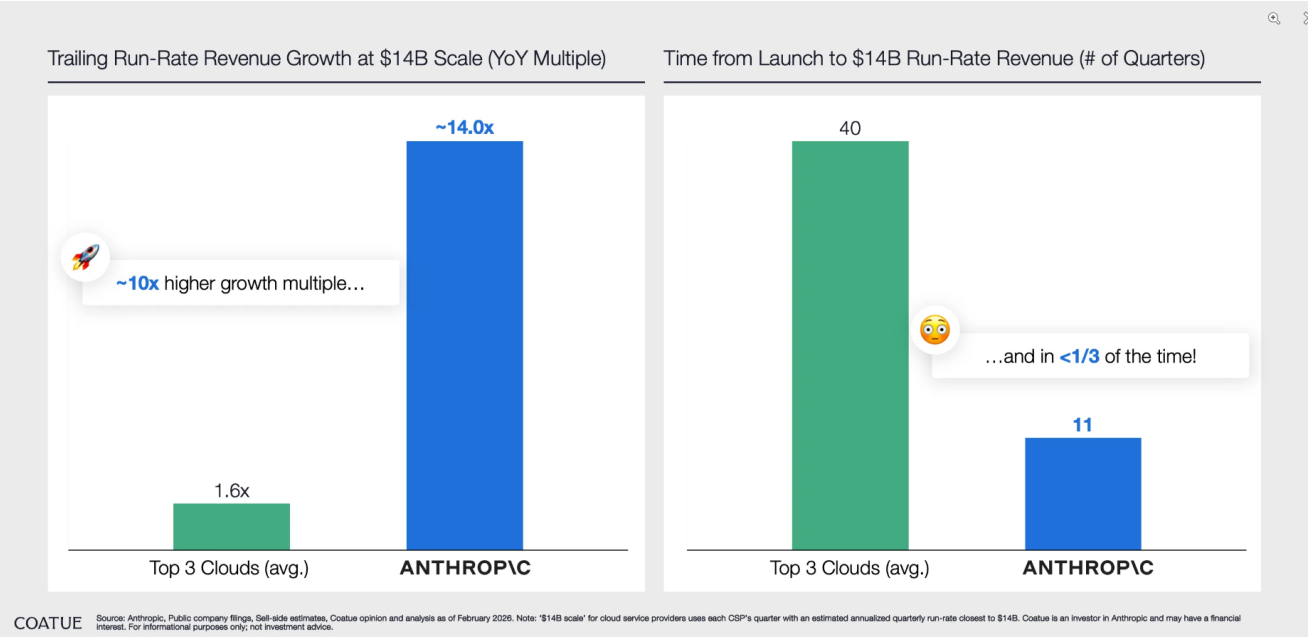
Nos últimos dois anos, essa limitação diminuiu rapidamente. Gerações sucessivas de modelos alcançaram avanços substanciais em raciocínio, programação, uso de ferramentas e redução de erros. Ao mesmo tempo, o surgimento dos agentes de IA - sistemas capazes de orquestrar múltiplas chamadas de modelos, ferramentas e fluxos de trabalho - começou a transformar a inteligência probabilística em processos confiáveis e repetíveis, adequados para aplicações reais.

O desenvolvimento de software foi o primeiro domínio a sofrer disrupção em larga escala. Ferramentas de programação assistida por IA, como Cursor e Claude Code, são descritas por desenvolvedores experientes como um ponto de inflexão em produtividade. Muitos engenheiros relatam que escrever código de produção sem assistência de IA tornou-se uma

raridade. Os ganhos de produtividade são frequentemente descritos como melhorias de ordens de magnitude, reduzindo drasticamente os prazos de desenvolvimento.

O impacto comercial veio logo em seguida. Empresas nativas de IA atingiram receitas anuais bilionárias em uma velocidade raramente vista na história da tecnologia. O crescimento acelerado da OpenAI, da Anthropic e de uma nova geração de startups "AI-first" ilustra a rapidez com que a criação de valor está se acelerando.

A velocidade dessa criação de valor é sem precedentes. Segundo análise da Coatue (fevereiro de 2026)[7], a Anthropic atingiu um *run-rate* de receita de US$ 14 bilhões em apenas 11 trimestres desde o lançamento - menos de um terço do tempo que os três maiores provedores de nuvem levaram para alcançar a mesma escala. Mais impressionante ainda: nesse patamar, a Anthropic cresce a um múltiplo de ~14x ano contra ano, quase 10 vezes o ritmo médio dos grandes provedores de nuvem na mesma escala de receita.



Acreditamos que o desenvolvimento de software representa apenas a primeira onda. A próxima fronteira é a economia global de serviços. À medida que os agentes de IA se tornam mais confiáveis e capazes de operar em fluxos de trabalho corporativos, a gama de tarefas automatizáveis se expande de forma exponencial. Essa transição deve impulsionar um forte aumento nas cargas de inferência, à medida que as organizações implantarem agentes em áreas como atendimento ao cliente, operações, finanças e pesquisa.

O início de 2026 parece marcar o ponto de inflexão em que essa transição deixa a fase de experimentação para entrar na adoção em larga escala.

Não se trata de uma bolha como a da internet

Qualquer discussão sobre inteligência artificial leva inevitavelmente à mesma pergunta: estamos diante de uma nova bolha? A comparação com o *boom* da internet no final dos anos 1990 é compreensível. Aquele período deixou marcas profundas nos investidores, e a escala dos investimentos atuais em IA naturalmente convida a paralelos. É importante

[7] Chart of the Day

começar com um aviso explícito: excessos quase certamente ocorrerão. Em certos momentos, as expectativas podem ultrapassar a realidade, o capital pode ser mal alocado e bolsões especulativos podem surgir. Dinâmicas como essas são comuns nas fases iniciais de qualquer tecnologia transformadora.

Ainda assim, apesar das semelhanças superficiais, as condições que normalmente dão origem a grandes bolhas de ativos não nos parecem estar presentes hoje.

Paradoxalmente, um dos sinais mais fortes de que ainda não estamos em uma bolha é a presença constante de alertas sobre o tema. Bolhas tendem a se formar quando o ceticismo desaparece e dá lugar à certeza; elas florescem quando os investidores param de questionar o otimismo das projeções e passam a assumir que resultados positivos são inevitáveis. O momento em que motoristas de táxi dão dicas de ações ou quando conversas casuais giram em torno de "vencedores garantidos" costuma estar mais próximo do pico do excesso especulativo. Hoje, a narrativa dominante permanece cautelosa. As manchetes frequentemente alertam para o superaquecimento e para investimentos irreais. No mercado financeiro, esse ceticismo generalizado atua como uma força estabilizadora: quando os investidores temem uma bolha, os *valuations* tendem a permanecer ancorados.

O ambiente de informação também mudou profundamente desde os anos 1990. A expansão das redes sociais acelerou drasticamente o alcance da informação. No passado, narrativas especulativas se espalhavam de forma lenta e desigual. Hoje, qualquer tese - seja otimista ou crítica - circula globalmente em poucas horas. Isso reduz a probabilidade de bolhas duradouras se formarem simultaneamente em grandes classes de ativos. Em vez disso, excessos têm surgido em comunidades menores e mais concentradas.

A história recente dos criptoativos e dos NFTs ilustra bem essa dinâmica. O *Bored Ape Yacht Club* tornou-se um fenômeno cultural quando celebridades pagaram somas extraordinárias por ilustrações digitais. Neymar teria comprado seu *Bored Ape* por cerca de US$ 1,1 milhão; Justin Bieber, por US$ 1,3 milhão. No auge, a narrativa de escassez digital permitiu racionalizar avaliações desconectadas de qualquer utilidade real. Até o meio acadêmico abraçou o fenômeno por um breve momento; em 2023, um estudo de caso da *Harvard Business School* apresentou o projeto como um exemplo legítimo de marketing comunitário. Meses depois, o entusiasmo desapareceu e os preços desabaram. O episódio demonstra como bolhas podem surgir rapidamente em nichos, enquanto o mercado amplo permanece cético.

Ao comparar o ciclo atual de IA com o *boom* da internet, três diferenças se destacam:

1. *Valuation* No auge da bolha da internet, as empresas mais celebradas eram negociadas a múltiplos extraordinários. A Cisco Systems, considerada a espinha dorsal da rede, atingiu um índice Preço/Lucro (P/L) superior a 100x. Hoje, as principais empresas de infraestrutura de IA negociam a múltiplos muito mais contidos. A NVIDIA, apesar do crescimento fenomenal, negocia em torno de 25x o lucro projetado (*forward P/L*), muito abaixo dos três dígitos observados em 2000. Microsoft, Alphabet e Meta negociam entre 22x e 32x o lucro projetado — patamares mais próximos de líderes tecnológicos maduros do que de startups especulativas.

Valuations: Bolha da Internet (2000) vs. Era da IA (2026)

Forward P/E Ratio — Comparativo de múltiplos no pico da bolha dot-com vs. avaliações atuais das principais empresas de tecnologia

	Empresa	Ticker	Fwd P/E Pico 2000	Papel no Boom da Internet	Fwd P/E Atual (2026)	Papel no Boom da IA
1	**Cisco Systems**	CSCO	~131x	Infraestrutura de rede / "espinha dorsal" da internet	~18x	Networking / infraestrutura legada
2	**Microsoft**	MSFT	~55x	Software / sistemas operacionais / Office	~24x	Cloud (Azure) / Copilot / infraestrutura de IA
4	**Qualcomm**	QCOM	~167x	Infraestrutura de telecom / wireless	~13x	Chips mobile / modems / edge AI
5	**Oracle**	ORCL	~153x	Bancos de dados / software corporativo	~20x	Cloud infrastructure / bancos de dados para IA
6	**Amazon**	AMZN	N/M[1]	E-commerce (prejuízo operacional em 2000)	~25x	AWS / cloud computing / infraestrutura de IA
7	**Broadcom**	AVGO	~35x[2]	Semicondutores (empresa menor na época)	~33x	Chips de networking / ASICs customizados para IA
8	**Micron**	MU	~18x	Commodity DRAM / memória para PCs	~12x	HBM / memória avançada para GPUs de IA
9	**SK Hynix**	000660.KS	~15x[3]	Commodity DRAM (Hyundai Electronics)	~8x	HBM / memória avançada para GPUs de IA
10	**NVIDIA**	NVDA	—	Empresa de GPUs para gaming (pequena)	~24x	GPUs para IA / plataforma de computação dominante
11	**Alphabet (Google)**	GOOGL	—	Fundada em 1998, IPO em 2004	~27x	Cloud (GCP) / TPUs / modelos Gemini
12	**Meta (Facebook)**	META	—	Não existia (fundada em 2004)	~22x	Infraestrutura de IA / Llama / Metaverso
13	**Apple**	AAPL	~25x	Quase falida / iMac (participação mínima)	~31x	Dispositivos / Apple Intelligence / edge AI
14	**Tesla**	TSLA	—	Não existia (fundada em 2003)	~130x	Veículos autônomos / robótica / energia
	Mediana (empresas com dados em ambas)		~55x		~22x	
	NASDAQ-100 (índice)		~60x		~27x	

Notas:

[1] N/M = Não Mensurável. Amazon operava com prejuízo em 2000, tornando o P/E indefinido.

[2] Broadcom existia como uma empresa muito menor em 2000 (pré-fusão com Avago).

[3] SK Hynix operava como Hyundai Electronics em 2000.

Fontes: Forward P/E 2000 baseado em dados compilados do Wall Street Journal (Siegel, Mar/2000), MacroTrends, e análises históricas.
Forward P/E atual baseado em estimativas de consenso de analistas (Yahoo Finance, FactSet, Bloomberg) em fevereiro de 2026.

Observação: Os múltiplos aproximados ("~") refletem estimativas de consenso e podem variar conforme a fonte.
Para Micron e SK Hynix, os forward P/E atuais refletem estimativas de lucro para os próximos 12 meses, incorporando o crescimento de HBM.

2. Origem do Financiamento O *boom* da internet foi impulsionado por capital especulativo e empresas financiadas por dívida e emissões de ações. O ciclo atual é liderado pelas empresas mais ricas e lucrativas do mundo. Elas geram fluxos de caixa livre massivos e possuem uma visibilidade privilegiada sobre a demanda global. Sua escala oferece uma vantagem

informacional que investidores externos simplesmente não possuem. Por meio de suas plataformas, serviços de nuvem e relações corporativas, essas empresas interagem com praticamente todos

os setores da economia. Suas decisões de investimento refletem, portanto, uma visão excepcionalmente ampla de como as ferramentas de IA já estão transformando a demanda.

Os anúncios recentes de Capex (investimentos em bens de capital) ilustram o tamanho desse compromisso para 2026:

- **Amazon:** ~US$ 200 bilhões.
- **Alphabet:** US$ 175–185 bilhões.
- **Meta:** US$ 115–135 bilhões.
- **Microsoft:** US$ 100–120 bilhões.

Em conjunto, essas quatro empresas devem investir até **US$ 700 bilhões em 2026**, um aumento de mais de 70% em relação a 2025. É o maior ciclo de investimento privado em infraestrutura da história moderna. Cerca de 75% desse valor é destinado diretamente à IA, incluindo GPUs, memória avançada e sistemas de energia.

3. Natureza Defensiva do Investimento Talvez a diferença mais importante: na era da internet, buscava-se criar novas receitas; hoje, o investimento é tanto ofensivo quanto defensivo. Gigantes de nuvem e software precisam integrar IA para proteger seus mercados atuais. O risco de não fazê-lo é existencial. Um provedor de nuvem que ignore a IA pode perder sua base de clientes em poucos anos. Portanto, o gasto não é apenas uma aposta no futuro, mas um passo necessário para defender as receitas bilionárias que essas empresas já possuem.

Alguns estudos de instituições como Bain e Goldman Sachs sugerem que o investimento pode exceder o retorno de curto prazo. No entanto, essas análises geralmente ignoram essa dimensão defensiva. Se considerarmos que as receitas de *cloud computing* e software estão em jogo, o ciclo de investimento parece muito mais estratégico do que especulativo.

Nada disso garante que não haverá excessos. Mas a combinação de múltiplos moderados, alocação de capital informada e incentivos defensivos sugere que a fase atual da IA difere fundamentalmente das dinâmicas do final dos anos 1990.

O que pode dar errado

Nenhuma tese de investimento que dependa de transformação tecnológica pode ignorar a incerteza. A história da computação está repleta de exemplos de arquiteturas dominantes que acabaram dando lugar a novos paradigmas. Embora acreditemos que as forças estruturais que sustentam a demanda por memória avançada sejam poderosas, é essencial examinar os riscos que poderiam desafiar essa visão.

Vemos três categorias amplas de risco: a disrupção tecnológica, o ritmo de adoção da IA e a possibilidade de que o progresso dos modelos desacelere materialmente.

1. Disrupção tecnológica

O primeiro e mais óbvio risco é que novas arquiteturas de software ou hardware reduzam drasticamente a necessidade de Memória de Alta Largura de Banda (HBM). O paradigma atual da IA é construído em torno de modelos *transformer*, que

exigem enormes janelas de contexto e a movimentação repetida de grandes volumes de dados entre a memória e o processamento. No entanto, a fronteira da pesquisa está explorando alternativas ativamente.

No lado do software, um número crescente de equipes de pesquisa e startups está trabalhando em arquiteturas projetadas para ir além dos *transformers*. Um dos exemplos mais proeminentes é a Liquid (https://www.liquid.ai/), que está desenvolvendo arquiteturas neurais recorrentes inspiradas em sistemas biológicos. Esses modelos visam aprender continuamente a partir da inferência, reduzindo a necessidade de recarregar janelas de contexto massivas para cada consulta. Resultados iniciais sugerem que modelos recorrentes relativamente pequenos podem alcançar um desempenho comparável ao de modelos *transformer* muito maiores. Se tais abordagens provarem ser bem-sucedidas em escala, elas poderiam reduzir significativamente os requisitos de memória por carga de trabalho.

Essa direção de pesquisa não é especulativa; já está em andamento na academia e na indústria. Contudo, seu impacto na indústria de memória dependeria do equilíbrio entre ganhos de eficiência e o crescimento da adoção. Mesmo melhorias substanciais na eficiência dos modelos precisariam superar o crescimento extraordinário na adoção da IA para reduzir materialmente a demanda por memória. Dado o quão cedo ainda estamos na curva de adoção, este cenário parece possível, mas improvável no curto prazo.

No lado do hardware, a pesquisa foca cada vez mais em colapsar a distância entre a memória e o processamento - a principal ineficiência da arquitetura de von Neumann. Diversas iniciativas exploram designs de "computação na memória" (*compute-in-memory*) e até o renascimento de abordagens de computação analógica. Um exemplo é a startup recentemente financiada Unconventional AI[8] (https://unconv.ai/), que captou aproximadamente US$ 475 milhões com uma avaliação de US$ 4,5 bilhões de empresas líderes de capital de risco, incluindo a Sequoia Capital. A empresa é liderada por um experiente executivo de semicondutores e busca arquiteturas de computação radicalmente diferentes, inspiradas na eficiência energética do cérebro humano. No entanto, a empresa emprega apenas algumas dezenas de pessoas e deve enfrentar uma longa jornada de pesquisa e desenvolvimento. Levar paradigmas de computação fundamentalmente novos à viabilidade comercial provavelmente exigirá uma década ou mais.

Esses desenvolvimentos destacam um ponto importante: a trajetória de longo prazo da computação quase certamente envolverá uma integração mais estreita entre memória e processamento. Paradoxalmente, isso pode aumentar a importância estratégica das empresas que já dominam as tecnologias de memória avançada.

2. O ritmo de adoção da IA

Um segundo risco diz respeito à velocidade com que a inteligência artificial é adotada em toda a economia global. O potencial de longo prazo para a IA permear softwares e serviços parece substancial, mas a adoção raramente segue uma trajetória linear. Se a implementação corporativa progredir mais lentamente do que o esperado, a oferta e a demanda por memória podem divergir temporariamente, criando volatilidade nos preços e nas margens. Isso se assemelharia aos ciclos passados de memória, mesmo que a trajetória de longo prazo permaneça intacta.

3. O risco que o progresso dos modelos desacelere

Este é, em nossa visão, o risco mais estrutural para a tese.

Uma desaceleração na adoção de IA (risco 2) poderia gerar ciclos temporários de excesso de oferta e volatilidade de preços. Já uma estagnação real da fronteira tecnológica teria implicações mais profundas: limitaria o grau de penetração da IA na economia e reduziria a necessidade de expansão contínua da infraestrutura de inferência - o verdadeiro motor da demanda estrutural por memória avançada.

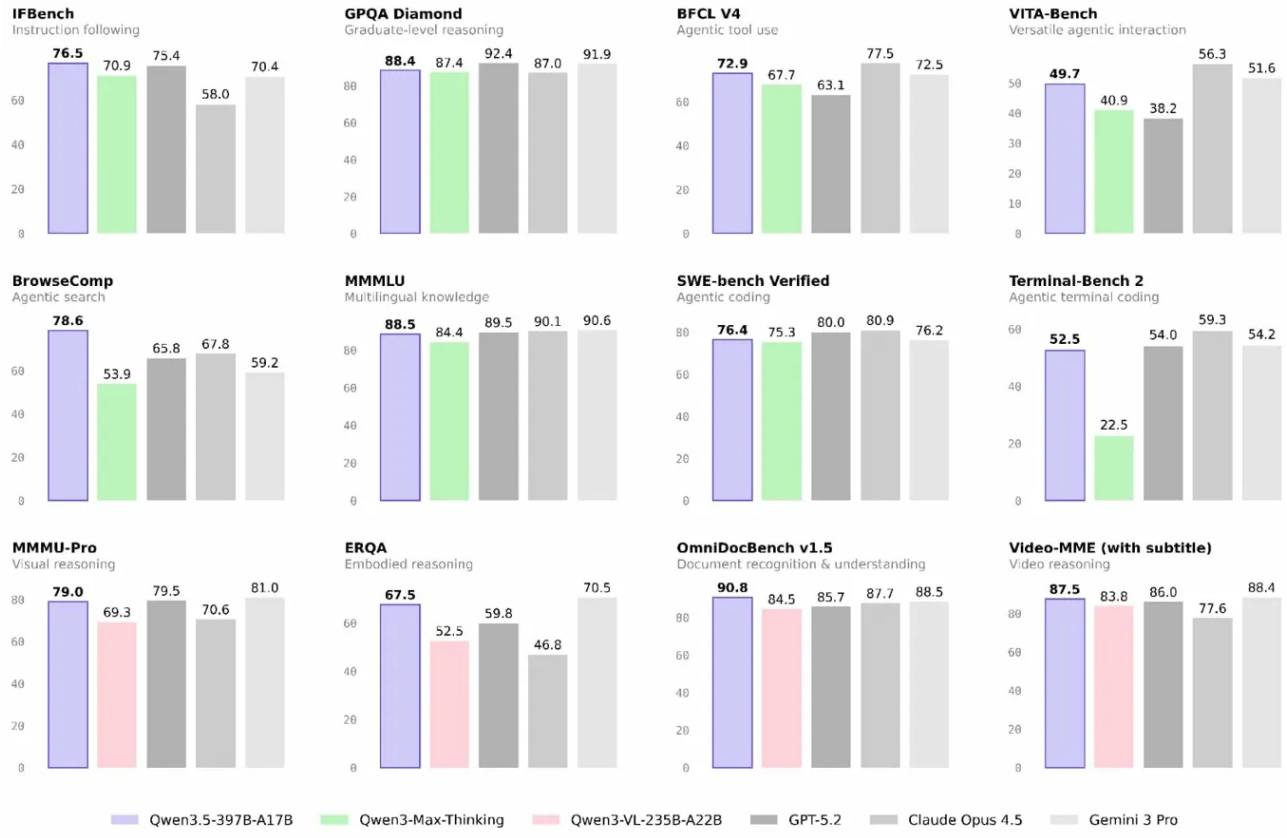
[8] https://www.youtube.com/watch?v=AgqdyrqpvDc

A boa notícia é que, até aqui, a evidência aponta na direção oposta.

Nos últimos dois anos, a fronteira de modelos continuou avançando em múltiplos eixos simultaneamente: raciocínio, coding, agentes, multimodalidade e eficiência de inferência. Esse progresso não está mais restrito a poucos laboratórios americanos; ele se tornou global e cada vez mais distribuído entre iniciativas open source e novos polos de pesquisa, particularmente na China.

O ritmo de evolução recente dos modelos chineses ilustra bem essa dinâmica. O lançamento do **GLM-5** pela Zhipu[9], no início deste mês, destaca ganhos relevantes em tarefas de coding e workflows agentic - exatamente os workloads que impulsionam o crescimento da inferência. A própria cadência de releases da empresa é reveladora: versões sucessivas vêm sendo lançadas em intervalos curtos, com melhorias consistentes de performance. Em sua documentação técnica, a Zhipu reportou que o modelo anterior, GLM-4.7, alcançou **73,8% no benchmark SWE-bench**, com ganho de **+5,8 pontos percentuais** em relação à versão precedente. Esse tipo de progresso incremental, repetido em ciclos curtos, sugere que a fronteira permanece dinâmica e competitiva.

Um exemplo ainda mais recente surgiu durante o Carnaval: em 16 de fevereiro, a Alibaba lançou o Qwen 3.5, um modelo *open source* com 397 bilhões de parâmetros que ativa apenas 17 bilhões por token. Nos benchmarks divulgados[10], o modelo iguala ou supera GPT-5.2, Claude Opus 4.5 e Gemini 3 Pro em cerca de 80% das categorias avaliadas, a um custo 60% menor. Um modelo aberto chinês competindo de igual para igual com os melhores modelos proprietários americanos reforça que o progresso segue acelerado, globalmente distribuído, e com múltiplos motores independentes de demanda por inferência.



[9] https://z.ai/blog/glm-5
[10] https://qwen.ai/blog?id=qwen3.5

Em paralelo, o ecossistema open source continua evoluindo rapidamente. Projetos como DeepSeek e outros modelos baseados em arquiteturas Mixture-of-Experts vêm introduzindo inovações voltadas à eficiência de inferência, como ativação parcial de parâmetros por token e novas formas de atenção latente. O significado econômico desse movimento é profundo: quanto mais distribuído e aberto se torna o progresso tecnológico, menor a probabilidade de uma estagnação abrupta da fronteira.

Ainda assim, o risco permanece real. O progresso dos modelos pode desacelerar por diversos motivos: limites na disponibilidade de dados de alta qualidade, retornos decrescentes de escala, custos de treinamento crescentes ou restrições regulatórias. Caso isso ocorra, a trajetória de adoção da IA poderia se tornar mais gradual. Em um cenário extremo, a IA deixaria de ser percebida como uma tecnologia generalista capaz de transformar amplamente a economia e passaria a ser vista como um conjunto de ferramentas úteis, porém com escopo mais limitado.

O impacto sobre a indústria de memória seria direto. A demanda estrutural por HBM depende da expansão contínua da inferência em larga escala. Se os modelos pararem de melhorar, a adoção desacelera. Se a adoção desacelera, o crescimento da inferência diminui. E, nesse cenário, o ciclo de expansão de capacidade poderia se antecipar à demanda, recriando dinâmicas de oferta e preço mais próximas dos ciclos históricos da indústria.

Por ora, entretanto, os sinais empíricos continuam apontando para progresso acelerado. O risco de estagnação deve ser monitorado como um possível *regime change*. Mas, à luz dos avanços recentes, ele permanece um risco potencial - não um cenário base.

Conclusão

Toda grande transição tecnológica cria momentos em que o mercado tem dificuldade para interpretar o que é uma mudança real e o que é meramente cíclico. Esses momentos raramente são confortáveis. Eles costumam parecer confusos, contraditórios e incertos. Mas são precisamente nesses momentos que surgem as oportunidades de longo prazo.

A indústria global de memória passou décadas sendo definida pela volatilidade, pela comoditização e por ciclos de expansão e retração (*boom-and-bust*). Esse histórico ainda molda a forma como o mercado percebe essas empresas hoje. No entanto, a história pode se tornar uma âncora poderosa que impede os investidores de reconhecerem quando os fundamentos econômicos de um setor começam a mudar.

Acreditamos que a indústria de memória está vivendo agora uma transição desse tipo.

A inteligência artificial não é simplesmente mais um ciclo tecnológico sobreposto aos motores de demanda existentes. Ela representa a reconstrução da infraestrutura de computação mundial. Essa reconstrução está sendo liderada não por novatos especulativos, mas pelas empresas mais lucrativas e tecnologicamente sofisticadas já criadas. A escala do investimento em curso é sem precedentes, o ritmo de adoção é extraordinário e o gargalo central deste novo paradigma de computação é cada vez mais claro: a movimentação de dados.

Pela primeira vez em décadas, a velocidade da memória tornou-se um determinante de primeira ordem para o desempenho, a eficiência energética e o custo. A Memória de Alta Largura de Banda (HBM) deixou de ser um produto de nicho para se tornar um componente fundamental da computação moderna. A demanda está acelerando, as barreiras de entrada estão subindo e o mix de produtos do setor está migrando para tecnologias que são muito mais difíceis de replicar.

E ainda assim, apesar dessa transformação, os múltiplos de *valuation* ainda refletem a narrativa do passado.

O mercado continua a ver a memória através das lentes dos ciclos históricos. Supõe-se que a oferta acabará alcançando a demanda, os preços se normalizarão e a indústria reverterá para padrões familiares. Isso pode, em última análise, provar-se correto. A mudança tecnológica nunca é linear e a incerteza permanece. Mas a ausência de uma reavaliação significativa (*re-rating*) sugere que os investidores ainda estão atribuindo uma baixa probabilidade à possibilidade de que este ciclo seja estruturalmente diferente.

Essa assimetria está no cerne do nosso investimento.

Se a indústria eventualmente reverter às suas dinâmicas históricas, as avaliações atuais já embutem grande parte desse risco. Se, no entanto, as forças estruturais descritas nesta carta continuarem a se desenrolar, as implicações de longo prazo para a lucratividade, intensidade de capital e posicionamento competitivo poderão ser profundas.

Esta não é uma aposta em um único trimestre, em um único ciclo de produto ou em um único marco tecnológico. É um investimento de longo prazo na infraestrutura necessária para alimentar uma das transformações tecnológicas mais significativas da história.

Identificamos esta oportunidade pela primeira vez no segundo semestre de 2024. Desde então, as primeiras evidências começaram a surgir, mas acreditamos que o setor permanece nos estágios iniciais de uma transição de vários anos. Os próximos anos provavelmente trarão volatilidade, ceticismo periódico e momentos em que a velha narrativa se reafirmará. Tais momentos são inevitáveis em longos ciclos tecnológicos.

Mas quando nos distanciamos e observamos o arco mais amplo da mudança, a direção parece clara.

O mundo está entrando em uma era na qual o custo da inteligência dependerá cada vez mais do custo de movimentação de dados. E, nesse mundo, a memória não é mais uma *commodity*.

DISCLAIMERS

Este material foi desenvolvido pela Drýs Capital Ltda. (“Drýs”) com caráter meramente informativo e, portanto, não deve ser entendido como oferta, recomendação ou análise de investimento ou ativo, nem tampouco constitui uma oferta de serviço e nem venda de cotas dos fundos sob gestão.

Parte do conteúdo contido nesse documento é relacionado a ativos financeiros investidos pelos fundos sob gestão da Drýs. Ressaltamos que as projeções ou estimativas apresentadas poderão ter origem em simulações ou modelos proprietários, com o risco de divergir significativamente dos resultados reais, sendo obtidas a partir de dados estatísticos, modelos probabilísticos e metodologias proprietárias, com base em fatos e resultados financeiros obtidos de fontes públicas, ou através de relatórios e análises contratados. Como tal, eventuais estimativas ou projeções contidas nesse documento servem somente para contextualizar o processo de decisão de investimentos da Gestora e não expressam, em nenhum momento, promessa ou garantia de retorno ou resultado do portfólio ou de ativos individuais.

Ainda, apesar do cuidado na obtenção e manuseio das informações apresentadas, a Drýs não declara ou garante a integridade, confiabilidade ou exatidão das informações, as quais podem inclusive serem modificadas sem comunicação, eximindo-se de quaisquer responsabilidades por prejuízos diretos ou indiretos que venham a ocorrer pelo seu uso.

O conteúdo deste documento não pode ser copiado, reproduzido, publicado, retransmitido ou distribuído, no todo ou em parte, por qualquer meio e modo, sem a prévia e expressa autorização, por escrito, da Drýs através de seus representantes.

Para dúvidas ou esclarecimentos adicionais sobre metodologia, modelos ou de métricas e estimativas relativas às empresas ou ativos investidos contidos nesse material, entrar em contato através do e-mail: **contato@dryscapital.com.br**

Para informações adicionais e acesso ao regulamento e lâminas, acesse o site

www.dryscapital.com.br

Autorregulação ANBIMA
Gestão de Recursos