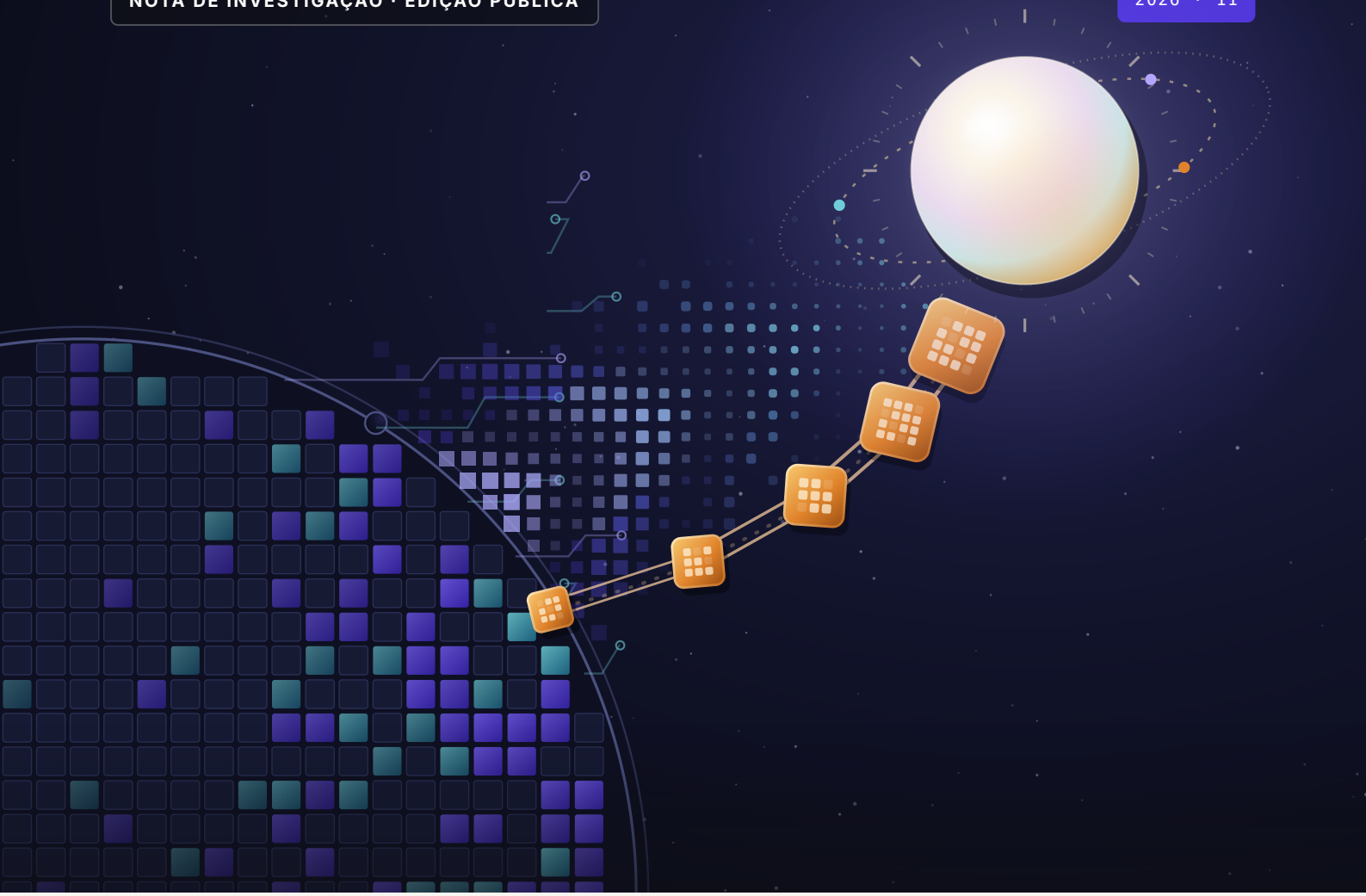
NOTA DE INVESTIGAÇÃO · EDIÇÃO PÚBLICA

2026 · 11



PROVA DE TRABALHO ÚTIL · CRIPTO × SILÍCIO

Computação como Consenso

IA × Cripto × Chips em 2026, a partir da prova de trabalho útil da Pearl

Pearl · Trabalho útil · DePIN e agentes · Economia dos ASIC · Nós de processo · Orçamentos por etapas

O VEREDICTO

A Pearl prova que a multiplicação de matrizes pode proteger uma blockchain, mas ainda não que o trabalho é útil. Em IA × cripto, a procura paga está na inferência, nos dados e na privacidade, e um ASIC só compensa quando o bolo é grande, as regras são estáveis e o chip tem um segundo mercado.

Tan Shuai

Tecnólogo e fundador · Investigador e programador de IA

tanshuai.com · i@tanshuai.com

Dados de 13 de novembro de 2026

Valores em dólares americanos

ÍNDICE

Índice



◆ **Sumário executivo** — O teorema mantém-se, o incentivo não — 04

01 **Âmbito e método** — Âmbito, método e quatro testes — 07

02 **Pearl · o mecanismo** — Pearl: a multiplicação de matrizes como consenso — 08

03 **Pearl · a economia** — Pearl: escala, emissão e o défice de utilidade — 10

04 **Pearl · as regras** — Pearl: risco de regras, aritmética numérica e governação — 12

05 **Cadeias de trabalho útil** — Cadeias de trabalho útil: nem matrizes nem conhecimento zero travam os ASIC — 13

06 **O panorama alargado** — O resto de IA × cripto: procura real mas pequena, e valor a passar para o capital — 16

07 **Economia dos ASIC** — A questão dos ASIC: uma fórmula, três medidas — 19

08 **Chip a chip** — Cinco tipos de chip, cinco respostas — 22

09 **Nós de processo** — Nós de processo: deixar que o maior custo ao longo da vida escolha — 26

10 **Orçamento por etapas** — Do modelo ao silício: um orçamento por etapas para um primeiro chip — 30

11 **Oportunidades** — Onde estão as oportunidades, e o que eu faria — 32

12 **Sinais a seguir** — O que seguir, de novembro de 2026 a 2027 — 35

13 **Perguntas a fazer** — Perguntas que não fez mas devia ter feito — 36

◆ **Sobre o autor** — Tan Shuai — 38

◆ **Fontes** — 39

FIGURAS

Figuras



FIG. 01	IA × cripto em seis setores: onde a procura paga é real	06
FIG. 02	No trabalho inteiro da Pearl, as GPU partilham um teto energético	09
FIG. 03	Os primeiros seis meses da Pearl: a taxa de hash segue o preço, as regras mudam	11
FIG. 04	O múltiplo de subsídio: quantas vezes a receita externa uma rede paga	15
FIG. 05	Porque a eficiência deixa de compensar: um ASIC retém no máximo 1 − 1/k da receita	20
FIG. 06	A maioria dos bolos é menor do que o orçamento de um só chip	21
FIG. 07	Cinco tipos de chip, cinco respostas	26
FIG. 08	Uma wafer de vanguarda custa dez vezes uma madura	27
FIG. 09	Comprar provas antes das máscaras: um orçamento por etapas para um primeiro chip	30

Etiquetas no texto

- **estimativa** cálculo ou inferência do autor a partir de dados citados
- **alegação do fornecedor** publicado pela empresa que vende o produto, sem verificação independente
- **análise** interpretação do autor sobre os dados

Esta é uma tradução da edição inglesa; em caso de divergência, prevalece a edição inglesa.

Dados de 13 de novembro de 2026 · Valores em dólares americanos

SUMÁRIO EXECUTIVO

O teorema mantém-se, o incentivo não

O VEREDICTO

A Pearl mostra que a «computação como consenso» funciona como engenharia, mas ainda não que é útil.

≈90 EH/s
rede da Pearl a 13 de novembro de 2026, em que um H é uma multiplicação-acumulação INT7

≈1%
dos blocos da Pearl transportam trabalho real de IA (Blockworks, 28 de outubro de 2026)

≈$70 M
por ano de receita externa verificável na Bittensor, contra mais de 300 milhões de dólares de emissões

175–180×
recompensas dos provadores da Aleo face às suas taxas no terceiro trimestre de 2025

R × (1 – 1/k)
o teto de lucro de um ASIC k vezes mais barato do que o minerador marginal estimativa

≈$1,5 M
para chegar ao silício medido em N7 através de um modelo, uma FPGA e uma MPW, antes de um conjunto de máscaras de ≈$3,5 M estimativa

Pearl

A 13 de novembro de 2026, a rede da Pearl funcionava a cerca de 90 EH/s. No seu código de consenso, um «hash» é uma multiplicação-acumulação inteira, pelo que a rede equivale ao débito inteiro de cerca de 122 000 H100. Contudo, apenas cerca de 1% dos blocos transporta trabalho real de IA, e a única parceria pública de inferência já não é oferecida na API serverless do seu fornecedor. A causa é económica, não criptográfica: uma GPU que minera enquanto serve inferência obtém apenas cerca de 1% a 27% da taxa de hash de um minerador dedicado, pelo que os mineradores racionais multiplicam matrizes aleatórias.

Procura paga

Em todo o domínio IA × cripto, a procura paga que se pode verificar surge em três lugares: a venda de inferência, a venda de dados e a venda de privacidade. A receita externa verificável da Bittensor é de cerca de 70 milhões de dólares por ano, cerca de quatro quintos dos quais provenientes de sub-redes de computação e infraestrutura. A Grass declara 17 milhões de dólares de receita de dados no primeiro semestre de 2026. A inferência confidencial em ambientes de execução confiáveis está em produção. Em contrapartida, os pagamentos efetivamente iniciados por agentes de IA rondam os 5 000–11 000 dólares por mês, e o zkML e o treino descentralizado ainda não têm procura paga divulgada.

Subsídio e valor

O múltiplo de subsídio é o teste de saúde mais honesto. As recompensas dos provadores da Aleo foram cerca de 175–180 vezes as suas taxas; as emissões da Bittensor são cerca de 4–11 vezes a sua receita verificável; o valor de mercado da Render é cerca de 360 vezes as suas taxas on-chain; e os cerca de 460 milhões de dólares por ano de rendimento dos mineradores da Pearl são sobretudo subsídio por bloco, quase sem ligação à procura paga de IA. O subsídio é combustível para um arranque a frio, não um modelo de negócio. Entretanto, o valor está a

passar dos tokens para o capital: quatro empresas de capital que usam infraestrutura centralizada e distribuição nativa cripto estão avaliadas em cerca de 8,5 mil milhões de dólares, aproximadamente o mesmo que os tokens de IA das camadas de protocolo e de aplicação juntos.

ASIC

Quanto aos chips, o teto de lucro de um ASIC é aproximadamente a receita por unidade de computação vezes (1 − 1/k), em que k é quantas vezes ele é mais barato do que o hardware marginal. Como esse termo é côncavo, a eficiência para lá de k = 5–10 quase não acrescenta lucro; o que decide o resultado é o tamanho do bolo, quanto tempo duram as regras e se o chip tem um segundo mercado. As minhas respostas por tipo de chip: não fazer agora o tape-out de um chip INT7 dedicado à Pearl e apostar antes num acelerador de inferência que também possa minerar; como recém-chegado, ficar fora do mercado de mineradores de bitcoin para venda; como equipa pequena, não construir um chip de treino autónomo; construir chips de inferência quando houver um cliente assinado; cobrir-se nos aceleradores ZK e posicionar-se cedo no FHE.

Nós e orçamentos

O nó de processo deve ser escolhido pela maior parcela do custo do chip ao longo da sua vida: a energia para chips que funcionam dia e noite, o encapsulamento e a memória para chips de treino, a decisão entre SRAM e HBM para chips de inferência. E um primeiro chip deve ser comprado por etapas, um modelo, uma FPGA, uma wafer multiprojeto e só depois um conjunto de máscaras, para que uma mudança no protocolo custe o preço de um protótipo e não o de um tape-out. As oportunidades estão onde os domínios se cruzam: a cripto como primeiro cliente, a IA como segundo mercado.

Pergunta	Juízo	Números-chave (fontes nos capítulos)
① **A Pearl funciona, e o seu trabalho é útil?**	Funciona: uma blockchain protegida por multiplicação de matrizes em grande escala. Ainda não é útil: quase toda a mineração é feita sobre matrizes aleatórias	Cerca de 90 EH/s a 13 de novembro de 2026; 1 H = 1 multiplicação-acumulação INT7; cerca de 1% dos blocos com trabalho de IA
② **Onde está a procura paga real em IA × cripto?**	Inferência, dados e privacidade. Os pagamentos de agentes, o zkML e o treino descentralizado ainda estão antes da procura	Bittensor com cerca de 70 milhões de dólares por ano de receita verificável; Grass com 17 milhões de dólares no 1.º semestre de 2026; pagamentos de agentes cerca de 5 000–11 000 dólares por mês
③ **Vale a pena construir um ASIC?**	Só onde o bolo é grande, as regras são estáveis e o chip tem um segundo mercado. Hoje nenhuma rede oferece as três coisas	Teto de lucro ≈ R × (1 − 1/k); bitcoin com cerca de 11,5 mil milhões de dólares por ano de receita dos mineradores, Pearl cerca de 0,46–0,49 mil milhões, Kaspa cerca de 18 milhões
④ **Que chips deve construir uma equipa pequena?**	Aceleradores de inferência que também possam minerar; chips de inferência para um cliente assinado; trabalho inicial em FHE. Não mineradores de bitcoin para venda nem chips de treino	O melhor ASIC da Kaspa rende $0,86 por dia; orçamentos de conceção em 3/2 nm de 60–400 milhões de dólares ou mais
⑤ **Que nó de processo?**	Decide o maior custo ao longo da vida: energia, encapsulamento e memória, ou a arquitetura de memória	TSMC N2 cerca de 30 000 dólares por wafer contra cerca de 3 000 em N28; HBM e CoWoS cerca de 63% do custo de uma H100
⑥ **Quanto custa um primeiro chip, e como deve ser comprado?**	Em quatro etapas: um modelo, uma FPGA, uma wafer multiprojeto, depois as máscaras. Parar antes das máscaras se as regras ou o comprador não estiverem definidos	Um lugar comercial em MPW 50 000–500 000 dólares; máscaras N7 cerca de 3,5 milhões de dólares; um orçamento de conceção em 7/6 nm 15–100 milhões de dólares

FIG. 01 **IA × cripto em seis setores: onde a procura paga é real**

Procura paga verificável face aos sinais de subsídio e de avaliação, novembro de 2026

Cadeias de trabalho útil 01

REAL MAS PEQUENO

≈$70 M por ano

receita externa verificável da Bittensor

Sinal: emissões 4–11× a receita; Pearl ≈1% blocos IA

Os mecanismos funcionam; a utilidade precisa de quem pague

Computação descentralizada 02

REAL MAS PEQUENO

$2,35 M por ano

taxas on-chain da Render, as maiores verificáveis

Sinal: valor de mercado ≈360× as taxas

Sobretudo revenda de inferência banalizada

Treino e inferência 03

LUCRO NO CAPITAL

$100 M+ por ano

Prime Intellect, empresa de capital em GPU centralizadas

Sinal: mainnet de treino da Gensyn por lançar

O progresso é real; o lucro vai para o capital

IA verificável 04

INICIAL

398

utilizadores pagantes da nuvem TEE da Phala, fim de 2025

Sinal: nenhum zkML pago divulgado

TEE em produção; zkML inicial

Pagamentos de agentes 05

ANTES DA PROCURA

$5–11 mil por mês

pagamentos x402 iniciados de facto por agentes

Sinal: ≈200 M transações, sobretudo sinalização

As vias vieram primeiro; a procura não

Redes de dados 06

COMPRADORES REAIS

$17 M

receita de dados da Grass, primeiro semestre de 2026 (autodeclarada)

Sinal: valor de mercado ≈10× a receita anualizada

Um dos poucos setores com compradores reais

Fonte: DefiLlama, SubConnect (via hoge.gg), TRM Labs, Phala, Dealroom, SiliconANGLE, CoinGecko, Blockworks; síntese do autor.

CAPÍTULO 01

01 Âmbito, método e quatro testes

ÂMBITO E MÉTODO

Porquê este relatório, e porquê agora

Em abril de 2026, uma nova blockchain, a Pearl, pôs em produção uma ideia antiga: deixar os mineradores competir pelos blocos fazendo as multiplicações de matrizes de que a IA vive, em vez de calcular hashes. Em seis meses, a sua rede chegou a cerca de 90 EH/s e o seu token a um valor de mercado de cerca de 500 milhões de dólares. Ao mesmo tempo, mercados de computação para IA, redes de treino descentralizado, vias de pagamento para agentes e redes de dados angariaram milhares de milhões, e os mineradores de bitcoin começaram a converter os seus locais com energia em centros de dados de IA. Quis saber quais destes têm clientes reais, se o silício à medida faz sentido em algum deles e que nó de processo e que foundry uma equipa de chips pode realmente usar.

Quatro testes para qualquer projeto IA × cripto

Um estudo de segurança de 2017 sobre a Gridcoin defendia que a computação útil só pode proteger uma blockchain se for difícil de resolver, fácil de verificar e fácil de recompensar. Acrescento um quarto teste. Em conjunto, decidem a maior parte do que se segue.

01 **Difícil de resolver.** A computação honesta tem de ser a forma mais barata de produzir uma prova válida; não pode haver atalho.

02 **Fácil de verificar.** Verificar o trabalho tem de custar muito menos do que fazê-lo, ou a rede paga duas vezes.

03 **Fácil de recompensar.** O protocolo tem de conseguir calcular a recompensa só a partir da prova, sem confiar num terceiro.

04 **Alguém paga o resultado.** Se nenhum cliente fora da rede pagar o resultado, o trabalho «útil» é um subsídio com uma narrativa.

Os puzzles de hash passam os três primeiros testes e falham o quarto. As tarefas reais de IA passam o quarto mas falham o segundo: provar uma computação de IA em conhecimento zero custa 100 a 10 000 vezes a própria computação, e já se demonstrou que os esquemas de prova de aprendizagem podem ser falsificados a baixo custo. A multiplicação de matrizes com ruído da Pearl é a primeira ponte credível que passa o teste da verificação mantendo a mesma forma que a computação de IA.

Como trabalhei

Esta é uma investigação documental, feita como eu reveria um chip ou um plano de negócio. Li o código da Pearl, as suas propostas de protocolo e o seu whitepaper; dados da cadeia de um nó completo independente; documentos oficiais e comunicados de resultados de empresas; artigos de criptografia, arquitetura de computadores e aprendizagem automática; dados de preços de foundries e de fornecedores; e imprensa especializada e de tecnologia. Cada número tem a sua data. Quando eu próprio derivo um número, digo-o e assinalo-o estimativa.

Os números que só o vendedor comunica estão assinalados alegação do fornecedor; a minha interpretação dos dados está assinalada análise. Os valores estão em dólares americanos. A situação é a de 13 de novembro de 2026.

CAPÍTULO 02

02

PEARL · O MECANISMO

Pearl: a multiplicação de matrizes como consenso

A ideia, do teorema à mainnet

Em abril de 2025, Ilan Komargodski e Omri Weinstein publicaram «Proofs of Useful Work from Arbitrary Matrix Multiplication». Constroem uma prova de trabalho útil para qualquer produto de matrizes com um sobrecusto de apenas 1 + o(1) vezes uma multiplicação ingénua; a sua segurança assenta na dificuldade de resolver lotes de equações lineares aleatórias de baixo rank, o que é uma conjetura, não uma prova. O artigo nomeia a verdadeira dificuldade do domínio: uma prova de trabalho útil com um sobrecusto de fator constante é trivial mas inútil; o difícil é manter a computação honesta como a forma mais barata de ganhar quando os mineradores escolhem as suas próprias entradas.

A mainnet da Pearl arrancou a 27 de abril de 2026, às 09:00 UTC. Os mineradores calculam um produto com ruído (A + E)(B + F), em que o ruído de baixo rank é derivado de um compromisso BLAKE3, e depois retiram o ruído para recuperar A·B. O estado do acumulador de cada tile de saída é transformado, por hash, num bilhete de lotaria; um tile vencedor é aberto e provado com uma prova recursiva de conhecimento zero Plonky2 de menos de 60 KB, dentro de um certificado de bloco limitado a 65 KB. A verificação é uma pequena prova de tamanho constante, independente do tamanho da matriz: tão barata de verificar como um puzzle de hash, enquanto a própria computação pode ser útil.

Uma multiplicação-acumulação é um «hash»

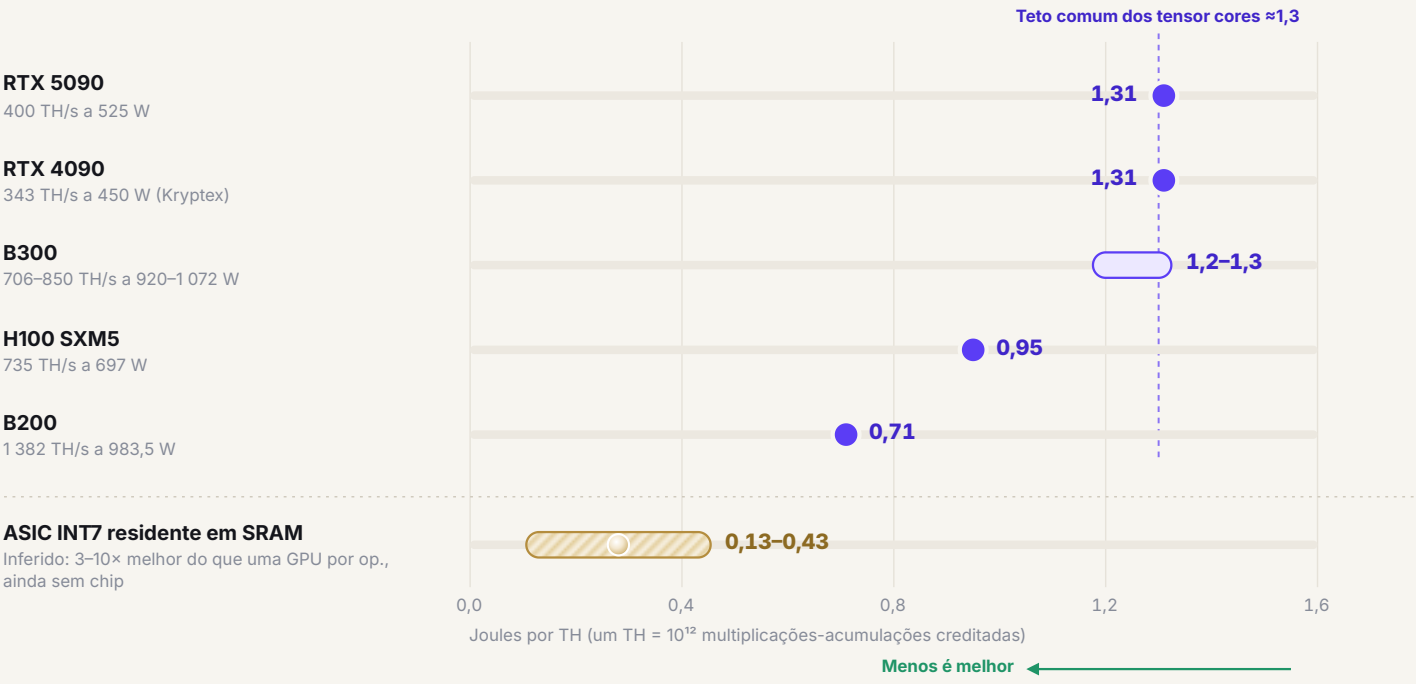
Para quem concebe chips, o facto mais importante é a unidade. O código de consenso da Pearl escala o alvo de cada tentativa pelo tamanho do tile vezes o comprimento do produto escalar, pelo que **1 H é exatamente uma multiplicação-acumulação (MAC) INT7 × INT7 → INT32 creditada**. Desde o bloco 96 251, uma penalização de rank exige um rank do ruído de pelo menos 128; no rank mínimo, 1 H continua a ser igual a 1 MAC. O «EH/s» da Pearl significa, portanto, literalmente 10^18 multiplicações-acumulações inteiras por segundo: **pela primeira vez, o orçamento de segurança de uma blockchain tem preço no débito tensorial dos aceleradores de IA.** As taxas de hash medidas nas GPU ficam entre 60% e 95% do pico de multiplicação-acumulação INT8 densa de cada placa, o que é coerente com esta leitura.

Hardware	Taxa de hash na Pearl	Potência	Joules por bilião de MAC (J/TH)	Fonte (data)
RTX 5090	400 TH/s	525 W	1,31 estimativa	Kryptex (12 de novembro)
RTX 4090	cerca de 305 TH/s (intervalo da comunidade 280–330); a Kryptex indica 343 TH/s	450 W (Kryptex)	1,31 (valores da Kryptex) estimativa	prlstats (3 de outubro); Kryptex
H100 SXM5	735 TH/s	697 W	0,95 estimativa	PeakMiner, via prlstats (19 de setembro)

Hardware	Taxa de hash na Pearl	Potência	Joules por bilião de MAC (J/TH)	Fonte (data)
B200	1,382 PH/s	983,5 W	0,71 estimativa	PeakMiner, via prlstats
B300	706–850 TH/s	920–1 072 W	1,2–1,3 estimativa	prlstats; SimeonOnSecurity

FIG. 02 **No trabalho inteiro da Pearl, as GPU partilham um teto energético**

Joules por bilião de multiplicações-acumulações INT7 (J/TH) à taxa de hash medida na Pearl; menos é melhor



Fonte: Kryptex (12 de novembro de 2026); benchmarks PeakMiner e SRBMiner via prlstats; SimeonOnSecurity; o intervalo do ASIC é inferência do autor, não uma medição.

Dois pormenores desta tabela importam a quem concebe chips. Primeiro, as placas de consumo RTX 4090 e 5090 e a B300 de centro de dados convergem em cerca de 1,3 J/TH, e mesmo a melhor placa, a B200, só chega a cerca de 0,71 J/TH. Uma GPU de centro de dados que custa dezenas de milhares de dólares gasta quase tanta energia por multiplicação-acumulação inteira como uma placa de jogos, o que aponta para um teto comum da energia dos tensor cores por MAC INT8, e esse teto é exatamente o que um chip dedicado poderia quebrar. Segundo, a B300 é mais lenta do que a B200: a NVIDIA cortou fortemente o débito INT8 na Blackwell Ultra, pelo que o trabalho inteiro tem de seguir um caminho mais lento nesse chip. O silício de IA dominante está a afastar-se do INT8 em direção ao FP8 e ao FP4. Quando o formato numérico escolhido por um consenso diverge do rumo do hardware, o protocolo é forçado a mudar as regras, e essa é a raiz do fork FP8 e do meu juízo sobre os ASIC mais adiante neste relatório.

CAPÍTULO 03

03

PEARL · A ECONOMIA

Pearl: escala, emissão e o défice de utilidade

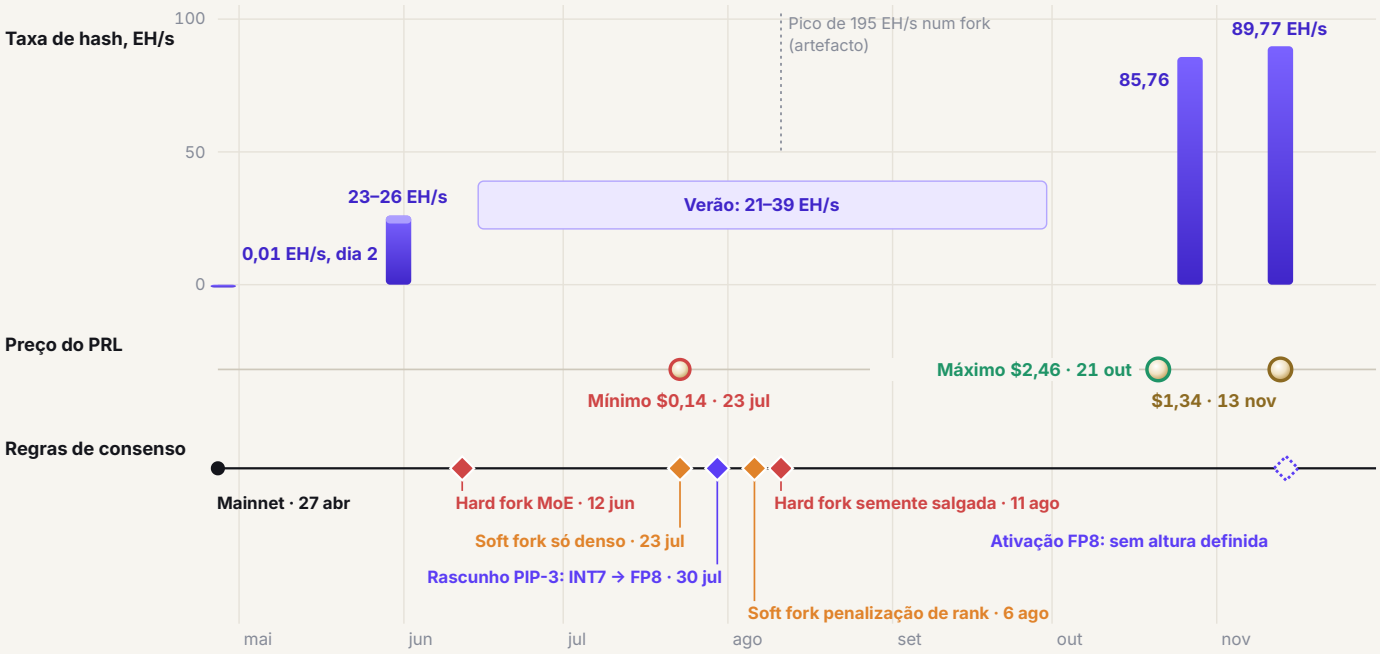
De 0,01 a 90 EH/s em seis meses

Métrica	Valor (13 de novembro de 2026)	Base e fonte
Taxa de hash da rede	Cerca de 90 EH/s (referência 89,77; intervalo do mesmo dia 88,20–93,22; máximo intradiário a 12 de novembro de 110–123)	Nó completo independente prlstats; Kryptex
Dificuldade e altura de bloco	58 841 982 e 142 530	prlstats
Hardware equivalente	Cerca de 122 000 H100 ou 225 000 RTX 5090	A 735 e 400 TH/s por placa estimativa
Preço, valor de mercado e valor totalmente diluído	$1,32–1,34; cerca de 500 milhões de dólares; cerca de 2,8 mil milhões de dólares (posição 101 na CoinGecko)	CoinGecko
Máximo e mínimo históricos	$2,46 (21 de outubro) e $0,1399 (23 de julho)	CoinGecko
Moedas mineradas	377,6 milhões de PRL, 17,98% do limite de 2,1 mil milhões	prlstats
Subsídio por bloco e emissão diária	Cerca de 2 173 PRL e cerca de 0,97–1,03 milhões de PRL	prlstats; Kryptex
Nova oferta nos próximos 12 meses	Cerca de 78% da oferta atual	prlstats
Rendimento dos mineradores	Cerca de 1,25 milhões de dólares por dia (modelo), cerca de 460 milhões de dólares por ano	prlstats
Receita por unidade de taxa de hash, R	Cerca de $14–15 por PH/s por dia; uma RTX 5090 rende cerca de $5–6 por dia	estimativa; Kryptex
Blocos com trabalho real de IA	Cerca de 1% (estimativa de 28 de outubro)	Blockworks
Concentração das pools	A Kryptex encontrou 46,4% dos blocos nas últimas 24 horas; as duas maiores pools cerca de 62%	prlstats
Concentração dos detentores	Coeficiente de Gini dos endereços 0,969; os 100 maiores endereços detêm 53,0%	prlstats

A taxa de hash da Pearl é quase uma função desfasada do seu preço. No dia a seguir ao lançamento, a rede funcionava a cerca de 0,01 EH/s; no fim de maio chegou a 23–26 EH/s. Durante o verão, o preço caiu cerca de 86% face ao máximo de maio e a taxa de hash andou de lado entre 21 e 39 EH/s. Depois de o preço ter subido cerca de dez vezes em setembro e outubro, a taxa de hash atingiu 85,76 EH/s a 27 de outubro e cerca de 90 EH/s em novembro. Um pico de 195 EH/s a 11 de agosto coincidiu com um fork e deve ser lido como um artefacto de estimativa. A energia cresceu a par: um estudo estimou cerca de 112 MW para toda a rede a 24 EH/s, e um modelo da prlstats situa-a hoje em cerca de 190–215 MW.

FIG. 03 **Os primeiros seis meses da Pearl: a taxa de hash segue o preço, as regras mudam**

Taxa de hash da rede, três marcas de preço e cada alteração de consenso, de 27 de abril a 13 de novembro de 2026



Fonte: Nó completo e séries da prlstats; Kryptex; CoinGecko; params.go e PIP-3 da Pearl. O pico de 11 de agosto é um artefacto de estimativa num fork.

A emissão amplifica a reflexividade

O limite de oferta é de 2,1 mil milhões de moedas e não há halvings: o subsídio por bloco decresce suavemente segundo E(t) = S·H / ((t + H)(t + H − 1)), com H = 650 226. Mas a emissão conta-se em blocos, não em tempo, e a dificuldade inicial começou demasiado baixa e ajustou-se devagar. **Cerca de 44,06 milhões de moedas foram mineradas no primeiro dia e cerca de 120 milhões nos primeiros seis dias**, cerca de 32% da oferta atual estimativa. Os blocos recentes, a cerca de 171–183 segundos, continuam a chegar mais depressa do que o alvo de 194 segundos. O resultado é que a oferta crescerá cerca de 78% nos próximos 12 meses, uma diluição que mineradores, fabricantes de hardware e detentores têm todos de enfrentar.

A economia dos mineradores decorre daqui. O rendimento dos mineradores em toda a rede é de cerca de 1,25 milhões de dólares por dia (modelo); a emissão diária da Kryptex vezes o preço atual dá cerca de 1,35 milhões de dólares por dia, pelo que **a receita por unidade de taxa de hash R ≈ $14–15 por PH/s por dia** estimativa. Uma RTX 5090 rende cerca de $5–6 por dia. Uma H100 ou H200 rende cerca de $11, enquanto o aluguer a pedido público mais barato é de cerca de $2,09 por placa por hora, cerca de $50 por dia, pelo que alugar placas só para minerar dá grandes prejuízos. R oscila muito: segundo a série da prlstats, foi de cerca de $10 em agosto e setembro e, brevemente, de cerca de $33 em meados de outubro estimativa. Poucas semanas após o lançamento, a receita diária estimada de uma RTX 5090 caiu para metade, de cerca de $33,80 para cerca de $17,19. Alugar GPU na nuvem para minerar só compensou nas primeiras semanas e durante o pico de preço de outubro.

O teorema mantém-se; o incentivo não

O verdadeiro problema é a procura, não a criptografia. A única integração comercial pública da Pearl é um endpoint Gemma-4-31B-it-pearl da Together AI, lançado a 15 de maio a mais de 25% abaixo do preço habitual, com o desconto financiado pelo valor futuro esperado da emissão. A 13 de novembro, a página do modelo indicava que **não está disponível na API serverless da Together**. Em junho, um recenseamento de 8 012 instâncias de mineração em linha, cerca de 21% da rede, concluiu que a rede não produzia qualquer computação de IA útil (o au-

tor considera os indícios sugestivos, não conclusivos), e mostrou que as pools aceitavam shares construídas a partir de matrizes puramente aleatórias. A 28 de outubro, a Blockworks estimou que **apenas cerca de 1% dos blocos transporta trabalho real de IA**; sem ele, escreveu, a Pearl é uma bitcoin mais pequena, menos segura e com inflação mais rápida.

A razão pode ser calculada. O sobrecusto do lado da inferência é genuinamente pequeno: a Pearl indica um aumento de 5,08% no tempo de ponta a ponta para o Llama 70B em quatro H200, e um teste independente mediu uma perda de débito de 9,2% numa GPU Hopper. O lado da mineração rende demasiado pouco. O plug-in oficial do vLLM só minera em lotes de prefill de pelo menos 1 024 tokens e nunca nos passos de descodificação. Uma placa a servir inferência obtém cerca de 27% da taxa de hash de um minerador dedicado com tráfego de prefill puro, cerca de 5,4% numa mistura de chat com 80% de descodificação e apenas cerca de 1% com tráfego dominado pela descodificação. Sem procura paga de inferência, um minerador racional multiplica matrizes aleatórias; o protocolo verifica que as matrizes comprometidas foram multiplicadas honestamente, não se eram úteis. **O teorema mantém-se; o incentivo não.** A utilidade da Pearl é um problema comercial fora da cadeia: alguém tem de pagar a inferência.

CAPÍTULO 04

PEARL · AS REGRAS

04 Pearl: risco de regras, aritmética numérica e governação

Quatro forks em dois meses

A outra face da Pearl é a rapidez com que as suas regras mudam. De junho a agosto de 2026, o laboratório impôs quatro alterações de consenso em alturas fixadas no código: um hard fork de mistura de especialistas (bloco 71 935, cerca de 12 de junho), um soft fork só para modelos densos (91 630, cerca de 23 de julho), um soft fork de penalização de rank (96 251, cerca de 6 de agosto) e um hard fork de semente salgada (99 000, cerca de 11 de agosto). Depois do fork de semente salgada, os mineradores antigos só conseguiam produzir shares inválidas; a penalização de rank cortou a taxa de hash efetiva das configurações de rank elevado em 50% a 87,5%.

O passo seguinte é o PIP-3, ainda em rascunho: passa a mineração de INT7 para FP8 × FP8 → FP32 e exige perfis de dispositivo exatos ao bit, **suportando apenas duas arquiteturas, Hopper e Blackwell**. O documento chama à dependência do hardware um mal necessário dos protocolos de vírgula flutuante e avisa que qualquer tolerância na verificação convida a ataques de grinding. A 12 de novembro, os parâmetros de mainnet do ramo fp8-scheme não definiam qualquer altura de ativação. Segundo o rascunho atual, a série RTX 40, a A100 e as placas AMD não são admitidas, enquanto um modelo de baixa confiança da prlstats estima que cerca de 82% da taxa de hash vem de placas de consumo. Forçar o fork arrisca uma divisão da cadeia ou uma revolta dos mineradores.

Um trilema do formato numérico

A Pearl enfrenta um trilema: um formato numérico pode ser útil para IA (FP8 ou FP4), verificável bit a bit entre hardware diferente e neutro quanto ao hardware, mas no máximo duas das três coisas. O esquema INT7 é verificável e neutro, mas diverge do rumo do silício de IA. O FP8 com lista branca é útil e verificável, mas amarra o consenso a um fornecedor. O FP8 com tolerância é útil e neutro, mas abre a porta a ataques de grinding. A aritmética de vírgula flutuante nos tensor cores em geral não segue uma norma numérica e não é reprodutível entre

gerações; na Hopper, por exemplo, a acumulação FP8 mantém apenas 13 bits fracionários e trunca. Do ponto de vista de quem concebe chips, um «perfil Hopper» é uma aritmética totalmente determinada que um ASIC poderia implementar bit a bit, talvez de forma mais eficiente do que a GPU. Mas fazê-lo seria um negócio numa zona cinzenta, contrário à intenção declarada do protocolo, com um risco de governação muito elevado análise.

Governação

A governação está concentrada. O laboratório redigiu todas as propostas de protocolo; os forks ativam-se em alturas fixadas no código sem sinalização dos mineradores; não existe uma segunda implementação de cliente; e, no início de outubro, não tinha sido publicada qualquer auditoria independente do nó nem dos circuitos de conhecimento zero. O jornal israelita TheMarker noticiou que o laboratório angariou cerca de 120–130 milhões de dólares numa ronda discreta (cerca de 55 milhões de dólares de capital mais um fundo de infraestrutura de GPU de cerca de 75 milhões de dólares); a Pearl não o confirmou por canais oficiais. «Nenhuma alocação a investidores no token» e «uma empresa com acionistas de capital de risco» podem ser ambas verdade, mas a tensão mantém-se: a receita de IA pode ir para o capital enquanto a diluição recai sobre os detentores do token.

No conjunto, a Pearl é a experiência de «computação como consenso» mais limpa até agora: um mecanismo novo, verificável na cadeia, com uma unidade que corresponde um a um ao hardware de IA. Mas a utilidade ainda não chegou, e o seu orçamento de segurança, a sua taxa de hash e o entusiasmo dos mineradores dependem todos de um token cuja procura principal ainda está por formar e cuja oferta crescerá cerca de 78% em 12 meses. Vale a pena seguir três sinais da Blockworks: se a atualização corre sem sobressaltos, se a parte dos blocos com trabalho de IA sobe de forma significativa e se algum fornecedor de inferência repassa o rendimento da mineração aos seus clientes.

CAPÍTULO 05



CADEIAS DE TRABALHO ÚTIL

05 Cadeias de trabalho útil: nem matrizes nem conhecimento zero travam os ASIC

Kaspa: uma matriz não é útil nem resistente a ASIC

A Kaspa é o caso de manual. O seu kHeavyHash coloca um produto matriz-vetor 64 × 64 sobre valores de 4 bits entre dois hashes, e o produto é deitado fora. Cerca de 16 meses depois do seu lançamento justo em novembro de 2021, a IceRiver anunciou os primeiros ASIC para Kaspa a 14 de março de 2023 e, no fim de 2023, a mineração com GPU tinha acabado. Comparar uma RTX 4090 a 3,8 GH/s e 300 W com a KS7 da Bitmain a 77 J/TH dá uma diferença de eficiência de cerca de **1 000 vezes** estimativa. Depois da fartura veio a compressão: a 13 de novembro, a rede funcionava a 358,9 PH/s, cerca de 64% abaixo do seu 1 EH/s de agosto de 2024, e a melhor KS7 rende apenas **$0,86 por dia** a $0,10 por kWh.

Aleo: o conhecimento zero também não trava o hardware especializado

O puzzle de prova da Aleo foi concebido para imitar o trabalho de gerar uma prova SNARK; não prova as transações dos utilizadores, e o projeto acolhe abertamente os ASIC de ZK. Cerca de cinco meses depois da sua mainnet de setembro de 2024, já se entregavam máquinas dedicadas. No terceiro trimestre de 2025, os provadores receberam 31,2 milhões de ALEO (cerca de 7,8 milhões de dólares), enquanto a rede cobrou apenas cerca de 43 000–44 000 dólares em taxas: **um múltiplo de subsídio de cerca de 175–180**. O preço caiu de cerca de $9

nas primeiras negociações para $0,22 no fim desse trimestre, e o ARC-46 exige agora um stake de 100 000 a 2,5 milhões de ALEO por solução, transformando a mineração num limiar de stake.

Qubic e Bittensor

A Qubic promove uma «prova de trabalho útil» que treina a sua rede neuronal Aigarth, mas o resultado mais visível da sua computação em 2025–2026 foram blocos de outras cadeias. Primeiro minerou Monero, subindo a sua parte de menos de 2% em meados de maio de 2025 para mais de 25% no fim de julho e reivindicando 51% em agosto, o que causou uma reorganização de seis blocos (outra fonte fala em 18). A 1 de abril de 2026 passou a minerar Dogecoin com ASIC, usando as receitas para recompras e queimas. Não há indícios independentes de que o seu treino de IA seja útil, e o seu valor de mercado caiu de cerca de 280–300 milhões de dólares para cerca de 50 milhões a 13 de novembro. Expôs também uma externalidade: uma cadeia que paga computação de uso geral cria uma reserva de taxa de hash grande, alugável e coordenada que pode ser virada contra cadeias favoráveis aos CPU.

A Bittensor não é prova de trabalho, mas pontuação subjetiva ponderada por stake: os validadores pontuam os mineradores, e os pesos são cortados na mediana ponderada por stake. É a única rede deste grupo com receita externa à escala. O índice SubConnect, que conta apenas clientes externos verificáveis, mostra cerca de **70 milhões de dólares por ano** no início de novembro de 2026 (cerca de 35 milhões a 1 de setembro), 78% a 82% dos quais de computação e infraestrutura; mas, aos preços do início de novembro, as suas emissões ultrapassam 300 milhões de dólares por ano, pelo que a receita cobre apenas 9% a 23%. As estimativas de receita variam muito com a base, de cerca de 17 milhões de dólares (Artemis) a 172 milhões anualizados a partir de totais autodeclarados. Em abril, 67 de 128 sub-redes queimavam mais de 99% dos seus incentivos e, quando a equipa principal Covenant saiu por divergências de governação, o TAO caiu 24% em poucas horas.

A história dá a mesma resposta

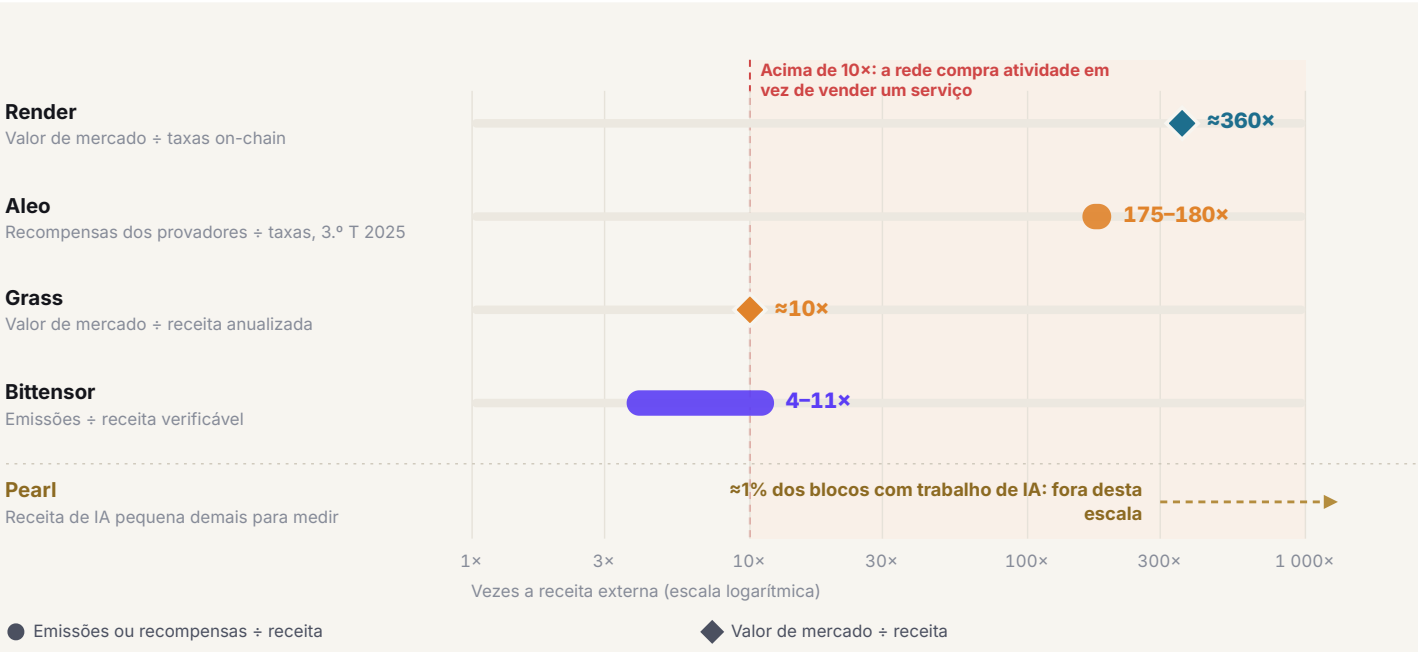
As cadeias de números primos da Primecoin verificavam-se depressa, mas ninguém precisava delas. A Gridcoin dependia de sites de estatísticas de terceiros para atestar as contribuições BOINC, e investigadores forjaram blocos sem fazer qualquer computação. As provas de armazenamento da Filecoin funcionam tecnicamente, mas no terceiro trimestre de 2025 a utilização era de apenas 36%, e 99,5% dos seus 793 000 dólares de taxas eram penalizações, com taxas de base de apenas 4 100 dólares.

Projeto	Trabalho ou consenso	Valor de mercado (data)	Percurso do hardware	Procura paga face ao subsídio	Juízo
Pearl	Multiplicação de matrizes com ruído (MAC INT7 mais certificado ZK)	Cerca de 500 milhões de dólares (13 de novembro)	GPU; o rascunho FP8 poria a Hopper e a Blackwell em lista branca	Cerca de 1% dos blocos com trabalho de IA; rendimento dos mineradores sobretudo subsídio	O mecanismo mais credível; procura por provar
Kaspa	kHeavyHash (um produto matriz-vetor 64 × 64 de 4 bits entre hashes)	Cerca de 840 milhões de dólares (13 de novembro)	ASIC cerca de 16 meses após o lançamento; hoje só ASIC	Não reivindica utilidade; receita dos mineradores cerca de 18 milhões de dólares por ano estimativa	Uma matriz não é útil nem resistente a ASIC
Aleo	Puzzle SNARK sintético mais prova de participação	Cerca de 118 milhões de dólares (fim do 3.º T 2025)	Máquinas dedicadas ao fim de cerca de cinco meses	Recompensas dos provadores cerca de 175–180 vezes as taxas	Trabalho sintético, movido a subsídio

Projeto	Trabalho ou consenso	Valor de mercado (data)	Percurso do hardware	Procura paga face ao subsídio	Juízo
Qubic	«uPoW» (diz treinar a Aigarth)	Cerca de 50 milhões de dólares (13 de novembro)	CPU e GPU; ASIC para Dogecoin desde abril de 2026	Sem indícios independentes de utilidade; computação usada para mineração externa e recompras	Agregação de computação embrulhada numa história de IA
Bittensor	Consenso Yuma (pontuação ponderada por stake)	Cerca de 2,55 mil milhões de dólares (11 de novembro)	GPU de uso geral; sem percurso para ASIC	Cerca de 70 milhões de dólares por ano de receita verificável contra mais de 300 milhões de dólares de emissões	A única com receita externa à escala; ainda assim um subsídio de 4–11 vezes
Filecoin (referência)	Provas de replicação e de espaço-tempo	Cerca de 1,5 mil milhões de dólares (fim do 3.º T 2025)	—	Utilização 36%; taxas de base trimestrais 4 100 dólares	Útil, mas sem procura suficiente

FIG. 04 **O múltiplo de subsídio: quantas vezes a receita externa uma rede paga**

Emissões ou valor de mercado divididos pela receita paga de fora da rede, escala logarítmica



Fonte: Blockworks (Aleo, 3.º T 2025), DefiLlama (taxas da Render), CoinGecko, hoge.gg (Bittensor), Dealroom (Grass), Blockworks 0xResearch (Pearl); cálculo do autor.

A lição para o hardware: um trabalho fixo e regular, pago num token líquido, recebe ASIC em cerca de 3 a 16 meses, muitas vezes de concorrentes antes dos incumbentes. As alegações de «utilidade» só se mantiveram onde um cliente externo paga o resultado. E os principais motores de hardware encalhado são um subsídio elevado face à receita e o controlo dos programadores sobre a função de trabalho.

CAPÍTULO 06

O PANORAMA ALARGADO

06 O resto de IA × cripto: procura real mas pequena, e valor a passar para o capital

Mercados de computação descentralizada

A computação descentralizada tem clientes pagantes, mas a procura é pequena, concentrada e sobretudo revenda de inferência de modelos abertos. As taxas verificáveis na cadeia são pequenas: a DefiLlama mostra cerca de 2,35 milhões de dólares de taxas para a Render no último ano e cerca de 680 000 dólares para a Livepeer, e todo o setor DePIN obteve apenas cerca de 72 milhões de dólares de receita on-chain em 2025. Os números autodeclarados são maiores, mas definidos de forma vaga: a Aethir declara «mais de 127,8 milhões de dólares» de receita no ano fiscal de 2025 sem dizer se se trata de faturação bruta ou da parte do protocolo. Os valores dos tokens e a receita separaram-se nos dois sentidos: o valor de mercado da Render é cerca de 360 vezes as suas taxas on-chain, enquanto o valor de mercado da Aethir, de cerca de 87,7 milhões de dólares, está abaixo da sua receita autodeclarada estimativa.

A alegação de ser 70% a 85% mais barato só se verifica face aos preços a pedido dos hiperescaladores. A 3 de novembro, o preço médio ponderado da H100 na Akash era de cerca de **$2,73 por hora de GPU** (intervalo $2,04–3,16), abaixo de cerca de $6,88 na AWS e $11,06 na Google Cloud, mas acima de cerca de $1,73 na Vast.ai, $1,99 na nuvem comunitária da RunPod e cerca de $2,44 numa reserva de um ano (preços das outras plataformas lidos a 12 de setembro; a Akash, que os publica, é parte interessada). A procura é sobretudo inferência banalizada distribuída através da OpenRouter, cerca de 1,7 mil milhões de tokens por dia na AkashML e mais de 4 mil milhões na io.net, em concorrência direta com todas as neoclouds. Os maiores montantes da «computação descentralizada» estão a recentralizar-se: a Axe Compute, que detém o token da Aethir como ativo de tesouraria, assinou um contrato de cinco anos no valor de 1,5 mil milhões de dólares para alugar o seu próprio cluster de 9 200 B300, contra uma receita de 35 000 dólares no primeiro trimestre de 2026.

Treino e inferência descentralizados

O treino pela internet foi demonstrado com 8 a 40 mil milhões de parâmetros. O Pluralis-8B treinou durante 40 dias em 330 nós, sobretudo GPU de consumo, e chegou a 63% da eficiência de uma referência centralizada em H100, embora o seu artigo admita que a execução não foi descentralizada e precisou dos nós da própria Pluralis e de uma autoridade de admissão. O Psyche da Nous Research concluiu a execução de pré-treino Consilience, de 40 mil milhões de parâmetros. À escala de fronteira, o trabalho recentralizou-se: o INTELLECT-3 da Prime Intellect, um modelo de mistura de especialistas com 106 mil milhões de parâmetros, foi treinado durante dois meses em 512 H200 ligadas por InfiniBand a 400 Gbps. As restrições são claras: os nós precisam de pelo menos 200 Mbit/s de largura de banda e de uma compressão da comunicação de cerca de 100 vezes, e a confiança e a verificação continuam a depender dos papéis privilegiados do operador.

O dinheiro foi para empresas de capital que vendem produtos centralizados. A Prime Intellect tinha mais de 100 milhões de dólares de receita anualizada antes de angariar 130 milhões de dólares com uma avaliação de 1 mil milhões de dólares em julho; a Venice, uma aplicação de inferência privada, declara mais de 70 milhões de dólares de receita anualizada e lucros desde o primeiro trimestre de 2026 (números da empresa), angariou 65 milhões de dólares com uma avaliação de 1 mil milhões e funciona com GPU alugadas. Do lado dos tokens, a rede

de treino central da Gensyn ainda está «por definir» e o seu token vale cerca de 17,2 milhões de dólares; o token da Ritual, uma layer-1 nativa de IA, caiu 72% face ao máximo após a mainnet (apenas o título da notícia).

IA verificável

A produção escolheu os ambientes de execução confiáveis (TEE). A computação confidencial da Hopper acrescenta menos de 7% de sobrecusto às consultas típicas de LLM e quase nada para modelos grandes; o sobrecusto em produção da Phala é de cerca de 0,5% a 5%, mas no fim de 2025 pagavam 398 dos seus 10 004 utilizadores. A raiz de confiança de um TEE é o fabricante do chip: o TEE.fail, divulgado em outubro de 2025, usou um interposer no barramento DDR5 que custa menos de 1 000 dólares para extrair chaves de atestação de máquinas Intel TDX e AMD SEV-SNP totalmente atualizadas, o que por sua vez compromete a computação confidencial das GPU da NVIDIA, e ambos os fabricantes dizem que os ataques físicos estão fora do seu modelo de ameaça. É exatamente o cenário DePIN, em que os operadores dos nós têm acesso físico. O zkML já consegue provar um LLM completo: o DeepProve da Lagrange prova o GPT-2 a cerca de 174 tokens por minuto e o Gemma 3 a cerca de 86, ordens de grandeza abaixo da inferência em produção e sem uso pago divulgado.

Pagamentos de agentes e dados

O x402 é o melhor exemplo de cautela. A TRM Labs contabiliza cerca de 198,9 milhões de transações e 52,7 milhões de dólares liquidados por facilitadores conhecidos desde maio de 2025, 99,6% em USDC; depois de retirar os autopagamentos, os fluxos em lote e os vendedores com menos de dez compradores, a parte efetivamente iniciada por agentes é de apenas 0,6% a 7,5%, **cerca de 5 000–11 000 dólares por mês**. A Artemis considera que mais de 95% da atividade é sinalização de protocolo e não compra e venda. Entretanto, as vias estão a ser normalizadas pelos incumbentes: a x402 Foundation foi criada no âmbito da Linux Foundation a 14 de julho, com Visa, Mastercard, Stripe, Google, AWS, Cloudflare e Coinbase como membros principais, e a Tempo, que coescreveu o Machine Payments Protocol com a Stripe, lançou a sua mainnet em março com uma avaliação de 5 mil milhões de dólares. As plataformas de tokens de agentes continuam a funcionar com a lógica das launchpads: a Virtuals declara 479 milhões de dólares de «aGDP», e o fundador da ElizaOS declarou o seu token morto após um acordo numa ação coletiva. O valor será mais provavelmente captado pelos emissores de stablecoins e pelos incumbentes dos pagamentos; o papel distintivo da cripto aqui pode ser a identidade e a responsabilização dos agentes.

Os dados são um dos poucos subsetores com compradores reais fora da cadeia. A Grass declara 17 milhões de dólares de receita de dados em 2025 e outros 17 milhões no primeiro semestre de 2026, aponta para 75 milhões no ano e diz ser rentável; o GRASS vale cerca de 328 milhões de dólares, cerca de dez vezes a sua receita anualizada do primeiro semestre estimativa. O seu fosso defensivo é a escala de recolha da largura de banda residencial distribuída, não a cripto em si. As DAO de dados e os mercados de dados não têm procura paga visível: a Vana vale cerca de 21,5 milhões de dólares e a Ocean cerca de 27,6 milhões, e a Ocean está em litígio com a Fetch.ai. A prova de pessoa está a ganhar integrações reais à medida que os agentes se multiplicam: a World diz ter verificado quase 18 milhões de pessoas, e a Zoom, a Docusign e o Tinder suportam partes do World ID.

O capital já votou

Setor	Procura paga verificável	Sinais de subsídio e de avaliação	Juízo
Cadeias de trabalho útil	Bittensor com cerca de 70 milhões de dólares por ano de receita externa verificável; Pearl com cerca de 1% dos blocos com trabalho de IA	Aleo 175–180 vezes; Bittensor 4–11 vezes; Pearl sobretudo subsídio	Os mecanismos têm valor; a utilidade precisa de clientes pagantes

Setor	Procura paga verificável	Sinais de subsídio e de avaliação	Juízo
Mercados de computação descentralizada	Taxas on-chain de cerca de 0,68–2,35 milhões de dólares por ano (Livepeer, Render); autodeclarado 4–128 milhões	Valor de mercado da Render cerca de 360 vezes as taxas	Real mas pequena, sobretudo revenda de inferência banalizada
Treino e inferência descentralizados	Empresas de capital: Prime Intellect mais de 100 milhões de dólares por ano, Venice mais de 70 milhões (infraestrutura centralizada)	Mainnet de treino da Gensyn por lançar; token da Ritual −72% face ao máximo	O progresso técnico é real; os lucros vão para o capital
IA verificável	TEE em produção (398 utilizadores pagantes na Phala); nenhum zkML pago divulgado	—	TEE úteis; zkML inicial
Pagamentos de agentes	Fluxos iniciados por agentes cerca de 5 000–11 000 dólares por mês	Cerca de 200 milhões de transações, sobretudo sinalização e especulação	Primeiro as vias, depois a procura
Redes de dados	Grass com 17 milhões de dólares no 1.º semestre de 2026 (autodeclarado)	GRASS cerca de 10 vezes a receita anualizada	Um dos poucos setores com compradores reais

Pondo os seis setores lado a lado, a conclusão é coerente: **o subsídio é combustível para um arranque a frio, não um modelo de negócio.** A 13 de novembro, a categoria de IA da CoinGecko valia no total cerca de 18,6 mil milhões de dólares, mas cerca de 56% disso são layer-1 de uso geral como NEAR e ICP, mais projetos ao nível do consenso como TAO, TAO em staking (possivelmente contado em duplicado com o TAO) e Pearl; os tokens de IA das camadas de protocolo e de aplicação somam cerca de 8 mil milhões de dólares ou menos estimativa. É aproximadamente a avaliação conjunta de quatro empresas de capital: Tempo (5 mil milhões de dólares), Nous (1,5 mil milhões, segundo notícias), Prime Intellect (1 mil milhões) e Venice (1 mil milhões). O financiamento de capital de risco em cripto está estagnado em cerca de 4,5–5 mil milhões de dólares por trimestre, com o aumento a vir da infraestrutura ligada à IA, cuja parte nos montantes divulgados subiu de 8,1% para 18,2% no terceiro trimestre de 2026; e 84,7% dos 118 tokens lançados em 2025 negoceiam abaixo do preço de lançamento. Numa frase: **a cripto faz a distribuição e a liquidação, o capital fica com os lucros**, pelo menos em 2026.

CAPÍTULO 07

07

ECONOMIA DOS ASIC

A questão dos ASIC: uma fórmula, três medidas

Uma fórmula: porque a eficiência deixa de compensar depois de k = 5

Seja R a receita diária por unidade de computação (rendimento dos mineradores da rede dividido pela taxa de hash da rede), c o custo diário de uma unidade de computação no hardware marginal (na Pearl, uma GPU) e k quantas vezes um ASIC é mais barato do que esse hardware marginal (c_ASIC = c/k). Enquanto as GPU continuarem a ser os mineradores marginais, a livre entrada e saída empurra R para c, pelo que um ASIC ganha R − c/k por unidade de computação:

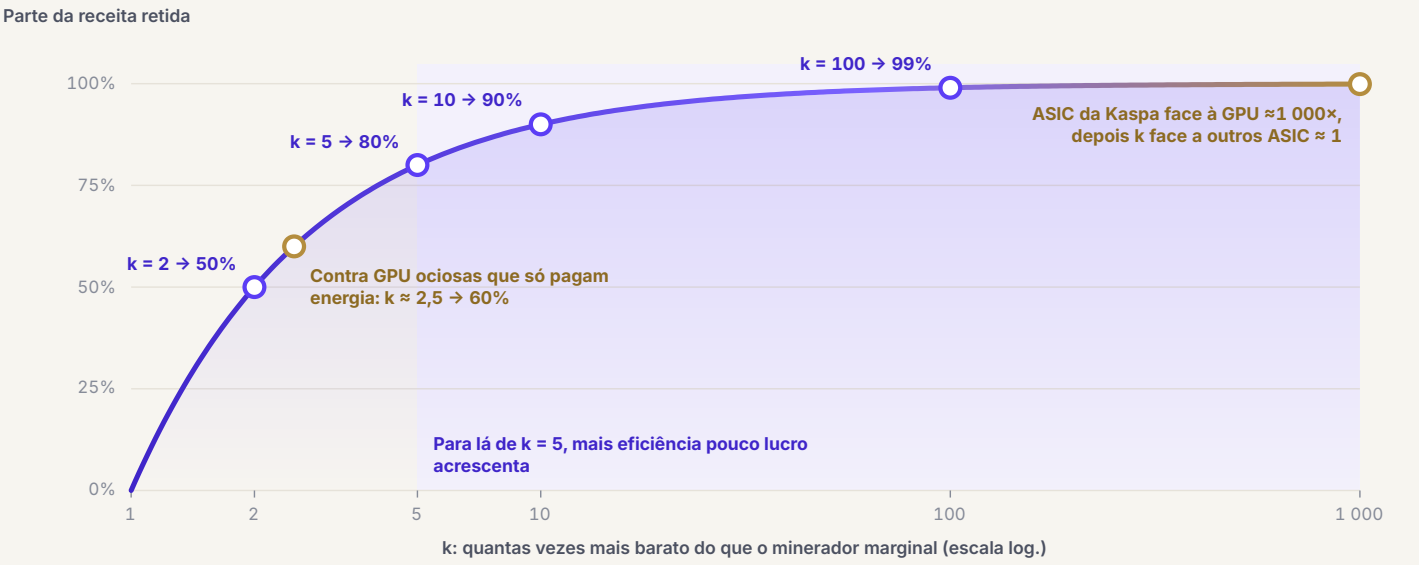
Teto de lucro de um ASIC ≈ R × (1 − 1/k)

Em toda a rede, o teto anual de lucro bruto de uma frota de ASIC é aproximadamente a receita dos mineradores da rede × (1 − 1/k), e apenas até os ASIC substituírem as GPU como mineradores marginais.

A fórmula tem três consequências. Primeiro, (1 − 1/k) é côncavo: com k = 2 um ASIC retém 50% da receita, com k = 5 80%, com k = 10 90%, com k = 100 99%. **Para lá de k = 5–10, mais eficiência quase não acrescenta lucro**; decidem o tamanho de R e o do bolo. Segundo, a referência recomeça: quando os ASIC são os mineradores marginais, k tem de ser medido face a outros ASIC. Os ASIC da Kaspa tinham uma vantagem de eficiência de cerca de 1 000 vezes sobre as GPU, que só comprou um ganho inesperado breve; hoje a melhor KS7 rende $0,86 por dia. Terceiro, numa cadeia favorável às GPU como a Pearl, a referência não é muitas vezes o preço de compra de uma GPU, mas **a eletricidade de uma GPU ociosa**: uma RTX 5090 própria a 525 W e $0,10 por kWh custa cerca de $1,26 por dia em energia e rende cerca de $5–6 estimativa. Face a GPU que são custos afundados, mesmo um ASIC que use um quinto da energia tem um k efetivo de apenas cerca de 2,5 se a sua amortização igualar a sua eletricidade, e um teto de lucro de cerca de 60% de R (um pressuposto ilustrativo).

FIG. 05 **Porque a eficiência deixa de compensar: um ASIC retém no máximo 1 − 1/k da receita**

Parte da receita que um ASIC pode reter face ao minerador marginal, segundo a sua vantagem de custo k



Fonte: Modelo do autor: quando o hardware marginal fixa o preço, a receita por unidade de computação tende para o seu custo; valores da Kaspa e da RTX 5090 do relatório.

Três medidas: o bolo, a escala de retorno e as regras

A primeira medida é **o bolo**. Anualizando cerca de 2 893 BTC de recompensas de bloco na primeira semana de novembro (cerca de 222 milhões de dólares), os mineradores de bitcoin ganham cerca de 11,5 mil milhões de dólares por ano estimativa. Os mineradores da Pearl ganham cerca de 0,46–0,49 mil milhões de dólares (aos preços atuais); as recompensas dos provadores da Aleo, anualizadas a partir do terceiro trimestre de 2025, cerca de 31 milhões de dólares; e os mineradores da Kaspa, com uma recompensa por bloco de 1,945 KAS e dez blocos por segundo, cerca de 18 milhões de dólares estimativa. **Um ano inteiro de receita dos mineradores da Kaspa é menos do que o limite inferior de um único orçamento de conceção de chip em 5/4 nm (cerca de 30 milhões de dólares).**

A segunda medida é **a taxa de hash necessária para recuperar o NRE**:

h* = NRE ÷ [R × (1 − 1/k) × T × p]

em que T é a janela útil em dias e p a probabilidade de as regras não mudarem nessa janela. Para a Pearl, tome-se R = $14 por PH/s por dia, k = 5, T = 365 dias e um NRE de 20 milhões de dólares (um chip de matriz repetitiva em N7/N6; no limite inferior dos intervalos de 5–30 milhões de dólares para nós intermédios da Silicon Analysts e de 15–100 milhões de dólares para ASIC de computação em 7/6 nm da AnySilicon, porque um chip de mineração usa pouco IP e uma estrutura repetida; pressuposto meu). Com p = 1, precisa de cerca de **4,9 EH/s** a funcionar durante um ano inteiro, cerca de 5% da rede; com p = 0,5, cerca de **9,8 EH/s**, cerca de 11%; e se R voltar ao nível do verão, de cerca de $10, mesmo p = 1 precisa de cerca de 6,9 EH/s estimativa. A escala não está fora de alcance. O problema são p e T.

A terceira medida é **o risco das regras**, e é muitas vezes decisiva. O fork anti-ASIC da Monero em 2018 deixou encalhada a X3 da Bitmain e custou brevemente à rede quase metade da sua taxa de hash. O fork da Sia no mesmo ano inutilizou os mineradores da Bitmain e da Innosilicon, enquanto o preço em segunda mão do minerador Obelisk, ligado aos programadores, subiu de 2 499 dólares para mais de 5 000. Pouco mais de um mês depois do Merge da Ethereum, menos de 20% dos antigos mineradores de Ethereum tinham encontrado outro trabalho. **Quem controla a função de trabalho controla o destino de quem investe em hardware**, e a Sia mostra

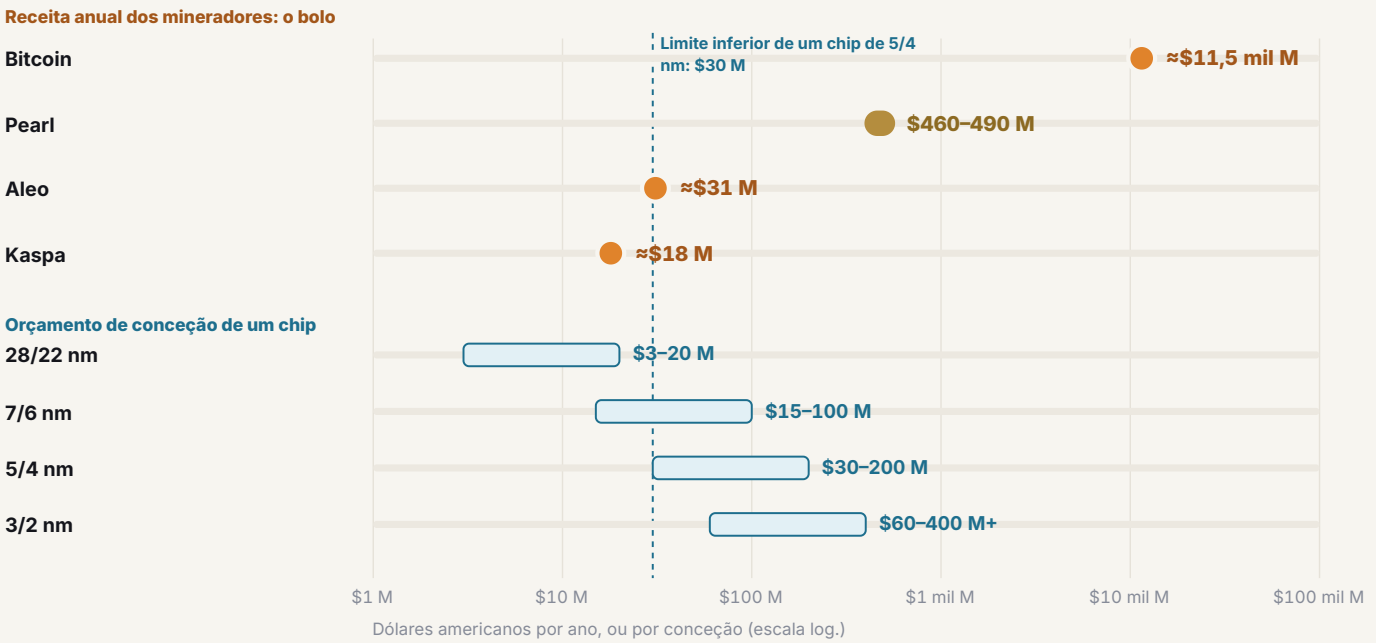
que a «resistência a ASIC» pode tornar-se o fosso defensivo dos próprios programadores. Sobrevive o hardware com segundo mercado: depois do Merge, as GPU passaram para a IA e para a nuvem, enquanto os ASIC de algoritmo único se amontoaram na Ethereum Classic. A janela também está a encolher: o primeiro ASIC da Kaspa demorou cerca de 16 meses, o da Aleo cerca de cinco, e o primeiro a chegar à fronteira da eficiência é muitas vezes um concorrente e não um incumbente.

Rede	Receita anual dos mineradores (aprox.)	Hardware marginal ou de referência	Estabilidade das regras	k disponível para um recém-chegado	Juízo
Bitcoin	11,5 mil milhões de dólares	Bitmain S23 XP Hyd (8,9 J/TH, $25,5 por TH)	Muito alta	≤1	Um bolo grande, mas o k pertence aos incumbentes
Pearl	0,46–0,49 mil milhões de dólares (aos preços atuais)	GPU próprias ou ociosas (custo marginal ≈ eletricidade) e GPU de centro de dados	Muito baixa: quatro forks em dois meses; lista branca FP8 pendente	Talvez 3–10 vezes em eficiência análise ; menos face a GPU de custo afundado	O bolo é suficientemente grande; o risco das regras veta-o
Aleo	Cerca de 31 milhões de dólares (3.º T 2025, anualizado)	IceRiver AE2 e outras máquinas dedicadas	Média: limiar de stake do ARC-46	≈1	Um bolo pequeno, e o limiar está a subir
Kaspa	Cerca de 18 milhões de dólares	KS7 (77 J/TH)	Alta	≈1	Receita anual abaixo do limite inferior de um orçamento de conceção em 5/4 nm

FIG. 06

A maioria dos bolos é menor do que o orçamento de um só chip

Receita anual dos mineradores de quatro redes face ao custo de conceção de um chip por nó de processo, escala logarítmica



Fonte: Hashrate Index (bitcoin, novembro de 2026), prlstats e Kryptex (Pearl), Blockworks (Aleo, 3.º T 2025), minerstat (Kaspa), AnySilicon (orçamentos de conceção, 2026); anualizado pelo autor.

O quadrante «bolo grande, regras estáveis e k > 1» está hoje vazio: a bitcoin tem o bolo e as regras estáveis, mas o seu k pertence aos incumbentes; o bolo da Pearl é suficientemente grande, mas as suas regras são as menos estáveis; a Kaspa e a Aleo têm regras suficientemente estáveis e bolos pequenos demais.

CAPÍTULO 08

08 CHIP A CHIP
Cinco tipos de chip, cinco respostas

PoUW ao estilo da Pearl: nenhum chip INT7 dedicado; apostar num minerador que possa servir inferência

Os argumentos a favor de um ASIC são reais. Hoje só cerca de 1% dos blocos transporta trabalho de IA, pelo que o argumento económico do whitepaper, de que um ASIC dedicado perde a capacidade de fazer trabalho útil e por isso ganha menos, está agora no seu ponto mais fraco. As partes mais caras de uma GPU, a HBM e o encapsulamento CoWoS, representam cerca de 63% do custo de fabrico de uma H100, e o die lógico apenas cerca de 9%. Um chip que não faz senão multiplicações-acumulações de inteiros de 7 bits, mantém os operandos em SRAM no chip e funciona a baixa tensão poderia plausivelmente chegar a 3 a 10 vezes a energia por operação de uma GPU do mesmo nó e a uma vantagem maior no custo por operação [análise; ainda sem medições]. Os argumentos contra pesam mais. Os prazos das foundries nos nós de ponta podem ultrapassar 50 semanas, e uma equipa nova precisa de cerca de 18 a 24 meses do arranque ao volume estimativa; quando o chip chegar, a Pearl pode bem ter passado para FP8 e depois para FP4. E, como o PIP-3 define a aritmética admitida através de uma lista branca de dispositivos, um chip INT7 deixaria simplesmente de funcionar assim que este fosse ativado.

Assim, a minha conclusão é: **não fazer agora o tape-out de um chip INT7 dedicado à Pearl.** Há duas apostas melhores. Uma é um acelerador de inferência que também possa minerar: construído em torno de GEMM INT8/FP8 geral e de operandos residentes em SRAM, para que, se o acesso à mineração for negado, o valor caia num mercado de inferência real. A outra é entrar no debate do protocolo e defender uma aritmética canónica neutra quanto ao fornecedor que um ASIC possa implementar com exatidão, juntamente com um compromisso de estabilidade do algoritmo, que é a condição prévia para que exista o quadrante «bolo grande, regras estáveis, k > 1». Até lá, uma arquitetura pode ser validada em FPGA ou em wafers multiprojeto (MPW) de nós maduros, mantendo o risco de NRE na ordem de um milhão de dólares. O capítulo 10 transforma isto num orçamento por etapas.

Mineradores ao estilo da bitcoin: os incumbentes são donos do k, e a energia é o ativo escasso

O risco das regras da bitcoin é o mais baixo, mas a concorrência empurra o k de um recém-chegado para baixo de 1. A S23 XP Hyd da Bitmain chega a 8,9 J/TH, vende-se a $25,5 por TH e é entregue a partir de novembro; a eficiência dos modelos topo de gama desceu de 17,5 J/TH da S21 em 2023 para 8,9 J/TH, cerca de 49% em três anos. Um recém-chegado a entregar em volume em 2027–2028 precisaria de cerca de 8 J/TH ou melhor para se aguentar estimativa. Os concorrentes estão a sair: a Canaan comunicou uma receita de 31,9 milhões de dólares no segundo trimestre de 2026 com um prejuízo bruto de 29,3 milhões e previu apenas 11–15 milhões para o terceiro trimestre; as vendas da Bitdeer das suas próprias máquinas a compradores externos caíram para 0,3 milhões de dólares no terceiro trimestre, com uma margem bruta do grupo de 3,4%; a Auradine mudou o nome para Velaura AI e virou-se para os chips de IA; e a Core Scientific cancelou uma encomenda de cerca de 15 EH/s de máquinas de 3 nm à unidade Proto da Block, registando um prejuízo de 41,9 milhões de dólares.

A procura também está a encolher. A 9 de novembro, o hashprice era de cerca de $31,89 por PH/s por dia, a curva a prazo até abril de 2027 tinha uma média de cerca de $32,7, e as taxas eram apenas 0,72% das recompensas de bloco. A $0,045 por kWh, a S23 XP Hyd paga-se em cerca de 3,1 anos estimativa, perto do próximo halving, esperado por volta de 2028. Mais importante é o custo de oportunidade da energia: os arrendamentos para IA e computação de alto desempenho rendem hoje cerca de 1,67–1,9 milhões de dólares por megawatt contratado por ano, o que, com 80% a 85% de resultado operacional líquido, é cerca de 1,5 a 2,9 vezes o lucro bruto da mineração por megawatt (cerca de 0,56–0,9 milhões de dólares) [estimativa; as duas bases por megawatt diferem]. A Core Scientific deixará de minerar no quarto trimestre; a Bitfarms, agora Keel, fechou todos os seus locais de mineração nos EUA; a IREN ganhou 200 milhões de dólares com a nuvem de IA e apenas 28 milhões com a mineração num trimestre. **A energia e a capacidade de centro de dados são os ativos; as máquinas de mineração são consumíveis.**

O cálculo do NRE é igualmente implacável. A eficiência de fronteira exige hoje processos da classe de 3 nm (segundo fontes secundárias, a MicroBT usa o Samsung SF3, e os chips da Auradine e da Proto são de 3 nm). As máscaras custam cerca de 10–20 milhões de dólares e o NRE completo 30–100 milhões de dólares ou mais; a $20–25 por TH e com 30% de margem bruta, cada 1 EH/s vendido rende cerca de 6–7,5 milhões de dólares de lucro bruto, pelo que cobrir uma I&D anual à escala da Bitdeer, cerca de 110–160 milhões de dólares, exige vender cerca de 20–25 EH/s por ano a compradores externos, cerca de 2% da rede estimativa. A conclusão: **uma equipa nova não deve entrar no mercado de mineradores de bitcoin para venda.** Três casos continuam a fazer sentido: a integração vertical com energia barata própria; formatos de nicho como os mineradores domésticos e de aquecimento (a Canaan também está a lançar produtos de aquecimento doméstico); e a participação nos roteiros dos incumbentes através de IP e serviços de conceção.

Chips de treino de IA: não para uma equipa pequena

Nos chips de treino, o k é o mais difícil de conquistar. Na mesma geração e precisão, o Ironwood da Google supera a B300 em eficiência FP8 ao nível do chip por apenas cerca de 1,1 a 1,8 vezes (a potência do Ironwood é uma estimativa de um jornalista) estimativa; os 30 a 80 vezes muitas vezes citados do artigo do TPU v1 foram medidos face a uma K80, que não tinha tensor cores. A verdadeira alavanca é o formato numérico: passar de FP8 para FP4 dá normalmente 2 a 3 vezes o débito no mesmo silício, e a NVIDIA é quem avança mais depressa nesse campo: reivindica 50 PF de inferência NVFP4 para a Rubin, em volume no segundo semestre de 2026. Os limiares de custo e de abastecimento são ainda mais altos: os orçamentos de conceção em 3/2 nm vão de 60–400 milhões de dólares ou mais; a capacidade CoWoS em 2026 é de cerca de 106 000–130 000 wafers por mês, e a TSMC estará a ponderar aumentos de preço de até 20%; a um preço contratual de $15,6 por GB para a HBM3E, só a HBM de um acelerador de 288 GB custa cerca de 4 500 dólares estimativa. Este mercado pertence ao codesenvolvimento com os hiperescaladores ao estilo da Broadcom (capítulo 11). A conclusão: **as equipas pequenas não devem construir chips de treino autónomos**; construam antes IP, chiplets ou serviços de conceção.

Chips de inferência de IA: trocar generalidade por k, e compensa

A inferência é onde os ASIC têm mais margem, porque o k pode ser comprado abdicando da generalidade. A Etched afirma que o seu servidor de oito chips executa o Llama-70B a mais de 500 000 tokens por segundo, cerca de 20 vezes oito H100; a Taalas grava os pesos do modelo em ROM de máscara e afirma 10 vezes menos consumo e 20 vezes menos custo de fabrico (ambos alegação do fornecedor). As arquiteturas centradas em SRAM evitam a HBM e o CoWoS, e os seus nós são mais acessíveis: a Taalas usa o TSMC 6 nm, e a Groq passou de 14 nm para o Samsung 4 nm. A distribuição cripto não ajuda os chips de inferência, já que as redes de computação descentralizada funcionam quase inteiramente com NVIDIA; a porta de entrada é a economia de compras das neoclouds e o grande mercado interno de inferência da China, onde os centros de dados financiados pelo Estado têm de usar chips de IA nacionais, as importações de H200 estão a ser restringidas e Pequim quer que as grandes empresas de internet façam a inferência em chips nacionais. A conclusão: **vale a pena construir**, desde que

haja um cliente claro e o grau de especialização em modelos e arquitetura seja uma escolha deliberada. É também o melhor segundo mercado para um chip de mineração PoUW.

Aceleradores ZK e FHE: cobrir-se no ZK, posicionar-se cedo no FHE

As GPU já alcançaram a prova ZK em tempo real: o SP1 Hypercube da Succinct provou 99,7% dos blocos da Ethereum em menos de 12 segundos em 16 RTX 5090. A Cysic afirma que o seu chip ZK C1 é 10 a 100 vezes mais eficiente do que os provadores em GPU e que será entregue em 2026, mas a sua procura vem em grande parte das recompensas do seu próprio token alegação do fornecedor; não encontrei indícios públicos de que os ASIC do puzzle da Aleo sejam usados para prova ZK real. A prova ZK é inteira e determinística, de forma próxima à do trabalho dos chips de mineração, mas os sistemas de prova mudam depressa (SP1, RISC Zero, Halo2, Plonky2 e outros). A conclusão: **esperar e cobrir-se**, com motores programáveis de MSM, NTT e hash, primeiro em FPGA, depois como ASIC.

O FHE é a primitiva mais dependente do hardware e com maior potencial. O Heracles da Intel, em Intel 3, com 197 mm², 176 W e 48 GB de HBM3, executa operações FHE 1 074 a 5 547 vezes mais depressa do que um Xeon de 24 núcleos, mas a Intel não tem planos comerciais e o responsável do projeto saiu em dezembro de 2025. A Niobium está a construir em Samsung 8LPU com a SemiFive como parceira de conceção (um contrato de desenvolvimento de cerca de 6,86 milhões de dólares) e angariou mais de 23 milhões de dólares; a Optalysys, com uma abordagem fotónica, angariou 31 milhões de dólares e diz que um chip comercial completo ainda está a dois ou três anos de distância; e o design FPGA de código aberto da Zama chega a cerca de 13 000 bootstraps programáveis por segundo a cerca de 200 W, a referência pública que qualquer ASIC tem de superar claramente. A conclusão: **posicionar-se cedo e validar com poucos ativos**, para uma janela por volta de 2027–2029. A procura deverá ancorar-se mais na IA cifrada para empresas e governos do que nas criptomoedas propriamente ditas, e o FHE precisa de HBM, pelo que a sua estrutura de custos se assemelha à de um chip de treino.

Tipo de chip	De onde vem o k	Procura e bolo	NRE e limiares	Risco de regras ou de especialização	Segundo mercado	Juízo
Chip de mineração PoUW tipo Pearl	Abdicar da HBM, do CoWoS e das unidades de uso geral: talvez 3–10 vezes em J/op análise	Rendimento dos mineradores cerca de 0,46–0,49 mil milhões de dólares por ano, volátil	Matriz repetitiva em N7/N6 cerca de 10–30 milhões de dólares análise	Muito alto: o protocolo decide o formato numérico e a lista branca de dispositivos	Inferência de IA, se construído como GEMM INT8/FP8 geral	Nenhum chip INT7 dedicado; construir para inferência e mineração, ou esperar pela aritmética canónica
Chip de mineração SHA-256 de bitcoin	k ≤ 1 face à Bitmain	Cerca de 11,5 mil milhões de dólares por ano, mercado de venda dominado por duas empresas	Máscaras da classe de 3 nm 10–20 milhões de dólares; NRE completo 30–100 milhões de dólares ou mais	Baixo (algoritmo estável); concorrência elevada e risco de halving	Quase nenhum (algoritmo único)	Os recém-chegados ficam fora do mercado de venda; só energia própria, formatos de nicho ou IP

Tipo de chip	De onde vem o k	Procura e bolo	NRE e limiares	Risco de regras ou de especialização	Segundo mercado	Juízo
Chip de treino de IA	Matrizes sistólicas da mesma geração só 1,1–1,8 vezes; o formato é a alavanca	O maior (Broadcom com cerca de 58 mil milhões de dólares de receita de semicondutores de IA no ano fiscal de 2026)	Conceção em 3/2 nm 60–400 milhões de dólares ou mais; CoWoS e HBM	Cadência anual da NVIDIA	—	As equipas pequenas fazem IP, chiplets e serviços de conceção, não chips autónomos
Chip de inferência de IA	Abdicar da generalidade: 10–20 vezes alegado alegação do fornecedor	Grande e fragmentada; o mercado interno da China continental é grande por si só	Arquiteturas SRAM em N6/N7–N4 são acessíveis	Especialização em modelos e arquitetura	É o mercado principal	Vale a pena construir: assinar o cliente, depois escolher a especialização
Acelerador ZK	10–100 vezes as GPU, alegado alegação do fornecedor	Pequena, movida sobretudo por recompensas em tokens	N7/N6 depois N5; FPGA primeiro	Os sistemas de prova mudam depressa	Limitado fora do ZK	Esperar e cobrir-se: motores MSM e NTT programáveis
Acelerador FHE	1 074–5 547 vezes um Xeon (Heracles)	Ainda por formar; ancorada em empresas e governos	Precisa de HBM; estrutura de custos próxima dos chips de treino	Normas e ecossistema indefinidos	Computação cifrada, IA privada	Posicionar-se cedo, validar com poucos ativos; janela 2027–2029

FIG. 07 **Cinco tipos de chip, cinco respostas**

De onde vem a vantagem de custo k, em que mercado o silício pode ser vendido se as regras mudarem, e o veredicto



Chip de mineração PoUW tipo Pearl

AGORA NÃO

k: sem HBM nem unidades gerais: 3–10× por op. (inferido)

Segundo mercado: inferência de IA, se GEMM INT8/FP8 geral

Apostar num chip de inferência que possa minerar, não numa peça INT7 dedicada



Chip de mineração SHA-256 de bitcoin

FICAR DE FORA

k: ≤1 face aos incumbentes

Segundo mercado: quase nenhum

Só com energia própria, um formato de nicho ou como IP



Chip de treino de IA

IP, NÃO CHIPS

k: 1,1–1,8× à mesma precisão

Segundo mercado: —

Decidem o encapsulamento e a HBM; fazer IP, chiplets, serviços



Chip de inferência de IA

CONSTRUIR

k: abdicar da generalidade: 10–20× alegado

Segundo mercado: é o mercado principal

Assinar o cliente, depois escolher a especialização



Acelerador de provas ZK

ESPERAR E COBRIR

k: 10–100× as GPU, alegado

Segundo mercado: limitado fora do ZK

Motores MSM e NTT programáveis; FPGA primeiro



Acelerador FHE

POSICIONAR CEDO

k: 1 074–5 547× um Xeon (Heracles)

Segundo mercado: computação cifrada, IA privada

Uma janela 2027–2029; validar com poucos ativos

Fonte: Síntese do autor dos capítulos 7 e 8; valores dos fornecedores assinalados como tal no relatório.

CAPÍTULO 09



NÓS DE PROCESSO

09 Nós de processo: deixar que o maior custo ao longo da vida escolha

A escada de custos: as wafers de vanguarda custam dez vezes as maduras

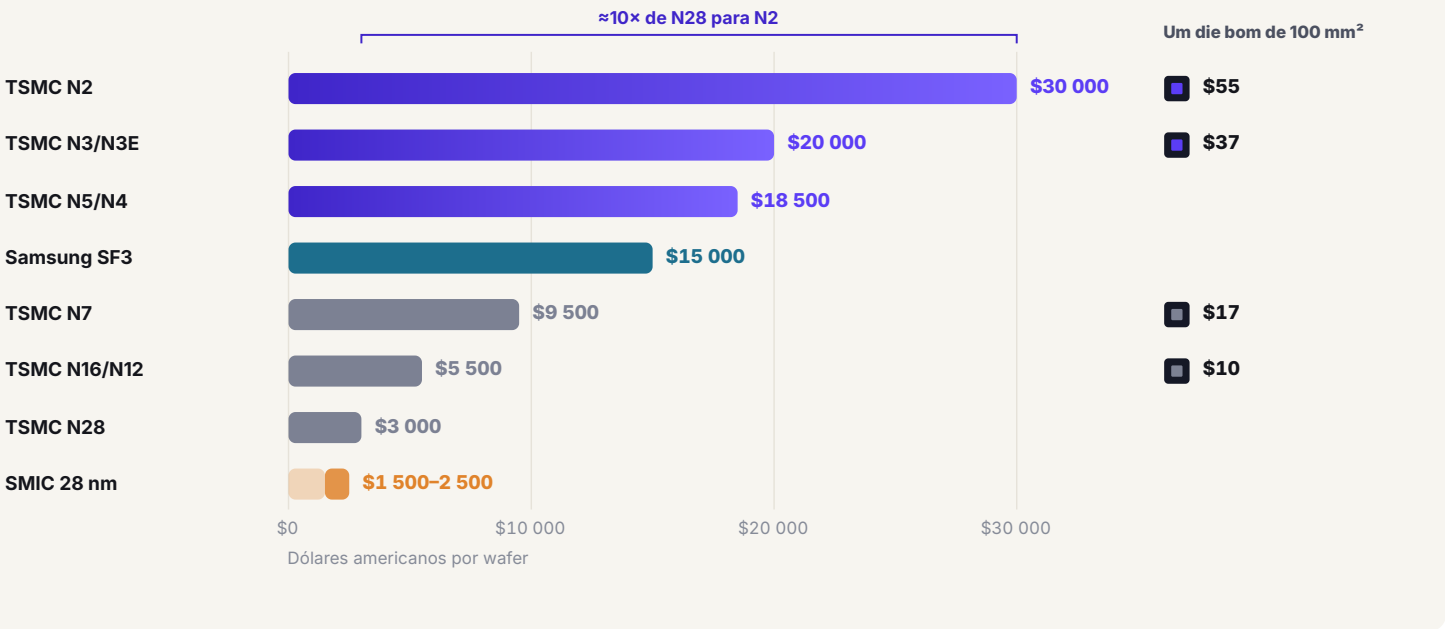
Comece-se pelas contas. Uma wafer de vanguarda custa hoje cerca de dez vezes uma madura: TSMC N2 cerca de 30 000 dólares, N3 cerca de 20 000, N7 cerca de 9 500 e N28 cerca de 3 000; Samsung SF3 cerca de 15 000, cerca de 25% abaixo do N3; e o custo de fabrico da SMIC na classe de 7 nm estará 40% a 50% acima do da TSMC. A TSMC deverá subir os preços mais 3% a 6% a partir de janeiro de 2027, sobretudo para 2 nm e 3 nm. Os custos de máscaras e de conceção diferem ainda mais, e as fontes divergem, pelo que a tabela mostra ambos.

Nó	Preço da wafer (estimativa)	Conjunto de máscaras (valor do modelo / intervalo de referência)	Orçamento de conceção (AnySilicon)	MPW (por projeto)
TSMC N2	Cerca de 30 000 dólares	—	3/2 nm: 60–400 milhões de dólares ou mais	—

Nó	Preço da wafer (estimativa)	Conjunto de máscaras (valor do modelo / intervalo de referência)	Orçamento de conceção (AnySilicon)	MPW (por projeto)
TSMC N3/N3E	Cerca de 20 000 dólares (18 000–27 000)	15 milhões de dólares / 10–20 milhões	Idem	—
Samsung SF3	Cerca de 15 000 dólares (13 000–17 000)	12 milhões de dólares	—	—
TSMC N5/N4	Cerca de 18 500 dólares (16 000–21 300, baixa confiança)	6,5 milhões de dólares / 10–20 milhões	5/4 nm: 30–200 milhões de dólares	—
TSMC N7	Cerca de 9 500 dólares (8 100–10 900)	3,5 milhões de dólares / 5–10 milhões	7/6 nm: 15–100 milhões de dólares	Comercial 100 000–500 000 dólares
TSMC N16/N12	Cerca de 5 500 dólares	1,8 milhões de dólares	16/12 nm: 6–40 milhões de dólares	Académico cerca de 35 000 dólares por mm²
TSMC N28	Cerca de 3 000 dólares	1 milhão de dólares / 1–3 milhões	28/22 nm: 3–20 milhões de dólares	Comercial 50 000–150 000 dólares; académico cerca de 12 000 dólares por mm²
SMIC 28 nm	1 500–2 500 dólares	—	—	—
SMIC classe de 7 nm (N+2/N+3)	40–50% acima da TSMC	—	—	—

FIG. 08 **Uma wafer de vanguarda custa dez vezes uma madura**

Preço estimado por wafer de 300 mm por nó, e o custo de um die bom de 100 mm² com 85% de rendimento



Fonte: Preços de wafers da Silicon Analysts (estimativas, N5 de baixa confiança), Wccftech (SMIC); custos por die calculados pelo autor com um rendimento assumido.

Depois, a física. As medições clássicas de Horowitz mostram que uma instrução programável custa cerca de 70 pJ, enquanto a aritmética em si custa alguns picojoules: **a aritmética é barata e o sobrecusto é caro**, e é por isso que existem as matrizes sistólicas e os chips dedicados. Reduzir para metade a dimensão mínima reduz aproximadamente para metade a energia por operação, mas a tensão de alimentação estagnou por volta de 0,9–1,1 V.

Uma wafer N2 custa 50% a 67% mais do que uma N3, pelo que, a menos que a densidade suba mais de cerca de 1,5 vezes, o custo por transístor sobe em vez de descer estimativa: os nós de ponta têm de se pagar pela eficiência e pelo desempenho. Para um die de 100 mm² com 85% de rendimento, o custo de cada die bom é de cerca de 55 dólares em N2, 37 em N3, 29–34 em N5, 17 em N7 e 10 em N16 [estimativa; o rendimento é um pressuposto].

Escolher um nó por tipo de chip e por equipa

A minha regra é: **deixar que a maior parcela do custo do chip ao longo da sua vida escolha o nó.** Para chips de computação que funcionam a plena carga dia e noite (mineração, ZK), a eletricidade domina o custo ao longo da vida, pelo que, quando o algoritmo estiver estável, devem passar para N5/N4 e N3/SF3; enquanto o algoritmo estiver por definir ou faltar dinheiro, valide-se primeiro em MPW de nós maduros. Para chips de treino, o estrangulamento é o encapsulamento e a memória (a HBM e o CoWoS representam cerca de 63% da lista de materiais de uma H100), pelo que o nó é secundário. Para chips de inferência, a primeira decisão é a arquitetura de memória: os designs centrados em SRAM podem ficar entre N6/N7 e N4, enquanto os designs com HBM herdam a estrutura de custos dos chips de treino. O FHE está limitado pela largura de banda da memória e não pode evitar a HBM.

Tipo de chip	Startup ou equipa pequena	Empresa de escala	Via da China continental	Pontos-chave
Chip de mineração PoUW por multiplicação de matrizes (residente em SRAM, sem HBM)	Protótipo em MPW de 28 nm ou 16/12 nm; primeiro volume em N7/N6; N5/N4 quando o algoritmo e a receita estiverem estáveis; dies de 50–150 mm² em encapsulamentos flip-chip normais	N3/N3E ou Samsung SF3	SMIC N+1/N+2 serve (não precisa de HBM nem de CoWoS), com cerca de 40–50% mais por die bom	Domina a energia; as regras do protocolo são o maior risco
Chip de mineração SHA-256	Não perseguir a vanguarda; os formatos de nicho podem usar nós maduros	Classe de 3 nm; os incumbentes passam do TSMC 5 nm para 2 nm e do Samsung SF3 para 2 nm (fontes secundárias)	Difícil igualar a eficiência dos nós de ponta	O preço da energia e o ciclo do halving decidem as margens
Chip de treino de IA	Nenhum chip autónomo: IP, chiplets, serviços de conceção	N3P e depois N2/A16; 2–8 chiplets; garantir CoWoS e HBM cedo	SMIC N+2/N+3 com HBM nacional (o modelo da Huawei); FP8 por chip cerca de 0,2–0,5 vezes o da NVIDIA, compensado pela escala do sistema	O encapsulamento e a HBM são o estrangulamento
Chip de inferência de IA	Centrado em SRAM ou com pesos gravados, N6/N7 ou N4/N5/SF4; só INT8, FP8 e FP4	N3/N2 com HBM (contexto longo, mistura de especialistas), ou SRAM multichiplet	SMIC N+2 com LPDDR ou GDDR; a memória da classe HBM é o ponto de estrangulamento	O grau de especialização da arquitetura é a decisão central
Acelerador ZK	Começar em FPGA ou MPW de 16/12 nm; depois N7/N6 e N5	N5/N3	SMIC N+1/N+2 serve	Inteiro e determinístico; fácil de verificar
Acelerador FHE	FPGA (face à referência de código aberto) ou MPW de 16/12 nm; o volume exige HBM	N5/N3 com HBM	Limitado pelo acesso à HBM	A largura de banda da memória, e não a aritmética, dita a velocidade

Avançado ou maduro, quatro tipos de foundry, e chiplets

Opção	Principais pontos fortes	Principais pontos fracos	Mais indicado para
Nós de ponta (N3/N2, SF3/SF2, Intel 18A)	Cada nó completo reduz aproximadamente para metade a energia por operação; SRAM densa	Wafers 3–10 vezes, máscaras 5–15 vezes e orçamentos de conceção 10–20 vezes os dos nós maduros; prazos acima de 50 semanas; em N2 o custo por transístor pode subir	Chips de computação com algoritmos estáveis a funcionar dia e noite; chips de treino
Nós maduros e intermédios (N7/N6, N16/12, N28)	Máscaras na ordem do milhão de dólares, MPW barato, iteração rápida, muitas foundries (TSMC, UMC, GlobalFoundries, SMIC)	Eficiência pouco competitiva a plena carga dia e noite	Protótipos, mercados de nicho, inferência na periferia
TSMC	O melhor rendimento e ecossistema	Os preços mais altos; capacidade N2/N3 apertada; longas filas para o CoWoS	Equipas que perseguem a fronteira da eficiência com compromissos de volume
Samsung Foundry	Wafers cerca de 20–25% mais baratas (SF3 cerca de 15 000 dólares contra cerca de 20 000 para N3); usada pela segunda geração da Groq e por outros	O negócio de foundry está a perder dinheiro, com o objetivo de atingir o ponto de equilíbrio no fim de 2027	Chips de computação sensíveis ao custo
Intel Foundry	Chips complexos como o Heracles comprovados em Intel 3	Sem preços externos públicos nem dados de clientes para o 18A; o negócio de foundry está a perder dinheiro	Avaliar ecossistema e apoio caso a caso
Foundries da China continental (SMIC e outras)	Perto de um grande mercado interno; wafers de 28 nm a 1 500–2 500 dólares	Capacidade da classe de 7 nm cerca de 45 000–60 000 wafers por mês, em grande parte absorvida pela Huawei; rendimentos cerca de um terço dos da TSMC; custos 40–50% mais altos; só DUV	Chips de inferência, mineração e ZK para o mercado interno
Chiplets	Superam os limites do retículo e do rendimento; padrão para sistemas com HBM (MI355X, Ironwood, Trainium3)	Custo de encapsulamento mais alto e dependência do CoWoS; o teste de known-good die acrescenta custo	Chips de treino e de inferência grandes; desnecessários para pequenos chips de mineração ou ZK

CAPÍTULO 10

ORÇAMENTO POR ETAPAS

10 Do modelo ao silício: um orçamento por etapas para um primeiro chip

Comprar provas antes das máscaras

Um chip full-custom é uma série de apostas, cada uma maior do que a anterior, e a disciplina consiste em comprar provas antes de comprar máscaras. Uso quatro etapas análise. A etapa 0 é um modelo: um estudo de arquitetura que estima J/op e $/op a partir da aritmética, da SRAM e da E/S, face ao hardware de referência do capítulo 7; custa semanas e dezenas de milhares de dólares. A etapa 1 é um protótipo em FPGA que executa a carga real de ponta a ponta, shares de mineração ou tokens de inferência, e prova o datapath e a integração com o protocolo; custa alguns meses e algumas centenas de milhares de dólares. A etapa 2 é uma wafer multiprojeto: um lugar comercial custa cerca de 50 000–150 000 dólares em N28 e 100 000–500 000 dólares em N7, e o silício medido substitui as estimativas de energia do modelo. A etapa 3 é o conjunto completo de máscaras e o primeiro lote de produção: máscaras de cerca de 1 milhão de dólares em N28, 3,5 milhões em N7 e 15 milhões em N3, dentro de um orçamento de conceção de 15–100 milhões de dólares em 7/6 nm. Cada etapa precisa de um número que pare o projeto, como um k abaixo de 3 face ao hardware marginal, ou uma taxa de hash de retorno h* que a janela não consiga atingir.

FIG. 09 **Comprar provas antes das máscaras: um orçamento por etapas para um primeiro chip**

Despesa acumulada em dólares americanos de um projeto N7/N6 em quatro etapas, escala logarítmica; cada etapa tem um número que para o projeto



Fonte: Preços de wafers, MPW e máscaras como no capítulo 9 (Silicon Analysts, AnySilicon, valores de modelo); calendário e despesa das etapas são pressupostos ilustrativos do autor.

A escada é íngreme. Num projeto em N7/N6, as três primeiras etapas custam no conjunto cerca de 1,5 milhões de dólares e cerca de onze meses; a quarta custa mais de dez vezes isso estimativa. É por isso que a ordem importa mais do que o nó: se o protocolo mudar a sua aritmética, ou se o comprador desistir, uma equipa que parou na etapa 2 perde o preço de um protótipo, não o de um tape-out.

Quanto custa cada chip, e o que tem de ser verdade

Projeto	Etapa 2 (protótipo)	Etapa 3 (nó e orçamento de conceção)	Um die bom de 100 mm²	O que tem de ser verdade para haver retorno
Acelerador de inferência que possa minerar	MPW de 16/12 nm ou N7, 0,1–0,5 milhões de dólares	N6/N7, cerca de 15–30 milhões de dólares estimativa	Cerca de 17 dólares em N7, antes do encapsulamento e da memória	Um cliente de inferência assinado suporta a maior parte do NRE; o rendimento da mineração é um extra, não o cenário base
Minerador INT7 dedicado à Pearl (não recomendado agora)	MPW de 28 nm, 0,05–0,15 milhões de dólares	N7/N6, cerca de 20 milhões de dólares	Cerca de 17 dólares em N7	Cerca de 4,9 EH/s durante um ano inteiro com p = 1, ou 9,8 EH/s com p = 0,5 (capítulo 7)
Acelerador ZK	FPGA, depois MPW de 16/12 nm	N7/N6, depois N5	Cerca de 10 dólares em N16, 17 em N7	Os sistemas de prova estabilizam em kernels MSM e NTT que um design fixo possa servir
Acelerador FHE	FPGA face à referência de código aberto	N5/N3 com HBM, 30 milhões de dólares ou mais	Dominado pela HBM e pelo encapsulamento	Um comprador de IA cifrada em volume, que ainda não existe

Daqui decorrem dois pontos comerciais. Primeiro, a eficiência mais barata é arquitetural: manter os operandos em SRAM e deixar de fora a HBM e o CoWoS elimina a maior parte do custo de uma GPU (cerca de 63% da lista de materiais de uma H100) antes de se escolher qualquer nó. Segundo, o tempo é um custo: uma equipa nova precisa de cerca de 18 a 24 meses do arranque ao volume e os prazos na vanguarda ultrapassam 50 semanas, pelo que um protocolo que muda a sua aritmética de poucos em poucos meses pode tornar um chip obsoleto antes de a etapa 3 se pagar.

Três regras para uma equipa pequena

No conjunto, estes números dão três regras análise. Primeiro, **ficar na etapa 1 ou 2 até haver um comprador identificado ou uma versão de protocolo congelada**; nem o preço do token nem a taxa de hash substituem qualquer um deles. Segundo, **comprar k primeiro com a arquitetura e só depois com o nó**: uma matriz residente em SRAM em N7 pode superar uma peça de uso geral em N3 para uma carga concreta, por uma fração do custo das máscaras. Terceiro, **escrever a taxa de hash ou o volume de retorno, com p, antes do tape-out**, e fixar os pagamentos por marcos face a uma versão de protocolo fixa, para que seja o cliente, e não a empresa de conceção, a suportar o risco das regras.

CAPÍTULO 11

11 OPORTUNIDADES
Onde estão as oportunidades, e o que eu faria

Chips à medida: a procura vem esmagadoramente da IA

Em 2026, a procura de chips à medida vem esmagadoramente da IA. A receita de semicondutores de IA da Broadcom no terceiro trimestre do seu ano fiscal de 2026 foi de 16,7 mil milhões de dólares, mais 221% em termos homólogos, cerca de 58 mil milhões de dólares no ano, 73% dos quais de XPU para seis clientes (a partir de uma transcrição automática da chamada de resultados). A receita da Global Unichip no segundo trimestre subiu 128% em termos homólogos, com a receita de produção chave na mão a subir 188% e uma margem bruta de 21,5%. A repartição da receita da Alchip por nó no primeiro trimestre foi de 21% em 3/2 nm, 63% em 7/5 nm e 16% em 12 nm e acima. As novas encomendas da VeriSilicon para 2026 chegaram a 8,24 mil milhões de RMB (cerca de 1,16 mil milhões de dólares) no fim de abril, 91,37% das quais para computação de IA. O mercado tem três níveis: no topo, codesenvolvimento plurianual no valor de milhares de milhões; no meio, projetos de 7 nm a 2 nm com NRE de alguns milhões a mais de cem milhões de dólares (5–30 milhões de dólares em nós intermédios, 30–100 milhões de dólares ou mais na vanguarda); e uma cauda longa de ASIC em nós maduros com NRE de 0,5–5 milhões de dólares. Os clientes de cripto e de mineração são um nicho pequeno e cíclico, mas são hoje o «primeiro cliente» mais rápido e sem necessidade de autorização a pagar computação.

As capacidades de uma equipa de chips de mineração que melhor se transferem, operação a baixa tensão próxima do limiar, matrizes de computação replicadas em massa, alimentação e refrigeração da classe dos quilowatts e tape-outs rápidos em nós de ponta, são exatamente o que precisam os chips de inferência e os aceleradores ZK e FHE. Os mercados de capitais atribuem-lhes preço: a Velaura AI (antiga Auradine) angariou 110 milhões de dólares em agosto com uma avaliação acima de 1 mil milhões de dólares. O contraexemplo é a unidade de chips de IA na periferia da Canaan, que teve uma receita de 0,9 milhões de dólares no ano fiscal de 2024 enquanto consumia 15% dos custos operacionais, e foi encerrada em junho de 2025. A diferença está numa via própria para o mercado, não na reutilização dos canais da mineração.

Desenvolvimento de IA: as horas são uma mercadoria; o valor está nas junções

Na gama baixa, os serviços de desenvolvimento de IA estão banalizados: os preços típicos na Clutch são de 25–49 dólares por hora, e 50–99 nos Estados Unidos, no Canadá e em mercados semelhantes; uma equipa de desenvolvimento de blockchain custa cerca de 15 000 a mais de 90 000 dólares por mês. Vender horas não tem fosso defensivo. O valor está nas junções entre domínios: kernels e compiladores para chips à medida, mineração embutida no serviço de inferência (o plug-in vLLM da Pearl é um exemplo), inferência verificável que combina TEE com verificações pontuais ZK, e pagamentos e identidade para agentes.

Sistemas IA × cripto: assinar o comprador de computação antes do token

Para quem concebe uma cadeia nativa de IA ou uma rede de computação, a caixa de ferramentas de verificação de 2026 já ganhou forma. Por ordem de custo por unidade de trabalho, do mais baixo para o mais alto: amostragem estatística (a prova de amostragem da Hyperbolic), delegação arbitrada (o Verde da Gensyn, que só precisa de um verificador honesto mas exige execução exata ao bit), TEE (velocidade quase nativa, mas confiança no fabricante do chip), redundância total e zkML geral. O design prático é em camadas: TEE ou amostragem para

baixa latência, jogos de disputa ou verificações pontuais ZK para a finalidade; um artigo de dezembro de 2025 simulou computação confidencial em H100 com provas de fraude e verificações pontuais ZK aleatórias a 99% do débito centralizado e um custo marginal de $0,07 por consulta. A filosofia de hardware decide a segurança e a distribuição: uma PoUW nativa de GPU é protegida, de início, por GPU de nuvem alugáveis, pelo que o custo de um ataque fica ligado aos preços de aluguer, e a Qubic mostra que uma prova de trabalho em hardware de uso geral pode ser capturada por grosso por incentivos externos.

O custo e o tempo não são a barreira. A Pearl passou do artigo (abril de 2025) à mainnet em cerca de 12 meses, num fork do btcd, com nó completo, carteira, cliente leve e minerador no lançamento. A maioria das auditorias DeFi custa 25 000–100 000 dólares, as auditorias empresariais ou multicadeia 150 000–250 000 dólares ou mais, os circuitos ZK 80% a 120% mais, e os projetos com valor real em risco orçamentam 150 000–500 000 dólares por ano para segurança. As verdadeiras barreiras são a distribuição e a liquidez. Fundadores relatam que a Binance pediu 0% a 3,5%, e alegadamente 8%, da oferta do token para uma listagem (a Binance nega lucrar com as listagens); todos os seis projetos analisados pela Cointelegraph que usaram acordos de market making de «empréstimo de tokens mais opção de compra» caíram, e o market maker da Movement vendeu cerca de 38 milhões de dólares de tokens no dia da listagem. As alternativas são a liquidez na cadeia, como na Hyperliquid, que ultrapassou 10 mil milhões de dólares de valor de mercado sem listagem na Binance e cujos mercados perpétuos HIP-3 ultrapassaram 3,2 mil milhões de dólares de interesse em aberto em junho, e as vendas públicas através de grandes bolsas, como quando a Monad angariou 269 milhões de dólares através da Coinbase. Um valor totalmente diluído elevado com uma oferta em circulação baixa foi o indicador mais claro de queda após o lançamento em 2025–2026.

O design do token faz parte do design do sistema. A emissão deve seguir a procura paga e não antecipá-la: as redes deste relatório com os múltiplos de subsídio mais saudáveis ligam as recompensas a trabalho que um comprador externo paga, e as piores pagam taxa de hash ou stake sem nada por trás. Uma regra prática é publicar, antes do lançamento, o múltiplo de subsídio que a rede pretende atingir no primeiro, segundo e terceiro ano, e cortar a emissão quando falhar a meta análise.

Direção	Quem compra, e disponibilidade para pagar	Janela	Capacidades-chave	Principal risco
Acelerador de inferência que possa minerar	Neoclouds, implementações privadas, redes PoUW	2027–2028 (consoante os forks FP8 e FP4)	GEMM INT8, FP8 e FP4; arquitetura SRAM; integração com o protocolo	Admissão pelo protocolo, especialização da arquitetura
Chip de inferência SRAM ou com pesos gravados	Neoclouds, centros de dados na China continental, periferia	Agora	Compiladores e adaptação de modelos; conceção em N6–N4	Os modelos mudam depressa demais
Chip de mineração (de nicho ou verticalmente integrado)	Mineradores com energia própria; mineradores domésticos e de aquecimento	Agora, limitado pelo halving e pelo hashprice	Matrizes de baixa tensão, eficiência, cadeia de abastecimento em volume	Guerras de preços dos incumbentes, preços da energia
Aceleração de prova ZK	Redes de provadores, rollups	Esperar até os sistemas de prova convergirem	MSM e NTT programáveis	A procura depende de recompensas em tokens
Aceleração FHE	IA cifrada para empresas e governos; cadeias de privacidade	2027–2029	NTT, sistemas com HBM, engenharia criptográfica	Procura ainda por formar; ciclos de venda longos

Direção	Quem compra, e disponibilidade para pagar	Janela	Capacidades-chave	Principal risco
Reforço de TEE e inferência verificável	Redes de inferência confidencial, DePIN	Agora	Cifra de memória com integridade, raízes de atestação independentes, verificações pontuais ZK	Dependência dos ecossistemas dos fabricantes
Cadeia nativa de IA ou rede de computação	Compradores de inferência, mineradores, programadores	Agora	Design da verificação, design da emissão, distribuição	Queda após o lançamento, liquidez do token

Três caminhos que eu seguiria, cada um com um primeiro passo

Se tivesse de escolher, seguiria no máximo três caminhos, cada um com um pequeno primeiro passo que o teste antes de comprometer dinheiro análise.

- **Caminho A: um acelerador de inferência que também possa minerar.** Primeiro passo: um estudo de arquitetura de quatro semanas que modele J/op e $/op para uma matriz GEMM INT8/FP8 residente em SRAM face à RTX 5090 e à H100, tanto no trabalho da Pearl como na inferência de transformers, seguido de validação em FPGA ou MPW num nó maduro com um NRE na ordem de um milhão de dólares.
- **Caminho B: serviços de silício à medida com a cripto como primeiro cliente e a IA como segundo mercado.** Primeiro passo: transformar o modelo de NRE e de ponto de equilíbrio deste relatório num estudo de viabilidade pago para projetos de chips de computação, com pagamentos por marcos e aceitação face a uma versão de protocolo fixa.
- **Caminho C: uma rede nativa de IA que assine o seu comprador de computação antes do seu token.** Primeiro passo: uma nota de conceção sobre verificação em camadas (TEE mais amostragem mais verificações pontuais ZK) e sobre uma aritmética canónica neutra quanto ao fornecedor, escrita com um comprador de inferência identificado antes de qualquer design de token.

CAPÍTULO 12



12

SINAIS A SEGUIR

O que seguir, de novembro de 2026 a 2027

Quando	O que seguir	Porque importa
A qualquer momento	A altura de ativação FP8 da Pearl	Decide se o hardware inteiro e as placas não NVIDIA continuam admitidos, e se a cadeia se divide
Mensalmente	A parte dos blocos da Pearl com trabalho de IA	O melhor teste isolado de se o «útil» se está a tornar verdade
Mensalmente	Se os fornecedores de inferência repassam o rendimento da mineração aos clientes	A procura paga a encontrar a emissão: o modelo de que a Pearl precisa
Trimestralmente	A receita verificável da Bittensor face às suas emissões	Se a maior rede de trabalho útil está a fechar o seu défice de subsídio
Trimestralmente	A parte dos fluxos x402 efetivamente iniciados por agentes	Se os pagamentos de agentes passam da sinalização ao comércio
Janeiro de 2027	A subida de preços de 3% a 6% da TSMC noticiada, sobretudo para 2 nm e 3 nm	O custo dos tape-outs em nós de ponta
Primeiro trimestre de 2027	FP4 na Pearl, segundo o AMA dos fundadores	Mais uma deslocação do trabalho para os formatos mais recentes da NVIDIA
Fim de 2027	O objetivo da Samsung Foundry de atingir o ponto de equilíbrio	Se uma segunda fonte mais barata se mantém saudável
Abril de 2028	O próximo halving da bitcoin	Reduz para metade o bolo dos mineradores SHA-256

CAPÍTULO 13

13 PERGUNTAS A FAZER
Perguntas que não fez mas devia ter feito

As perguntas abaixo são as que eu faria antes de comprometer dinheiro ou silício em qualquer projeto IA × cripto.

Quem controla a função de trabalho? Se uma pequena equipa pode mudar a aritmética numa altura fixada no código, a vida do seu hardware é decisão dela. Peça um compromisso de estabilidade por escrito, ou conceba hardware com um segundo mercado.

Qual é o múltiplo de subsídio? Divida as emissões (ou o valor de mercado) pela receita paga de fora da rede. Um múltiplo acima de 10 significa que a rede está a comprar atividade, não a vender um serviço.

Quanto custa a próxima etapa, e que número o faz parar? Uma wafer multiprojeto custa dezenas ou centenas de milhares de dólares; um conjunto de máscaras custa milhões. Decida o número de paragem antes de gastar.

Qual é a taxa de hash de retorno, com p? Divida o NRE por R × (1 − 1/k) × T × p. Se a resposta for uma grande parte da rede, o chip é uma aposta nas regras, não no silício.

Qual é o segundo mercado do chip? Se o protocolo fizer um fork ou o token cair, em que outro mercado pode o silício ser vendido? As GPU sobreviveram ao Merge da Ethereum porque tinham um; os ASIC de algoritmo único não.

O valor fica no token ou no capital? Em 2026, as empresas IA × cripto rentáveis são sobretudo empresas de capital com distribuição cripto; decida qual das duas está a construir antes de angariar dinheiro.

Quem é o primeiro cliente pagante, com nome? Todos os casos duradouros deste relatório, da Grass às sub-redes de computação da Bittensor, começaram com um comprador identificado fora da rede.

São áreas que gostaria de explorar em conjunto com outros, através de intercâmbio e colaboração técnicos: chips à medida para mineração, treino e inferência de IA, e aceleradores; desenvolvimento de IA; e sistemas IA × cripto, da cadeia ao nó e ao chip. Pode contactar-me em i@tanshuai.com.

Limitações deste relatório

Esta é uma investigação documental baseada em documentos públicos até 13 de novembro de 2026. Não fiz entrevistas primárias nem medições independentes para esta edição. Vários números-chave são alegações de empresas ou de fornecedores, como a receita da Aethir, a receita da Grass, o desempenho da Etched e da Taalas e a eficiência da Cysic; estão assinalados como tal. As métricas da cadeia da Pearl vêm de um nó completo independente e das pools, e mudam de hora a hora; usei as leituras por volta das 07:13 UTC de 13 de novembro e indico os intervalos do mesmo dia. Os custos de wafers, máscaras e conceção são estimativas de terceiros e valo-

res de modelos que divergem entre si; mostro intervalos. O meu cálculo do ponto de equilíbrio dos ASIC e o orçamento por etapas usam pressupostos ilustrativos (k, T, p, NRE e custos das etapas) que o leitor deve substituir pelos seus.

Sobre o autor

Tan Shuai é tecnólogo e fundador, com mais de vinte anos de trabalho em chips, hardware e software de sistema, e 20 patentes concedidas. Tem atrás de si muitos anos de I&D e gestão de chips fabless, incluindo a conceção de datapaths ASIC full-custom para chips de mineração de criptomoedas e chips multimédia para câmaras, que é a lente deste relatório: energia por operação, custo por wafer, o que um protocolo permite fixar em silício, e quem paga. O seu trabalho levou-o a Silicon Valley, Nova Iorque, Chicago, Taiwan, à Grande Área da Baía (Shenzhen e Hong Kong), ao Vietname e à Malásia. Hoje trabalha como investigador e programador de IA.

Discussão técnica: i@tanshuai.com · tanshuai.com

Notas e exoneração de responsabilidade

Este relatório é uma investigação documental independente baseada em documentos públicos disponíveis até 13 de novembro de 2026. É fornecido apenas para informação e aprendizagem e não constitui aconselhamento de investimento ou de engenharia. Considera apenas tecnologia e negócio. Os preços, as métricas das cadeias e as regras dos protocolos mudam; verifique na fonte primária tudo aquilo em que se apoiar. Os valores assinalados alegação do fornecedor, estimativa ou análise devem ser lidos como tal. Os preços dos criptoativos são muito voláteis.

© 2026 Tan Shuai. Pode partilhar este relatório sem alterações, com atribuição.

Fontes

Todas consultadas entre 10 e 13 de novembro de 2026. Agrupadas por capítulo.

Pearl: protocolo, código e dados da cadeia

- Komargodski e Weinstein, "Proofs of Useful Work from Arbitrary Matrix Multiplication", arXiv 2504.09971: https://arxiv.org/abs/2504.09971
- Pearl Research Labs, whitepaper INT: https://pearlresearch.ai/research/int-whitepaper
- Código-fonte da Pearl: bloco génese: https://github.com/pearl-research-labs/pearl/blob/main/node/chaincfg/genesis.go ; parâmetros de consenso e alturas dos forks: https://github.com/pearl-research-labs/pearl/blob/main/node/chaincfg/params.go ; contabilização da taxa de hash: https://github.com/pearl-research-labs/pearl/blob/main/zk-pow/src/api/sanity_checks.rs ; guia de atualização do fork de semente salgada: https://github.com/pearl-research-labs/pearl/blob/main/docs/salted-seed-fork-upgrade-guide.md ; parâmetros do ramo fp8-scheme: https://github.com/pearl-research-labs/pearl/blob/fp8-scheme/node/chaincfg/params.go
- Pearl Improvement Proposals: https://github.com/pearl-research-labs/pips ; PIP-3, mineração em vírgula flutuante: https://github.com/pearl-research-labs/pips/blob/main/PIPS/pip-0003.md
- prlstats (nó completo independente): rede: https://prlstats.com/api/network ; séries: https://prlstats.com/api/series ; página inicial e modelo de rendimento dos mineradores: https://prlstats.com/api/home ; oferta minerada: https://prlstats.com/api/mined ; pools: https://prlstats.com/api/pools ; detentores: https://prlstats.com/api/holders ; pressupostos de hardware e de sobrecusto: https://prlstats.com/api/assumptions
- Kryptex Pool: rede da Pearl: https://pool.kryptex.com/prl/about-coin ; taxa de hash: https://pool.kryptex.com/prl/hashrate ; RTX 5090: https://pool.kryptex.com/de/device/gpu/nvidia/nvidia-geforce-rtx-5090 ; RTX 4090: https://pool.kryptex.com/device/gpu/nvidia/nvidia-geforce-rtx-4090 ; "Pearl FP8 fork": https://pool.kryptex.com/articles/pearl-fp8-fork-en
- CoinGecko, Pearl: https://api.coingecko.com/api/v3/coins/pearl-2 ; https://www.coingecko.com/en/coins/pearl-research ; histórico de preços: https://api.coingecko.com/api/v3/coins/pearl-2/market_chart?vs_currency=usd&days=365&interval=daily ; pesquisa do ticker «Perle»: https://api.coingecko.com/api/v3/search?query=perle
- SimeonOnSecurity, "Mining Pearl (PRL) in 2026": https://simeononsecurity.com/articles/pearl-prl-mining-guide-2026/
- Hashrate Index, "Pearl (PRL): Inside the AI-Compute Cryptocurrency", 2 de junho de 2026: https://hashrateindex.com/blog/pearl-prl-ai-compute-cryptocurrency/
- Tom's Hardware, "New AI compute cryptocurrency Pearl sparks a GPU mining rush, but profitability is sliding": https://www.tomshardware.com/tech-industry/cryptomining/new-ai-compute-cryptocurrency-pearl-sparks-a-gpu-mining-rush-but-profitability-is-sliding
- Together AI, parceria com a Pearl Research Labs: https://together.ai/blog/together-ai-partners-with-pearl-research-labs ; página do modelo: https://together.ai/models/gemma-4-31b-it-pearl
- Basu, recenseamento das instâncias de mineração da Pearl, arXiv 2606.04819: https://arxiv.org/html/2606.04819v2
- Blockworks, newsletter 0xResearch, 28 de outubro de 2026: https://blockworks.com/newsletter/0xresearch/issue/post_daf0d74a-2bf4-41e4-86d3-2a1275475a24
- Datawallet, "Pearl Explained": https://www.datawallet.com/crypto/pearl-explained
- Coingabbar sobre a notícia do TheMarker acerca do financiamento da Pearl Research Labs: https://www.coingabbar.com/en/pearl-research-labs-funding-130m-stealth-round-themarker

Cadeias de trabalho útil e história

- Grothe et al., análise de segurança da Gridcoin, USENIX WOOT 2017: https://www.usenix.org/system/files/conference/woot17/woot17-paper-grothe.pdf
- Falsificação da prova de aprendizagem, arXiv 2208.03567: https://arxiv.org/abs/2208.03567
- hoge.gg, verificação e confiança na inferência descentralizada: https://hoge.gg/decentralized-inference-2026-verification-trust/ ; receita e tesouraria da Bittensor: https://hoge.gg/bittensor-quantum-wallet-revenue-tao-treasury-2026/
- Kaspa: código da matriz do kHeavyHash: https://github.com/kaspanet/rusty-kaspa/blob/master/consensus/pow/src/matrix.rs ; lançamento: https://coinranking.com/coin/V8GxkwWow+kaspa-kas/about ; anúncio das IceRiver KS1/KS2: https://www.iceriver.io/2023/03/14/1677644580/ ; história da passagem das GPU aos ASIC: https://captainmining.com/blogs/crypto-news/the-evolution-of-kaspa-mining-from-gpus-to-asic-powerhouses ; taxas de hash das GPU: https://coincalculators.io/algorithms/kheavyhash ; KS7: https://www.asicminervalue.com/miners/bitmain/antminer-ks7-40th ; rede: https://minerstat.com/coin/kas ; 1 EH/s em 2024: https://ourcryptotalk.com/news/kaspa-reaches-1-ehs-in-hashrate ; Crescendo: https://ourcryptotalk.com/news/kaspa-crescendo-live
- Aleo: FAQ da mainnet: https://aleo.org/post/aleo-mainnet-faq/ ; prova descentralizada: https://aleo.org/post/decentralized-proving-advantages/ ; ASIC: https://whattomine.com/coins/432-aleo-zksnark/asics ; Blockworks, "State of Aleo Q3 2025": https://app.blockworks.com/report/state-of-aleo-q3-2025 ; ARC-46: https://chaincatcher.com/en/article/2144326
- Qubic: The Block sobre a reorganização da Monero: https://www.theblock.co/post/366535/monero-faces-chain-reorganization-fears-after-qubic-says-it-controls-51-of-hashrate ; Halborn: https://www.halborn.com/blog/post/explained-the-monero-51-percent-attack-august-2025 ; mineração de Dogecoin: https://hackernoon.com/qubic-is-now-mining-dogecoin-on-its-ai-compute-infrastructure ; CoinGecko: https://www.coingecko.com/en/coins/qubic

- Documentação do consenso da Bittensor: https://www.bittensor.com/llms.mdx/docs/internals/consensus/content.md ; FalconX, "State of Bittensor": https://www.falconx.io/newsroom/state-of-bittensor-subnet-adoption-trends-network-mechanics-and-covenants-departure
- Primecoin: https://en.wikipedia.org/wiki/Primecoin ; Blockworks, "State of Filecoin Q3 2025": https://app.blockworks.com/report/state-of-filecoin-q3-2025
- Monero e forks de resistência a ASIC: https://lbry.com/news/forking-for-asic-resistance ; fork da Sia: https://hotfix.coindesk.com/markets/2018/11/01/sia-completes-hard-fork-as-bricked-miners-form-rival-blockchain ; preços da Obelisk: https://www.cryptoglobe.com/latest/2018/11/obelisk-hardware-skyrockets-in-selling-price-after-sia-hard-fork/ ; mineradores depois do Merge: https://hotfix.coindesk.com/business/2022/11/11/ex-ethereum-miners-token-hop-to-stay-alive-after-the-merge

Setores de IA × cripto

- DefiLlama: Render: https://defillama.com/protocol/render-network-bme ; Livepeer: https://defillama.com/protocol/livepeer
- Cointelegraph sobre a receita DePIN segundo a Messari: https://cointelegraph.com/news/depin-revenues-grow-despite-token-losses
- Aethir, balanço de 2025: https://aethir.com/blog-posts/aethirs-2025-wrap-up-decentralized-gpu-cloud-milestones
- CoinGecko, categoria de IA: https://www.coingecko.com/en/categories/artificial-intelligence
- Akash, preços de aluguer da H100: https://akash.network/the-bid/h100-rental-price-2026-cost-per-hour/ ; relatório do 1.º T 2026: https://akash.network/blog/akash-network-q1-2026-report/
- io.net, contrato de 8 milhões de dólares e tokens diários: https://lifestyle.kbew98country.com/story/271004/io-net-to-burn-up-to-12m-tokens-as-network-closes-8m-deal-and-hits-4-billion-daily-ai-tokens/
- DCD, contrato da Axe Compute: https://www.datacenterdynamics.com/en/news/axe-compute-secures-15bn-compute-contract-with-unnamed-customer/
- Pluralis-8B: https://www.alphaxiv.org/abs/2607.13332.md ; Nous Research, Psyche: https://nousresearch.com/the-next-phase-of-psyche ; Implicator sobre o INTELLECT-3: https://www.implicator.ai/prime-intellects-intellect-3-open-source-ambition-meets-centralized-reality/
- SiliconANGLE, financiamento da Prime Intellect: https://siliconangle.com/2026/07/08/prime-intellect-raises-130m-1b-valuation-ai-training-platform/ ; The Block, Venice: https://www.theblock.co/post/406934 ; RuntimeWire, financiamento da Nous: https://runtimewire.com/article/nous-research-75-million-round-1-5-billion-valuation
- ownyourmind, Gensyn: https://ownyourmind.ai/projects/gensyn/ ; Phemex, Ritual (apenas o título): https://phemex.com/news/article/ritual-launches-ai-layer-1-mainnet-and-ritual-token-price-dives-72-from-high-101259
- Sobrecusto da computação confidencial, arXiv 2409.03992: https://arxiv.org/abs/2409.03992v2 ; relatório de 2025 da Phala: https://phala.com/posts/phala-2025-report ; TEE.fail: https://thehackernews.com/2025/10/new-teefail-side-channel-attack.html ; DeepProve, IACR ePrint 2026/1112: https://eprint.iacr.org/2026/1112
- TRM Labs, medição dos pagamentos de agentes de IA: https://www.trmlabs.com/trm-tech-blog/whos-actually-paying-measuring-ai-agent-payments-onchain ; Forkast sobre a x402 Foundation: https://forkast.news/the-x402-foundation-activated-a-27-year-old-internet-code-200-million-transactions-later-the-real-volume-is-still-tiny/ ; The Block, mainnet da Tempo: https://www.theblock.co/amp/post/394131/tempo-mainnet-goes-live-with-machine-payments-protocol-for-agents ; BlockEden sobre o aGDP da Virtuals: https://blockeden.xyz/blog/2026/04/21/agdp-agent-gdp-virtuals-protocol-ai-blockchain-valuation-primitive-tvl-displacement/ ; The Block sobre a ElizaOS: https://theblock.co/post/410774/eliza-labs-native-token-dead
- Dealroom, receita da Grass: https://dealroom.co/news/157794-grass-backed-by-multicoin-projects-75m-in-2026-data-revenue/ ; CryptoSlate, Fetch.ai e Ocean: https://cryptoslate.com/cryptos-flagship-ai-pact-fracture-fetch-sues-ocean-over-263m-fet-community-sales/ ; Fortune sobre a World: https://fortune.com/2026/06/01/internet-according-to-sam-altman-seeing-orb/
- Galaxy, capital de risco em cripto: https://www.galaxy.com/insights/research/crypto-blockchain-venture-capital-q3 ; Crypto Briefing (apenas o título): https://cryptobriefing.com/crypto-vc-investment-q3-2026-galaxy/ ; Blockhead sobre a Tiger Research: https://www.blockhead.co/2026/10/14/bitcoin-rose-43-in-q3-crypto-venture-funding-did-not-follow/ ; ForkLog sobre os lançamentos de tokens de 2025: https://forklog.com/en/85-of-tokens-launched-in-2025-fell-below-initial-prices/

Economia dos ASIC e tipos de chip

- Hashrate Index, resumo de 9 de novembro de 2026: https://hashrateindex.com/blog/hashrate-index-roundup-november-9-2026/
- AnySilicon, custo de conceção de ASIC por nó em 2026: https://anysilicon.com/news/how-much-does-asic-design-cost-in-2026-real-price-ranges-by-process-node/ ; Silicon Analysts, guia de custos de tape-out para fabless: https://siliconanalysts.com/analysis/fabless-startup-tapeout-cost-guide ; dados de mercado: https://siliconanalysts.com/market-data
- MillionMiner, Bitmain S23 XP Hyd: https://millionminer.com/news/bitmain-wdms-2026-s23-xp-hyd-launch
- Khattak e Mikaitis, acumulação FP8 nos tensor cores, arXiv 2512.07004: https://arxiv.org/html/2512.07004v2
- Canaan, resultados do segundo trimestre de 2026: https://finviz.com/news/389337/canaan-inc-reports-unaudited-second-quarter-2026-financial-results ; Bitdeer, resultados do terceiro trimestre de 2026: https://greenstocknews.com/news/nasdaq/btdr/bitdeer-reports-unaudited-third-quarter-2026-results-record-revenue-of-us-320-7-million-adjusted-ebitda-up-173-to-us-69-6-million
- TheEnergyMag: Velaura AI: https://theenergymag.com/news/market-news/velaura-ai-series-a-bitcoin-pivot ; Core Scientific e Block: https://theenergymag.com/news/2026-07-29/core-scientific-exit-block-bitcoin-deal-ai ; economia dos arrendamentos para IA: https://theenergymag.com/news/2026-08-06/miner-weekly-ai-pumpability-out-fuel
- Blockworks, Cipher e Fluidstack: https://blockworks.co/news/cipher-mining-inks-3b-ai-hosting-deal-with-fluidstack ; Cointelegraph, Core Scientific: https://cointelegraph.com/news/core-scientific-to-quit-bitcoin-mining-by-year-end-as-ai-business-surges ; ForkLog, Bitfarms: https://forklog.com/en/former-bitfarms-shuts-down-all-mining-sites-in-the-us/ ; IREN, resultados do 1.º T FY27: https://globenewswire.com/news-release/2026/11/05/3397020/0/en/iren-reports-q1-fy27-results.html

- Spark, concentração do mercado de ASIC de bitcoin: https://www.spark.money/research/bitcoin-asic-manufacturer-market-concentration
- Google, documentação do TPU7x: https://docs.cloud.google.com/tpu/docs/tpu7x?hl=es ; The Register, Ironwood: https://www.theregister.com/2025/04/10/googles_7thgen_ironwood_tpus_debut/ ; SemiAnalysis, B300: https://inferencex.semianalysis.com/chips/b300 ; artigo do TPU v1, arXiv 1704.04760: https://arxiv.org/abs/1704.04760 ; Glenn Lockwood, MI355X: https://glennklockwood.com/garden/processors/mi355x
- Tom's Hardware, Vera Rubin: https://www.tomshardware.com/pc-components/gpus/nvidia-launches-vera-rubin-nvl72-ai-supercomputer-at-ces-promises-up-to-5x-greater-inference-performance-and-10x-lower-cost-per-token-than-blackwell-coming-2h-2026 ; TrendForce, preços do CoWoS: https://trendforce.com/news/2026/10/27/news-tsmc-reportedly-eyes-20-cowos-price-hike-skips-equipment-price-renegotiations-amid-ai-demand-surge
- Menon Lab, comparação de aceleradores de inferência de IA: https://themenonlab.blog/blog/ai-inference-accelerators-compared ; Introl, aceleradores de IA para lá das GPU: https://introl.com/blog/ai-accelerators-beyond-gpus-tpu-trainium-gaudi-cerebras-2025
- StratNews (Reuters), chips de IA nacionais em centros de dados financiados pelo Estado: https://stratnewsglobal.com/team-sng/china-orders-state-funded-data-centres-to-use-domestic-ai-chips/ ; TNW, importações de H200: https://thenextweb.com/news/nvidia-h200-chips-hong-kong-china-import-approval
- Succinct, prova em tempo real em 16 GPU: https://blog.succinct.xyz/real-time-proving-16-gpus/ ; produtos ZK ASIC da Cysic: https://cysic-docs.pages.dev/hardware-products/zk-asic-products
- Tom's Hardware, Intel Heracles: https://www.tomshardware.com/tech-industry/cyber-security/intels-heracles-chip-computes-fully-encrypted-data-without-decrypting-it-chip-is-1-074-to-5-547-times-faster-than-a-24-core-intel-xeon-in-fhe-math-operations ; IEEE Spectrum: https://spectrum.ieee.org/fhe-intel ; Evertiq, Niobium e SemiFive: https://evertiq.com/design/2026-02-20-semifive-secures-design-win-with-niobium-for-fhe-accelerator ; The Quantum Insider, financiamento da Niobium: https://thequantuminsider.com/2025/12/03/niobium-23m-fhe-funding/ ; Optica OPN, Optalysys: https://www.optica-opn.org/home/industry/2026/february/optalysys_raises_us_$31_million_to_push_photonics_computing_toward_cloud_infrastructure/ ; Zama, HPU em FPGA: https://www.zama.org/post/announcing-hpu-on-fpga-the-first-open-source-hardware-accelerator-for-fhe

Nós de processo

- Silicon Analysts, preços de wafers: https://siliconanalysts.com/data/wafer-pricing
- Tom's Hardware, rumores sobre os preços do TSMC A16: https://www.tomshardware.com/tech-industry/semiconductors/tsmc-could-charge-up-to-usd45-000-for-1-6nm-wafers-rumors-allege-a-50-percent-increase-in-pricing-over-prior-gen-wafers ; Wccftech, preços da SMIC: https://wccftech.com/smic-5nm-chip-prices-up-to-50-percent-more-expensive/amp/
- EUROPRACTICE, calendário MPW da TSMC para 2026: https://europractice-ic.com/wp-content/uploads/2025/11/TSMC-Europractice-MPW-schedule-2026.pdf
- Horowitz, "Computing's Energy Problem", ISSCC 2014: https://www.Gwern.net/doc/cs/hardware/2014-horowitz-2.pdf
- Tom's Hardware, roteiro do Huawei Ascend: https://www.tomshardware.com/tech-industry/artificial-intelligence/huawei-ascend-npu-roadmap-examined-company-targets-4-zettaflops-fp4-performance-by-2028-amid-manufacturing-constraints ; The Substrate, cadeia de abastecimento de chips de IA da China: https://www.the-substrate.net/p/where-chinas-ai-chip-supply-chain ; Tom's Hardware, Trainium3: https://www.tomshardware.com/tech-industry/artificial-intelligence/amazon-launches-trainium3-ai-accelerator-competing-directly-against-blackwell-ultra-in-fp8-performance-new-trn3-gen2-ultraserver-takes-vertical-scaling-notes-from-nvidias-playbook

Oportunidades e economia dos lançamentos

- Webull (Benzinga), chamada de resultados da Broadcom: https://www.webull.com/news/15511095937917952 ; Quartr, Global Unichip 2.º T 2026: https://quartr.com/events/global-unichip-corp-3443-q2-2026_F9×6NJQY ; Taiwan News, Alchip: https://www.taiwannews.com.tw/news/6370946 ; Design & Reuse, VeriSilicon: https://www.design-reuse.com/news/202530457-surge-in-custom-chip-business-propels-verisilicon-s-q1-revenue-to-double-with-new-orders-skyrocketing-to-8-2-billion-yuan-increased-r-d-investment-leads-to-wider-net-loss-earnings-insight/ ; The Block, unidade de IA da Canaan: https://www.theblock.co/post/359250/crypto-miner-canaan-to-shutter-non-core-ai-semiconductor-business-unit-amid-strategic-realignment
- Clutch, preços de desenvolvimento de IA: https://clutch.co/developers/artificial-intelligence/pricing ; PixelPlex, custo de desenvolvimento de blockchain: https://pixelplex.io/blog/blockchain-development-cost/
- Hyperbolic, prova de amostragem: https://hyperbolic.ai/blog/deep-dive-into-hyperbolic-proof-of-sampling ; Gensyn, Verde: https://www.gensyn.ai/research/verde-a-verification-system-for-machine-learning-over-untrusted-nodes ; verificação em camadas, arXiv 2512.20176: https://arxiv.org/abs/2512.20176v1
- Sherlock, referência de preços de auditorias para 2026: https://sherlock.xyz/post/smart-contract-audit-pricing-a-market-reference-for-2026
- Blockworks, critérios de listagem da Binance: https://blockworks.com/news/binance-listing-standards-scrutiny ; Cointelegraph, acordos de market making: https://cointelegraph.com/magazine/market-maker-deals-quietly-killing-crypto-projects ; Blockworks, Movement: https://blockworks.co/news/movement-rushi-manche-suspension-market-maker-web3port ; Nansen, HIP-3: https://nansen.ai/post/what-is-hip-3-hyperliquid ; The Block, Monad: https://www.theblock.co/post/380094/monad-mainnet-launches

PROVA DE TRABALHO ÚTIL · CRIPTO × SILÍCIO

Computação como Consenso

IA × Cripto × Chips em 2026, a partir da prova de trabalho útil da Pearl

Cada número-chave deste relatório indica a sua fonte e data. Quando o autor deriva um número, este é assinalado como estimativa; quando só o vendedor o comunica, é assinalado como alegação do fornecedor. Considera apenas tecnologia e negócio.

Tan Shuai

Tecnólogo e fundador · Investigador e programador de IA

tanshuai.com · i@tanshuai.com

EDIÇÃO PÚBLICA · Dados de 13 de novembro de 2026

