Nota de investigação · Novembro de 2026

Computação como consenso

IA × Cripto × Chips em 2026, a partir da Pearl

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

Tan Shuai
tanshuai.com · i@tanshuai.com



O teorema mantém-se; o incentivo não

01 A Pearl funciona como engenharia, ainda não como trabalho útil

Cerca de 90 EH/s de multiplicação de matrizes inteiras protegem a cadeia, mas só cerca de 1% dos blocos transporta trabalho real de IA.

02 A procura paga está na inferência, nos dados e na privacidade

Bittensor com cerca de 70 milhões de dólares por ano de receita verificável, Grass com 17 milhões no 1.º semestre de 2026, TEE em produção; pagamentos de agentes cerca de 5 000–11 000 dólares por mês.

03 Um ASIC só compensa com um bolo grande, regras estáveis e um segundo mercado

O seu teto de lucro é R × (1 − 1/k). Hoje nenhuma rede oferece as três coisas análise .

≈90 EH/s

Pearl a 13 de novembro de 2026: o débito inteiro de cerca de 122 000 H100

≈1%

dos blocos da Pearl transportam trabalho real de IA (Blockworks)

4–11×

emissões da Bittensor face à sua receita verificável

IA × cripto em seis setores: onde a procura paga é real

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

> O subsídio é combustível para um arranque a frio, não um modelo de negócio. A cripto faz a distribuição e a liquidação; o capital fica com os lucros, pelo menos em 2026.

Uma multiplicação-acumulação é um hash

Do teorema à mainnet

Komargodski e Weinstein (abril de 2025): uma prova de trabalho útil para qualquer produto de matrizes com sobrecusto 1 + o(1). A mainnet da Pearl arrancou a 27 de abril de 2026.

Como funciona

Os mineradores multiplicam matrizes com ruído, fazem o hash do acumulador de cada tile num bilhete de lotaria e provam o tile vencedor com uma prova Plonky2 de menos de 60 KB.

A unidade

1 H é uma multiplicação-acumulação INT7 × INT7 → INT32 creditada: o orçamento de segurança de uma blockchain com preço em débito tensorial.

PEARL A 13 DE NOVEMBRO DE 2026

89,77 EH/s

taxa de hash da rede (nó completo independente)

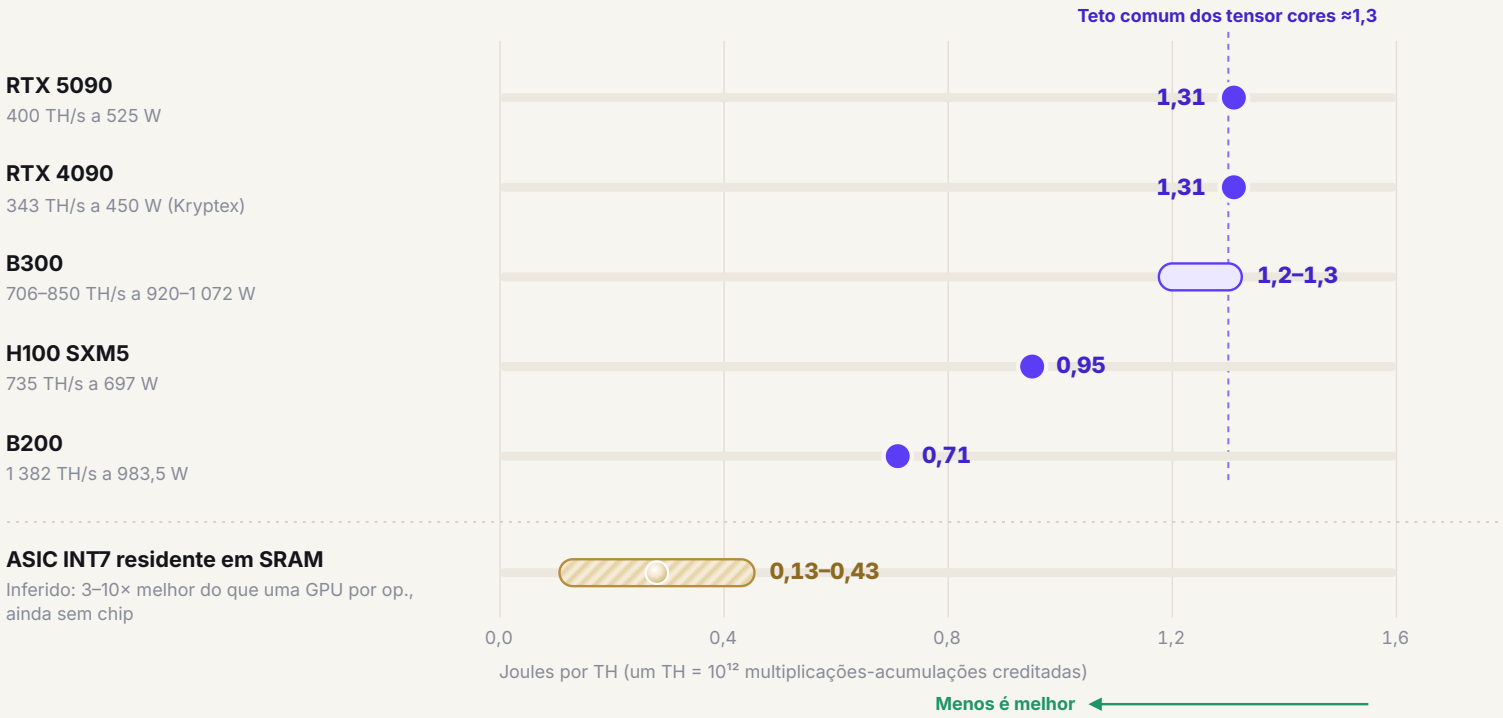
≈$500 M

valor de mercado; totalmente diluído cerca de 2,8 mil milhões de dólares

+78%

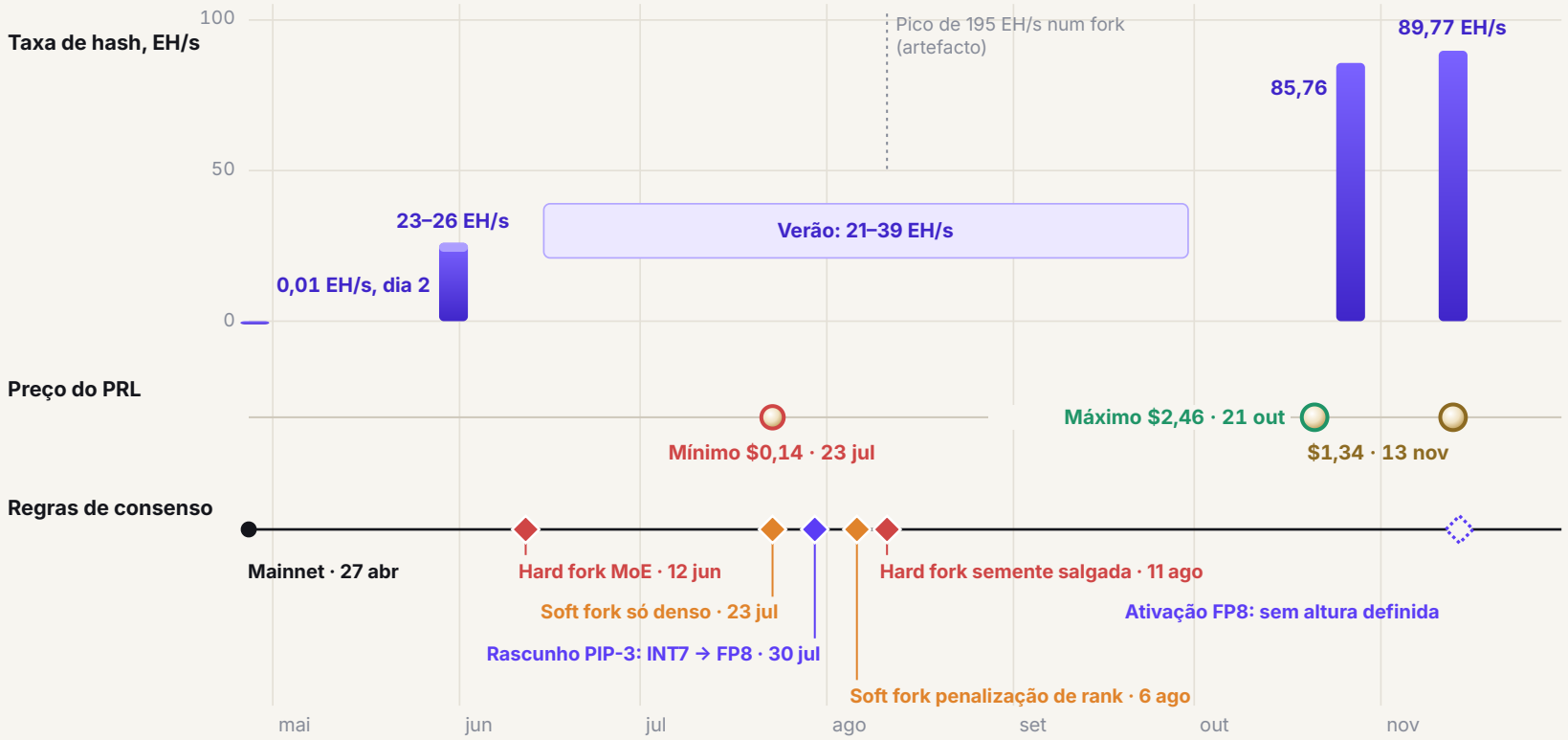
de nova oferta nos próximos 12 meses

No trabalho inteiro da Pearl, as GPU partilham um teto energético



As placas de jogos e a B300 encontram-se em cerca de 1,3 J/TH; mesmo a B200 só chega a 0,71. Um chip que apenas multiplique inteiros de 7 bits a partir de SRAM poderia plausivelmente ser 3–10 vezes melhor **análise**.

A taxa de hash segue o preço; as regras continuam a mudar



Cerca de 120 milhões de moedas, 32% da oferta atual, foram mineradas nos primeiros seis dias. Uma RTX 5090 rende hoje cerca de 5–6 dólares por dia; alugar placas só para minerar dá prejuízo.

O teorema mantém-se; o incentivo não

O QUE CUSTA O TRABALHO ÚTIL

Lado da inferência: pouco sobrecusto

A Pearl indica mais 5,08% de tempo para o Llama 70B em quatro H200; um teste independente mediu menos 9,2% de débito numa GPU Hopper.

O QUE RENDE O TRABALHO ÚTIL

Lado da mineração: taxa de hash insuficiente

Uma placa a servir inferência minera a cerca de 27% de um minerador dedicado no prefill, 5,4% em chat, cerca de 1% quando domina a descodificação.

Por isso, os mineradores racionais multiplicam matrizes aleatórias. A utilidade é um problema comercial fora da cadeia: alguém tem de pagar a inferência.

Quatro forks em dois meses, e a seguir FP8

12 de junho
Hard fork de mistura de especialistas (bloco 71 935)

23 de julho
Soft fork só para modelos densos (bloco 91 630)

6 de agosto
Soft fork de penalização de rank: algumas configurações perdem 50–87,5% da taxa de hash efetiva

11 de agosto
Hard fork de semente salgada: os mineradores antigos só produzem shares inválidas

Por definir
PIP-3: de INT7 para FP8, só Hopper e Blackwell admitidas

2 de 3

útil para IA, verificável bit a bit, neutro quanto ao hardware: um formato numérico consegue no máximo dois

13 bits

mantidos pela acumulação FP8 da Hopper, com truncagem: não reprodutível entre gerações

≈82%

da taxa de hash em placas de consumo que o PIP-3 não admitiria (modelo de baixa confiança)

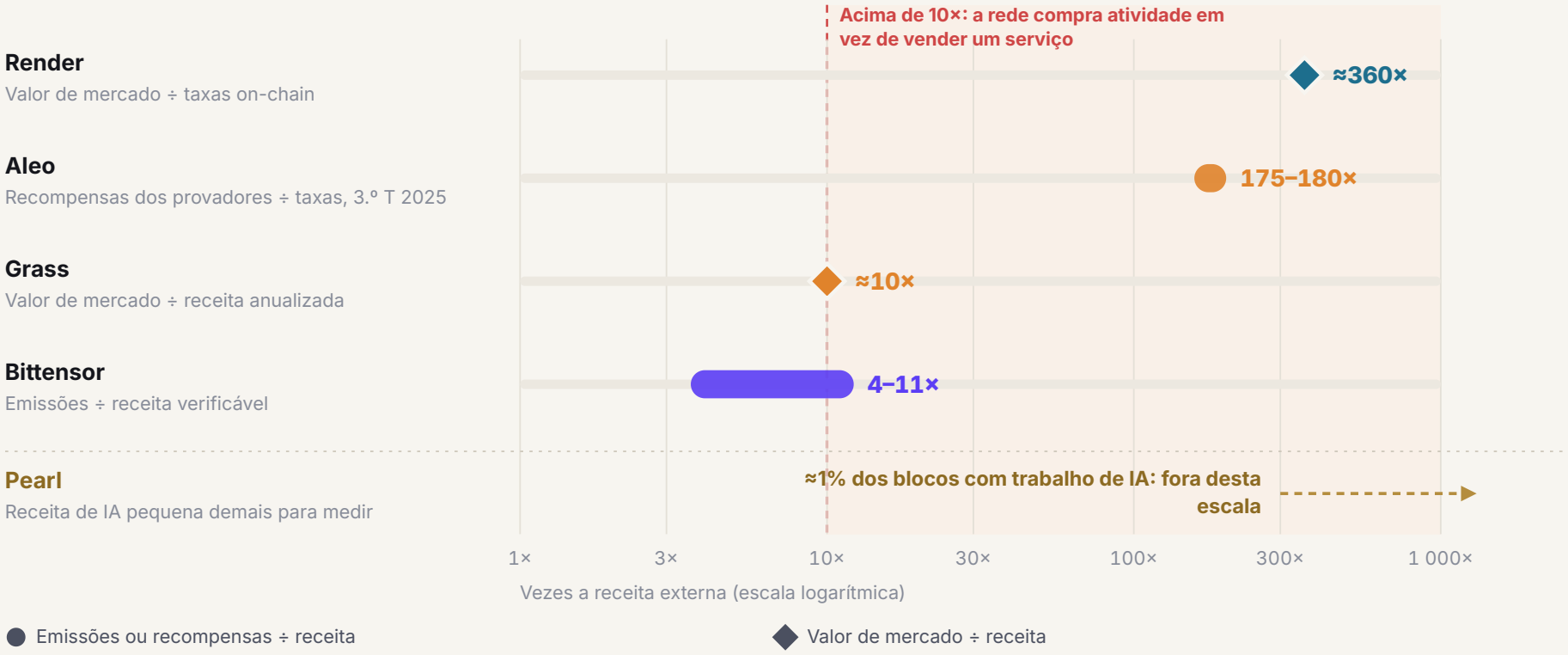
Quem controla a função de trabalho controla o destino de quem investe em hardware análise.

Nem matrizes nem conhecimento zero travam os ASIC

Rede	O que aconteceu	O que ensina
Kaspa	Um produto de matrizes 64 × 64 entre hashes; ASIC ao fim de 16 meses, cerca de 1 000× a eficiência das GPU; o melhor rende hoje $0,86 por dia	Uma matriz não é útil nem resistente a ASIC
Aleo	Um puzzle SNARK sintético; máquinas dedicadas ao fim de cinco meses; recompensas dos provadores 175–180× as taxas	O trabalho sintético vive de subsídio
Qubic	Diz fazer treino de IA; a sua computação minerou Monero (51% alegado) e agora Dogecoin	Uma PoW de uso geral pode ser capturada de fora
Bittensor	Pontuação ponderada por stake; cerca de 70 milhões de dólares por ano de receita verificável contra mais de 300 milhões de emissões	A única com receita externa à escala
Monero, Sia, Ethereum	Os forks inutilizaram ASIC; depois do Merge, menos de 20% dos mineradores encontraram outro trabalho	Sobrevive o hardware com segundo mercado

SUBSÍDIO

O múltiplo de subsídio é o teste de saúde mais honesto



Acima de cerca de 10 vezes a receita externa, uma rede está a comprar atividade em vez de vender um serviço. A receita de IA da Pearl é pequena demais para caber na escala.

Procura real mas pequena, e valor a passar para o capital

Mercados de computação

Taxas on-chain da Render cerca de 2,35 milhões de dólares por ano; a H100 da Akash a $2,73 por hora bate a AWS e a Google, mas não a Vast.ai nem a RunPod.

Treino

8–40 mil milhões de parâmetros treinados pela internet; o INTELLECT-3 voltou a 512 H200 em InfiniBand.

IA verificável

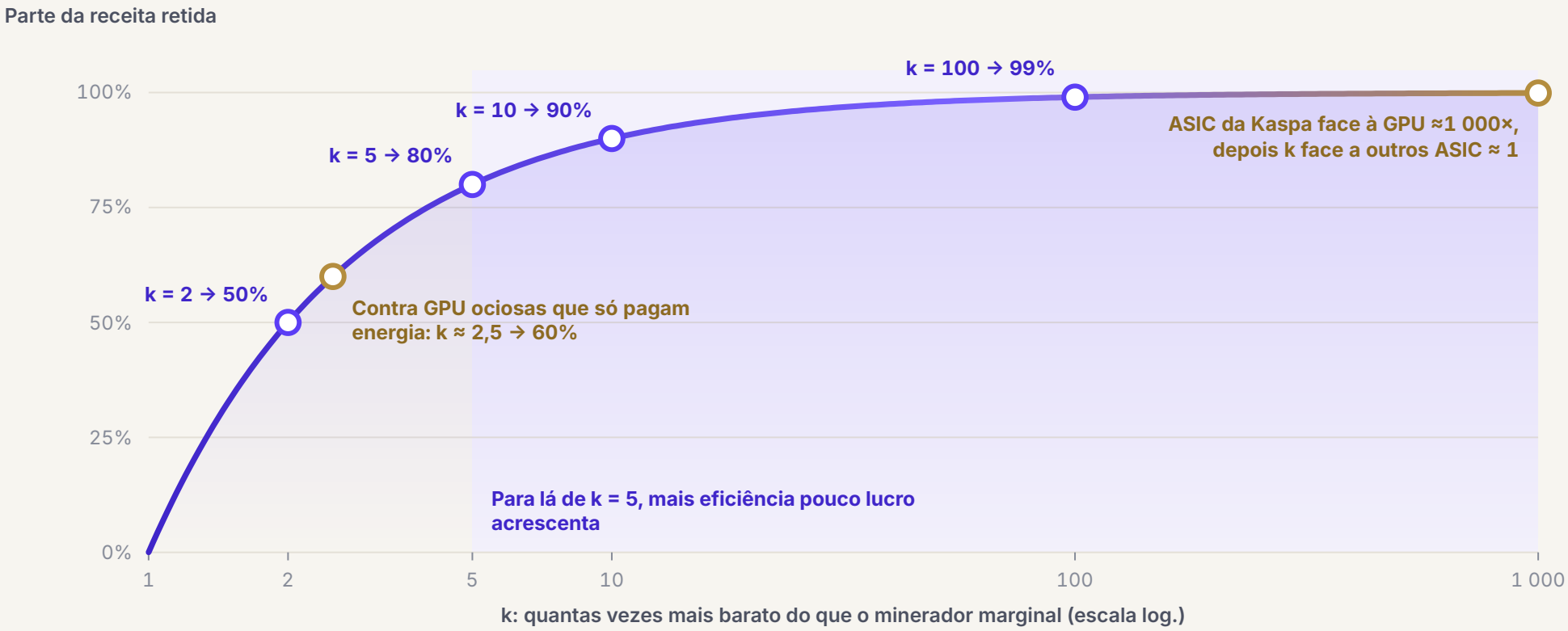
Os TEE acrescentam menos de 7% e funcionam em produção; o TEE.fail quebrou a atestação com um interposer de 1 000 dólares; o zkML prova o GPT-2 a 174 tokens por minuto.

Agentes e dados

Fluxos x402 iniciados por agentes cerca de 5 000–11 000 dólares por mês; Grass com 17 milhões de dólares de receita de dados no 1.º semestre de 2026.

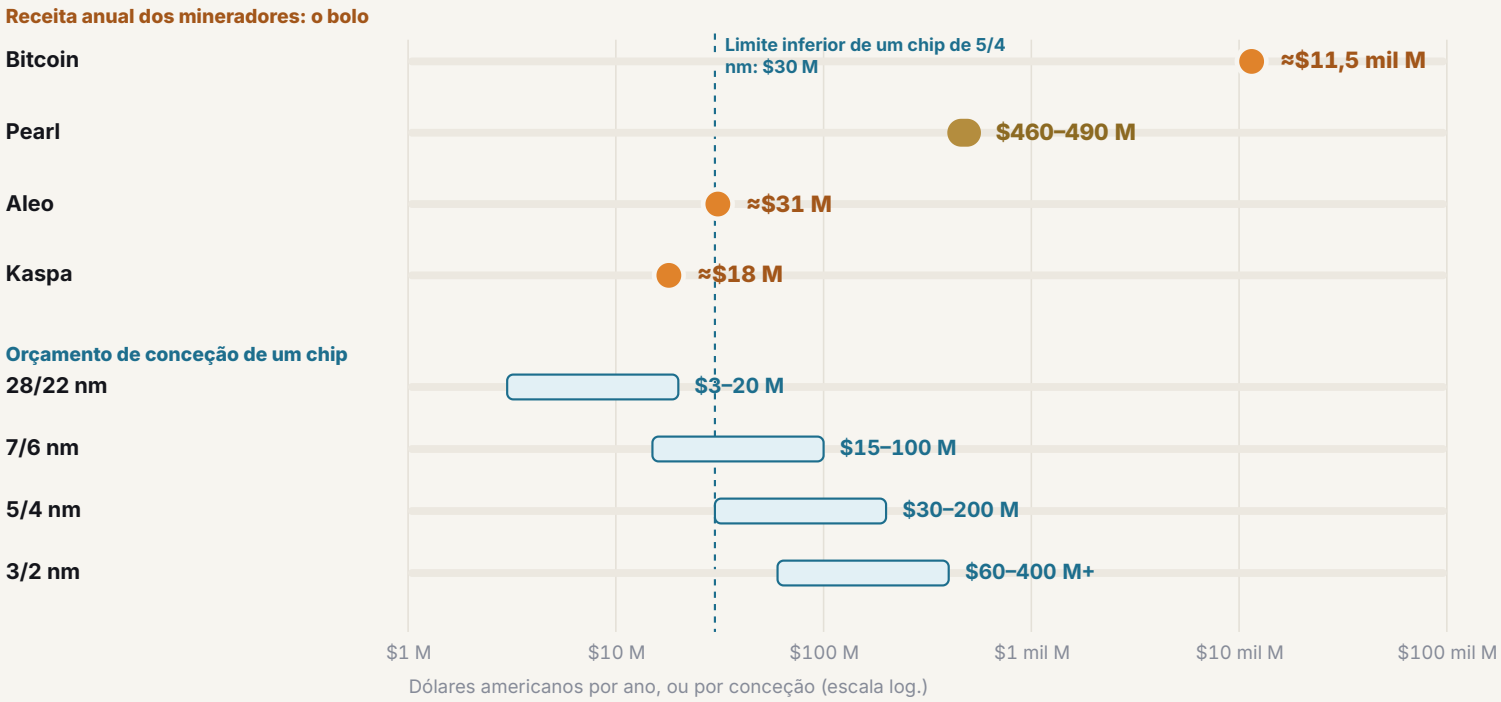
Quatro empresas de capital em infraestrutura centralizada valem cerca de 8,5 mil milhões de dólares, aproximadamente tanto como todos os tokens de IA das camadas de protocolo e de aplicação.

Porque a eficiência deixa de compensar depois de k = 5



Teto de lucro ≈ R × (1 − 1/k). Quando os ASIC passam a ser os mineradores marginais, o k recomeça face a outros ASIC; face a GPU ociosas que só pagam a energia, k é cerca de 2,5 **estimativa**.

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



Taxa de hash de retorno h* = NRE ÷ [R × (1 − 1/k) × T × p]. Para a Pearl, com 20 milhões de dólares de NRE e k = 5: cerca de 4,9 EH/s durante um ano se as regras se mantiverem, 9,8 EH/s se houver apenas metade das hipóteses de se manterem **estimativa**.

Cinco tipos de chip, cinco respostas

 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

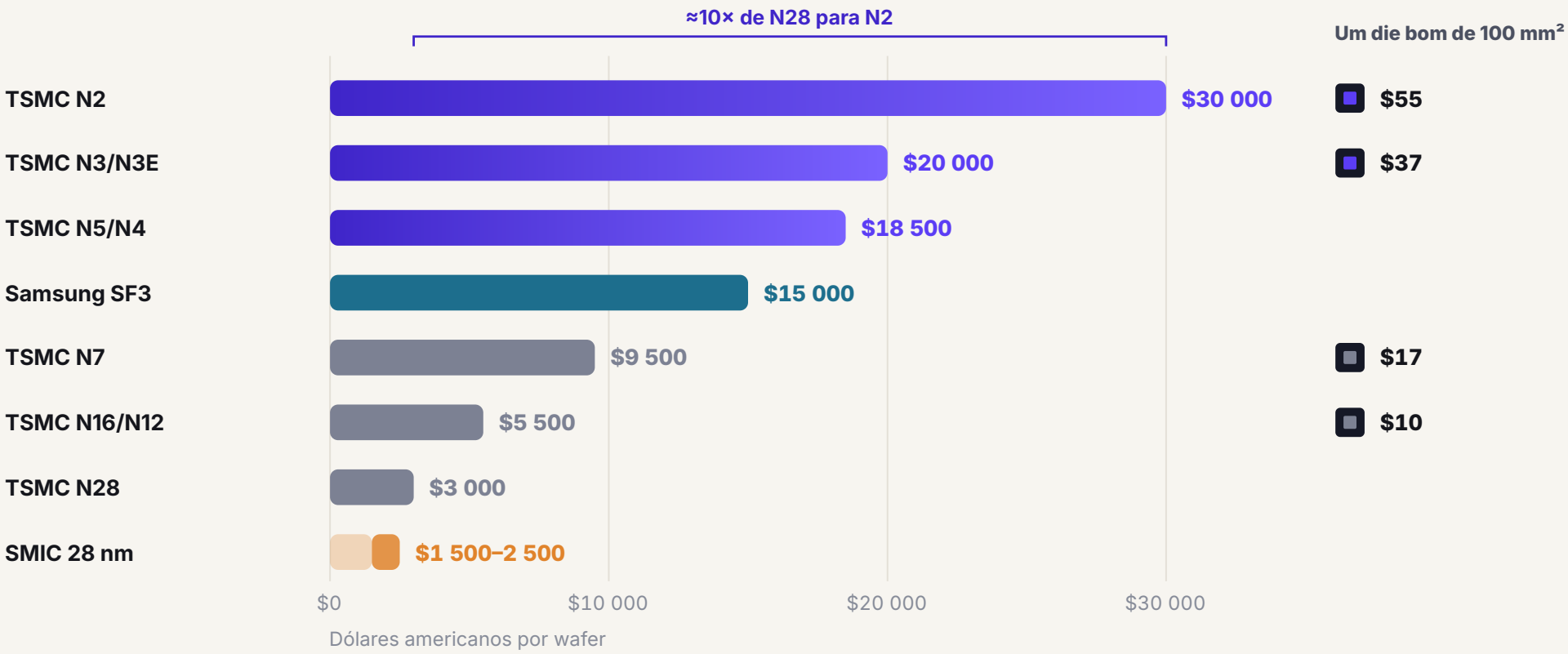
k: 1 074–5 547× um Xeon (Heracles)

Segundo mercado: computação cifrada, IA privada

Uma janela 2027–2029; validar com poucos ativos

> Nenhuma rede oferece hoje, em simultâneo, um bolo grande, regras estáveis e k acima de 1: o k da bitcoin pertence aos incumbentes, as regras da Pearl são as menos estáveis, os bolos da Kaspa e da Aleo são pequenos demais.

Uma wafer de vanguarda custa dez vezes uma madura



> Deixe que o maior custo ao longo da vida do chip escolha: a energia para chips de computação contínua, o encapsulamento e a HBM para chips de treino (63% do custo de uma H100), a escolha entre SRAM e HBM para chips de inferência.

NÓS DE PROCESSO

Um nó para cada chip e cada equipa

Chip	Equipa pequena	Empresa de escala	Via da China continental
Chip de mineração PoUW	MPW de 28 ou 16/12 nm; primeiro volume N7/N6	N3/N3E ou Samsung SF3	SMIC N+1/N+2 serve, cerca de 40–50% mais por die
Chip de mineração SHA-256	Não perseguir a vanguarda	Classe de 3 nm	Difícil igualar a eficiência dos nós de ponta
Chip de treino de IA	IP, chiplets, serviços de conceção	N3P e depois N2/A16 com CoWoS e HBM	SMIC N+2/N+3 com HBM nacional
Chip de inferência de IA	Centrado em SRAM em N6/N7 ou N4/N5/SF4	N3/N2 com HBM, ou chiplets SRAM	SMIC N+2 com LPDDR ou GDDR
Aceleradores ZK e FHE	FPGA ou MPW de 16/12 nm primeiro	N5/N3; FHE com HBM	ZK viável; FHE limitado pela HBM

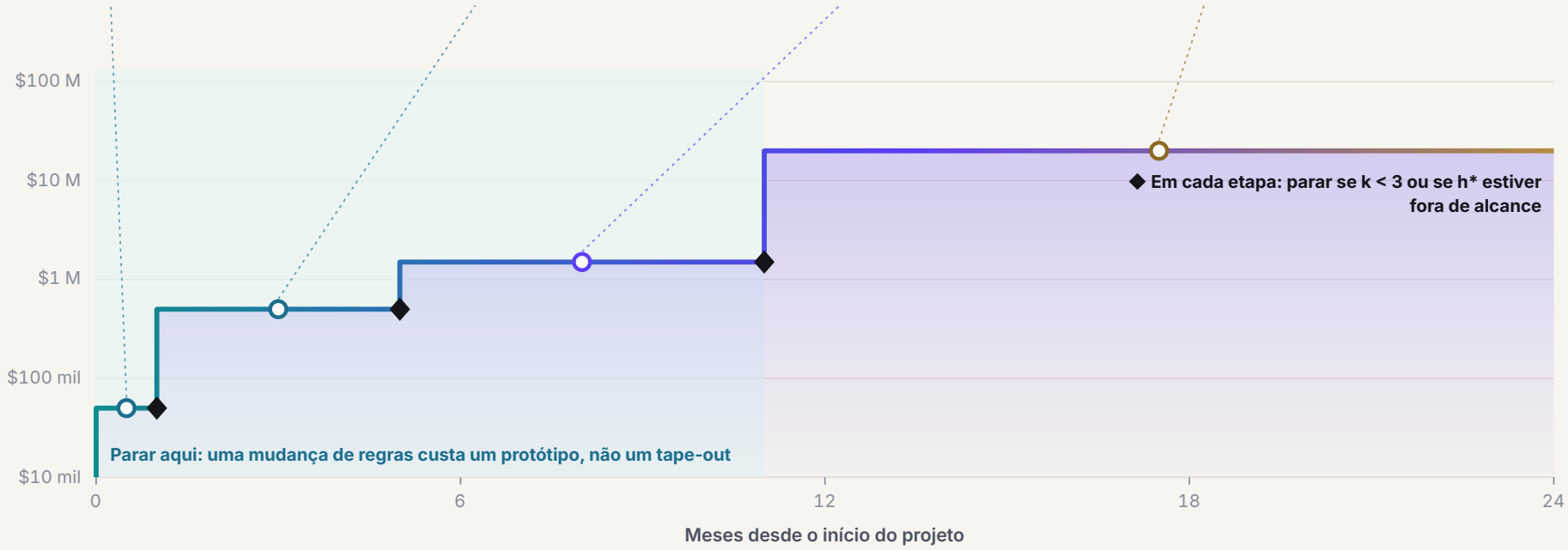
Comprar provas antes das máscaras

Etapa 0 · Modelo
≈$50 mil · semanas
J/op e $/op a partir da aritmética, SRAM e E/S

Etapa 1 · FPGA
≈$0,5 M · meses
A carga real de ponta a ponta: shares ou tokens

Etapa 2 · MPW
≈$1,5 M · mês 11
O silício medido substitui o modelo

Etapa 3 · Máscaras e primeiro lote
≈$20 M · mês 24
Máscaras N7 ≈$3,5 M num projeto de $15–100 M



Um modelo, uma FPGA e uma MPW chegam ao silício medido por cerca de 1,5 milhões de dólares em onze meses; o conjunto de máscaras e o primeiro lote custam mais de dez vezes isso. Parar antes das máscaras até o comprador ou a versão do protocolo estarem fixados.

A cripto como primeiro cliente, a IA como segundo mercado

Chips à medida

- A procura é esmagadoramente de IA: a Broadcom com cerca de 58 mil milhões de dólares em semicondutores de IA no ano fiscal de 2026
- As competências de chips de mineração transferem-se: matrizes de baixa tensão, alimentação de quilowatts, tape-outs rápidos
- Ganhar com uma via própria para o mercado, não com os canais da mineração

Desenvolvimento de IA

- As horas são uma mercadoria: 25–49 dólares por hora na Clutch
- O valor está nas junções: kernels para chips à medida, mineração dentro do serviço de inferência
- Inferência verificável: TEE mais verificações pontuais ZK

Sistemas IA × cripto

- Verificação em camadas: TEE ou amostragem, depois disputas ou verificações pontuais ZK
- As barreiras são a distribuição e a liquidez, não o custo
- Assinar o comprador de computação antes do token

Três caminhos, cada um com um pequeno primeiro passo

Caminho	Primeiro passo
A. Um acelerador de inferência que também possa minerar	Um estudo de quatro semanas de J/op e $/op para uma matriz INT8/FP8 residente em SRAM face à RTX 5090 e à H100, depois FPGA ou MPW num nó maduro
B. Serviços de silício à medida: cripto primeiro, IA depois	Transformar o modelo de NRE e de ponto de equilíbrio deste relatório em estudos de viabilidade pagos, com marcos e uma versão de protocolo fixa
C. Uma rede nativa de IA com um comprador antes de um token	Uma nota de conceção sobre verificação em camadas e uma aritmética canónica neutra quanto ao fornecedor, escrita com um comprador de inferência identificado

No máximo três caminhos, cada um testado antes de comprometer dinheiro análise **.**

O que seguir, e perguntas a fazer

SINAIS A SEGUIR

A qualquer momento	A altura de ativação FP8 da Pearl, e se a cadeia se divide
Mensalmente	A parte dos blocos da Pearl com trabalho de IA
Trimestralmente	Receita verificável da Bittensor face às emissões; fluxos x402 iniciados por agentes
Janeiro de 2027	Subida de preços de 3–6% da TSMC noticiada para 2 nm e 3 nm
1.º T 2027	FP4 na Pearl, segundo o AMA dos fundadores
Abril de 2028	O próximo halving da bitcoin

PERGUNTAS A FAZER

01 Quem controla a função de trabalho?

02 Qual é o múltiplo de subsídio?

03 Quanto custa a próxima etapa, e o que o faz parar?

04 Qual é a taxa de hash de retorno, com p?

05 Qual é o segundo mercado do chip?

06 Quem é o primeiro cliente pagante, com nome?

São áreas que podemos explorar em conjunto, através de intercâmbio e colaboração técnicos: chips à medida, desenvolvimento de IA e sistemas IA × cripto.

i@tanshuai.com · tanshuai.com