GHI CHÚ NGHIÊN CỨU · BẢN CÔNG KHAI

2026 · 11



BẰNG CHỨNG CÔNG VIỆC HỮU ÍCH · CRYPTO × SILICON

Tính toán là đồng thuận

AI × Crypto × Chip năm 2026, khởi đầu từ bằng chứng công việc hữu ích của Pearl

Pearl · Công việc hữu ích · DePIN và tác tử · Kinh tế học ASIC · Tiến trình · Ngân sách theo giai đoạn

KẾT LUẬN

Pearl chứng minh rằng phép nhân ma trận có thể bảo vệ một blockchain, nhưng chưa chứng minh được rằng công việc ấy hữu ích. Trong toàn bộ lĩnh vực AI × crypto, nhu cầu trả tiền nằm ở suy luận, dữ liệu và quyền riêng tư; và một ASIC chỉ sinh lời khi quỹ thưởng lớn, luật lệ ổn định và con chip có thị trường thứ hai.

Tan Shuai

Nhà công nghệ và nhà sáng lập · Nhà nghiên cứu và phát triển AI

tanshuai.com · i@tanshuai.com

Số liệu tính đến ngày 13 tháng 11 năm 2026

Tiền tệ tính bằng đô la Mỹ

MỤC LỤC

Mục lục



◆ **Tóm tắt** — Định lý đứng vững, động lực thì không — 04

01 **Phạm vi và phương pháp** — Phạm vi, phương pháp và bốn phép thử — 07

02 **Pearl · cơ chế** — Pearl: phép nhân ma trận làm đồng thuận — 08

03 **Pearl · kinh tế** — Pearl: quy mô, phát hành và khoảng cách hữu ích — 10

04 **Pearl · luật lệ** — Pearl: rủi ro luật lệ, số học và quản trị — 12

05 **Chuỗi công việc hữu ích** — Chuỗi công việc hữu ích: cả ma trận lẫn zero knowledge đều không chặn được ASIC — 13

06 **Bức tranh rộng hơn** — Phần còn lại của AI × crypto: nhu cầu thật nhưng nhỏ, và giá trị đang chuyển sang cổ phần — 16

07 **Kinh tế học ASIC** — Câu hỏi ASIC: một công thức, ba thước đo — 18

08 **Từng loại chip** — Năm loại chip, năm câu trả lời — 21

09 **Tiến trình** — Tiến trình: hãy để chi phí vòng đời lớn nhất quyết định — 24

10 **Ngân sách theo giai đoạn** — Từ mô hình đến silicon: ngân sách theo giai đoạn cho con chip đầu tiên — 27

11 **Cơ hội** — Cơ hội nằm ở đâu, và tôi sẽ làm gì — 29

12 **Dấu hiệu cần theo dõi** — Điều cần theo dõi, tháng 11 năm 2026 đến năm 2027 — 32

13 **Câu hỏi nên đặt ra** — Những câu hỏi bạn chưa hỏi nhưng nên hỏi — 33

◆ **Về tác giả** — Tan Shuai — 34

◆ **Nguồn** — 35

HÌNH

Hình



HÌNH 01	AI × crypto qua sáu mảng: nhu cầu trả tiền thật nằm ở đâu	06
HÌNH 02	Với công việc số nguyên của Pearl, các GPU chung một trần năng lượng	09
HÌNH 03	Sáu tháng đầu của Pearl: hashrate theo giá, luật lệ liên tục đổi	11
HÌNH 04	Bội số trợ cấp: mạng trả gấp bao nhiêu lần doanh thu bên ngoài	15
HÌNH 05	Vì sao hiệu quả thôi sinh lời: ASIC giữ tối đa 1 − 1/k doanh thu	19
HÌNH 06	Phần lớn quỹ thường nhỏ hơn ngân sách thiết kế một con chip	20
HÌNH 07	Năm loại chip, năm câu trả lời	24
HÌNH 08	Một wafer tiên tiến nhất đắt gấp mười wafer trưởng thành	25
HÌNH 09	Mua bằng chứng trước mặt nạ: ngân sách theo giai đoạn cho chip đầu tiên	28

Nhãn trong văn bản

- **ước tính** — phép tính hoặc suy luận của tác giả từ dữ liệu được trích dẫn
- **hãng tự công bố** — do chính công ty bán sản phẩm công bố, chưa được kiểm chứng độc lập
- **phân tích** — cách diễn giải bằng chứng của tác giả

Đây là bản dịch từ ấn bản tiếng Anh; nếu có khác biệt, ấn bản tiếng Anh được ưu tiên.

Số liệu tính đến ngày 13 tháng 11 năm 2026 · Tiền tệ tính bằng đô la Mỹ

TÓM TẮT

Định lý đứng vững, động lực thì không

KẾT LUẬN

Pearl cho thấy "tính toán là đồng thuận" vận hành được về mặt kỹ thuật, nhưng chưa cho thấy nó hữu ích.

≈90 EH/s
mạng Pearl ngày 13 tháng 11 năm 2026, trong đó một H là một phép nhân-cộng INT7

≈1%
số khối Pearl chứa công việc AI thật (Blockworks, ngày 28 tháng 10 năm 2026)

≈70 tr USD
doanh thu bên ngoài có thể kiểm chứng mỗi năm của Bittensor, so với hơn 300 triệu USD phát hành

175–180×
phần thưởng prover của Aleo so với phí giao dịch trong quý 3 năm 2025

R × (1 – 1/k)
trần lợi nhuận của một ASIC rẻ hơn k lần so với máy đào biên ước tính

≈1,5 tr USD
để đến silicon đo được trên N7 qua mô hình, FPGA và MPW, trước một bộ mặt nạ ≈3,5 triệu USD ước tính

Pearl

Ngày 13 tháng 11 năm 2026, mạng Pearl chạy ở khoảng 90 EH/s. Trong mã đồng thuận của nó, một "hash" là một phép nhân-cộng số nguyên, nên mạng này tương đương thông lượng số nguyên của khoảng 122.000 chiếc H100. Thế nhưng chỉ khoảng 1% số khối chứa công việc AI thật, và mối hợp tác suy luận công khai duy nhất nay không còn được cung cấp trên API serverless của nhà cung cấp. Nguyên nhân là kinh tế, không phải mật mã: một GPU vừa đào vừa phục vụ suy luận chỉ đạt khoảng 1% đến 27% hashrate của một máy đào chuyên dụng, nên thợ đào duy lý nhân các ma trận ngẫu nhiên.

Nhu cầu trả tiền

Trong toàn bộ AI × crypto, nhu cầu trả tiền có thể kiểm chứng xuất hiện ở ba nơi: bán suy luận, bán dữ liệu và bán quyền riêng tư. Doanh thu bên ngoài có thể kiểm chứng của Bittensor khoảng 70 triệu USD mỗi năm, khoảng bốn phần năm trong đó đến từ các subnet điện toán và hạ tầng. Grass báo cáo 17 triệu USD doanh thu dữ liệu trong nửa đầu năm 2026. Suy luận bảo mật trong môi trường thực thi tin cậy đã vào vận hành thực tế. Ngược lại, các khoản thanh toán thực sự do tác tử AI khởi tạo chỉ khoảng 5.000–11.000 USD mỗi tháng, còn zkML và huấn luyện phi tập trung chưa có nhu cầu trả tiền nào được công bố.

Trợ cấp và giá trị

Bội số trợ cấp là phép kiểm tra sức khỏe trung thực nhất. Phần thưởng prover của Aleo gấp khoảng 175–180 lần phí của nó; lượng phát hành của Bittensor gấp khoảng 4–11 lần doanh thu kiểm chứng được; vốn hóa của Render gấp khoảng 360 lần phí on-chain; và khoảng 460 triệu USD thu nhập thợ đào mỗi năm của Pearl phần lớn là trợ cấp khối, gần như không liên quan đến nhu cầu AI trả tiền. Trợ cấp là nhiên liệu để khởi động, không phải mô hình kinh doanh. Trong khi đó giá trị đang chuyển từ token sang cổ phần: bốn công ty cổ phần dùng hạ tầng tập trung và kênh phân phối gốc crypto được định giá khoảng 8,5 tỷ USD, xấp xỉ toàn bộ token AI tầng giao thức và tầng ứng dụng cộng lại.

ASIC

Về chip, trần lợi nhuận của một ASIC xấp xỉ doanh thu trên mỗi đơn vị tính toán nhân với (1 − 1/k), trong đó k là số lần nó rẻ hơn phần cứng biên. Vì số hạng này lõm, hiệu quả vượt quá k = 5–10 gần như không thêm lợi nhuận; điều quyết định kết cục là quy mô quỹ thưởng, luật lệ kéo dài bao lâu và con chip có thị trường thứ hai hay không. Câu trả lời của tôi theo từng loại chip: đừng tape-out một chip Pearl INT7 chuyên dụng lúc này, thay vào đó đặt cược vào một bộ tăng tốc ưu tiên suy luận và cũng đào được; là người mới thì đứng ngoài thị trường máy đào bitcoin thương mại; là đội nhỏ thì đừng làm chip huấn luyện độc lập; làm chip suy luận khi đã ký được khách hàng; phòng hộ với bộ tăng tốc ZK và đón đầu sớm với FHE.

Tiến trình và ngân sách

Tiến trình nên được chọn theo khoản lớn nhất trong chi phí vòng đời của con chip: năng lượng với chip chạy suốt ngày đêm, đóng gói và bộ nhớ với chip huấn luyện, quyết định SRAM hay HBM với chip suy luận. Và con chip đầu tiên nên được mua theo từng giai đoạn, một mô hình, một FPGA, một wafer đa dự án và chỉ sau đó mới đến bộ mặt nạ, để một thay đổi của giao thức chỉ tốn giá một mẫu thử chứ không phải giá một lần tape-out. Cơ hội nằm ở nơi các lĩnh vực giao nhau: crypto là khách hàng đầu tiên, AI là thị trường thứ hai.

Câu hỏi	Nhận định	Con số chính (nguồn trong các chương)
① **Pearl có vận hành được không, và công việc của nó có hữu ích không?**	Vận hành được: một blockchain được bảo vệ bằng phép nhân ma trận ở quy mô lớn. Chưa hữu ích: gần như toàn bộ việc đào là trên ma trận ngẫu nhiên	Khoảng 90 EH/s ngày 13 tháng 11 năm 2026; 1 H = 1 phép nhân-cộng INT7; khoảng 1% số khối có công việc AI
② **Nhu cầu trả tiền thật trong AI × crypto nằm ở đâu?**	Suy luận, dữ liệu và quyền riêng tư. Thanh toán của tác tử, zkML và huấn luyện phi tập trung vẫn chưa có nhu cầu	Bittensor khoảng 70 triệu USD doanh thu kiểm chứng được mỗi năm; Grass 17 triệu USD trong nửa đầu 2026; thanh toán của tác tử khoảng 5.000–11.000 USD mỗi tháng
③ **Có đáng làm một ASIC không?**	Chỉ khi quỹ thưởng lớn, luật lệ ổn định và con chip có thị trường thứ hai. Hiện chưa mạng nào hội đủ cả ba	Trần lợi nhuận ≈ R × (1 − 1/k); bitcoin khoảng 11,5 tỷ USD doanh thu thợ đào mỗi năm, Pearl khoảng 0,46–0,49 tỷ USD, Kaspa khoảng 18 triệu USD
④ **Đội nhỏ nên làm chip gì?**	Bộ tăng tốc ưu tiên suy luận và cũng đào được; chip suy luận cho một khách hàng đã ký; công việc FHE sớm. Không làm máy đào bitcoin thương mại hay chip huấn luyện	ASIC tốt nhất của Kaspa kiếm 0,86 USD mỗi ngày; ngân sách thiết kế 3/2 nm từ 60–400 triệu USD trở lên
⑤ **Tiến trình nào?**	Chi phí vòng đời lớn nhất quyết định: năng lượng, đóng gói và bộ nhớ, hoặc kiến trúc bộ nhớ	TSMC N2 khoảng 30.000 USD một wafer so với khoảng 3.000 USD cho N28; HBM và CoWoS khoảng 63% chi phí một H100
⑥ **Con chip đầu tiên tốn bao nhiêu, và nên mua nó thế nào?**	Theo bốn giai đoạn: mô hình, FPGA, wafer đa dự án, rồi mặt nạ. Dừng trước mặt nạ nếu luật lệ hoặc người mua chưa chắc chắn	Một suất MPW thương mại 50.000–500.000 USD; mặt nạ N7 khoảng 3,5 triệu USD; ngân sách thiết kế 7/6 nm 15–100 triệu USD

HÌNH 01 **AI × crypto qua sáu mảng: nhu cầu trả tiền thật nằm ở đâu**

Nhu cầu trả tiền có thể kiểm chứng so với tín hiệu trợ cấp và định giá, tháng 11 năm 2026

Chuỗi công việc hữu ích — 01

THẬT NHƯNG NHỎ

≈70 tr USD/năm

doanh thu bên ngoài kiểm chứng được của Bittensor

Tín hiệu: phát hành gấp 4–11× doanh thu; Pearl ≈1% khối AI

Cơ chế chạy được; tính hữu ích cần người trả tiền

Điện toán phi tập trung — 02

THẬT NHƯNG NHỎ

2,35 tr USD/năm

phí on-chain của Render, lớn nhất kiểm chứng được

Tín hiệu: vốn hóa ≈360× phí

Chủ yếu bán lại suy luận phổ thông

Huấn luyện và suy luận — 03

LÃI VỀ CỔ PHẦN

100+ tr USD/năm

Prime Intellect, công ty cổ phần dùng GPU tập trung

Tín hiệu: mainnet huấn luyện của Gensyn chưa chạy

Tiến bộ là thật; lợi nhuận về cổ phần

AI kiểm chứng được — 04

CÒN SỚM

398

người dùng trả tiền của đám mây TEE Phala, cuối 2025

Tín hiệu: chưa công bố zkML trả phí

TEE đã vào vận hành; zkML còn sớm

Thanh toán tác tử — 05

CHƯA CÓ NHU CẦU

5–11 nghìn USD/tháng

thanh toán x402 thực sự do tác tử khởi tạo

Tín hiệu: ≈200 tr giao dịch, phần lớn là tín hiệu

Hạ tầng có trước; nhu cầu thì chưa

Mạng dữ liệu — 06

CÓ NGƯỜI MUA

17 tr USD

doanh thu dữ liệu của Grass, nửa đầu 2026 (tự báo cáo)

Tín hiệu: vốn hóa ≈10× doanh thu năm hóa

Một trong số ít mảng có người mua thật

Nguồn: DefiLlama, SubConnect (qua hoge.gg), TRM Labs, Phala, Dealroom, SiliconANGLE, CoinGecko, Blockworks; tổng hợp của tác giả.

01

PHẠM VI VÀ PHƯƠNG PHÁP

Phạm vi, phương pháp và bốn phép thử

Vì sao có báo cáo này, và vì sao là lúc này

Tháng 4 năm 2026, một blockchain mới, Pearl, đưa một ý tưởng lâu đời vào vận hành thực tế: để thợ đào cạnh tranh giành khối bằng cách làm chính các phép nhân ma trận mà AI chạy trên đó, thay vì băm. Trong vòng sáu tháng, mạng của nó đạt khoảng 90 EH/s và token của nó đạt vốn hóa khoảng 500 triệu USD. Cùng lúc đó, các chợ điện toán AI, mạng huấn luyện phi tập trung, hạ tầng thanh toán cho tác tử và mạng dữ liệu huy động hàng tỷ đô la, và các thợ đào bitcoin bắt đầu chuyển các địa điểm có nguồn điện của họ thành trung tâm dữ liệu AI. Tôi muốn biết những thứ nào trong số này có khách hàng thật, silicon tùy biến có hợp lý ở chỗ nào trong đó không, và một đội chip thực sự có thể dùng tiến trình nào và xưởng đúc nào.

Bốn phép thử cho mọi dự án AI × crypto

Một nghiên cứu bảo mật năm 2017 về Gridcoin lập luận rằng tính toán hữu ích chỉ có thể bảo vệ một blockchain nếu nó khó giải, dễ kiểm chứng và dễ thưởng. Tôi thêm phép thử thứ tư. Cùng nhau, chúng quyết định phần lớn những gì tiếp theo.

01 **Khó giải.** Tính toán trung thực phải là cách rẻ nhất để tạo ra một bằng chứng hợp lệ; không được có đường tắt.

02 **Dễ kiểm chứng.** Kiểm tra công việc phải tốn ít hơn nhiều so với làm nó, nếu không mạng phải trả tiền hai lần.

03 **Dễ thưởng.** Giao thức phải tính được phần thưởng chỉ từ bằng chứng, không cần tin một bên thứ ba.

04 **Có người trả tiền cho đầu ra.** Trừ khi một khách hàng bên ngoài mạng trả tiền cho kết quả, công việc "hữu ích" chỉ là trợ cấp kèm một câu chuyện.

Câu đố băm vượt qua ba phép thử đầu và trượt phép thứ tư. Các tác vụ AI thật vượt qua phép thứ tư nhưng trượt phép thứ hai: chứng minh một phép tính AI bằng zero knowledge tốn gấp 100 đến 10.000 lần chính phép tính đó, và các cơ chế proof-of-learning đã bị chứng minh là có thể giả mạo với chi phí thấp. Phép nhân ma trận có nhiễu của Pearl là cây cầu đáng tin đầu tiên vượt qua phép thử kiểm chứng mà vẫn giữ nguyên hình dạng của tính toán AI.

Tôi đã làm như thế nào

Đây là nghiên cứu tại bàn, được làm theo cách tôi vẫn thẩm định một con chip hay một kế hoạch kinh doanh. Tôi đã đọc mã nguồn, các đề xuất giao thức và whitepaper của Pearl; dữ liệu chuỗi từ một full node độc lập; hồ sơ công bố và báo cáo kết quả kinh doanh của các công ty; các bài báo về mật mã học, kiến trúc máy tính và học máy; dữ liệu giá của xưởng đúc và nhà cung cấp; cùng báo chí chuyên ngành và công nghệ. Mọi con số đều kèm ngày. Con số nào tôi tự suy ra, tôi nói rõ và đánh dấu ước tính. Con số chỉ do bên bán công bố được đánh dấu hãng tự công bố; cách tôi diễn giải bằng chứng được đánh dấu phân tích. Tiền tệ tính bằng đô la Mỹ. Tình trạng tính đến ngày 13 tháng 11 năm 2026.

CHƯƠNG 02

02 PEARL · CƠ CHẾ Pearl: phép nhân ma trận làm đồng thuận

Ý tưởng, từ định lý đến mainnet

Tháng 4 năm 2025, Ilan Komargodski và Omri Weinstein công bố "Proofs of Useful Work from Arbitrary Matrix Multiplication". Họ xây dựng một bằng chứng công việc hữu ích cho mọi tích ma trận với chi phí phụ trội chỉ 1 + o(1) lần một phép nhân ngây thơ; tính bảo mật của nó dựa trên độ khó của việc giải các lô phương trình tuyến tính hạng thấp ngẫu nhiên, đây là một giả thuyết, không phải một chứng minh. Bài báo gọi tên khó khăn thật sự của lĩnh vực: một bằng chứng công việc hữu ích với chi phí phụ trội hệ số hằng thì tầm thường nhưng vô dụng; phần khó là giữ cho tính toán trung thực luôn là cách rẻ nhất để thắng khi thợ đào tự chọn đầu vào.

Mainnet của Pearl khởi chạy ngày 27 tháng 4 năm 2026, lúc 09:00 UTC. Thợ đào tính một tích có nhiễu (A + E)(B + F), trong đó nhiễu hạng thấp được suy ra từ một cam kết BLAKE3, rồi loại bỏ nhiễu để khôi phục A·B. Trạng thái bộ tích lũy của mỗi ô đầu ra được băm thành một vé số; một ô trúng được mở ra và chứng minh bằng một bằng chứng zero-knowledge Plonky2 đệ quy dưới 60 KB, bên trong một chứng chỉ khối giới hạn ở 65 KB. Việc kiểm chứng là một bằng chứng nhỏ có kích thước không đổi, độc lập với kích thước ma trận: rẻ để kiểm tra như một câu đố băm, trong khi bản thân phép tính có thể hữu ích.

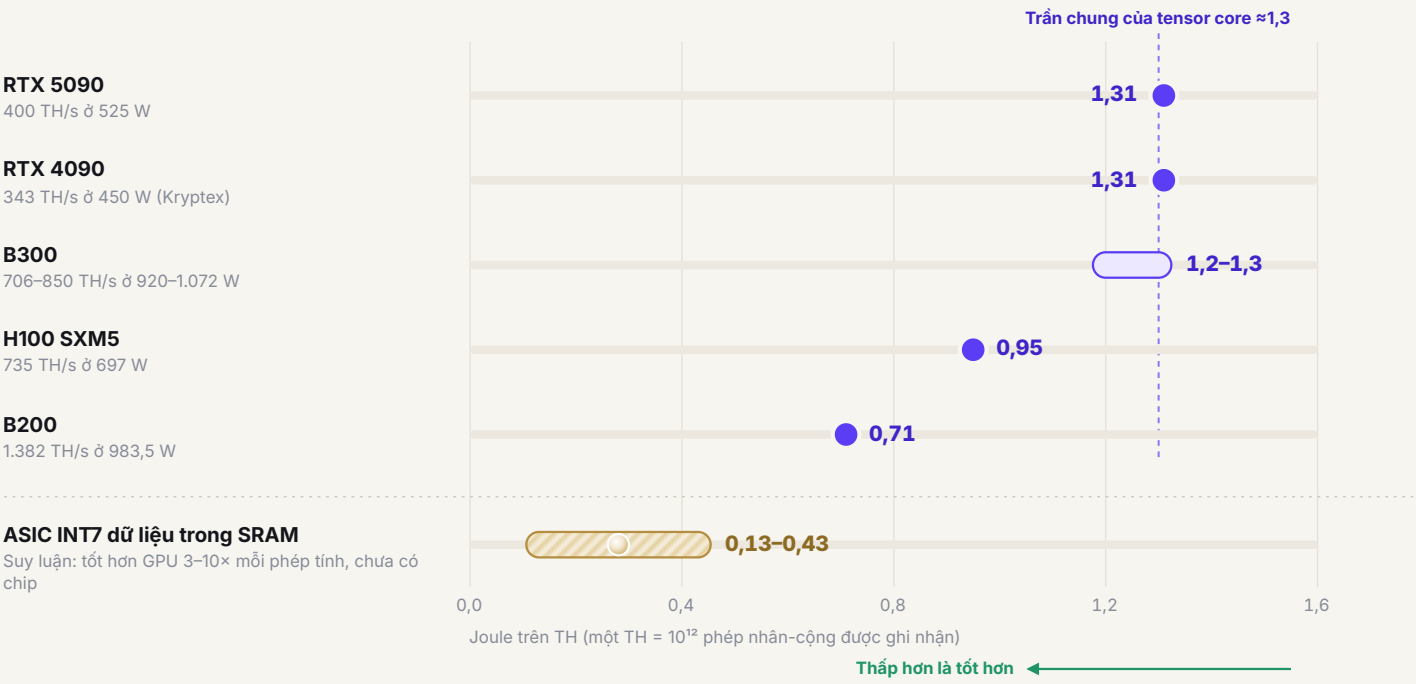
Một phép nhân-cộng là một "hash"

Với người thiết kế chip, sự thật quan trọng nhất là đơn vị. Mã đồng thuận của Pearl co giãn mục tiêu của mỗi lần thử theo kích thước ô nhân với độ dài tích vô hướng, nên **1 H chính xác là một phép nhân-cộng (MAC) INT7 × INT7 → INT32 được ghi nhận**. Từ khối 96.251, một hình phạt hạng yêu cầu hạng nhiễu tối thiểu 128; ở hạng tối thiểu, 1 H vẫn bằng 1 MAC. Do đó "EH/s" của Pearl theo đúng nghĩa đen là 10^18 phép nhân-cộng số nguyên mỗi giây: **lần đầu tiên, ngân sách bảo mật của một blockchain được định giá bằng thông lượng tensor của các bộ tăng tốc AI.** Hashrate GPU đo được rơi vào khoảng 60% đến 95% đỉnh nhân-cộng INT8 dense của từng card, khớp với cách đọc này.

Phần cứng	Hashrate Pearl	Công suất	Joule trên nghìn tỷ MAC (J/TH)	Nguồn (ngày)
RTX 5090	400 TH/s	525 W	1,31 ước tính	Kryptex (12 tháng 11)
RTX 4090	khoảng 305 TH/s (dải cộng đồng 280–330); Kryptex ghi 343 TH/s	450 W (Kryptex)	1,31 (số liệu Kryptex) ước tính	prlstats (3 tháng 10); Kryptex
H100 SXM5	735 TH/s	697 W	0,95 ước tính	PeakMiner, qua prlstats (19 tháng 9)
B200	1,382 PH/s	983,5 W	0,71 ước tính	PeakMiner, qua prlstats
B300	706–850 TH/s	920–1.072 W	1,2–1,3 ước tính	prlstats; SimeonOnSecurity

HÌNH 02 **Với công việc số nguyên của Pearl, các GPU chung một trần năng lượng**

Joule trên mỗi nghìn tỷ phép nhân-cộng INT7 (J/TH) ở hashrate Pearl đo được; thấp hơn là tốt hơn



Nguồn: Kryptex (12/11/2026); benchmark PeakMiner và SRBMiner qua prlstats; SimeonOnSecurity; dải ASIC là suy luận của tác giả, không phải số đo.

Hai chi tiết trong bảng này quan trọng với người thiết kế chip. Thứ nhất, các card tiêu dùng RTX 4090 và 5090 và chiếc B300 cho trung tâm dữ liệu hội tụ ở khoảng 1,3 J/TH, và ngay cả card tốt nhất, B200, cũng chỉ đạt khoảng 0,71 J/TH. Một GPU trung tâm dữ liệu giá hàng chục nghìn đô la tiêu tốn năng lượng cho mỗi phép nhân-cộng số nguyên gần bằng một card chơi game, điều này chỉ ra một trần chung về năng lượng của tensor core trên mỗi MAC INT8, và trần đó chính là thứ một con chip chuyên dụng có thể phá vỡ. Thứ hai, B300 chậm hơn B200: NVIDIA đã cắt mạnh thông lượng INT8 trong Blackwell Ultra, nên công việc số nguyên phải đi một đường chậm hơn trên con chip đó. Silicon AI chủ lưu đang rời khỏi INT8 để hướng tới FP8 và FP4. Khi định dạng số mà một cơ chế đồng thuận chọn đi chệch khỏi hướng phần cứng đang đi, giao thức buộc phải thay đổi luật, và đó là gốc rễ của fork FP8 cũng như của nhận định về ASIC của tôi ở phần sau báo cáo.

CHƯƠNG 03

03 PEARL · KINH TẾ
Pearl: quy mô, phát hành và khoảng cách hữu ích

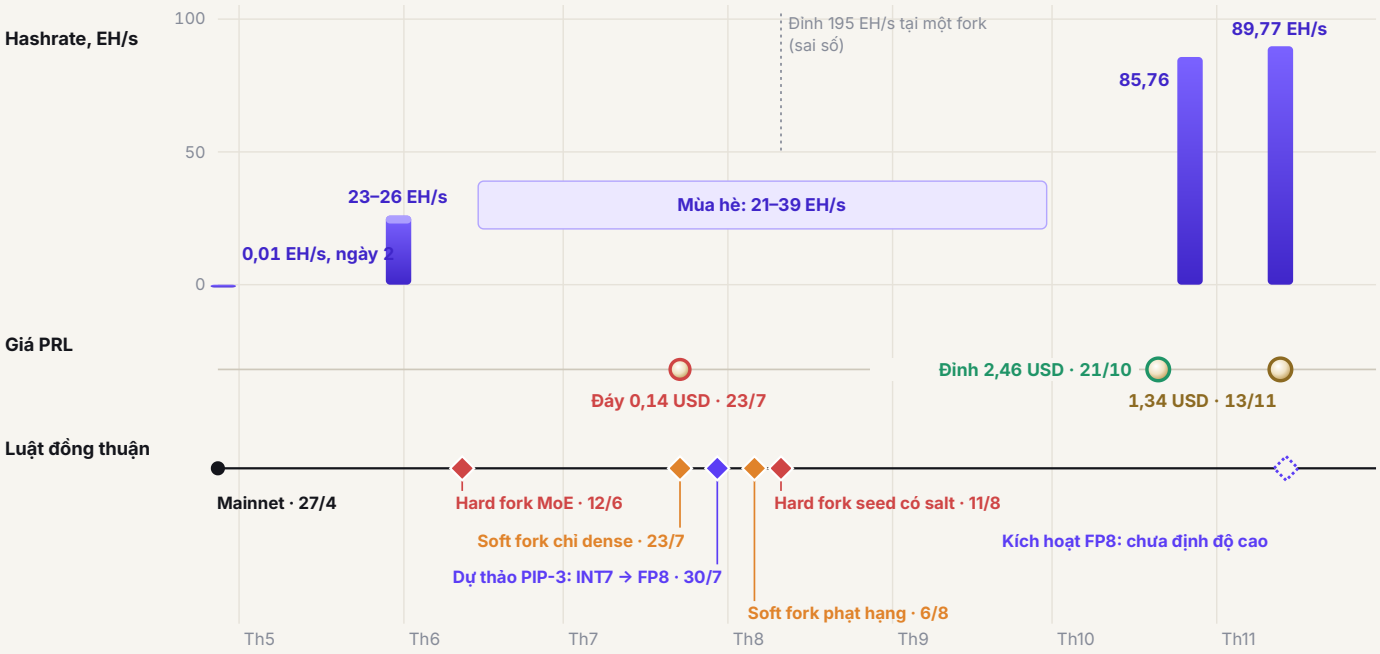
Từ 0,01 lên 90 EH/s trong sáu tháng

Chỉ số	Giá trị (13 tháng 11 năm 2026)	Cơ sở và nguồn
Hashrate mạng	Khoảng 90 EH/s (mốc cơ sở 89,77; dải trong ngày 88,20–93,22; đỉnh trong ngày 12 tháng 11 là 110–123)	Full node độc lập prlstats; Kryptex
Độ khó và độ cao khối	58.841.982 và 142.530	prlstats
Phần cứng tương đương	Khoảng 122.000 chiếc H100 hoặc 225.000 chiếc RTX 5090	Ở 735 và 400 TH/s mỗi card ước tính
Giá, vốn hóa và giá trị pha loãng hoàn toàn	1,32–1,34 USD; khoảng 500 triệu USD; khoảng 2,8 tỷ USD (hạng 101 trên CoinGecko)	CoinGecko
Đỉnh và đáy mọi thời đại	2,46 USD (21 tháng 10) và 0,1399 USD (23 tháng 7)	CoinGecko
Số coin đã đào	377,6 triệu PRL, 17,98% mức trần 2,1 tỷ	prlstats
Trợ cấp khối và phát hành hằng ngày	Khoảng 2.173 PRL và khoảng 0,97–1,03 triệu PRL	prlstats; Kryptex
Nguồn cung mới trong 12 tháng tới	Khoảng 78% nguồn cung hiện tại	prlstats
Thu nhập thợ đào	Khoảng 1,25 triệu USD mỗi ngày (mô hình), khoảng 460 triệu USD mỗi năm	prlstats
Doanh thu trên mỗi đơn vị hashrate, R	Khoảng 14–15 USD mỗi PH/s mỗi ngày; một RTX 5090 kiếm khoảng 5–6 USD mỗi ngày	ước tính ; Kryptex
Số khối chứa công việc AI thật	Khoảng 1% (ước tính ngày 28 tháng 10)	Blockworks
Mức tập trung của pool	Kryptex tìm được 46,4% số khối trong 24 giờ qua; hai pool lớn nhất khoảng 62%	prlstats
Mức tập trung của người nắm giữ	Hệ số Gini theo địa chỉ 0,969; 100 địa chỉ lớn nhất nắm 53,0%	prlstats

Hashrate của Pearl gần như là một hàm trễ của giá. Một ngày sau khi ra mắt, mạng chạy ở khoảng 0,01 EH/s; đến cuối tháng 5 đã đạt 23–26 EH/s. Suốt mùa hè, giá giảm khoảng 86% so với đỉnh tháng 5 và hashrate đi ngang ở 21–39 EH/s. Sau khi giá tăng khoảng mười lần trong tháng 9 và tháng 10, hashrate đạt 85,76 EH/s ngày 27 tháng 10 và khoảng 90 EH/s trong tháng 11. Một đỉnh vọt lên 195 EH/s ngày 11 tháng 8 trùng với một fork và nên được hiểu là sai số ước lượng. Năng lượng tăng theo: một nghiên cứu ước tính khoảng 112 MW cho toàn mạng ở 24 EH/s, và một mô hình của prlstats đặt con số hiện nay ở khoảng 190–215 MW.

HÌNH 03 **Sáu tháng đầu của Pearl: hashrate theo giá, luật lệ liên tục đổi**

Hashrate mạng, ba mốc giá và mọi thay đổi đồng thuận, từ 27/4 đến 13/11/2026



Nguồn: Full node và chuỗi số liệu prlstats; Kryptex; CoinGecko; params.go và PIP-3 của Pearl. Đỉnh ngày 11/8 là sai số ước lượng tại một fork.

Phát hành khuếch đại tính phản thân

Nguồn cung tối đa là 2,1 tỷ coin và không có halving: trợ cấp khối giảm dần mượt mà theo E(t) = S·H / ((t + H)(t + H − 1)) với H = 650.226. Nhưng phát hành được tính theo khối, không theo thời gian, và độ khó ban đầu khởi điểm quá thấp và điều chỉnh chậm. **Khoảng 44,06 triệu coin được đào trong ngày đầu tiên và khoảng 120 triệu trong sáu ngày đầu**, khoảng 32% nguồn cung hiện nay ước tính. Các khối gần đây, khoảng 171–183 giây, vẫn đến nhanh hơn mục tiêu 194 giây. Kết quả là nguồn cung sẽ tăng khoảng 78% trong 12 tháng tới, một mức pha loãng mà thợ đào, nhà sản xuất phần cứng và người nắm giữ đều phải đối mặt.

Kinh tế của thợ đào đi theo đó. Thu nhập thợ đào toàn mạng khoảng 1,25 triệu USD mỗi ngày (mô hình); lượng phát hành hằng ngày của Kryptex nhân với giá hiện tại cho khoảng 1,35 triệu USD mỗi ngày, nên **doanh thu trên mỗi đơn vị hashrate R ≈ 14–15 USD mỗi PH/s mỗi ngày** ước tính. Một RTX 5090 kiếm khoảng 5–6 USD mỗi ngày. Một H100 hoặc H200 kiếm khoảng 11 USD, trong khi giá thuê theo yêu cầu công khai rẻ nhất khoảng 2,09 USD mỗi card mỗi giờ, khoảng 50 USD mỗi ngày, nên thuê card chỉ để đào thì lỗ nặng. R dao động rất mạnh: theo chuỗi số liệu của prlstats, nó vào khoảng 10 USD trong tháng 8 và tháng 9 và có lúc khoảng 33 USD vào giữa tháng 10 ước tính. Chỉ vài tuần sau khi ra mắt, doanh thu hằng ngày ước tính của một RTX 5090 đã giảm một nửa, từ khoảng 33,80 USD xuống khoảng 17,19 USD. Thuê GPU đám mây để đào chỉ có lãi trong vài tuần đầu và trong đợt tăng giá tháng 10.

Định lý đứng vững; động lực thì không

Vấn đề thật là nhu cầu, không phải mật mã. Tích hợp thương mại công khai duy nhất của Pearl là endpoint Gemma-4-31B-it-pearl của Together AI, ra mắt ngày 15 tháng 5 với giá thấp hơn mức thông thường hơn 25%, phần chiết khấu được tài trợ bằng giá trị tương lai kỳ vọng của lượng phát hành. Ngày 13 tháng 11, trang mô hình cho thấy nó **không còn có trên API serverless của Together**. Vào tháng 6, một cuộc khảo sát 8.012 phiên đào trực tuyến, khoảng 21% mạng, kết luận rằng mạng hoàn toàn không tạo ra phép tính AI hữu ích nào (tác giả gọi bằng chứng là có tính gợi ý, chưa phải kết luận), và cho thấy các share dựng từ ma trận hoàn toàn ngẫu nhiên vẫn

được pool chấp nhận. Ngày 28 tháng 10, Blockworks ước tính rằng **chỉ khoảng 1% số khối chứa công việc AI thật**; nếu không có nó, theo bài viết, Pearl là một bitcoin nhỏ hơn, kém an toàn hơn và lạm phát nhanh hơn.

Lý do có thể tính ra được. Chi phí phụ trội phía suy luận thực sự nhỏ: Pearl báo cáo thời gian đầu-cuối tăng 5,08% cho Llama 70B trên bốn H200, và một thử nghiệm độc lập đo được mức giảm thông lượng 9,2% trên một GPU Hopper. Phía đào thì kiếm được quá ít. Plug-in vLLM chính thức chỉ đào trên các lô prefill từ 1.024 token trở lên và không bao giờ đào ở các bước decode. Một card phục vụ suy luận nhận được khoảng 27% hashrate của máy đào chuyên dụng với lưu lượng thuần prefill, khoảng 5,4% với hỗn hợp chat có 80% là decode, và chỉ khoảng 1% với lưu lượng do decode chiếm ưu thế. Không có nhu cầu suy luận trả tiền, thợ đào duy lý sẽ nhân các ma trận ngẫu nhiên; giao thức kiểm chứng rằng các ma trận đã cam kết được nhân một cách trung thực, chứ không kiểm chứng chúng có hữu ích hay không. **Định lý đứng vững; động lực thì không.** Tính hữu ích của Pearl là một bài toán thương mại ngoài chuỗi: phải có người trả tiền cho suy luận.

CHƯƠNG 04



04 PEARL · LUẬT LỆ
Pearl: rủi ro luật lệ, số học và quản trị

Bốn fork trong hai tháng

Mặt kia của Pearl là tốc độ thay đổi luật lệ. Từ tháng 6 đến tháng 8 năm 2026, phòng thí nghiệm đã đẩy ra bốn thay đổi đồng thuận ở các độ cao được mã hóa cứng: một hard fork mixture-of-experts (khối 71.935, khoảng ngày 12 tháng 6), một soft fork chỉ dense (91.630, khoảng ngày 23 tháng 7), một soft fork phạt hạng (96.251, khoảng ngày 6 tháng 8) và một hard fork seed có salt (99.000, khoảng ngày 11 tháng 8). Sau fork seed có salt, máy đào cũ chỉ có thể tạo ra share không hợp lệ; hình phạt hạng cắt giảm hashrate hiệu dụng của các cấu hình hạng cao từ 50% đến 87,5%.

Bước tiếp theo là PIP-3, vẫn còn là dự thảo: nó chuyển việc đào từ INT7 sang FP8 × FP8 → FP32 và yêu cầu hồ sơ thiết bị chính xác từng bit, **chỉ hỗ trợ hai kiến trúc, Hopper và Blackwell**. Tài liệu gọi sự phụ thuộc phần cứng là một điều ác cần thiết của các giao thức dấu phẩy động và cảnh báo rằng bất kỳ dung sai nào trong kiểm chứng cũng mời gọi tấn công grinding. Ngày 12 tháng 11, tham số mainnet của nhánh fp8-scheme chưa đặt độ cao kích hoạt. Theo dự thảo hiện tại, dòng RTX 40, A100 và card AMD không được chấp nhận, trong khi một mô hình độ tin cậy thấp của prlstats ước tính khoảng 82% hashrate đến từ card tiêu dùng. Ép fork này thông qua có nguy cơ gây tách chuỗi hoặc phản ứng dữ dội từ thợ đào.

Thế tam nan của định dạng số

Pearl đối mặt một thế tam nan: một định dạng số có thể hữu ích cho AI (FP8 hoặc FP4), kiểm chứng được chính xác từng bit trên mọi phần cứng, và trung lập phần cứng, nhưng chỉ đạt tối đa hai trong ba. Cơ chế INT7 kiểm chứng được và trung lập nhưng đi chệch hướng silicon AI. FP8 với danh sách trắng thì hữu ích và kiểm chứng được nhưng buộc đồng thuận vào một nhà cung cấp. FP8 với dung sai thì hữu ích và trung lập nhưng mở cửa cho tấn công grinding. Số học dấu phẩy động trong tensor core nhìn chung không tuân theo một chuẩn số học và không tái lập được giữa các thế hệ; chẳng hạn trên Hopper, phép tích lũy FP8 chỉ giữ 13 bit phần phân số và cắt bớt phần còn lại. Từ góc nhìn của người thiết kế chip, một "hồ sơ Hopper" là một phép số học được xác định hoàn toàn mà một ASIC có thể hiện thực chính xác từng bit, có lẽ còn hiệu quả hơn GPU. Nhưng làm vậy sẽ là kinh doanh trong vùng xám, đi ngược ý định đã tuyên bố của giao thức, với rủi ro quản trị rất cao phân tích.

Quản trị

Quyền quản trị tập trung. Phòng thí nghiệm viết mọi đề xuất giao thức; các fork kích hoạt ở độ cao mã hóa cứng mà không có tín hiệu từ thợ đào; không có bản hiện thực client thứ hai; và tính đến đầu tháng 10 chưa có báo cáo kiểm toán bên thứ ba nào về node hay các mạch zero-knowledge được công bố. Tờ báo Israel TheMarker đưa tin phòng thí nghiệm đã huy động khoảng 120–130 triệu USD trong một vòng gọi vốn kín (khoảng 55 triệu USD vốn cổ phần cộng một quỹ hạ tầng GPU khoảng 75 triệu USD); Pearl chưa xác nhận qua kênh chính thức. "Không có phân bổ token cho nhà đầu tư" và "một công ty có cổ đông mạo hiểm" có thể cùng đúng, nhưng mâu thuẫn vẫn còn: doanh thu AI có thể chảy về cổ phần trong khi pha loãng đổ lên người nắm giữ token.

Tổng hợp lại, Pearl là thí nghiệm "tính toán là đồng thuận" sạch sẽ nhất cho đến nay: một cơ chế mới, kiểm chứng được trên chuỗi, với một đơn vị ánh xạ một-một lên phần cứng AI. Nhưng tính hữu ích chưa đến, và ngân sách bảo mật, hashrate cũng như sự hăng hái của thợ đào đều treo trên một token mà nhu cầu chính chưa hình thành và nguồn cung sẽ tăng khoảng 78% trong 12 tháng. Ba tín hiệu từ Blockworks đáng theo dõi: bản nâng cấp có diễn ra suôn sẻ không, tỷ lệ khối có công việc AI có tăng đáng kể không, và có nhà cung cấp suy luận nào chuyển thu nhập đào cho khách hàng của mình không.

CHƯƠNG 05

CHUỖI CÔNG VIỆC HỮU ÍCH

05 Chuỗi công việc hữu ích: cả ma trận lẫn zero knowledge đều không chặn được ASIC

Kaspa: ma trận không hữu ích, cũng không chống được ASIC

Kaspa là trường hợp kinh điển. kHeavyHash của nó đặt một tích ma trận-vector 64 × 64 trên các giá trị 4 bit vào giữa hai lần băm, và tích đó bị vứt bỏ. Khoảng 16 tháng sau đợt ra mắt công bằng vào tháng 11 năm 2021, IceRiver công bố các ASIC Kaspa đầu tiên ngày 14 tháng 3 năm 2023, và đến cuối năm 2023 thì việc đào bằng GPU chấm dứt. So sánh một RTX 4090 ở 3,8 GH/s và 300 W với KS7 của Bitmain ở 77 J/TH cho khoảng cách hiệu quả khoảng **1.000 lần** ước tính. Sau bữa tiệc là sự co lại: ngày 13 tháng 11, mạng chạy ở 358,9 PH/s, thấp hơn khoảng 64% so với mức 1 EH/s của tháng 8 năm 2024, và chiếc KS7 tốt nhất chỉ kiếm **0,86 USD mỗi ngày** với giá điện 0,10 USD mỗi kWh.

Aleo: zero knowledge cũng không chặn được phần cứng chuyên dụng

Câu đố prover của Aleo được thiết kế để mô phỏng công việc tạo bằng chứng SNARK; nó không chứng minh giao dịch của người dùng, và dự án công khai chào đón ASIC ZK. Khoảng năm tháng sau mainnet tháng 9 năm 2024, máy chuyên dụng đã được giao hàng. Trong quý 3 năm 2025, các prover nhận 31,2 triệu ALEO (khoảng 7,8 triệu USD) trong khi mạng chỉ thu khoảng 43.000–44.000 USD phí: **bội số trợ cấp khoảng 175–180**. Giá giảm từ khoảng 9 USD lúc mới giao dịch xuống 0,22 USD vào cuối quý đó, và ARC-46 nay yêu cầu stake từ 100.000 đến 2,5 triệu ALEO cho mỗi lời giải, biến việc đào thành một ngưỡng stake.

Qubic và Bittensor

Qubic quảng bá "bằng chứng công việc hữu ích" để huấn luyện mạng nơ-ron Aigarth của mình, nhưng đầu ra dễ thấy nhất của sức tính của nó trong 2025–2026 lại là khối của các chuỗi khác. Đầu tiên nó đào Monero, nâng thị phần từ dưới 2% vào giữa tháng 5 năm 2025 lên hơn 25% vào cuối tháng 7 và tuyên bố đạt 51% vào tháng 8, gây

ra một đợt tái tổ chức sáu khối (một nguồn khác nói 18). Ngày 1 tháng 4 năm 2026, nó chuyển sang đào Dogecoin bằng ASIC, dùng tiền thu được để mua lại và đốt token. Không có bằng chứng độc lập nào cho thấy việc huấn luyện AI của nó hữu ích, và vốn hóa của nó giảm từ khoảng 280–300 triệu USD xuống khoảng 50 triệu USD ngày 13 tháng 11. Nó cũng phơi bày một ngoại ứng: một chuỗi trả tiền cho sức tính đa dụng tạo ra một khối hashrate lớn, cho thuê được và được điều phối, có thể bị quay sang tấn công các chuỗi thân thiện với CPU.

Bittensor không phải bằng chứng công việc mà là chấm điểm chủ quan theo trọng số stake: validator chấm điểm thợ đào, và trọng số bị cắt ở trung vị theo trọng số stake. Đây là mạng duy nhất trong nhóm có doanh thu bên ngoài ở quy mô lớn. Chỉ số SubConnect, vốn chỉ đếm khách hàng bên ngoài có thể kiểm chứng, cho thấy khoảng **70 triệu USD mỗi năm** vào đầu tháng 11 năm 2026 (khoảng 35 triệu USD ngày 1 tháng 9), 78% đến 82% trong đó đến từ điện toán và hạ tầng; nhưng ở mức giá đầu tháng 11, lượng phát hành của nó vượt 300 triệu USD mỗi năm, nên doanh thu chỉ bù được 9% đến 23%. Các ước tính doanh thu chênh lệch lớn tùy cơ sở tính, từ khoảng 17 triệu USD (Artemis) đến 172 triệu USD năm hóa từ tổng số tự báo cáo. Vào tháng 4, 67 trên 128 subnet đã đốt hơn 99% phần khuyến khích của mình, và khi đội dẫn đầu Covenant rời đi vì bất đồng quản trị, TAO giảm 24% trong vài giờ.

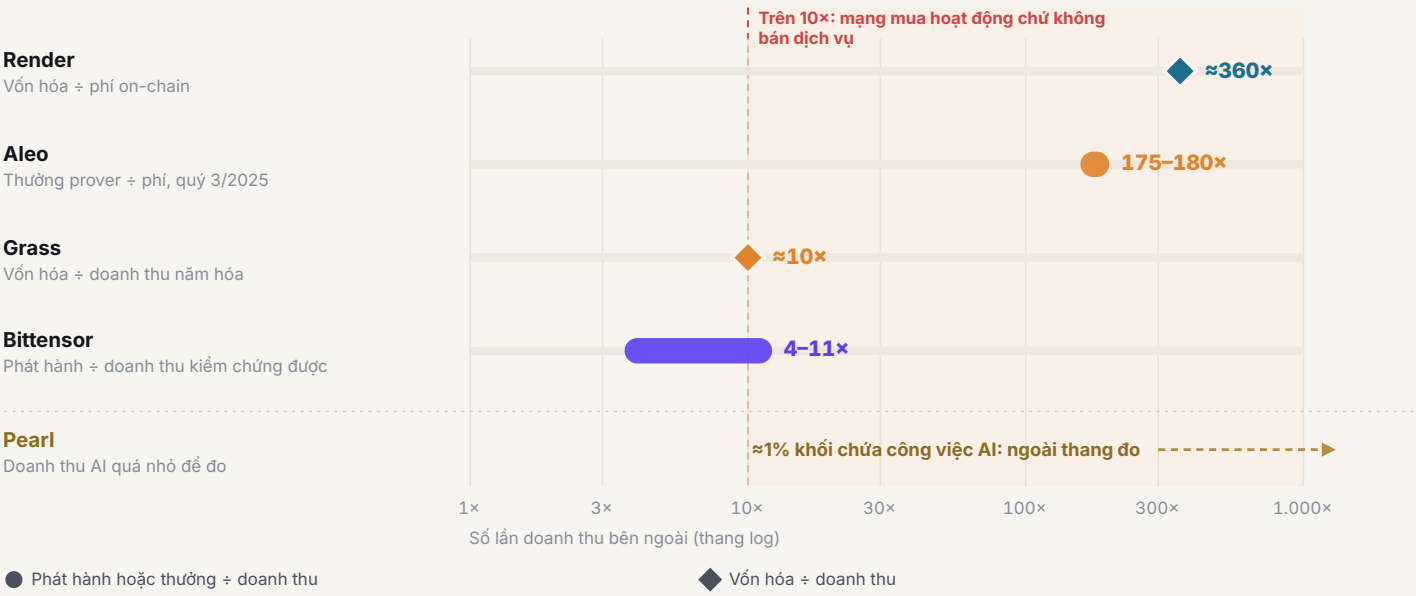
Lịch sử cho cùng một câu trả lời

Chuỗi số nguyên tố của Primecoin kiểm chứng nhanh, nhưng không ai cần đến chúng. Gridcoin dựa vào các trang thống kê bên thứ ba để xác nhận đóng góp BOINC, và các nhà nghiên cứu đã làm giả khối mà không cần tính toán gì. Bằng chứng lưu trữ của Filecoin hoạt động về mặt kỹ thuật, nhưng trong quý 3 năm 2025 mức sử dụng chỉ đạt 36%, và 99,5% trong số 793.000 USD phí của nó là tiền phạt, với phí cơ sở chỉ 4.100 USD.

Dự án	Công việc hoặc đồng thuận	Vốn hóa (ngày)	Lộ trình phần cứng	Nhu cầu trả tiền so với trợ cấp	Nhận định
Pearl	Phép nhân ma trận có nhiễu (MAC INT7 cộng chứng chỉ ZK)	Khoảng 500 triệu USD (13 tháng 11)	GPU; dự thảo FP8 sẽ đưa Hopper và Blackwell vào danh sách trắng	Khoảng 1% số khối có công việc AI; thu nhập thợ đào phần lớn là trợ cấp	Cơ chế đáng tin nhất; nhu cầu chưa được chứng minh
Kaspa	kHeavyHash (tích ma trận-vector 64 × 64, 4 bit giữa các lần băm)	Khoảng 840 triệu USD (13 tháng 11)	ASIC khoảng 16 tháng sau khi ra mắt; nay hoàn toàn là ASIC	Không tuyên bố tính hữu ích; doanh thu thợ đào khoảng 18 triệu USD mỗi năm ước tính	Ma trận không hữu ích, cũng không chống được ASIC
Aleo	Câu đố SNARK tổng hợp cộng bằng chứng cổ phần	Khoảng 118 triệu USD (cuối quý 3 năm 2025)	Máy chuyên dụng sau khoảng năm tháng	Phần thưởng prover gấp khoảng 175–180 lần phí	Công việc tổng hợp, chạy bằng trợ cấp
Qubic	"uPoW" (tuyên bố huấn luyện Aigarth)	Khoảng 50 triệu USD (13 tháng 11)	CPU và GPU; ASIC cho Dogecoin từ tháng 4 năm 2026	Không có bằng chứng độc lập về tính hữu ích; sức tính dùng để đào bên ngoài và mua lại token	Gom sức tính khoác câu chuyện AI
Bittensor	Đồng thuận Yuma (chấm điểm theo trọng số stake)	Khoảng 2,55 tỷ USD (11 tháng 11)	GPU đa dụng; không có lộ trình ASIC	Khoảng 70 triệu USD doanh thu kiểm chứng được mỗi năm so với hơn 300 triệu USD phát hành	Mạng duy nhất có doanh thu bên ngoài ở quy mô lớn; vẫn là trợ cấp gấp 4–11 lần
Filecoin (tham chiếu)	Bằng chứng sao chép và không-thời gian	Khoảng 1,5 tỷ USD (cuối quý 3 năm 2025)	—	Mức sử dụng 36%; phí cơ sở hằng quý 4.100 USD	Hữu ích, nhưng thiếu nhu cầu

HÌNH 04 **Bội số trợ cấp: mạng trả gấp bao nhiêu lần doanh thu bên ngoài**

Phát hành hoặc vốn hóa chia cho doanh thu trả từ bên ngoài mạng, thang log



Nguồn: Blockworks (Aleo quý 3/2025), DefiLlama (phí Render), CoinGecko, hoge.gg (Bittensor), Dealroom (Grass), Blockworks 0xResearch (Pearl); phép tính của tác giả.

Bài học cho phần cứng: công việc cố định, đều đặn, được trả bằng một token thanh khoản sẽ có ASIC trong khoảng 3 đến 16 tháng, thường đến từ những kẻ thách thức trước cả các hãng đương nhiệm. Tuyên bố "hữu ích" chỉ đứng vững khi có khách hàng bên ngoài trả tiền cho đầu ra. Và các nguyên nhân chính khiến phần cứng bị mắc kẹt là trợ cấp cao so với doanh thu và quyền kiểm soát của nhà phát triển đối với hàm công việc.

CHƯƠNG 06

BỨC TRANH RỘNG HƠN

06 Phần còn lại của AI × crypto: nhu cầu thật nhưng nhỏ, và giá trị đang chuyển sang cổ phần

Thị trường điện toán phi tập trung

Điện toán phi tập trung có khách hàng trả tiền, nhưng nhu cầu nhỏ, tập trung và phần lớn là bán lại suy luận mô hình mở. Phí kiểm chứng được trên chuỗi rất nhỏ: DefiLlama cho thấy khoảng 2,35 triệu USD phí của Render trong năm qua và khoảng 680.000 USD của Livepeer, và toàn bộ mảng DePIN chỉ thu được khoảng 72 triệu USD doanh thu on-chain trong năm 2025. Các con số tự báo cáo lớn hơn nhưng được định nghĩa lỏng lẻo: Aethir báo cáo doanh thu "hơn 127,8 triệu USD" cho năm tài chính 2025 mà không nói đó là tổng hóa đơn hay phần giao thức giữ lại. Giá trị token và doanh thu đã tách rời theo cả hai chiều: vốn hóa của Render gấp khoảng 360 lần phí on-chain, trong khi vốn hóa khoảng 87,7 triệu USD của Aethir thấp hơn doanh thu tự báo cáo của nó ước tính.

Tuyên bố rẻ hơn 70% đến 85% chỉ đúng khi so với giá theo yêu cầu của các hyperscaler. Ngày 3 tháng 11, giá H100 bình quân gia quyền trên Akash khoảng **2,73 USD mỗi GPU-giờ** (dải 2,04–3,16 USD), thấp hơn mức khoảng 6,88 USD trên AWS và 11,06 USD trên Google Cloud, nhưng cao hơn mức khoảng 1,73 USD trên Vast.ai, 1,99 USD trên community cloud của RunPod và khoảng 2,44 USD cho gói đặt trước một năm (giá trên các nền tảng khác đọc ngày 12 tháng 9; Akash, bên công bố, là một bên có lợi ích liên quan). Nhu cầu chủ yếu là suy luận phổ thông được phân phối qua OpenRouter, khoảng 1,7 tỷ token mỗi ngày trên AkashML và hơn 4 tỷ trên io.net, cạnh tranh trực diện với mọi neocloud. Những đồng đô la "điện toán phi tập trung" lớn nhất đang tái tập trung: Axe Compute, công ty nắm giữ token của Aethir như một tài sản ngân quỹ, đã ký hợp đồng năm năm trị giá 1,5 tỷ USD để cho thuê cụm 9.200 chiếc B300 của riêng mình, trong khi doanh thu quý 1 năm 2026 là 35.000 USD.

Huấn luyện và suy luận phi tập trung

Huấn luyện qua internet đã được chứng minh ở quy mô 8 đến 40 tỷ tham số. Pluralis-8B được huấn luyện trong 40 ngày trên 330 node, phần lớn là GPU tiêu dùng, và đạt 63% hiệu quả của một mốc cơ sở H100 tập trung, dù bài báo của họ thừa nhận đợt chạy này không phi tập trung và cần các node riêng của Pluralis cùng một cơ quan cấp quyền tham gia. Psyche của Nous Research đã hoàn tất đợt tiền huấn luyện Consilience 40 tỷ tham số. Ở quy mô tiên phong, công việc lại tái tập trung: INTELLECT-3 của Prime Intellect, một mô hình mixture-of-experts 106 tỷ tham số, được huấn luyện trong hai tháng trên 512 chiếc H200 kết nối bằng InfiniBand 400 Gbps. Các ràng buộc rất rõ: node cần băng thông ít nhất 200 Mbit/s và nén truyền thông khoảng 100 lần, còn niềm tin và kiểm chứng vẫn phụ thuộc vào các vai trò đặc quyền của bên vận hành.

Tiền đã chảy vào các công ty cổ phần bán sản phẩm tập trung. Prime Intellect có hơn 100 triệu USD doanh thu năm hóa trước khi gọi được 130 triệu USD ở mức định giá 1 tỷ USD vào tháng 7; Venice, một ứng dụng suy luận riêng tư, báo cáo hơn 70 triệu USD doanh thu năm hóa và có lãi từ quý 1 năm 2026 (số liệu của công ty), đã gọi 65 triệu USD ở mức định giá 1 tỷ USD và chạy trên GPU đi thuê. Ở phía token, mạng huấn luyện cốt lõi của Gensyn vẫn "chưa xác định" và token của nó trị giá khoảng 17,2 triệu USD; token của Ritual, một layer-1 gốc AI, đã giảm 72% so với đỉnh sau mainnet (chỉ theo tiêu đề tin).

AI kiểm chứng được

Vận hành thực tế đã chọn môi trường thực thi tin cậy (TEE). Điện toán bảo mật của Hopper thêm dưới 7% chi phí phụ trội cho các truy vấn LLM điển hình và gần như bằng không với các mô hình lớn; chi phí phụ trội vận hành của Phala khoảng 0,5% đến 5%, nhưng đến cuối năm 2025 chỉ 398 trong số 10.004 người dùng của nó trả tiền. Gốc tin cậy của một TEE là nhà sản xuất chip: TEE.fail, công bố tháng 10 năm 2025, dùng một interposer bus DDR5 giá dưới 1.000 USD để trích xuất khóa attestation từ các máy Intel TDX và AMD SEV-SNP đã cập nhật đầy đủ, điều này đến lượt nó làm suy yếu điện toán bảo mật GPU của NVIDIA, và cả hai nhà sản xuất đều nói tấn công vật lý nằm ngoài mô hình mối đe dọa của họ. Đó chính là bối cảnh của DePIN, nơi bên vận hành node có quyền truy cập vật lý. zkML nay đã có thể chứng minh một LLM hoàn chỉnh: DeepProve của Lagrange chứng minh GPT-2 ở khoảng 174 token mỗi phút và Gemma 3 ở khoảng 86, cách suy luận thực tế nhiều bậc độ lớn và chưa có ứng dụng trả tiền nào được công bố.

Thanh toán của tác tử và dữ liệu

x402 là ví dụ cảnh tỉnh rõ nhất. TRM Labs đếm được khoảng 198,9 triệu giao dịch và 52,7 triệu USD được các facilitator đã biết thanh toán kể từ tháng 5 năm 2025, 99,6% bằng USDC; sau khi loại bỏ tự thanh toán, các dòng theo lô và người bán có ít hơn mười người mua, phần thực sự do tác tử khởi tạo chỉ còn 0,6% đến 7,5%, **khoảng 5.000–11.000 USD mỗi tháng**. Artemis đánh giá hơn 95% hoạt động là tín hiệu giao thức chứ không phải mua bán. Trong khi đó hạ tầng đang được các hãng đương nhiệm chuẩn hóa: x402 Foundation được thành lập dưới Linux Foundation ngày 14 tháng 7, với Visa, Mastercard, Stripe, Google, AWS, Cloudflare và Coinbase là thành viên cốt lõi, và Tempo, đơn vị cùng Stripe soạn Machine Payments Protocol, ra mắt mainnet vào tháng 3 ở mức định giá 5 tỷ USD. Các nền tảng token tác tử vẫn vận hành theo logic launchpad: Virtuals tự báo cáo 479 triệu USD "aGDP", và nhà sáng lập ElizaOS tuyên bố token của nó đã chết sau một vụ dàn xếp kiện tập thể. Giá trị nhiều khả năng sẽ được các nhà phát hành stablecoin và các hãng thanh toán đương nhiệm nắm giữ; vai trò riêng của crypto ở đây có thể là danh tính và trách nhiệm giải trình cho tác tử.

Dữ liệu là một trong số ít phân mảng có người mua thật ngoài chuỗi. Grass báo cáo 17 triệu USD doanh thu dữ liệu trong năm 2025 và thêm 17 triệu USD trong nửa đầu năm 2026, đặt mục tiêu 75 triệu USD cho cả năm và nói rằng đang có lãi; GRASS trị giá khoảng 328 triệu USD, khoảng mười lần doanh thu nửa đầu năm đã năm hóa ước tính. Hào lũy của nó là quy mô thu thập từ băng thông dân dụng phân tán, không phải bản thân crypto. Các DAO dữ liệu và chợ dữ liệu thiếu nhu cầu trả tiền nhìn thấy được: Vana trị giá khoảng 21,5 triệu USD và Ocean khoảng 27,6 triệu USD, và Ocean đang kiện tụng với Fetch.ai. Bằng chứng nhân thân đang có những tích hợp thật khi tác tử ngày càng nhiều: World nói đã xác minh gần 18 triệu người, và Zoom, Docusign và Tinder hỗ trợ một phần World ID.

Dòng vốn đã bỏ phiếu

Mảng	Nhu cầu trả tiền kiểm chứng được	Tín hiệu trợ cấp và định giá	Nhận định
Chuỗi công việc hữu ích	Bittensor khoảng 70 triệu USD doanh thu bên ngoài kiểm chứng được mỗi năm; Pearl khoảng 1% số khối có công việc AI	Aleo 175–180 lần; Bittensor 4–11 lần; Pearl phần lớn là trợ cấp	Cơ chế có giá trị; tính hữu ích cần khách hàng trả tiền
Thị trường điện toán phi tập trung	Phí on-chain khoảng 0,68–2,35 triệu USD mỗi năm (Livepeer, Render); tự báo cáo 4–128 triệu USD	Vốn hóa của Render khoảng 360 lần phí	Thật nhưng nhỏ, phần lớn là bán lại suy luận phổ thông
Huấn luyện và suy luận phi tập trung	Công ty cổ phần: Prime Intellect trên 100 triệu USD mỗi năm, Venice trên 70 triệu USD (hạ tầng tập trung)	Mainnet huấn luyện của Gensyn chưa chạy; token Ritual −72% so với đỉnh	Tiến bộ kỹ thuật là thật; lợi nhuận về cổ phần

Mảng	Nhu cầu trả tiền kiểm chứng được	Tín hiệu trợ cấp và định giá	Nhận định
AI kiểm chứng được	TEE đã vào vận hành (398 người dùng trả tiền ở Phala); chưa công bố zkML trả phí	—	TEE hữu ích; zkML còn sớm
Thanh toán của tác tử	Dòng do tác tử khởi tạo khoảng 5.000–11.000 USD mỗi tháng	Khoảng 200 triệu giao dịch, phần lớn là tín hiệu và đầu cơ	Hạ tầng trước, nhu cầu sau
Mạng dữ liệu	Grass 17 triệu USD trong nửa đầu 2026 (tự báo cáo)	GRASS khoảng 10 lần doanh thu năm hóa	Một trong số ít mảng có người mua thật

Đặt sáu mảng cạnh nhau, kết luận nhất quán: **trợ cấp là nhiên liệu để khởi động, không phải mô hình kinh doanh.** Ngày 13 tháng 11, danh mục AI trên CoinGecko có tổng giá trị khoảng 18,6 tỷ USD, nhưng khoảng 56% trong đó là các layer-1 đa dụng như NEAR và ICP, cộng các dự án ở tầng đồng thuận như TAO, TAO đã stake (có thể bị đếm trùng với TAO) và Pearl; token AI ở tầng giao thức và tầng ứng dụng tổng cộng khoảng 8 tỷ USD hoặc ít hơn ước tính. Con số đó xấp xỉ tổng định giá của bốn công ty cổ phần: Tempo (5 tỷ USD), Nous (được đưa tin 1,5 tỷ USD), Prime Intellect (1 tỷ USD) và Venice (1 tỷ USD). Vốn đầu tư mạo hiểm vào crypto dậm chân ở khoảng 4,5–5 tỷ USD mỗi quý, phần tăng thêm đến từ hạ tầng liên quan đến AI, với tỷ trọng trong các khoản đã công bố tăng từ 8,1% lên 18,2% trong quý 3 năm 2026; và 84,7% trong số 118 token ra mắt năm 2025 đang giao dịch dưới giá ra mắt. Tóm lại một câu: **crypto lo phân phối và thanh toán, cổ phần hưởng lợi nhuận**, ít nhất là trong năm 2026.

CHƯƠNG 07

07

KINH TẾ HỌC ASIC

Câu hỏi ASIC: một công thức, ba thước đo

Một công thức: vì sao hiệu quả thôi sinh lời sau k = 5

Gọi R là doanh thu hằng ngày trên mỗi đơn vị tính toán (thu nhập thợ đào toàn mạng chia cho hashrate toàn mạng), c là chi phí hằng ngày của một đơn vị tính toán trên phần cứng biên (với Pearl là GPU), và k là số lần một ASIC rẻ hơn phần cứng biên đó (c_ASIC = c/k). Chừng nào GPU còn là máy đào biên, việc tự do gia nhập và rút lui đẩy R về gần c, nên một ASIC kiếm được R − c/k trên mỗi đơn vị tính toán:

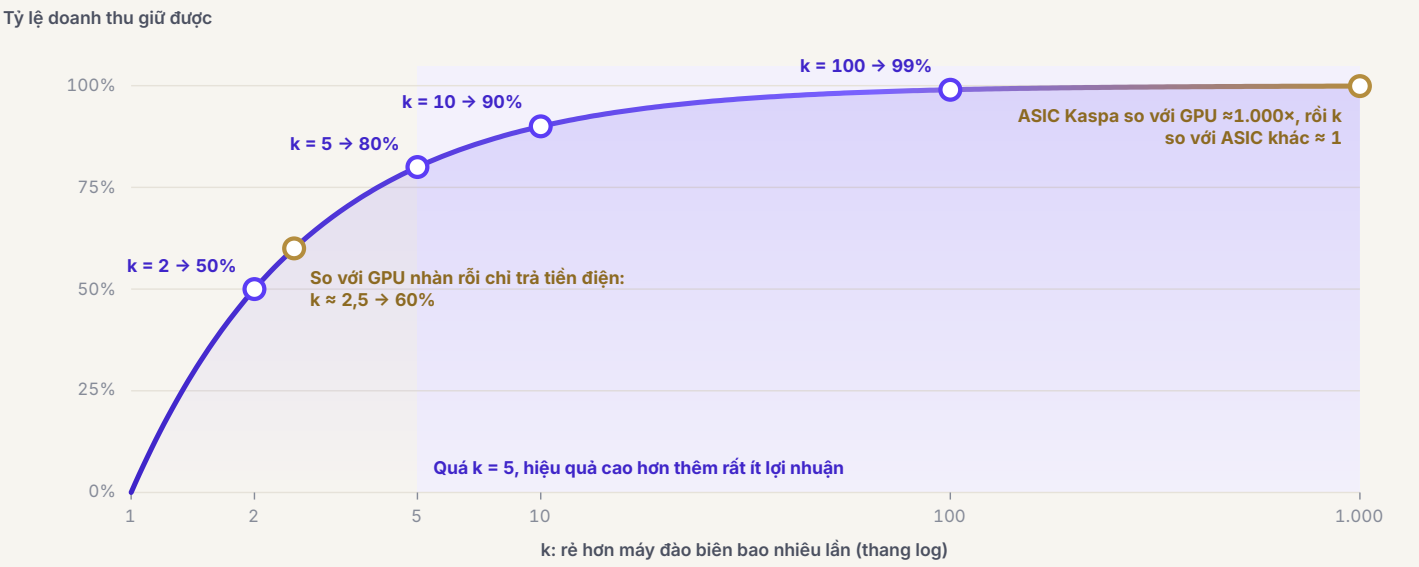
Trần lợi nhuận của ASIC ≈ R × (1 − 1/k)

Trên toàn mạng, trần lợi nhuận gộp hằng năm của một đội ASIC xấp xỉ doanh thu thợ đào toàn mạng × (1 − 1/k), và chỉ cho đến khi ASIC thay GPU trở thành máy đào biên.

Công thức có ba hệ quả. Thứ nhất, (1 − 1/k) là hàm lõm: ở k = 2 một ASIC giữ 50% doanh thu, ở k = 5 là 80%, ở k = 10 là 90%, ở k = 100 là 99%. **Quá k = 5–10, hiệu quả cao hơn gần như không thêm lợi nhuận**; quy mô của R và của quỹ thưởng mới quyết định. Thứ hai, mốc tham chiếu được đặt lại: khi ASIC đã là máy đào biên, k phải được đo so với các ASIC khác. ASIC của Kaspa có lợi thế hiệu quả khoảng 1.000 lần so với GPU, nhưng chỉ mua được một khoản lời bất ngờ ngắn ngủi; ngày nay chiếc KS7 tốt nhất kiếm 0,86 USD mỗi ngày. Thứ ba, trên một chuỗi thân thiện với GPU như Pearl, mốc tham chiếu thường không phải giá mua GPU mà là **tiền điện của một GPU nhàn rỗi**: một RTX 5090 đã sở hữu, chạy 525 W với giá 0,10 USD mỗi kWh, tốn khoảng 1,26 USD tiền điện mỗi ngày và kiếm khoảng 5–6 USD ước tính. So với các GPU là chi phí chìm, ngay cả một ASIC chỉ dùng một phần năm năng lượng cũng chỉ có k hiệu dụng khoảng 2,5 nếu khấu hao của nó bằng tiền điện, và trần lợi nhuận khoảng 60% của R (một giả định minh họa).

HÌNH 05 **Vì sao hiệu quả thôi sinh lời: ASIC giữ tối đa 1 – 1/k doanh thu**

Tỷ lệ doanh thu ASIC giữ được so với máy đào biên, theo lợi thế chi phí k



Nguồn: Mô hình của tác giả: khi phần cứng biên định giá, doanh thu trên mỗi đơn vị tính toán tiến về chi phí của nó; số liệu Kaspa và RTX 5090 lấy từ báo cáo.

Ba thước đo: quỹ thưởng, quy mô hoàn vốn và luật lệ

Thước đo thứ nhất là **quỹ thưởng**. Năm hóa khoảng 2.893 BTC phần thưởng khối trong tuần đầu tháng 11 (khoảng 222 triệu USD), thợ đào bitcoin kiếm khoảng 11,5 tỷ USD mỗi năm ước tính. Thợ đào Pearl kiếm khoảng 0,46–0,49 tỷ USD (theo giá hiện tại); phần thưởng prover của Aleo, năm hóa từ quý 3 năm 2025, khoảng 31 triệu USD; và thợ đào Kaspa, với phần thưởng khối 1,945 KAS và mười khối mỗi giây, khoảng 18 triệu USD ước tính. **Doanh thu thợ đào Kaspa của cả một năm thấp hơn mức thấp của ngân sách thiết kế một con chip 5/4 nm (khoảng 30 triệu USD).**

Thước đo thứ hai là **hashrate cần thiết để hoàn vốn NRE**:

h* = NRE ÷ [R × (1 – 1/k) × T × p]

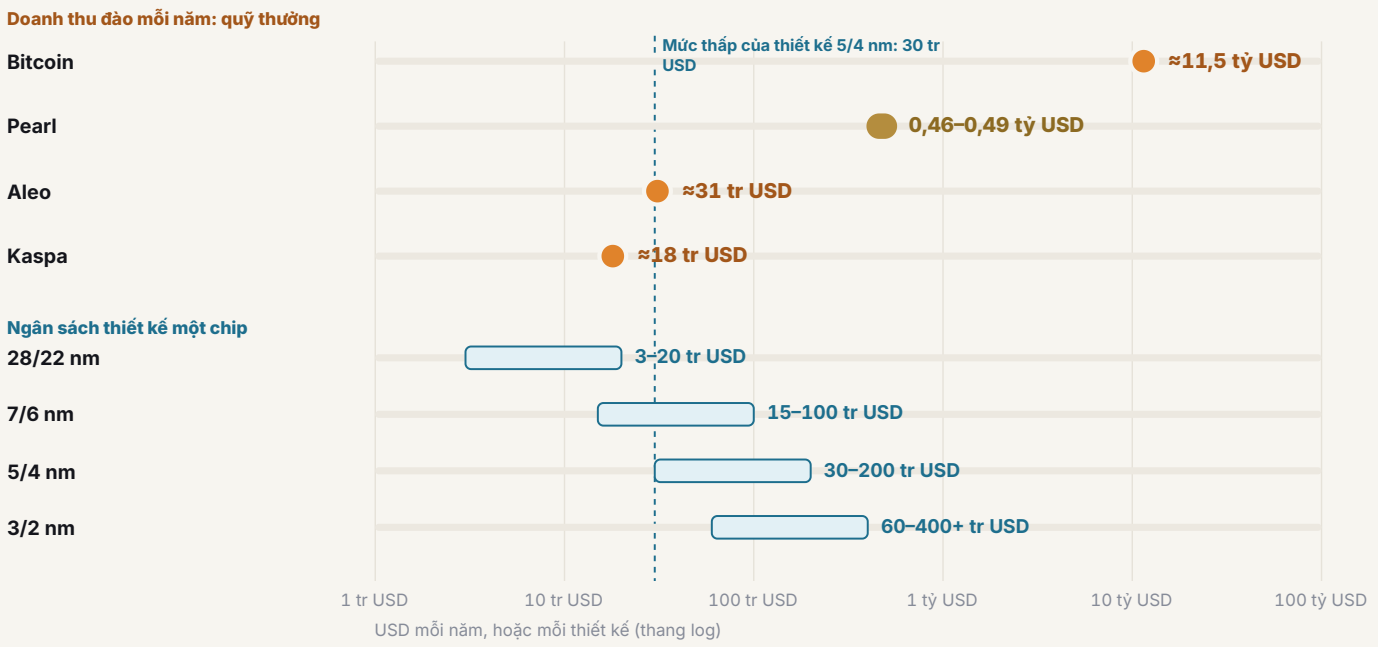
trong đó T là cửa sổ hữu dụng tính bằng ngày và p là xác suất luật lệ giữ nguyên trong cửa sổ đó. Với Pearl, lấy R = 14 USD mỗi PH/s mỗi ngày, k = 5, T = 365 ngày và NRE 20 triệu USD (một chip mảng lặp N7/N6; ở mức thấp của các dải 5–30 triệu USD cho tiến trình tầm trung theo Silicon Analysts và 15–100 triệu USD cho ASIC tính toán 7/6 nm theo AnySilicon, vì chip đào dùng ít IP và có cấu trúc lặp lại; đây là giả định của tôi). Với p = 1, cần khoảng **4,9 EH/s** chạy trọn một năm, khoảng 5% mạng; với p = 0,5, khoảng **9,8 EH/s**, khoảng 11%; và nếu R lùi về mức mùa hè khoảng 10 USD, ngay cả p = 1 cũng cần khoảng 6,9 EH/s ước tính. Quy mô này không ngoài tầm với. Vấn đề là p và T.

Thước đo thứ ba là **rủi ro luật lệ**, và nó thường mang tính quyết định. Fork chống ASIC của Monero năm 2018 khiến máy X3 của Bitmain bị mắc kẹt và có lúc làm mạng mất gần một nửa hashrate. Fork của Sia cùng năm biến máy đào của Bitmain và Innosilicon thành phế liệu, trong khi giá máy cũ của Obelisk, hãng máy đào có liên hệ với các nhà phát triển, tăng từ 2.499 USD lên hơn 5.000 USD. Chỉ hơn một tháng sau The Merge của Ethereum, chưa đến 20% thợ đào Ethereum cũ tìm được việc mới. **Ai kiểm soát hàm công việc thì nắm số phận của nhà đầu tư phần cứng**, và Sia cho thấy "kháng ASIC" có thể trở thành hào lũy của chính các nhà phát triển. Phần cứng có thị trường thứ hai thì sống sót: sau The Merge, GPU chuyển sang AI và đám mây, còn ASIC đơn thuật toán dồn về Ethereum Classic. Cửa sổ cũng đang thu hẹp: ASIC đầu tiên của Kaspa mất khoảng 16 tháng, của Aleo khoảng năm tháng, và bên đầu tiên chạm đến biên hiệu quả thường là kẻ thách thức chứ không phải hãng đương nhiệm.

Mạng	Doanh thu thợ đào hằng năm (khoảng)	Phần cứng biên hoặc tham chiếu	Độ ổn định của luật lệ	k dành cho người mới	Nhận định
Bitcoin	11,5 tỷ USD	Bitmain S23 XP Hyd (8,9 J/TH, 25,5 USD mỗi TH)	Rất cao	≤1	Quỹ thưởng lớn, nhưng k thuộc về các hãng đương nhiệm
Pearl	0,46–0,49 tỷ USD (theo giá hiện tại)	GPU đã sở hữu hoặc nhàn rỗi (chi phí biên ≈ tiền điện) và GPU trung tâm dữ liệu	Rất thấp: bốn fork trong hai tháng; danh sách trắng FP8 đang chờ	Có thể 3–10 lần về hiệu quả phân tích ; thấp hơn khi so với GPU chi phí chìm	Quỹ thưởng đủ lớn; rủi ro luật lệ phủ quyết
Aleo	Khoảng 31 triệu USD (quý 3 năm 2025, năm hóa)	IceRiver AE2 và các máy chuyên dụng khác	Trung bình: ngưỡng stake ARC-46	≈1	Quỹ thưởng nhỏ, và ngưỡng đang tăng
Kaspa	Khoảng 18 triệu USD	KS7 (77 J/TH)	Cao	≈1	Doanh thu hằng năm thấp hơn mức thấp của ngân sách một thiết kế 5/4 nm

HÌNH 06 **Phần lớn quỹ thưởng nhỏ hơn ngân sách thiết kế một con chip**

Doanh thu thợ đào hằng năm của bốn mạng so với chi phí thiết kế một chip theo tiến trình, thang log



Nguồn: Hashrate Index (bitcoin, 11/2026), prlstats và Kryptex (Pearl), Blockworks (Aleo, quý 3/2025), minerstat (Kaspa), AnySilicon (ngân sách thiết kế, 2026); tác giả quy ra năm.

Góc phần tư "quỹ thưởng lớn, luật lệ ổn định và k > 1" hiện đang trống: bitcoin có quỹ thưởng và luật lệ ổn định, nhưng k của nó thuộc về các hãng đương nhiệm; quỹ thưởng của Pearl đủ lớn nhưng luật lệ kém ổn định nhất; Kaspa và Aleo có luật lệ đủ ổn định nhưng quỹ thưởng quá nhỏ.

CHƯƠNG 08

08

TỪNG LOẠI CHIP

Năm loại chip, năm câu trả lời

PoUW kiểu Pearl: không làm chip INT7 chuyên dụng; đặt cược vào máy đào có thể phục vụ suy luận

Lý lẽ ủng hộ ASIC là có thật. Hiện chỉ khoảng 1% số khối chứa công việc AI, nên lập luận kinh tế của whitepaper, rằng một ASIC chuyên dụng mất khả năng làm công việc hữu ích và do đó kiếm ít hơn, đang ở lúc yếu nhất. Những phần đắt nhất của một GPU, HBM và đóng gói CoWoS, chiếm khoảng 63% chi phí sản xuất một H100, trong khi die logic chỉ khoảng 9%. Một con chip không làm gì ngoài phép nhân-cộng số nguyên 7 bit, giữ toán hạng trong SRAM trên chip và chạy ở điện áp thấp có thể đạt mức năng lượng trên mỗi phép tính tốt hơn 3 đến 10 lần so với một GPU cùng tiến trình, và lợi thế còn lớn hơn về chi phí trên mỗi phép tính [phân tích; chưa có số đo]. Lý lẽ phản đối thì nặng ký hơn. Thời gian chờ xưởng đúc ở tiến trình tiên tiến có thể vượt 50 tuần, và một đội mới cần khoảng 18 đến 24 tháng từ lúc bắt đầu đến khi sản xuất hàng loạt ước tính; đến khi chip ra đời, Pearl rất có thể đã chuyển sang FP8 rồi FP4. Và vì PIP-3 định nghĩa phép số học được chấp nhận thông qua một danh sách trắng thiết bị, một chip INT7 sẽ đơn giản là ngừng hoạt động khi nó được kích hoạt.

Vì vậy kết luận của tôi là: **đừng tape-out một chip Pearl INT7 chuyên dụng lúc này.** Có hai đặt cược tốt hơn. Một là bộ tăng tốc ưu tiên suy luận và cũng đào được: xây quanh GEMM INT8/FP8 đa dụng với toán hạng nằm trong SRAM, để nếu cửa đào bị đóng thì giá trị vẫn rơi vào một thị trường suy luận thật. Hai là tham gia cuộc tranh luận về giao thức và thúc đẩy một phép số học chuẩn trung lập với nhà cung cấp mà ASIC có thể hiện thực chính xác, cùng với cam kết ổn định thuật toán, vốn là điều kiện tiên quyết để góc phần tư "quỹ thưởng lớn, luật lệ ổn định, k > 1" tồn tại. Cho đến lúc đó, có thể kiểm chứng kiến trúc trên FPGA hoặc wafer đa dự án (MPW) ở tiến trình trưởng thành, giữ rủi ro NRE ở mức khoảng một triệu đô la. Chương 10 biến điều này thành một ngân sách theo giai đoạn.

Máy đào kiểu bitcoin: các hãng đương nhiệm nắm k, và điện là tài sản khan hiếm

Rủi ro luật lệ của bitcoin là thấp nhất, nhưng cạnh tranh đẩy k của người mới xuống dưới 1. S23 XP Hyd của Bitmain đạt 8,9 J/TH, bán với giá 25,5 USD mỗi TH và giao hàng từ tháng 11; hiệu quả của dòng chủ lực giảm từ 17,5 J/TH của S21 năm 2023 xuống 8,9 J/TH, khoảng 49% trong ba năm. Một người mới giao hàng số lượng lớn trong 2027–2028 sẽ cần khoảng 8 J/TH hoặc tốt hơn để trụ được ước tính. Những kẻ thách thức đang rời đi: Canaan báo cáo doanh thu quý 2 năm 2026 là 31,9 triệu USD với lỗ gộp 29,3 triệu USD và chỉ dự báo 11–15 triệu USD cho quý 3; doanh số máy tự sản xuất bán cho người mua bên ngoài của Bitdeer trong quý 3 giảm xuống 0,3 triệu USD, với biên lợi nhuận gộp toàn tập đoàn 3,4%; Auradine đổi tên thành Velaura AI và chuyển sang chip AI; và Core Scientific hủy một đơn hàng khoảng 15 EH/s máy 3 nm từ đơn vị Proto của Block, ghi nhận khoản lỗ 41,9 triệu USD.

Nhu cầu cũng đang co lại. Ngày 9 tháng 11, hashprice khoảng 31,89 USD mỗi PH/s mỗi ngày, đường kỳ hạn đến tháng 4 năm 2027 bình quân khoảng 32,7 USD, và phí giao dịch chỉ bằng 0,72% phần thưởng khối. Với giá điện 0,045 USD mỗi kWh, S23 XP Hyd hoàn vốn trong khoảng 3,1 năm ước tính, sát với kỳ halving tiếp theo dự kiến khoảng năm 2028. Quan trọng hơn là chi phí cơ hội của nguồn điện: các hợp đồng thuê cho AI và điện toán hiệu năng cao nay mang về khoảng 1,67–1,9 triệu USD mỗi megawatt theo hợp đồng mỗi năm, với biên thu nhập hoạt động ròng 80% đến 85% là khoảng 1,5 đến 2,9 lần lợi nhuận gộp của việc đào trên mỗi megawatt (khoảng 0,56–0,9 triệu USD) [ước tính; hai cơ sở megawatt khác nhau]. Core Scientific sẽ ngừng đào trong quý 4; Bitfarms, nay

là Keel, đã đóng toàn bộ các địa điểm đào tại Mỹ; IREN kiếm 200 triệu USD từ đám mây AI và chỉ 28 triệu USD từ việc đào trong một quý. **Nguồn điện và công suất trung tâm dữ liệu mới là tài sản; máy đào chỉ là vật tư tiêu hao.**

Phép tính NRE cũng khắc nghiệt không kém. Hiệu quả ở biên hiện nay cần quy trình lớp 3 nm (theo nguồn thứ cấp, MicroBT dùng Samsung SF3, còn chip của Auradine và Proto là 3 nm). Mặt nạ tốn khoảng 10–20 triệu USD và NRE đầy đủ từ 30–100 triệu USD trở lên; ở mức 20–25 USD mỗi TH và biên lợi nhuận gộp 30%, mỗi 1 EH/s bán ra mang về khoảng 6–7,5 triệu USD lợi nhuận gộp, nên để bù chi phí R&D hằng năm ở quy mô của Bitdeer, khoảng 110–160 triệu USD, cần bán khoảng 20–25 EH/s mỗi năm cho người mua bên ngoài, khoảng 2% mạng ước tính. Kết luận: **một đội mới không nên tham gia thị trường máy đào bitcoin thương mại.** Ba trường hợp vẫn hợp lý: tích hợp dọc với nguồn điện rẻ của riêng mình; các dạng sản phẩm ngách như máy đào gia đình và máy đào sưởi ấm (Canaan cũng đang ra mắt sản phẩm sưởi ấm gia đình); và tham gia lộ trình của các hãng đương nhiệm thông qua IP và dịch vụ thiết kế.

Chip huấn luyện AI: không dành cho đội nhỏ

Với chip huấn luyện, k là thứ khó kiếm nhất. Ở cùng thế hệ và độ chính xác, Ironwood của Google chỉ dẫn trước B300 khoảng 1,1 đến 1,8 lần về hiệu quả FP8 cấp chip (công suất của Ironwood là ước tính của một phóng viên) ước tính; con số 30 đến 80 lần thường được trích từ bài báo TPU v1 là đo so với K80, vốn không có tensor core. Đòn bẩy thật sự là định dạng số: chuyển từ FP8 sang FP4 thường cho thông lượng gấp 2 đến 3 lần trên cùng silicon, và NVIDIA đi nhanh nhất ở đó: hãng tuyên bố 50 PF suy luận NVFP4 cho Rubin, sản xuất hàng loạt trong nửa cuối năm 2026. Ngưỡng chi phí và nguồn cung còn cao hơn nữa: ngân sách thiết kế 3/2 nm từ 60–400 triệu USD trở lên; công suất CoWoS năm 2026 khoảng 106.000–130.000 wafer mỗi tháng, và TSMC được cho là đang cân nhắc tăng giá tới 20%; với giá hợp đồng 15,6 USD mỗi GB cho HBM3E, riêng HBM trong một bộ tăng tốc 288 GB đã tốn khoảng 4.500 USD ước tính. Thị trường này thuộc về mô hình đồng phát triển với hyperscaler kiểu Broadcom (chương 11). Kết luận: **đội nhỏ không nên làm chip huấn luyện độc lập**; thay vào đó hãy làm IP, chiplet hoặc dịch vụ thiết kế.

Chip suy luận AI: đổi tính đa dụng lấy k, và điều đó sinh lời

Suy luận là nơi ASIC có nhiều dư địa nhất, vì k có thể mua được bằng cách từ bỏ tính đa dụng. Etched tuyên bố máy chủ tám chip của họ chạy Llama-70B ở hơn 500.000 token mỗi giây, khoảng 20 lần tám chiếc H100; Taalas khắc cứng trọng số mô hình vào mask ROM và tuyên bố công suất thấp hơn 10 lần và chi phí sản xuất thấp hơn 20 lần (cả hai đều hãng tự công bố). Kiến trúc lấy SRAM làm trung tâm tránh được HBM và CoWoS, và tiến trình của chúng dễ quản lý hơn: Taalas dùng TSMC 6 nm, còn Groq chuyển từ 14 nm sang Samsung 4 nm. Kênh phân phối crypto không giúp được chip suy luận, vì các mạng điện toán phi tập trung gần như chạy hoàn toàn trên NVIDIA; lối vào là kinh tế học mua sắm của các neocloud và thị trường suy luận nội địa rộng lớn ở Trung Quốc, nơi các trung tâm dữ liệu dùng vốn nhà nước phải dùng chip AI nội địa, nhập khẩu H200 đang bị siết lại, và Bắc Kinh muốn các công ty internet lớn chạy suy luận trên chip nội địa. Kết luận: **đáng làm**, miễn là có khách hàng rõ ràng và mức độ khóa chặt vào mô hình và kiến trúc là một lựa chọn có chủ đích. Đây cũng là thị trường thứ hai tốt nhất cho một chip đào PoUW.

Bộ tăng tốc ZK và FHE: phòng hộ với ZK, đón đầu sớm với FHE

GPU đã bắt kịp việc chứng minh ZK theo thời gian thực: SP1 Hypercube của Succinct đã chứng minh 99,7% số khối Ethereum trong chưa đến 12 giây trên 16 chiếc RTX 5090. Cysic tuyên bố chip ZK C1 của họ hiệu quả gấp 10 đến 100 lần prover dùng GPU và sẽ giao hàng trong năm 2026, nhưng nhu cầu của nó phần lớn đến từ phần thưởng token của chính họ hãng tự công bố; tôi không tìm thấy bằng chứng công khai nào cho thấy ASIC giải câu đố Aleo được dùng để chứng minh ZK thật. Chứng minh ZK là phép tính số nguyên và tất định, gần về hình dạng với việc chip đào làm, nhưng các hệ chứng minh thay đổi nhanh (SP1, RISC Zero, Halo2, Plonky2 và nhiều hệ khác). Kết luận: **chờ và phòng hộ**, với các engine MSM, NTT và băm lập trình được, trước tiên trên FPGA, sau đó là ASIC.

FHE là nguyên thủy mật mã phụ thuộc phần cứng nhiều nhất và có tiềm năng lớn nhất. Heracles của Intel, trên Intel 3 với 197 mm², 176 W và 48 GB HBM3, chạy các phép toán FHE nhanh hơn 1.074 đến 5.547 lần so với một Xeon 24 nhân, nhưng Intel không có kế hoạch thương mại và trường dự án đã rời đi vào tháng 12 năm 2025. Niobium đang xây dựng trên Samsung 8LPU với SemiFive là đối tác thiết kế (hợp đồng phát triển khoảng 6,86 triệu USD) và đã gọi được hơn 23 triệu USD; Optalysys, với cách tiếp cận quang tử, đã gọi 31 triệu USD và nói rằng một con chip thương mại hoàn chỉnh vẫn còn cách hai đến ba năm; và thiết kế FPGA mã nguồn mở của Zama đạt khoảng 13.000 lần bootstrap lập trình được mỗi giây ở khoảng 200 W, mốc cơ sở công khai mà bất kỳ ASIC nào cũng phải vượt rõ ràng. Kết luận: **đón đầu sớm và kiểm chứng với ít tài sản**, cho một cửa sổ khoảng 2027–2029. Nhu cầu nhiều khả năng sẽ neo vào AI mã hóa cho doanh nghiệp và chính phủ hơn là vào bản thân tiền điện tử, và FHE cần HBM, nên cấu trúc chi phí của nó giống chip huấn luyện.

Loại chip	k đến từ đâu	Nhu cầu và quỹ thưởng	NRE và ngưỡng	Rủi ro luật lệ hoặc khóa chặt	Thị trường thứ hai	Nhận định
Chip đào PoUW kiểu Pearl	Bỏ HBM, CoWoS và khối đa dụng: có thể 3–10 lần về J/op phân tích	Thu nhập thợ đào khoảng 0,46–0,49 tỷ USD mỗi năm, biến động	Mảng lặp N7/N6 khoảng 10–30 triệu USD phân tích	Rất cao: giao thức quyết định định dạng số và danh sách trắng thiết bị	Suy luận AI, nếu xây như GEMM INT8/FP8 đa dụng	Không làm chip INT7 chuyên dụng; làm chip ưu tiên suy luận và đào được, hoặc chờ phép số học chuẩn
Chip đào Bitcoin SHA-256	k ≤ 1 so với Bitmain	Khoảng 11,5 tỷ USD mỗi năm, thị trường thương mại do hai hãng nắm	Mặt nạ lớp 3 nm 10–20 triệu USD; NRE đầy đủ từ 30–100 triệu USD trở lên	Thấp (thuật toán ổn định); cạnh tranh cao và rủi ro halving	Gần như không có (đơn thuật toán)	Người mới đứng ngoài thị trường thương mại; chỉ khi có điện riêng, dạng ngách hoặc làm IP
Chip huấn luyện AI	Mảng systolic cùng thế hệ chỉ 1,1–1,8 lần; định dạng mới là đòn bẩy	Lớn nhất (Broadcom khoảng 58 tỷ USD doanh thu bán dẫn AI trong năm tài chính 2026)	Thiết kế 3/2 nm từ 60–400 triệu USD trở lên; CoWoS và HBM	Nhịp ra mắt hằng năm của NVIDIA	—	Đội nhỏ làm IP, chiplet và dịch vụ thiết kế, không làm chip độc lập
Chip suy luận AI	Từ bỏ tính đa dụng: tuyên bố 10–20 lần hãng tự công bố	Lớn và phân mảnh; bản thân thị trường nội địa Trung Quốc đại lục đã rất lớn	Kiến trúc SRAM ở N6/N7–N4 dễ quản lý	Khóa chặt vào mô hình và kiến trúc	Đây là thị trường chính	Đáng làm: ký khách hàng trước, rồi mới chọn mức khóa chặt
Bộ tăng tốc ZK	Tuyên bố 10–100 lần GPU hãng tự công bố	Nhỏ, chủ yếu do phần thưởng token thúc đẩy	N7/N6 rồi N5; FPGA trước	Hệ chứng minh thay đổi nhanh	Hạn chế ngoài ZK	Chờ và phòng hộ: engine MSM và NTT lập trình được
Bộ tăng tốc FHE	1.074–5.547 lần một Xeon (Heracles)	Chưa hình thành; neo vào doanh nghiệp và chính phủ	Cần HBM; cấu trúc chi phí gần chip huấn luyện	Tiêu chuẩn và hệ sinh thái chưa định hình	Điện toán mã hóa, AI riêng tư	Đón đầu sớm, kiểm chứng với ít tài sản; cửa sổ 2027–2029

HÌNH 07 **Năm loại chip, năm câu trả lời**

Lợi thế chi phí k đến từ đâu, silicon có thể bán vào đâu nếu luật đổi, và kết luận



Chip đào PoUW kiểu Pearl

CHƯA NÊN

k: không HBM, không khối đa dụng: 3–10× mỗi phép tính (suy luận)

Thị trường thứ hai: suy luận AI, nếu là GEMM INT8/FP8 đa dụng

Đặt cược vào chip ưu tiên suy luận có thể đào, không phải chip INT7 chuyên dụng



Chip đào Bitcoin SHA-256

ĐỨNG NGOÀI

k: ≤1 so với các hãng đương nhiệm

Thị trường thứ hai: gần như không

Chỉ khi có điện riêng, dạng sản phẩm ngách, hoặc làm IP



Chip huấn luyện AI

IP, KHÔNG LÀM CHIP

k: 1,1–1,8× ở cùng độ chính xác

Thị trường thứ hai: —

Đóng gói và HBM quyết định; làm IP, chiplet, dịch vụ



Chip suy luận AI

NÊN LÀM

k: bỏ tính đa dụng: tuyên bố 10–20×

Thị trường thứ hai: đây là thị trường chính

Ký khách hàng trước, rồi mới chọn mức khóa chặt



Bộ tăng tốc chứng minh ZK

CHỜ VÀ PHÒNG HỘ

k: tuyên bố 10–100× GPU

Thị trường thứ hai: hạn chế ngoài ZK

Engine MSM và NTT lập trình được; FPGA trước



Bộ tăng tốc FHE

ĐÓN ĐẦU SỚM

k: 1.074–5.547× một Xeon (Heracles)

Thị trường thứ hai: điện toán mã hóa, AI riêng tư

Cửa sổ 2027–2029; kiểm chứng với ít tài sản

Nguồn: Tổng hợp của tác giả từ chương 7 và 8; số liệu của hãng được ghi nhãn như trong báo cáo.

CHƯƠNG 09

09

TIẾN TRÌNH

Tiến trình: hãy để chi phí vòng đời lớn nhất quyết định

Bậc thang chi phí: wafer tiên tiến nhất đắt gấp mười wafer trưởng thành

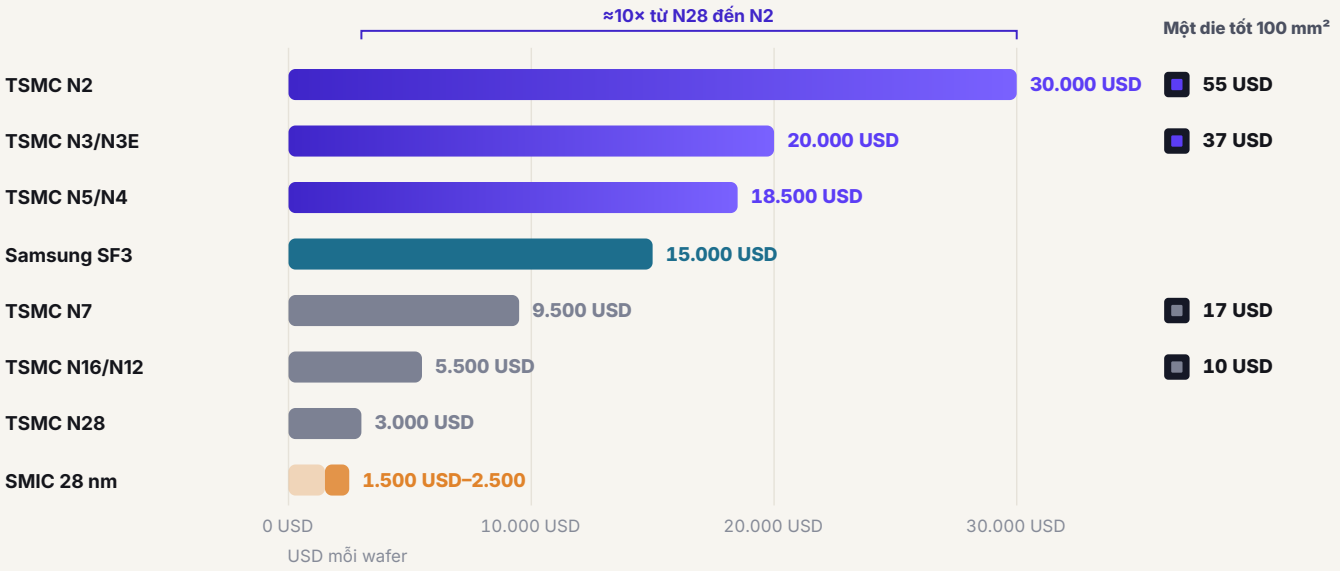
Bắt đầu từ sổ sách. Một wafer tiên tiến nhất nay đắt khoảng mười lần một wafer trưởng thành: TSMC N2 khoảng 30.000 USD, N3 khoảng 20.000 USD, N7 khoảng 9.500 USD và N28 khoảng 3.000 USD; Samsung SF3 khoảng 15.000 USD, thấp hơn N3 khoảng 25%; và chi phí sản xuất lớp 7 nm của SMIC được cho là cao hơn TSMC 40% đến 50%. TSMC được cho là sẽ tăng giá thêm 3% đến 6% từ tháng 1 năm 2027, nhiều nhất ở 2 nm và 3 nm. Chi phí mặt nạ và thiết kế còn chênh lệch hơn nữa, và các nguồn không thống nhất, nên bảng dưới đây ghi cả hai.

Tiến trình	Giá wafer (ước tính)	Bộ mặt nạ (giá trị mô hình / dải hướng dẫn)	Ngân sách thiết kế (AnySilicon)	MPW (mỗi dự án)
TSMC N2	Khoảng 30.000 USD	—	3/2 nm: từ 60–400 triệu USD trở lên	—
TSMC N3/N3E	Khoảng 20.000 USD (18.000–27.000 USD)	15 triệu USD / 10–20 triệu USD	Như trên	—

Tiến trình	Giá wafer (ước tính)	Bộ mặt nạ (giá trị mô hình / dải hướng dẫn)	Ngân sách thiết kế (AnySilicon)	MPW (mỗi dự án)
Samsung SF3	Khoảng 15.000 USD (13.000–17.000 USD)	12 triệu USD	—	—
TSMC N5/N4	Khoảng 18.500 USD (16.000–21.300 USD, độ tin cậy thấp)	6,5 triệu USD / 10–20 triệu USD	5/4 nm: 30–200 triệu USD	—
TSMC N7	Khoảng 9.500 USD (8.100–10.900 USD)	3,5 triệu USD / 5–10 triệu USD	7/6 nm: 15–100 triệu USD	Thương mại 100.000–500.000 USD
TSMC N16/N12	Khoảng 5.500 USD	1,8 triệu USD	16/12 nm: 6–40 triệu USD	Học thuật khoảng 35.000 USD mỗi mm²
TSMC N28	Khoảng 3.000 USD	1 triệu USD / 1–3 triệu USD	28/22 nm: 3–20 triệu USD	Thương mại 50.000–150.000 USD; học thuật khoảng 12.000 USD mỗi mm²
SMIC 28 nm	1.500–2.500 USD	—	—	—
SMIC lớp 7 nm (N+2/N+3)	Cao hơn TSMC 40–50%	—	—	—

HÌNH 08 **Một wafer tiên tiến nhất đắt gấp mười wafer trưởng thành**

Giá ước tính mỗi wafer 300 mm theo tiến trình, và chi phí một die tốt 100 mm² ở yield 85%



Nguồn: Bảng giá wafer của Silicon Analysts (ước tính, N5 độ tin cậy thấp), Wccftech (SMIC); chi phí die là phép tính của tác giả với yield giả định.

Rồi đến vật lý. Các phép đo kinh điển của Horowitz cho thấy một lệnh lập trình được tốn khoảng 70 pJ trong khi bản thân phép tính số học chỉ tốn vài picojoule: **số học thì rẻ, chi phí phụ trội mới đắt**, đó là lý do mảng systolic và chip chuyên dụng tồn tại. Giảm một nửa kích thước đặc trưng gần như giảm một nửa năng lượng mỗi phép tính, nhưng điện áp cấp đã chững lại quanh 0,9–1,1 V. Một wafer N2 đắt hơn N3 từ 50% đến 67%, nên trừ khi mật độ tăng hơn khoảng 1,5 lần, chi phí mỗi transistor sẽ tăng chứ không giảm **ước tính**: các tiến trình tiên tiến phải tự trả giá cho mình bằng hiệu quả và hiệu năng. Với một die 100 mm² ở yield 85%, chi phí mỗi die tốt khoảng 55 USD ở N2, 37 USD ở N3, 29–34 USD ở N5, 17 USD ở N7 và 10 USD ở N16 [ước tính; yield là giả định].

Chọn tiến trình theo loại chip và đội ngũ

Quy tắc của tôi là: **hãy để khoản lớn nhất trong chi phí vòng đời của con chip chọn tiến trình.** Với chip tính toán chạy đầy tải suốt ngày đêm (đào, ZK), tiền điện chiếm ưu thế trong chi phí vòng đời, nên khi thuật toán đã ổn định chúng nên chuyển sang N5/N4 và N3/SF3; khi thuật toán chưa định hình hoặc thiếu tiền, hãy kiểm chứng trên MPW tiến trình trưởng thành trước. Với chip huấn luyện, điểm nghẽn là đóng gói và bộ nhớ (HBM và CoWoS chiếm khoảng 63% danh mục vật tư của một H100), nên tiến trình là thứ yếu. Với chip suy luận, quyết định đầu tiên là kiến trúc bộ nhớ: thiết kế lấy SRAM làm trung tâm có thể ở lại N6/N7 đến N4, trong khi thiết kế dùng HBM thừa hưởng cấu trúc chi phí của chip huấn luyện. FHE bị giới hạn bởi băng thông bộ nhớ và không thể tránh HBM.

Loại chip	Startup hoặc đội nhỏ	Công ty quy mô lớn	Lộ trình Trung Quốc đại lục	Điểm chính
Chip đào PoUW dùng phép nhân ma trận (dữ liệu trong SRAM, không HBM)	Làm mẫu trên MPW 28 nm hoặc 16/12 nm; sản lượng đầu trên N7/N6; N5/N4 khi thuật toán và doanh thu ổn định; die 50–150 mm² trong gói flip-chip tiêu chuẩn	N3/N3E hoặc Samsung SF3	SMIC N+1/N+2 dùng được (không cần HBM hay CoWoS), đắt hơn khoảng 40–50% mỗi die tốt	Năng lượng chiếm ưu thế; luật giao thức là rủi ro lớn nhất
Chip đào SHA-256	Đừng chạy theo tiến trình đầu bảng; dạng ngách có thể dùng tiến trình trưởng thành	Lớp 3 nm; các hãng đương nhiệm chuyển từ TSMC 5 nm sang 2 nm và từ Samsung SF3 sang 2 nm (nguồn thứ cấp)	Khó bắt kịp hiệu quả của tiến trình tiên tiến nhất	Giá điện và chu kỳ halving quyết định biên lợi nhuận
Chip huấn luyện AI	Không làm chip độc lập: IP, chiplet, dịch vụ thiết kế	N3P rồi N2/A16; 2–8 chiplet; giữ chỗ CoWoS và HBM sớm	SMIC N+2/N+3 với HBM nội địa (mô hình Huawei); FP8 mỗi chip khoảng 0,2–0,5 lần của NVIDIA, bù bằng quy mô hệ thống	Đóng gói và HBM là điểm nghẽn
Chip suy luận AI	Lấy SRAM làm trung tâm hoặc khắc cứng, N6/N7 hoặc N4/N5/SF4; chỉ INT8, FP8 và FP4	N3/N2 với HBM (ngữ cảnh dài, mixture of experts), hoặc SRAM đa chiplet	SMIC N+2 với LPDDR hoặc GDDR; bộ nhớ lớp HBM là điểm thắt	Mức độ khóa chặt kiến trúc là quyết định trung tâm
Bộ tăng tốc ZK	Bắt đầu trên FPGA hoặc MPW 16/12 nm; rồi N7/N6 và N5	N5/N3	SMIC N+1/N+2 dùng được	Số nguyên và tất định; dễ kiểm chứng
Bộ tăng tốc FHE	FPGA (so với mốc cơ sở mã nguồn mở) hoặc MPW 16/12 nm; sản xuất hàng loạt cần HBM	N5/N3 với HBM	Bị giới hạn bởi khả năng tiếp cận HBM	Băng thông bộ nhớ, chứ không phải số học, quyết định tốc độ

Tiên tiến hay trưởng thành, bốn loại xưởng đúc, và chiplet

Lựa chọn	Điểm mạnh chính	Điểm yếu chính	Phù hợp nhất cho
Tiến trình tiên tiến (N3/N2, SF3/SF2, Intel 18A)	Mỗi nút đầy đủ giảm khoảng một nửa năng lượng mỗi phép tính; SRAM mật độ cao	Wafer đắt 3–10 lần, mặt nạ 5–15 lần và ngân sách thiết kế 10–20 lần so với tiến trình trưởng thành; thời gian chờ trên 50 tuần; chi phí mỗi transistor ở N2 có thể tăng	Chip tính toán có thuật toán ổn định chạy suốt ngày đêm; chip huấn luyện

Lựa chọn	Điểm mạnh chính	Điểm yếu chính	Phù hợp nhất cho
Tiến trình trưởng thành và tầm trung (N7/N6, N16/12, N28)	Mặt nạ ở mức triệu đô la, MPW rẻ, lặp nhanh, nhiều xưởng đúc (TSMC, UMC, GlobalFoundries, SMIC)	Hiệu quả không cạnh tranh khi chạy đầy tải suốt ngày đêm	Mẫu thử, thị trường ngách, suy luận biên
TSMC	Yield và hệ sinh thái tốt nhất	Giá cao nhất; công suất N2/N3 eo hẹp; hàng chờ CoWoS dài	Đội ngũ theo đuổi biên hiệu quả với cam kết sản lượng
Samsung Foundry	Wafer rẻ hơn khoảng 20–25% (SF3 khoảng 15.000 USD so với khoảng 20.000 USD của N3); được thế hệ thứ hai của Groq và các bên khác sử dụng	Mảng đúc chip đang thua lỗ, đặt mục tiêu hòa vốn vào cuối năm 2027	Chip tính toán nhạy cảm về chi phí
Intel Foundry	Chip phức tạp như Heracles đã được chứng minh trên Intel 3	Chưa có giá công khai cho khách bên ngoài hay dữ liệu khách hàng của 18A; mảng đúc chip đang thua lỗ	Đánh giá hệ sinh thái và hỗ trợ theo từng trường hợp
Xưởng đúc ở Trung Quốc đại lục (SMIC và các bên khác)	Gần một thị trường nội địa lớn; wafer 28 nm giá 1.500–2.500 USD	Công suất lớp 7 nm khoảng 45.000–60.000 wafer mỗi tháng, phần lớn do Huawei hấp thụ; yield khoảng một phần ba của TSMC; chi phí cao hơn 40–50%; chỉ có DUV	Chip suy luận, chip đào và chip ZK cho thị trường nội địa
Chiplet	Vượt giới hạn reticle và yield; tiêu chuẩn cho hệ thống HBM (MI355X, Ironwood, Trainium3)	Chi phí đóng gói cao hơn và phụ thuộc CoWoS; kiểm thử die tốt đã biết làm tăng chi phí	Chip huấn luyện và chip suy luận lớn; không cần cho chip đào hay chip ZK nhỏ

CHƯƠNG 10

NGÂN SÁCH THEO GIAI ĐOẠN

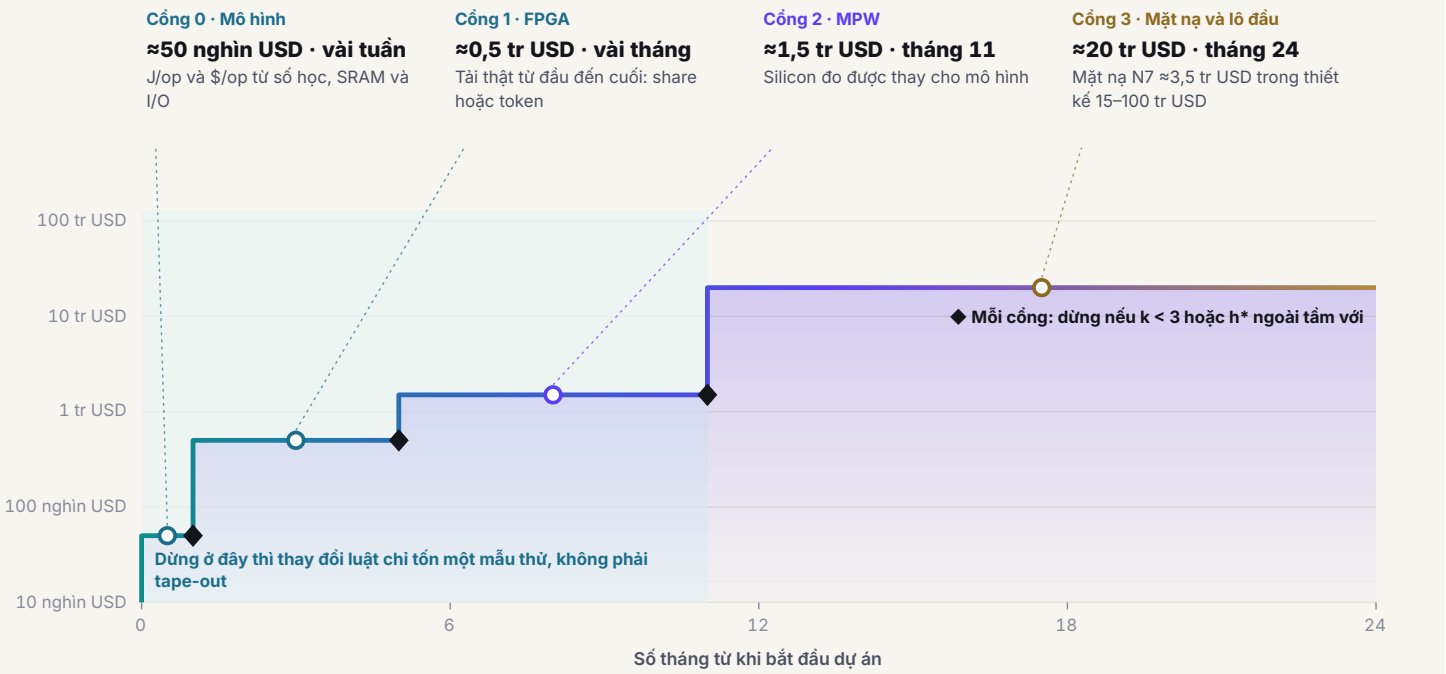
10 Từ mô hình đến silicon: ngân sách theo giai đoạn cho con chip đầu tiên

Mua bằng chứng trước khi mua mặt nạ

Một con chip full-custom là một chuỗi các canh bạc, canh sau lớn hơn canh trước, và kỷ luật ở đây là mua bằng chứng trước khi mua mặt nạ. Tôi dùng bốn cổng phân tích. Cổng 0 là một mô hình: một nghiên cứu kiến trúc ước tính J/op và $/op từ phần số học, SRAM và I/O, so với phần cứng tham chiếu của chương 7; nó tốn vài tuần và vài chục nghìn đô la. Cổng 1 là một mẫu thử FPGA chạy khối lượng công việc thật từ đầu đến cuối, tạo share đào hoặc token suy luận, và chứng minh đường dữ liệu cùng việc tích hợp giao thức; nó tốn vài tháng và vài trăm nghìn đô la. Cổng 2 là một wafer đa dự án: một suất thương mại tốn khoảng 50.000–150.000 USD ở N28 và 100.000–500.000 USD ở N7, và silicon đo được thay thế các ước tính năng lượng của mô hình. Cổng 3 là bộ mặt nạ đầy đủ và lô sản xuất đầu tiên: mặt nạ khoảng 1 triệu USD ở N28, 3,5 triệu USD ở N7 và 15 triệu USD ở N3, nằm trong ngân sách thiết kế 15–100 triệu USD ở 7/6 nm. Mỗi cổng cần một con số khiến dự án dừng lại, chẳng hạn k dưới 3 so với phần cứng biên, hoặc một hashrate hoàn vốn h* mà cửa sổ thời gian không thể đạt tới.

HÌNH 09 **Mua bằng chứng trước mặt nạ: ngân sách theo giai đoạn cho chip đầu tiên**

Chi tiêu lũy kế (USD) của một dự án N7/N6 qua bốn cổng, thang log; mỗi cổng có một con số để dừng dự án



Nguồn: Giá wafer, MPW và mặt nạ như ở chương 9 (Silicon Analysts, AnySilicon, giá trị mô hình); thời gian và chi tiêu từng cổng là giả định minh họa của tác giả.

Bậc thang này rất dốc. Trong một dự án N7/N6, ba cổng đầu cộng lại tốn khoảng 1,5 triệu USD và khoảng mười một tháng; cổng thứ tư tốn gấp hơn mười lần ước tính. Đó là lý do thứ tự quan trọng hơn tiến trình: nếu giao thức thay đổi phép số học, hoặc người mua bỏ đi, một đội đã dừng ở Cổng 2 chỉ mất giá một mẫu thử, chứ không phải giá một lần tape-out.

Mỗi con chip tốn bao nhiêu, và điều gì phải đúng

Dự án	Cổng 2 (mẫu thử)	Cổng 3 (tiến trình và ngân sách thiết kế)	Một die tốt 100 mm²	Điều phải đúng để hoàn vốn
Bộ tăng tốc ưu tiên suy luận có thể đào	MPW 16/12 nm hoặc N7, 0,1–0,5 triệu USD	N6/N7, khoảng 15–30 triệu USD ước tính	Khoảng 17 USD ở N7, chưa tính đóng gói và bộ nhớ	Một khách hàng suy luận đã ký gánh phần lớn NRE; thu nhập đào là phần cộng thêm, không phải kịch bản cơ sở
Máy đào Pearl INT7 chuyên dụng (hiện chưa khuyến nghị)	MPW 28 nm, 0,05–0,15 triệu USD	N7/N6, khoảng 20 triệu USD	Khoảng 17 USD ở N7	Khoảng 4,9 EH/s trong trọn một năm với p = 1, hoặc 9,8 EH/s với p = 0,5 (chương 7)
Bộ tăng tốc ZK	FPGA, rồi MPW 16/12 nm	N7/N6, rồi N5	Khoảng 10 USD ở N16, 17 USD ở N7	Các hệ chứng minh hội tụ về kernel MSM và NTT mà một thiết kế cố định có thể phục vụ
Bộ tăng tốc FHE	FPGA so với mốc cơ sở mã nguồn mở	N5/N3 với HBM, từ 30 triệu USD trở lên	Do HBM và đóng gói chi phối	Một người mua AI mã hóa ở quy mô lớn, hiện chưa tồn tại

Từ đó rút ra hai điểm thương mại. Thứ nhất, hiệu quả rẻ nhất đến từ kiến trúc: giữ toán hạng trong SRAM và bỏ HBM cùng CoWoS đã loại bỏ phần lớn chi phí của một GPU (khoảng 63% danh mục vật tư của một H100) trước khi chọn bất kỳ tiến trình nào. Thứ hai, thời gian là chi phí: một đội mới cần khoảng 18 đến 24 tháng từ lúc bắt đầu đến khi sản xuất hàng loạt và thời gian chờ ở tiến trình tiên tiến vượt 50 tuần, nên một giao thức thay đổi phép số học vài tháng một lần có thể khiến con chip lỗi thời trước khi Cổng 3 kịp hoàn vốn.

Ba quy tắc cho một đội nhỏ

Gộp lại, các con số này cho ba quy tắc phân tích. Thứ nhất, **hãy dừng ở Cổng 1 hoặc 2 cho đến khi có một người mua cụ thể hoặc một phiên bản giao thức được đóng băng**; cả giá token lẫn hashrate đều không thay thế được hai điều đó. Thứ hai, **mua k bằng kiến trúc trước, bằng tiến trình sau**: một mảng dữ liệu trong SRAM trên N7 có thể đánh bại một linh kiện đa dụng trên N3 cho một khối lượng công việc, với một phần nhỏ chi phí mặt nạ. Thứ ba, **ghi ra hashrate hoặc sản lượng hoàn vốn, có tính p, trước khi tape-out**, và định giá các khoản thanh toán theo mốc dựa trên một phiên bản giao thức cố định, để khách hàng, chứ không phải công ty thiết kế, gánh rủi ro luật lệ.

CHƯƠNG 11

11 CƠ HỘI
Cơ hội nằm ở đâu, và tôi sẽ làm gì

Chip tùy biến: nhu cầu áp đảo đến từ AI

Năm 2026, nhu cầu chip tùy biến áp đảo đến từ AI. Doanh thu bán dẫn AI của Broadcom trong quý 3 năm tài chính 2026 là 16,7 tỷ USD, tăng 221% so với cùng kỳ, khoảng 58 tỷ USD cho cả năm, 73% trong đó từ XPU cho sáu khách hàng (theo bản ghi tự động của cuộc gọi công bố kết quả kinh doanh). Doanh thu quý 2 của Global Unichip tăng 128% so với cùng kỳ, với doanh thu sản xuất trọn gói tăng 188% và biên lợi nhuận gộp 21,5%. Cơ cấu doanh thu quý 1 của Alchip theo tiến trình là 21% ở 3/2 nm, 63% ở 7/5 nm và 16% ở 12 nm trở lên. Đơn hàng mới năm 2026 của VeriSilicon đạt 8,24 tỷ nhân dân tệ (khoảng 1,16 tỷ USD) tính đến cuối tháng 4, 91,37% trong đó cho điện toán AI. Thị trường có ba tầng: trên cùng là đồng phát triển nhiều năm trị giá hàng tỷ đô la; ở giữa là các dự án từ 7 nm đến 2 nm với NRE từ vài triệu đến hơn một trăm triệu đô la (5–30 triệu USD ở tiến trình tầm trung, từ 30–100 triệu USD trở lên ở tiến trình tiên tiến nhất); và một cái đuôi dài các ASIC tiến trình trưởng thành với NRE 0,5–5 triệu USD. Khách hàng crypto và khai thác là một thị trường ngách nhỏ, có tính chu kỳ, nhưng họ là "khách hàng đầu tiên" nhanh nhất, ít rào cản nhất đang trả tiền cho sức tính hiện nay.

Những năng lực của một đội chip đào chuyển giao tốt nhất, vận hành điện áp thấp gần ngưỡng, mảng tính toán nhân bản quy mô lớn, cấp nguồn và tản nhiệt lớp kilowatt, và tape-out nhanh ở tiến trình tiên tiến, chính là những gì chip suy luận và bộ tăng tốc ZK, FHE cần. Thị trường vốn định giá điều này: Velaura AI (trước đây là Auradine) đã gọi 110 triệu USD vào tháng 8 ở mức định giá trên 1 tỷ USD. Phản ví dụ là đơn vị chip AI biên của Canaan, có doanh thu 0,9 triệu USD trong năm tài chính 2024 trong khi ngốn 15% chi phí hoạt động, và đã bị đóng cửa vào tháng 6 năm 2025. Khác biệt nằm ở một kênh ra thị trường độc lập, không phải ở việc tái sử dụng kênh đào coin.

Phát triển AI: giờ công là hàng phổ thông; giá trị nằm ở các khớp nối

Ở phân khúc thấp, dịch vụ phát triển AI đã bị phổ thông hóa: mức giá điển hình trên Clutch là 25–49 USD mỗi giờ, và 50–99 USD ở Mỹ, Canada và các thị trường tương tự; một đội phát triển blockchain tốn khoảng 15.000 đến hơn 90.000 USD mỗi tháng. Bán giờ công không có hào lũy. Giá trị nằm ở các khớp nối giữa các lĩnh vực: kernel và

trình biên dịch cho chip tùy biến, đào nhúng trong dịch vụ suy luận (plug-in vLLM của Pearl là một ví dụ), suy luận kiểm chứng được kết hợp TEE với kiểm tra ngẫu nhiên bằng ZK, và thanh toán cùng danh tính cho tác tử.

Hệ thống AI × crypto: ký với người mua sức tính trước khi làm token

Với bất kỳ ai thiết kế một chuỗi hay mạng điện toán gốc AI, bộ công cụ kiểm chứng năm 2026 đã định hình. Theo thứ tự chi phí trên mỗi đơn vị công việc, từ thấp đến cao: lấy mẫu thống kê (proof of sampling của Hyperbolic), ủy thác có trọng tài (Verde của Gensyn, chỉ cần một bên kiểm chứng trung thực nhưng đòi hỏi thực thi chính xác từng bit), TEE (tốc độ gần như gốc, nhưng phải tin nhà sản xuất chip), dư thừa toàn phần và zkML tổng quát. Thiết kế thực tế là phân tầng: TEE hoặc lấy mẫu cho độ trễ thấp, trò chơi tranh chấp hoặc kiểm tra ZK ngẫu nhiên cho tính chung cuộc; một bài báo tháng 12 năm 2025 đã mô phỏng điện toán bảo mật H100 với fraud proof và kiểm tra ZK ngẫu nhiên đạt 99% thông lượng tập trung và chi phí biên 0,07 USD mỗi truy vấn. Triết lý phần cứng quyết định bảo mật và phân phối: PoUW gốc GPU lúc đầu được bảo vệ bởi GPU đám mây cho thuê được, nên chi phí tấn công gắn với giá thuê, và Qubic cho thấy bằng chứng công việc trên phần cứng đa dụng có thể bị các động lực bên ngoài chiếm trọn.

Chi phí và thời gian không phải là rào cản. Pearl đi từ bài báo (tháng 4 năm 2025) đến mainnet trong khoảng 12 tháng, fork từ btcd, với full node, ví, light client và phần mềm đào ngay khi ra mắt. Phần lớn kiểm toán DeFi tốn 25.000–100.000 USD, kiểm toán doanh nghiệp hoặc đa chuỗi từ 150.000–250.000 USD trở lên, mạch ZK tốn thêm 80% đến 120%, và các dự án có giá trị thật đang chịu rủi ro dành 150.000–500.000 USD mỗi năm cho bảo mật. Rào cản thật là phân phối và thanh khoản. Các nhà sáng lập cho biết Binance đã đòi từ 0% đến 3,5%, và bị cáo buộc là 8%, nguồn cung token để niêm yết (Binance phủ nhận việc trục lợi từ niêm yết); cả sáu dự án mà Cointelegraph xem xét có dùng thỏa thuận tạo lập thị trường kiểu "cho vay token cộng quyền chọn mua" đều giảm giá, và bên tạo lập thị trường của Movement đã bán khoảng 38 triệu USD token ngay ngày niêm yết. Các lựa chọn thay thế là thanh khoản on-chain, như Hyperliquid, đã vượt 10 tỷ USD vốn hóa mà không niêm yết trên Binance và có các thị trường hợp đồng vĩnh cửu HIP-3 vượt 3,2 tỷ USD hợp đồng mở vào tháng 6, và bán công khai qua các sàn lớn, như khi Monad huy động 269 triệu USD qua Coinbase. Giá trị pha loãng hoàn toàn cao với lượng lưu hành thấp là chỉ báo rõ nhất cho sự sụt giảm sau ra mắt trong giai đoạn 2025–2026.

Thiết kế token là một phần của thiết kế hệ thống. Phát hành nên đi theo nhu cầu trả tiền chứ không đi trước nó: các mạng trong báo cáo này có bội số trợ cấp lành mạnh nhất gắn phần thưởng với công việc mà một người mua bên ngoài trả tiền, còn các mạng tệ nhất trả cho hashrate hay stake mà không có gì đứng sau. Một quy tắc thực tế là công bố, trước khi ra mắt, bội số trợ cấp mà mạng nhắm tới trong năm thứ nhất, thứ hai và thứ ba, và cắt giảm phát hành khi không đạt phân tích.

Hướng đi	Ai mua, và mức sẵn lòng chi trả	Cửa sổ	Năng lực chính	Rủi ro chính
Bộ tăng tốc ưu tiên suy luận có thể đào	Neocloud, triển khai riêng, mạng PoUW	2027–2028 (tùy fork FP8 và FP4)	GEMM INT8, FP8 và FP4; kiến trúc SRAM; tích hợp giao thức	Quyền tham gia giao thức, khóa chặt kiến trúc
Chip suy luận SRAM hoặc khắc cứng	Neocloud, trung tâm dữ liệu ở Trung Quốc đại lục, biên	Ngay bây giờ	Trình biên dịch và thích ứng mô hình; thiết kế N6–N4	Mô hình thay đổi quá nhanh
Chip đào (ngách hoặc tích hợp dọc)	Thợ đào có điện riêng; máy đào gia đình và sưởi ấm	Ngay bây giờ, bị giới hạn bởi halving và hashprice	Mảng điện áp thấp, hiệu quả, chuỗi cung ứng sản lượng lớn	Cuộc chiến giá của các hãng đương nhiệm, giá điện
Tăng tốc chứng minh ZK	Mạng prover, rollup	Chờ đến khi các hệ chứng minh hội tụ	MSM và NTT lập trình được	Nhu cầu phụ thuộc vào phần thưởng token

Hướng đi	Ai mua, và mức sẵn lòng chi trả	Cửa sổ	Năng lực chính	Rủi ro chính
Tăng tốc FHE	AI mã hóa cho doanh nghiệp và chính phủ; chuỗi riêng tư	2027–2029	NTT, hệ thống HBM, kỹ thuật mật mã	Nhu cầu chưa hình thành; chu kỳ bán hàng dài
Gia cố TEE và suy luận kiểm chứng được	Mạng suy luận bảo mật, DePIN	Ngay bây giờ	Mã hóa bộ nhớ có toàn vẹn, gốc attestation độc lập, kiểm tra ZK ngẫu nhiên	Phụ thuộc hệ sinh thái của nhà sản xuất
Chuỗi hoặc mạng điện toán gốc AI	Người mua suy luận, thợ đào, nhà phát triển	Ngay bây giờ	Thiết kế kiểm chứng, thiết kế phát hành, phân phối	Sụt giảm sau ra mắt, thanh khoản token

Ba hướng đi tôi sẽ chọn, mỗi hướng một bước đầu

Nếu phải chọn, tôi sẽ đi tối đa ba hướng, mỗi hướng có một bước đầu nhỏ để kiểm tra trước khi cam kết tiền phân tích.

- **Hướng A: một bộ tăng tốc ưu tiên suy luận và cũng đào được.** Bước đầu: một nghiên cứu kiến trúc bốn tuần mô hình hóa J/op và $/op cho một mảng GEMM INT8/FP8 dữ liệu trong SRAM so với RTX 5090 và H100, cả trên công việc của Pearl lẫn trên suy luận transformer, tiếp theo là kiểm chứng trên FPGA hoặc MPW tiến trình trưởng thành với NRE ở mức khoảng một triệu đô la.
- **Hướng B: dịch vụ silicon tùy biến với crypto là khách hàng đầu tiên và AI là thị trường thứ hai.** Bước đầu: biến mô hình NRE và hòa vốn của báo cáo này thành một nghiên cứu khả thi có thu phí cho các dự án chip tính toán, với thanh toán theo mốc và nghiệm thu dựa trên một phiên bản giao thức cố định.
- **Hướng C: một mạng gốc AI ký được người mua sức tính trước khi làm token.** Bước đầu: một ghi chú thiết kế về kiểm chứng phân tầng (TEE cộng lấy mẫu cộng kiểm tra ZK ngẫu nhiên) và về một phép số học chuẩn trung lập với nhà cung cấp, viết cùng một người mua suy luận cụ thể trước mọi thiết kế token.

CHƯƠNG 12

12

DẤU HIỆU CẦN THEO DÕI

Điều cần theo dõi, tháng 11 năm 2026 đến năm 2027

Khi nào	Cần theo dõi gì	Vì sao quan trọng
Bất cứ lúc nào	Độ cao kích hoạt FP8 của Pearl	Quyết định phần cứng số nguyên và card không phải NVIDIA có còn được chấp nhận không, và chuỗi có tách đôi không
Hằng tháng	Tỷ lệ khối Pearl chứa công việc AI	Phép thử tốt nhất cho việc "hữu ích" có đang thành sự thật không
Hằng tháng	Nhà cung cấp suy luận có chuyển thu nhập đào cho khách hàng không	Nhu cầu trả tiền gặp phát hành: mô hình mà Pearl cần
Hằng quý	Doanh thu kiểm chứng được của Bittensor so với lượng phát hành	Mạng công việc hữu ích lớn nhất có đang thu hẹp khoảng cách trợ cấp không
Hằng quý	Tỷ lệ dòng x402 thực sự do tác tử khởi tạo	Thanh toán của tác tử có chuyển từ tín hiệu sang thương mại không
Tháng 1 năm 2027	Mức tăng giá 3% đến 6% của TSMC theo tin đồn, nhiều nhất ở 2 nm và 3 nm	Chi phí tape-out ở tiến trình tiên tiến
Quý 1 năm 2027	FP4 trên Pearl, theo buổi AMA của các nhà sáng lập	Công việc dịch chuyển thêm về phía các định dạng mới nhất của NVIDIA
Cuối năm 2027	Mục tiêu hòa vốn của Samsung Foundry	Liệu một nguồn cung thứ hai rẻ hơn có còn khỏe mạnh
Tháng 4 năm 2028	Kỳ halving tiếp theo của bitcoin	Giảm một nửa quỹ thưởng của máy đào SHA-256

CHƯƠNG 13

13

CÂU HỎI NÊN ĐẶT RA

Những câu hỏi bạn chưa hỏi nhưng nên hỏi

Dưới đây là những câu hỏi tôi sẽ đặt ra trước khi cam kết tiền hay silicon cho bất kỳ dự án AI × crypto nào.

Ai kiểm soát hàm công việc? Nếu một đội nhỏ có thể thay đổi phép số học ở một độ cao mã hóa cứng, tuổi thọ phần cứng của bạn là quyết định của họ. Hãy yêu cầu một cam kết ổn định bằng văn bản, hoặc thiết kế phần cứng có thị trường thứ hai.

Bội số trợ cấp là bao nhiêu? Chia lượng phát hành (hoặc vốn hóa) cho doanh thu trả từ bên ngoài mạng. Bội số trên 10 nghĩa là mạng đang mua hoạt động, chứ không bán dịch vụ.

Cổng tiếp theo tốn bao nhiêu, và con số nào khiến bạn dừng lại? Một wafer đa dự án tốn hàng chục hoặc hàng trăm nghìn đô la; một bộ mặt nạ tốn hàng triệu. Hãy quyết định con số dừng trước khi chi tiền.

Hashrate hoàn vốn là bao nhiêu, khi tính cả p? Chia NRE cho R × (1 − 1/k) × T × p. Nếu kết quả là một phần lớn của mạng, con chip là một canh bạc vào luật lệ, không phải vào silicon.

Thị trường thứ hai của con chip là gì? Nếu giao thức fork hoặc token sụt giá, silicon còn bán được vào đâu? GPU sống sót qua The Merge của Ethereum vì có thị trường thứ hai; ASIC đơn thuật toán thì không.

Giá trị rơi vào token hay vào cổ phần? Năm 2026, các công ty AI × crypto có lãi phần lớn là công ty cổ phần có kênh phân phối crypto; hãy quyết định bạn đang xây cái nào trong hai thứ đó trước khi gọi vốn.

Khách hàng trả tiền đầu tiên là ai, tên cụ thể? Mọi trường hợp bền vững trong báo cáo này, từ Grass đến các subnet điện toán của Bittensor, đều bắt đầu với một người mua cụ thể bên ngoài mạng.

Đây là những lĩnh vực tôi muốn cùng khám phá với những người khác, qua trao đổi kỹ thuật và hợp tác kỹ thuật: chip tùy biến cho khai thác, huấn luyện và suy luận AI, và các bộ tăng tốc; phát triển AI; và hệ thống AI × crypto từ chuỗi đến node đến chip. Bạn có thể liên hệ tôi qua i@tanshuai.com.

Giới hạn của báo cáo này

Đây là nghiên cứu tại bàn dựa trên tài liệu công khai đến ngày 13 tháng 11 năm 2026. Tôi không thực hiện phỏng vấn sơ cấp hay phép đo độc lập nào cho ấn bản này. Một số con số chính là tuyên bố của công ty hoặc nhà cung cấp, chẳng hạn doanh thu của Aethir, doanh thu của Grass, hiệu năng của Etched và Taalas và hiệu quả của Cysic; chúng được ghi nhãn tương ứng. Các chỉ số chuỗi của Pearl đến từ một full node độc lập và từ các pool, và thay đổi theo giờ; tôi dùng các số đọc quanh 07:13 UTC ngày 13 tháng 11 và đưa ra dải trong ngày. Chi phí wafer, mặt nạ và thiết kế là ước tính của bên thứ ba và giá trị mô hình không thống nhất với nhau; tôi trình bày dưới dạng dải. Phép tính hòa vốn ASIC và ngân sách theo giai đoạn của tôi dùng các giả định minh họa (k, T, p, NRE và chi phí từng cổng) mà người đọc nên thay bằng giả định của riêng mình.



Về tác giả

Tan Shuai là nhà công nghệ và nhà sáng lập với hơn hai mươi năm làm việc trong lĩnh vực chip, phần cứng và phần mềm hệ thống, cùng 20 bằng sáng chế đã được cấp. Ông có nhiều năm nghiên cứu, phát triển và quản lý chip fabless, bao gồm thiết kế đường dữ liệu ASIC full-custom cho chip đào tiền điện tử và chip camera đa phương tiện, và đó là lăng kính của báo cáo này: năng lượng mỗi phép tính, chi phí mỗi wafer, giao thức cho phép bạn cố định những gì trong silicon, và ai là người trả tiền. Công việc đã đưa ông qua Thung lũng Silicon, New York, Chicago, Đài Loan, Vùng Vịnh Lớn (Thâm Quyến và Hồng Kông), Việt Nam và Malaysia. Hiện ông là nhà nghiên cứu và phát triển AI.

Trao đổi kỹ thuật: i@tanshuai.com · tanshuai.com

Ghi chú và tuyên bố miễn trừ

Báo cáo này là nghiên cứu tại bàn độc lập dựa trên tài liệu công khai có đến ngày 13 tháng 11 năm 2026. Nó chỉ nhằm mục đích cung cấp thông tin và học tập, không phải tư vấn đầu tư hay kỹ thuật. Báo cáo chỉ xem xét công nghệ và kinh doanh. Giá cả, chỉ số chuỗi và luật giao thức đều thay đổi; hãy đối chiếu mọi thông tin bạn dựa vào với nguồn gốc. Các số liệu được đánh dấu hãng tự công bố, ước tính hoặc phân tích cần được hiểu đúng như vậy. Giá tài sản crypto biến động rất mạnh.

© 2026 Tan Shuai. Bạn có thể chia sẻ nguyên vẹn báo cáo này kèm ghi nguồn.

Nguồn

Tất cả được truy cập trong khoảng từ ngày 10 đến ngày 13 tháng 11 năm 2026. Nhóm theo chương.

Pearl: giao thức, mã nguồn và dữ liệu chuỗi

- Komargodski and Weinstein, "Proofs of Useful Work from Arbitrary Matrix Multiplication", arXiv 2504.09971: https://arxiv.org/abs/2504.09971
- Pearl Research Labs, INT whitepaper: https://pearlresearch.ai/research/int-whitepaper
- Mã nguồn Pearl: khối genesis: https://github.com/pearl-research-labs/pearl/blob/main/node/chaincfg/genesis.go ; tham số đồng thuận và độ cao fork: https://github.com/pearl-research-labs/pearl/blob/main/node/chaincfg/params.go ; cách tính hashrate: https://github.com/pearl-research-labs/pearl/blob/main/zk-pow/src/api/sanity_checks.rs ; hướng dẫn nâng cấp fork seed có salt: https://github.com/pearl-research-labs/pearl/blob/main/docs/salted-seed-fork-upgrade-guide.md ; tham số nhánh 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, đào bằng dấu phẩy động: https://github.com/pearl-research-labs/pips/blob/main/PIPS/pip-0003.md
- prlstats (full node độc lập): mạng: https://prlstats.com/api/network ; chuỗi số liệu: https://prlstats.com/api/series ; trang chủ và mô hình thu nhập thợ đào: https://prlstats.com/api/home ; nguồn cung đã đào: https://prlstats.com/api/mined ; pool: https://prlstats.com/api/pools ; người nắm giữ: https://prlstats.com/api/holders ; giả định phần cứng và chi phí phụ trội: https://prlstats.com/api/assumptions
- Kryptex Pool: mạng Pearl: https://pool.kryptex.com/prl/about-coin ; hashrate: 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 ; lịch sử giá: https://api.coingecko.com/api/v3/coins/pearl-2/market_chart?vs_currency=usd&days=365&interval=daily ; tìm mã "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", ngày 2 tháng 6 năm 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, hợp tác với Pearl Research Labs: https://together.ai/blog/together-ai-partners-with-pearl-research-labs ; trang mô hình: https://together.ai/models/gemma-4-31b-it-pearl
- Basu, khảo sát các phiên đào Pearl, arXiv 2606.04819: https://arxiv.org/html/2606.04819v2
- Blockworks, bản tin 0xResearch, ngày 28 tháng 10 năm 2026: https://blockworks.com/newsletter/0xresearch/issue/post_daf0d74a-2bf4-41e4-86d3-2a1275475a24
- Datawallet, "Pearl Explained": https://www.datawallet.com/crypto/pearl-explained
- Coingabbar về bài của TheMarker về vòng gọi vốn của Pearl Research Labs: https://www.coingabbar.com/en/pearl-research-labs-funding-130m-stealth-round-themarker

Chuỗi công việc hữu ích và lịch sử

- Grothe và cộng sự, phân tích bảo mật Gridcoin, USENIX WOOT 2017: https://www.usenix.org/system/files/conference/woot17/woot17-paper-grothe.pdf
- Giả mạo proof-of-learning, arXiv 2208.03567: https://arxiv.org/abs/2208.03567
- hoge.gg, kiểm chứng và niềm tin trong suy luận phi tập trung: https://hoge.gg/decentralized-inference-2026-verification-trust/ ; doanh thu và ngân quỹ Bittensor: https://hoge.gg/bittensor-quantum-wallet-revenue-tao-treasury-2026/
- Kaspa: mã ma trận kHeavyHash: https://github.com/kaspanet/rusty-kaspa/blob/master/consensus/pow/src/matrix.rs ; ra mắt: https://coinranking.com/coin/V8GxkwWow+kaspa-kas/about ; thông báo IceRiver KS1/KS2: https://www.iceriver.io/2023/03/14/1677644580/ ; lịch sử từ GPU đến ASIC: https://captainmining.com/blogs/crypto-news/the-evolution-of-kaspa-mining-from-gpus-to-asic-powerhouses ; hashrate GPU: https://coincalculators.io/algorithms/kheavyhash ; KS7: https://www.asicminervalue.com/miners/bitmain/antminer-ks7-40th ; mạng: https://minerstat.com/coin/kas ; 1 EH/s năm 2024: https://ourcryptotalk.com/news/kaspa-reaches-1-ehs-in-hashrate ; Crescendo: https://ourcryptotalk.com/news/kaspa-crescendo-live
- Aleo: FAQ về mainnet: https://aleo.org/post/aleo-mainnet-faq/ ; chứng minh phi tập trung: https://aleo.org/post/decentralized-proving-advantages/ ; ASICs: 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 về đợt tái tổ chức 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 ; đào Dogecoin: https://hackernoon.com/qubic-is-now-mining-dogecoin-on-its-ai-compute-infrastructure ; CoinGecko: https://www.coingecko.com/en/coins/qubic

- Tài liệu đồng thuận 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 và các fork kháng ASIC: https://lbry.com/news/forking-for-asic-resistance ; fork Sia: https://hotfix.coindesk.com/markets/2018/11/01/sia-completes-hard-fork-as-bricked-miners-form-rival-blockchain ; giá Obelisk: https://www.cryptoglobe.com/latest/2018/11/obelisk-hardware-skyrockets-in-selling-price-after-sia-hard-fork/ ; thợ đào sau The Merge: https://hotfix.coindesk.com/business/2022/11/11/ex-ethereum-miners-token-hop-to-stay-alive-after-the-merge

Các mảng AI × crypto

- DefiLlama: Render: https://defillama.com/protocol/render-network-bme ; Livepeer: https://defillama.com/protocol/livepeer
- Cointelegraph về số liệu doanh thu DePIN của Messari: https://cointelegraph.com/news/depin-revenues-grow-despite-token-losses
- Aethir, tổng kết 2025: https://aethir.com/blog-posts/aethirs-2025-wrap-up-decentralized-gpu-cloud-milestones
- CoinGecko, danh mục AI: https://www.coingecko.com/en/categories/artificial-intelligence
- Akash, giá thuê H100: https://akash.network/the-bid/h100-rental-price-2026-cost-per-hour/ ; báo cáo quý 1/2026: https://akash.network/blog/akash-network-q1-2026-report/
- io.net, hợp đồng 8 triệu USD và lượng token hằng ngày: 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, hợp đồng của 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 về INTELLECT-3: https://www.implicator.ai/prime-intellects-intellect-3-open-source-ambition-meets-centralized-reality/
- SiliconANGLE, vòng gọi vốn của 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, vòng gọi vốn của Nous: https://runtimewire.com/article/nous-research-75-million-round-1-5-billion-valuation
- ownyourmind, Gensyn: https://ownyourmind.ai/projects/gensyn/ ; Phemex, Ritual (chỉ tiêu đề): https://phemex.com/news/article/ritual-launches-ai-layer-1-mainnet-and-ritual-token-price-dives-72-from-high-101259
- Chi phí phụ trội của điện toán bảo mật, arXiv 2409.03992: https://arxiv.org/abs/2409.03992v2 ; báo cáo 2025 của 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, đo lường thanh toán của tác tử AI: https://www.trmlabs.com/trm-tech-blog/whos-actually-paying-measuring-ai-agent-payments-onchain ; Forkast về 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 Tempo: https://www.theblock.co/amp/post/394131/tempo-mainnet-goes-live-with-machine-payments-protocol-for-agents ; BlockEden về aGDP của Virtuals: https://blockeden.xyz/blog/2026/04/21/agdp-agent-gdp-virtuals-protocol-ai-blockchain-valuation-primitive-tvl-displacement/ ; The Block về ElizaOS: https://theblock.co/post/410774/eliza-labs-native-token-dead
- Dealroom, doanh thu Grass: https://dealroom.co/news/157794-grass-backed-by-multicoin-projects-75m-in-2026-data-revenue/ ; CryptoSlate, Fetch.ai and Ocean: https://cryptoslate.com/cryptos-flagship-ai-pact-fracture-fetch-sues-ocean-over-263m-fet-community-sales/ ; Fortune về World: https://fortune.com/2026/06/01/internet-according-to-sam-altman-seeing-orb/
- Galaxy, vốn mạo hiểm crypto: https://www.galaxy.com/insights/research/crypto-blockchain-venture-capital-q3 ; Crypto Briefing (chỉ tiêu đề): https://cryptobriefing.com/crypto-vc-investment-q3-2026-galaxy/ ; Blockhead về Tiger Research: https://www.blockhead.co/2026/10/14/bitcoin-rose-43-in-q3-crypto-venture-funding-did-not-follow/ ; ForkLog về các token ra mắt năm 2025: https://forklog.com/en/85-of-tokens-launched-in-2025-fell-below-initial-prices/

Kinh tế học ASIC và các loại chip

- Hashrate Index, bản tổng hợp ngày 9 tháng 11 năm 2026: https://hashrateindex.com/blog/hashrate-index-roundup-november-9-2026/
- AnySilicon, chi phí thiết kế ASIC theo tiến trình năm 2026: https://anysilicon.com/news/how-much-does-asic-design-cost-in-2026-real-price-ranges-by-process-node/ ; Silicon Analysts, hướng dẫn chi phí tape-out cho fabless: https://siliconanalysts.com/analysis/fabless-startup-tapeout-cost-guide ; dữ liệu thị trường: https://siliconanalysts.com/market-data
- MillionMiner, Bitmain S23 XP Hyd: https://millionminer.com/news/bitmain-wdms-2026-s23-xp-hyd-launch
- Khattak và Mikaitis, tích lũy FP8 trong tensor core, arXiv 2512.07004: https://arxiv.org/html/2512.07004v2
- Canaan, kết quả quý 2 năm 2026: https://finviz.com/news/389337/canaan-inc-reports-unaudited-second-quarter-2026-financial-results ; Bitdeer, kết quả quý 3 năm 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 và Block: https://theenergymag.com/news/2026-07-29/core-scientific-exit-block-bitcoin-deal-ai ; kinh tế học cho thuê AI: https://theenergymag.com/news/2026-08-06/miner-weekly-ai-pumpability-out-fuel
- Blockworks, Cipher và 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, kết quả quý 1 năm tài chính 2027: https://globenewswire.com/news-release/2026/11/05/3397020/0/en/iren-reports-q1-fy27-results.html
- Spark, mức tập trung thị trường ASIC bitcoin: https://www.spark.money/research/bitcoin-asic-manufacturer-market-concentration

- Google, tài liệu 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 ; bài báo 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, giá 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, so sánh các bộ tăng tốc suy luận AI: https://themenonlab.blog/blog/ai-inference-accelerators-compared ; Introl, bộ tăng tốc AI ngoài GPU: https://introl.com/blog/ai-accelerators-beyond-gpus-tpu-trainium-gaudi-cerebras-2025
- StratNews (Reuters), chip AI nội địa trong trung tâm dữ liệu dùng vốn nhà nước: https://stratnewsglobal.com/team-sng/china-orders-state-funded-data-centres-to-use-domestic-ai-chips/ ; TNW, nhập khẩu H200: https://thenextweb.com/news/nvidia-h200-chips-hong-kong-china-import-approval
- Succinct, chứng minh thời gian thực trên 16 GPU: https://blog.succinct.xyz/real-time-proving-16-gpus/ ; sản phẩm ASIC ZK của 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 và SemiFive: https://evertiq.com/design/2026-02-20-semifive-secures-design-win-with-niobium-for-fhe-accelerator ; The Quantum Insider, vòng gọi vốn của 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 trên FPGA: https://www.zama.org/post/announcing-hpu-on-fpga-the-first-open-source-hardware-accelerator-for-fhe

Tiến trình

- Silicon Analysts, giá wafer: https://siliconanalysts.com/data/wafer-pricing
- Tom's Hardware, tin đồn giá 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, giá SMIC: https://wccftech.com/smic-5nm-chip-prices-up-to-50-percent-more-expensive/amp/
- EUROPRACTICE, lịch MPW TSMC 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, lộ trình 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, chuỗi cung ứng chip AI của Trung Quốc: 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

Cơ hội và kinh tế học ra mắt

- Webull (Benzinga), cuộc gọi công bố kết quả của Broadcom: https://www.webull.com/news/15511095937917952 ; Quartr, Global Unichip Q2 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, đơn vị AI của Canaan: https://www.theblock.co/post/359250/crypto-miner-canaan-to-shutter-non-core-ai-semiconductor-business-unit-amid-strategic-realignment
- Clutch, giá phát triển AI: https://clutch.co/developers/artificial-intelligence/pricing ; PixelPlex, chi phí phát triển blockchain: https://pixelplex.io/blog/blockchain-development-cost/
- Hyperbolic, proof of sampling: 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 ; kiểm chứng phân tầng, arXiv 2512.20176: https://arxiv.org/abs/2512.20176v1
- Sherlock, tham chiếu giá kiểm toán năm 2026: https://sherlock.xyz/post/smart-contract-audit-pricing-a-market-reference-for-2026
- Blockworks, tiêu chuẩn niêm yết của Binance: https://blockworks.com/news/binance-listing-standards-scrutiny ; Cointelegraph, các thỏa thuận tạo lập thị trường: 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

BẰNG CHỨNG CÔNG VIỆC HỮU ÍCH · CRYPTO × SILICON

Tính toán là đồng thuận

AI × Crypto × Chip năm 2026, khởi đầu từ bằng chứng công việc hữu ích của Pearl

Mọi con số quan trọng trong báo cáo này đều kèm nguồn và ngày. Con số do tác giả tự suy ra được đánh dấu là ước tính; con số chỉ do bên bán công bố được đánh dấu là hãng tự công bố. Báo cáo chỉ xem xét công nghệ và kinh doanh.

Tan Shuai

Nhà công nghệ và nhà sáng lập · Nhà nghiên cứu và phát triển AI

tanshuai.com · i@tanshuai.com

BẢN CÔNG KHAI · Số liệu tính đến ngày 13 tháng 11 năm 2026

