NOTA DI RICERCA · EDIZIONE PUBBLICA

2026 · 11



PROOF OF USEFUL WORK · CRYPTO × SILICIO

Il calcolo come consenso

AI × crypto × chip nel 2026, a partire dalla proof of useful work di Pearl

Pearl · Lavoro utile · DePIN e agenti · Economia degli ASIC · Nodi di processo · Budget a tappe

IL VERDETTO

Pearl dimostra che la moltiplicazione di matrici può rendere sicura una blockchain, ma non ancora che il lavoro sia utile. In tutto il campo AI × crypto la domanda pagata sta nell'inferenza, nei dati e nella privacy, e un ASIC ripaga solo dove il piatto è grande, le regole sono stabili e il chip ha un secondo mercato.

Tan Shuai

Tecnologo e fondatore · Ricercatore e sviluppatore di AI

tanshuai.com · i@tanshuai.com

Dati aggiornati al 13 novembre 2026

Importi in dollari USA

INDICE

Indice

◆ **Sintesi** — Il teorema regge, l'incentivo no — 04

01 **Ambito e metodo** — Ambito, metodo e quattro test — 07

02 **Pearl · il meccanismo** — Pearl: la moltiplicazione di matrici come consenso — 08

03 **Pearl · l'economia** — Pearl: scala, emissione e divario di utilità — 10

04 **Pearl · le regole** — Pearl: rischio di regole, numerica e governance — 12

05 **Catene a lavoro utile** — Catene a lavoro utile: né le matrici né la conoscenza zero fermano gli ASIC — 13

06 **Il panorama più ampio** — Il resto di AI × crypto: domanda reale ma piccola, e il valore che passa all'equity — 16

07 **Economia degli ASIC** — La questione degli ASIC: una formula, tre metri di misura — 18

08 **Chip per chip** — Cinque tipi di chip, cinque risposte — 21

09 **Nodi di processo** — Nodi di processo: lasciare che scelga il costo più grande sull'intera vita — 25

10 **Un budget a tappe** — Dal modello al silicio: un budget a tappe per un primo chip — 29

11 **Opportunità** — Dove sono le opportunità, e che cosa farei — 31

12 **Segnali da seguire** — Che cosa seguire, da novembre 2026 al 2027 — 33

13 **Domande da porre** — Le domande che non avete fatto ma avreste dovuto fare — 34

◆ **L'autore** — Tan Shuai — 36

◆ **Fonti** — 37

FIGURE

Figure

FIG. 01	AI × crypto in sei settori: dove la domanda pagata è reale	06
FIG. 02	Sul lavoro intero di Pearl le GPU hanno un tetto energetico comune	09
FIG. 03	I primi sei mesi di Pearl: l'hashrate segue il prezzo, le regole cambiano	11
FIG. 04	Il multiplo del sussidio: quante volte i ricavi esterni paga una rete	15
FIG. 05	Perché l'efficienza smette di pagare: un ASIC tiene al massimo 1 − 1/k dei ricavi	19
FIG. 06	La maggior parte dei piatti è più piccola del budget di un solo chip	21
FIG. 07	Cinque tipi di chip, cinque risposte	25
FIG. 08	Un wafer all'avanguardia costa dieci volte uno maturo	26
FIG. 09	Prove prima delle maschere: un budget a tappe per un primo chip	29

Etichette nel testo

- **stima** — calcolo o deduzione dell'autore a partire da dati citati
- **dichiarato dal fornitore** — pubblicato dall'azienda che vende il prodotto, non verificato in modo indipendente
- **analisi** — interpretazione delle evidenze da parte dell'autore

Questa è una traduzione dell'edizione inglese; in caso di differenze prevale l'edizione inglese.

Dati aggiornati al 13 novembre 2026 · Importi in dollari USA

SINTESI

Il teorema regge, l'incentivo no

IL VERDETTO

Pearl dimostra che il "calcolo come consenso" funziona come ingegneria, ma non ancora che sia utile.

≈90 EH/s
la rete di Pearl il 13 novembre 2026, dove un H è una moltiplicazione-accumulo INT7

≈1%
dei blocchi di Pearl contiene lavoro di AI reale (Blockworks, 28 ottobre 2026)

≈70 mln $
l'anno di ricavi esterni verificabili di Bittensor, contro più di 300 milioni di dollari di emissioni

175–180×
le ricompense dei prover di Aleo rispetto alle sue commissioni nel terzo trimestre del 2025

R × (1 – 1/k)
il tetto di profitto di un ASIC k volte più economico del miner marginale stima

≈1,5 mln $
per arrivare al silicio misurato su N7 con un modello, un FPGA e un MPW, prima di un set di maschere da ≈3,5 mln $ stima

Pearl

Il 13 novembre 2026 la rete di Pearl girava a circa 90 EH/s. Nel suo codice di consenso un "hash" è una moltiplicazione-accumulo intera, quindi la rete equivale al throughput intero di circa 122.000 H100. Eppure solo circa l'1% dei blocchi contiene lavoro di AI reale, e l'unica partnership pubblica di inferenza non è più offerta sull'API serverless del suo fornitore. La causa è economica, non crittografica: una GPU che mina mentre serve inferenza guadagna solo dall'1% al 27% circa dell'hashrate di un miner dedicato, quindi i miner razionali moltiplicano matrici casuali.

Domanda pagata

In tutto il campo AI × crypto la domanda pagata che si può verificare compare in tre punti: vendere inferenza, vendere dati e vendere privacy. I ricavi esterni verificabili di Bittensor sono di circa 70 milioni di dollari l'anno, per circa quattro quinti da subnet di calcolo e infrastruttura. Grass dichiara 17 milioni di dollari di ricavi dati nel primo semestre 2026. L'inferenza riservata negli ambienti di esecuzione affidabili (TEE) è in produzione. Al contrario, i pagamenti avviati davvero da agenti AI arrivano a circa 5.000–11.000 dollari al mese, e zkML e addestramento decentralizzato non hanno ancora alcuna domanda pagante dichiarata.

Sussidio e valore

Il multiplo del sussidio è il controllo di salute più onesto. Le ricompense dei prover di Aleo erano circa 175–180 volte le sue commissioni; le emissioni di Bittensor sono circa 4–11 volte i suoi ricavi verificabili; il valore di mercato di Render è circa 360 volte le sue commissioni on-chain; e i circa 460 milioni di dollari l'anno di reddito dei miner di Pearl sono per lo più sussidio di blocco, quasi slegato dalla domanda pagata di AI. Il sussidio è carburante per l'avvio, non un modello di business. Nel frattempo il valore si sposta dai token all'equity: quattro società con equity che usano infrastruttura centralizzata e distribuzione nativa crypto sono valutate circa 8,5 miliardi di dollari, più o meno quanto i token AI dei livelli protocollo e applicazione messi insieme.

ASIC

Quanto ai chip, il tetto di profitto di un ASIC è all'incirca il ricavo per unità di calcolo per (1 − 1/k), dove k indica quante volte è più economico dell'hardware marginale. Poiché quel termine è concavo, l'efficienza oltre k = 5–10 non aggiunge quasi alcun profitto; a decidere l'esito sono la dimensione del piatto, quanto durano le regole e se il chip ha un secondo mercato. Le mie risposte per tipo di chip: non fare ora il tape-out di un chip Pearl INT7 dedicato, e puntare invece su un acceleratore pensato per l'inferenza che sappia anche minare; stare fuori, da nuovo entrante, dal mercato dei miner bitcoin venduti a terzi; non costruire, da team piccolo, un chip di addestramento autonomo; costruire chip di inferenza quando c'è un cliente firmato; coprirsi sugli acceleratori ZK e posizionarsi presto sull'FHE.

Nodi e budget

Il nodo di processo va scelto in base alla voce più grande nel costo del chip sull'intera vita: l'energia per i chip che girano giorno e notte, packaging e memoria per i chip di addestramento, la decisione tra SRAM e HBM per i chip di inferenza. E un primo chip va comprato a tappe, un modello, un FPGA, un wafer multiprogetto e solo dopo un set di maschere, in modo che un cambiamento del protocollo costi il prezzo di un prototipo e non quello di un tape-out. Le opportunità stanno dove i campi si incrociano: la crypto come primo cliente, l'AI come secondo mercato.

Domanda	Giudizio	Numeri chiave (fonti nei capitoli)
① **Pearl funziona, e il suo lavoro è utile?**	Funziona: una blockchain protetta su larga scala dalla moltiplicazione di matrici. Utile non ancora: quasi tutto il mining è su matrici casuali	Circa 90 EH/s il 13 novembre 2026; 1 H = 1 moltiplicazione-accumulo INT7; circa l'1% dei blocchi con lavoro di AI
② **Dov'è la vera domanda pagata in AI × crypto?**	Inferenza, dati e privacy. Pagamenti tra agenti, zkML e addestramento decentralizzato sono ancora prima della domanda	Bittensor circa 70 milioni di dollari l'anno di ricavi verificabili; Grass 17 milioni di dollari nel primo semestre 2026; pagamenti tra agenti circa 5.000–11.000 dollari al mese
③ **Vale la pena costruire un ASIC?**	Solo dove il piatto è grande, le regole sono stabili e il chip ha un secondo mercato. Oggi nessuna rete offre tutte e tre le cose	Tetto di profitto ≈ R × (1 − 1/k); bitcoin circa 11,5 miliardi di dollari l'anno di ricavi dei miner, Pearl circa 0,46–0,49 miliardi, Kaspa circa 18 milioni
④ **Quali chip dovrebbe costruire un team piccolo?**	Acceleratori pensati per l'inferenza che sanno anche minare; chip di inferenza per un cliente firmato; lavoro precoce sull'FHE. Non miner bitcoin da vendere a terzi né chip di addestramento	Il miglior ASIC di Kaspa rende 0,86 dollari al giorno; budget di progetto a 3/2 nm da 60 a 400 milioni di dollari o più
⑤ **Quale nodo di processo?**	Decide il costo più grande sull'intera vita: energia, packaging e memoria, o l'architettura della memoria	TSMC N2 circa 30.000 dollari a wafer contro circa 3.000 per N28; HBM e CoWoS circa il 63% del costo di una H100
⑥ **Quanto costa un primo chip, e come va comprato?**	In quattro tappe: un modello, un FPGA, un wafer multiprogetto, poi le maschere. Fermarsi prima delle maschere se le regole o l'acquirente non sono definiti	Uno slot MPW commerciale 50.000–500.000 dollari; maschere N7 circa 3,5 milioni di dollari; un budget di progetto a 7/6 nm 15–100 milioni di dollari

FIG. 01 **AI × crypto in sei settori: dove la domanda pagata è reale**

Domanda pagata verificabile contro i segnali di sussidio e di valutazione, novembre 2026

Catene a lavoro utile 01

REALE MA PICCOLO

≈70 mln $ l'anno

ricavi esterni verificabili di Bittensor

Segnale: emissioni 4–11× i ricavi; Pearl ≈1% blocchi AI

I meccanismi funzionano; l'utilità richiede chi paga

Calcolo decentralizzato 02

REALE MA PICCOLO

2,35 mln $ l'anno

commissioni on-chain di Render, le più alte verificabili

Segnale: valore di mercato ≈360× le commissioni

Soprattutto rivendita di inferenza standard

Addestramento e inferenza 03

PROFITTO ALL'EQUITY

100+ mln $ l'anno

Prime Intellect, società con equity su GPU centralizzate

Segnale: mainnet di addestramento di Gensyn non attiva

Il progresso è reale; il profitto va all'equity

AI verificabile 04

AGLI INIZI

398

utenti paganti del cloud TEE di Phala, fine 2025

Segnale: nessun zkML pagato dichiarato

TEE in produzione; zkML agli inizi

Pagamenti tra agenti 05

PRIMA DELLA DOMANDA

5–11 mila $ al mese

pagamenti x402 avviati davvero da agenti

Segnale: ≈200 mln di transazioni, per lo più segnali

Prima sono arrivati i binari; la domanda no

Reti di dati 06

ACQUIRENTI REALI

17 mln $

ricavi dati di Grass, primo semestre 2026 (autodichiarati)

Segnale: valore di mercato ≈10× i ricavi annualizzati

Uno dei pochi settori con acquirenti reali

Fonte: DefiLlama, SubConnect (via hoge.gg), TRM Labs, Phala, Dealroom, SiliconANGLE, CoinGecko, Blockworks; sintesi dell'autore.

01 AMBITO E METODO
Ambito, metodo e quattro test

Perché questo rapporto, e perché ora

Nell'aprile 2026 una nuova blockchain, Pearl, ha portato in produzione un'idea di vecchia data: far competere i miner per i blocchi eseguendo le moltiplicazioni di matrici su cui gira l'AI, invece dell'hashing. In sei mesi la sua rete ha raggiunto circa 90 EH/s e il suo token un valore di mercato di circa 500 milioni di dollari. Nello stesso periodo marketplace di calcolo AI, reti di addestramento decentralizzato, binari di pagamento per agenti e reti di dati hanno raccolto miliardi, e i miner di bitcoin hanno cominciato a convertire i loro siti con energia in data center AI. Volevo sapere quali di questi hanno clienti reali, se il silicio custom abbia senso da qualche parte al loro interno, e quale nodo di processo e quale fonderia un team di chip possa davvero usare.

Quattro test per qualsiasi progetto AI × crypto

Uno studio di sicurezza del 2017 su Gridcoin sosteneva che un calcolo utile può proteggere una blockchain solo se è difficile da risolvere, facile da verificare e facile da ricompensare. Io aggiungo un quarto test. Insieme decidono gran parte di ciò che segue.

01 **Difficile da risolvere.** Il calcolo onesto deve essere il modo più economico per produrre una prova valida; non devono esistere scorciatoie.

02 **Facile da verificare.** Controllare il lavoro deve costare molto meno che farlo, altrimenti la rete paga due volte.

03 **Facile da ricompensare.** Il protocollo deve poter calcolare la ricompensa dalla sola prova, senza fidarsi di una terza parte.

04 **Qualcuno paga per il risultato.** Se un cliente esterno alla rete non paga per il risultato, il lavoro "utile" è un sussidio con una narrazione.

I puzzle di hash superano i primi tre test e falliscono il quarto. I compiti di AI reali superano il quarto ma falliscono il secondo: provare a conoscenza zero un calcolo di AI costa da 100 a 10.000 volte il calcolo stesso, e si è dimostrato che gli schemi di proof of learning si possono falsificare a basso costo. La moltiplicazione di matrici rumorose di Pearl è il primo ponte credibile che supera il test di verifica pur mantenendo la stessa forma del calcolo AI.

Come ho lavorato

Questa è una ricerca a tavolino, fatta come esaminerei un chip o un business plan. Ho letto il codice di Pearl, le sue proposte di protocollo e il suo whitepaper; i dati della catena da un nodo completo indipendente; documenti societari e comunicati sui risultati; articoli di crittografia, architettura dei calcolatori e machine learning; dati sui prezzi di fonderie e fornitori; e la stampa di settore e di tecnologia. Ogni numero riporta la sua data. Dove ricavo un numero io stesso, lo dico e lo segno stima. I numeri riportati solo dal venditore sono segnati dichiarato dal fornitore; la mia interpretazione delle evidenze è segnata analisi. Gli importi sono in dollari USA. La situazione è aggiornata al 13 novembre 2026.

CAPITOLO 02

02

PEARL · IL MECCANISMO

Pearl: la moltiplicazione di matrici come consenso

L'idea, dal teorema alla mainnet

Nell'aprile 2025 Ilan Komargodski e Omri Weinstein hanno pubblicato "Proofs of Useful Work from Arbitrary Matrix Multiplication". Costruiscono una proof of useful work per qualsiasi prodotto di matrici con un overhead di sole 1 + o(1) volte una moltiplicazione ingenua; la sua sicurezza poggia sulla difficoltà di risolvere lotti di equazioni lineari casuali di rango basso, che è una congettura, non una dimostrazione. L'articolo indica la vera difficoltà del campo: una proof of useful work con un overhead a fattore costante è banale ma inutile; la parte difficile è fare in modo che il calcolo onesto resti il modo più economico per vincere quando i miner scelgono i propri input.

La mainnet di Pearl è partita il 27 aprile 2026 alle 09:00 UTC. I miner calcolano un prodotto rumoroso (A + E)(B + F), dove il rumore di rango basso è derivato da un commitment BLAKE3, e poi rimuovono il rumore per recuperare A·B. Lo stato dell'accumulatore di ogni tile di output viene trasformato con un hash in un biglietto della lotteria; un tile vincente viene aperto e provato con una prova ricorsiva a conoscenza zero Plonky2 di meno di 60 KB, dentro un certificato di blocco limitato a 65 KB. La verifica è una piccola prova di dimensione costante, indipendente dalla dimensione delle matrici: economica da controllare quanto un puzzle di hash, mentre il calcolo in sé può essere utile.

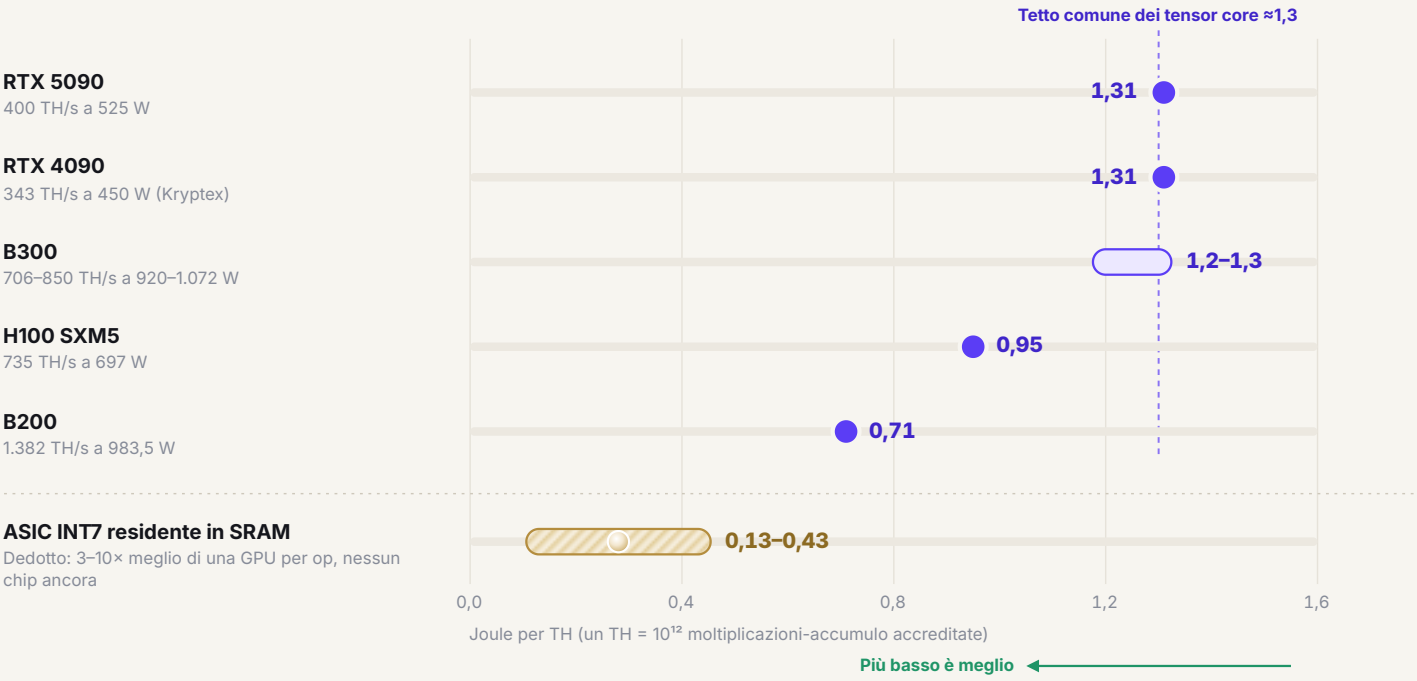
Una moltiplicazione-accumulo è un "hash"

Per un progettista di chip il fatto più importante è l'unità. Il codice di consenso di Pearl scala il target di ogni tentativo per la dimensione del tile moltiplicata per la lunghezza del prodotto scalare, quindi **1 H è esattamente una moltiplicazione-accumulo (MAC) INT7 × INT7 → INT32 accreditata**. Dal blocco 96.251 una penalità di rango richiede un rango del rumore di almeno 128; al rango minimo, 1 H equivale ancora a 1 MAC. Gli "EH/s" di Pearl significano quindi letteralmente 10^18 moltiplicazioni-accumulo intere al secondo: **per la prima volta il budget di sicurezza di una blockchain è prezzato nel throughput tensoriale degli acceleratori AI.** Gli hashrate misurati sulle GPU si collocano tra il 60% e il 95% del picco di moltiplicazione-accumulo INT8 denso di ciascuna scheda, il che è coerente con questa lettura.

Hardware	Hashrate su Pearl	Potenza	Joule per mille miliardi di MAC (J/TH)	Fonte (data)
RTX 5090	400 TH/s	525 W	1,31 stima	Kryptex (12 novembre)
RTX 4090	circa 305 TH/s (intervallo della comunità 280–330); Kryptex indica 343 TH/s	450 W (Kryptex)	1,31 (dati Kryptex) stima	prlstats (3 ottobre); Kryptex
H100 SXM5	735 TH/s	697 W	0,95 stima	PeakMiner, via prlstats (19 settembre)
B200	1,382 PH/s	983,5 W	0,71 stima	PeakMiner, via prlstats
B300	706–850 TH/s	920–1.072 W	1,2–1,3 stima	prlstats; SimeonOnSecurity

FIG. 02 **Sul lavoro intero di Pearl le GPU hanno un tetto energetico comune**

Joule per mille miliardi di moltiplicazioni-accumulo INT7 (J/TH) all'hashrate misurato su Pearl; più basso è meglio



Fonte: Kryptex (12 novembre 2026); benchmark PeakMiner e SRBMiner via prlstats; SimeonOnSecurity; l'intervallo dell'ASIC è una deduzione dell'autore, non una misura.

Due dettagli di questa tabella contano per i progettisti di chip. Primo, le schede consumer RTX 4090 e 5090 e la B300 da data center convergono a circa 1,3 J/TH, e anche la scheda migliore, la B200, arriva solo a circa 0,71 J/TH. Una GPU da data center che costa decine di migliaia di dollari spende per ogni moltiplicazione-accumulo intera quasi la stessa energia di una scheda da gaming, il che indica un tetto comune dell'energia dei tensor core per MAC INT8, e quel tetto è proprio ciò che un chip dedicato potrebbe infrangere. Secondo, la B300 è più lenta della B200: NVIDIA ha tagliato drasticamente il throughput INT8 in Blackwell Ultra, quindi su quel chip il lavoro intero deve seguire un percorso più lento. Il silicio AI di punta si sta allontanando da INT8 verso FP8 e FP4. Quando il formato numerico scelto da un consenso diverge dalla direzione dell'hardware, il protocollo è costretto a cambiare le regole, ed è questa la radice del fork FP8 e del mio giudizio sugli ASIC più avanti in questo rapporto.

CAPITOLO 03

03 PEARL · L'ECONOMIA
Pearl: scala, emissione e divario di utilità

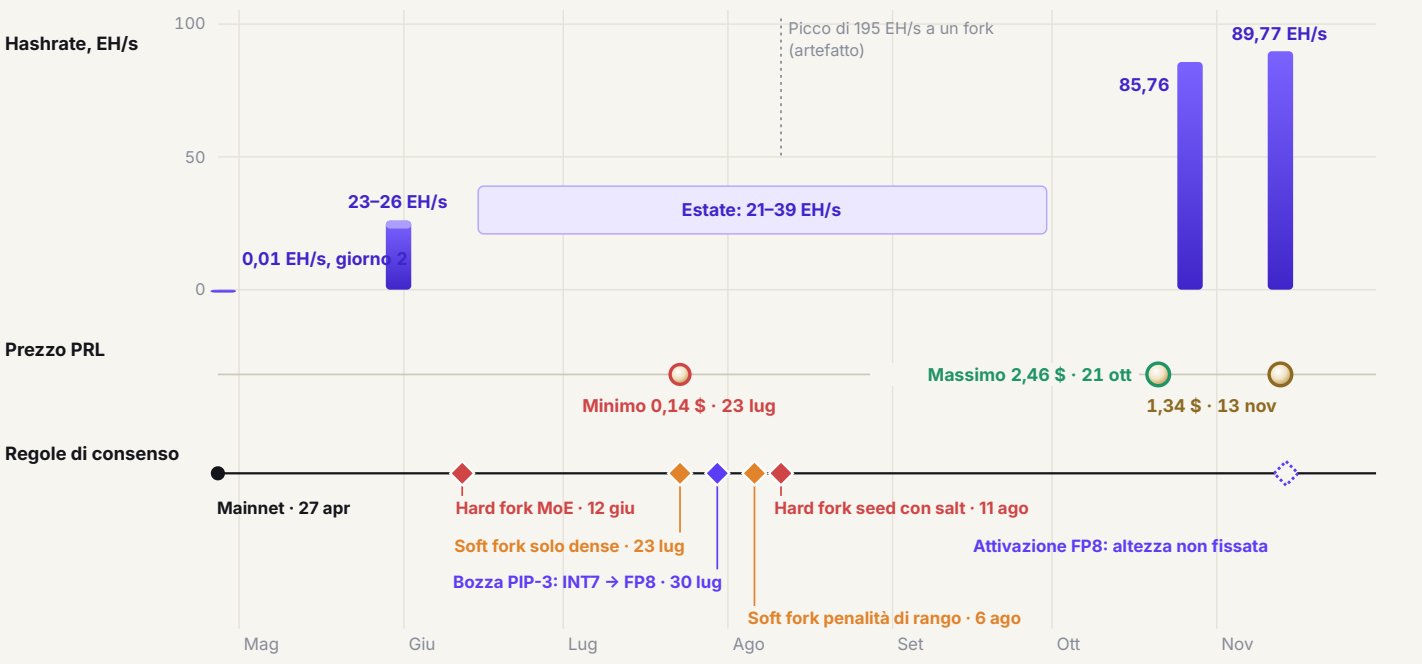
Da 0,01 a 90 EH/s in sei mesi

Metrica	Valore (13 novembre 2026)	Base e fonte
Hashrate di rete	Circa 90 EH/s (riferimento 89,77; intervallo nella stessa giornata 88,20–93,22; massimo intraday del 12 novembre 110–123)	Nodo completo indipendente prlstats; Kryptex
Difficoltà e altezza del blocco	58.841.982 e 142.530	prlstats
Hardware equivalente	Circa 122.000 H100 o 225.000 RTX 5090	A 735 e 400 TH/s per scheda stima
Prezzo, valore di mercato e valore completamente diluito	1,32–1,34 dollari; circa 500 milioni di dollari; circa 2,8 miliardi di dollari (101° posto su CoinGecko)	CoinGecko
Massimo e minimo storici	2,46 dollari (21 ottobre) e 0,1399 dollari (23 luglio)	CoinGecko
Monete minate	377,6 milioni di PRL, il 17,98% del tetto di 2,1 miliardi	prlstats
Sussidio di blocco ed emissione giornaliera	Circa 2.173 PRL e circa 0,97–1,03 milioni di PRL	prlstats; Kryptex
Nuova offerta nei prossimi 12 mesi	Circa il 78% dell'offerta attuale	prlstats
Reddito dei miner	Circa 1,25 milioni di dollari al giorno (modello), circa 460 milioni di dollari l'anno	prlstats
Ricavo per unità di hashrate, R	Circa 14–15 dollari per PH/s al giorno; una RTX 5090 rende circa 5–6 dollari al giorno	stima; Kryptex
Blocchi con lavoro di AI reale	Circa l'1% (stima del 28 ottobre)	Blockworks
Concentrazione dei pool	Kryptex ha trovato il 46,4% dei blocchi nelle ultime 24 ore; i primi due pool circa il 62%	prlstats
Concentrazione dei detentori	Coefficiente di Gini degli indirizzi 0,969; i primi 100 indirizzi detengono il 53,0%	prlstats

L'hashrate di Pearl è quasi una funzione ritardata del suo prezzo. Il giorno dopo il lancio la rete girava a circa 0,01 EH/s; a fine maggio aveva raggiunto 23–26 EH/s. Durante l'estate il prezzo è sceso di circa l'86% dal massimo di maggio e l'hashrate si è mosso lateralmente tra 21 e 39 EH/s. Dopo che a settembre e ottobre il prezzo si è moltiplicato per circa dieci, l'hashrate ha raggiunto 85,76 EH/s il 27 ottobre e circa 90 EH/s a novembre. Un picco a 195 EH/s l'11 agosto è coinciso con un fork e va letto come un artefatto di stima. L'energia è cresciuta di pari passo: uno studio ha stimato circa 112 MW per l'intera rete a 24 EH/s, e un modello di prlstats la colloca oggi a circa 190–215 MW.

FIG. 03 **I primi sei mesi di Pearl: l'hashrate segue il prezzo, le regole cambiano**

Hashrate di rete, tre livelli di prezzo e ogni modifica del consenso, dal 27 aprile al 13 novembre 2026



Fonte: Nodo completo e serie di prlstats; Kryptex; CoinGecko; params.go e PIP-3 di Pearl. Il picco dell'11 agosto è un artefatto di stima durante un fork.

L'emissione amplifica la riflessività

Il tetto dell'offerta è di 2,1 miliardi di monete e non ci sono halving: il sussidio di blocco cala gradualmente secondo E(t) = S·H / ((t + H)(t + H − 1)) con H = 650.226. Ma l'emissione si conta in blocchi, non nel tempo, e la difficoltà iniziale è partita troppo bassa e si è adeguata lentamente. **Circa 44,06 milioni di monete sono state minate il primo giorno e circa 120 milioni nei primi sei giorni**, circa il 32% dell'offerta attuale stima. I blocchi recenti, a circa 171–183 secondi, arrivano ancora più in fretta del target di 194 secondi. Il risultato è che l'offerta crescerà di circa il 78% nei prossimi 12 mesi, una diluizione con cui miner, produttori di hardware e detentori devono tutti fare i conti.

Ne discende l'economia dei miner. Il reddito dei miner sull'intera rete è di circa 1,25 milioni di dollari al giorno (modello); l'emissione giornaliera di Kryptex per il prezzo attuale dà circa 1,35 milioni di dollari al giorno, quindi **il ricavo per unità di hashrate R ≈ 14–15 dollari per PH/s al giorno** stima. Una RTX 5090 rende circa 5–6 dollari al giorno. Una H100 o H200 rende circa 11 dollari, mentre il noleggio on-demand pubblico più economico costa circa 2,09 dollari per scheda all'ora, circa 50 dollari al giorno, quindi noleggiare schede solo per minare fa perdere molto. R oscilla ampiamente: secondo la serie di prlstats era di circa 10 dollari ad agosto e settembre e per breve tempo di circa 33 dollari a metà ottobre stima. A poche settimane dal lancio il ricavo giornaliero stimato di una RTX 5090 si è dimezzato, da circa 33,80 a circa 17,19 dollari. Noleggiare GPU in cloud per minare ha reso solo nelle prime settimane e durante l'impennata del prezzo di ottobre.

Il teorema regge; l'incentivo no

Il vero problema è la domanda, non la crittografia. L'unica integrazione commerciale pubblica di Pearl è un endpoint Gemma-4-31B-it-pearl di Together AI, lanciato il 15 maggio a più del 25% sotto il prezzo abituale, con lo sconto finanziato dal valore futuro atteso dell'emissione. Il 13 novembre la pagina del modello indicava che **non è disponibile sull'API serverless di Together**. A giugno un censimento di 8.012 istanze di mining online, circa il 21% della rete, ha concluso che la rete non produceva alcun calcolo di AI utile (l'autore definisce le evidenze indica-

tive, non conclusive), e ha mostrato che le share costruite da matrici puramente casuali venivano accettate dai pool. Il 28 ottobre Blockworks ha stimato che **solo circa l'1% dei blocchi contiene lavoro di AI reale**; senza di esso, ha scritto, Pearl è un bitcoin più piccolo, meno sicuro e con un'inflazione più rapida.

La ragione si può calcolare. L'overhead lato inferenza è davvero piccolo: Pearl dichiara un aumento del 5,08% del tempo end-to-end per Llama 70B su quattro H200, e un test indipendente ha misurato una perdita di throughput del 9,2% su una GPU Hopper. Il lato mining rende troppo poco. Il plug-in ufficiale per vLLM mina solo su batch di prefill di almeno 1.024 token e mai sui passi di decodifica. Una scheda che serve inferenza ottiene circa il 27% dell'hashrate di un miner dedicato con traffico di solo prefill, circa il 5,4% con un mix di chat all'80% di decodifica, e solo circa l'1% con traffico dominato dalla decodifica. Senza domanda pagata di inferenza, un miner razionale moltiplica matrici casuali; il protocollo verifica che le matrici impegnate siano state moltiplicate onestamente, non che fossero utili. **Il teorema regge; l'incentivo no.** L'utilità di Pearl è un problema commerciale fuori catena: qualcuno deve pagare l'inferenza.

CAPITOLO 04



04 PEARL · LE REGOLE
Pearl: rischio di regole, numerica e governance

Quattro fork in due mesi

L'altra faccia di Pearl è la velocità con cui cambiano le sue regole. Da giugno ad agosto 2026 il laboratorio ha introdotto quattro modifiche del consenso ad altezze fissate nel codice: un hard fork mixture-of-experts (blocco 71.935, intorno al 12 giugno), un soft fork solo dense (91.630, intorno al 23 luglio), un soft fork con penalità di rango (96.251, intorno al 6 agosto) e un hard fork del seed con salt (99.000, intorno all'11 agosto). Dopo il fork del seed con salt, i vecchi miner potevano produrre solo share non valide; la penalità di rango ha tagliato l'hashrate effettivo delle configurazioni a rango alto dal 50% all'87,5%.

Il passo successivo è PIP-3, ancora in bozza: sposta il mining da INT7 a FP8 × FP8 → FP32 e richiede profili di dispositivo esatti al bit, **supportando solo due architetture, Hopper e Blackwell**. Il documento definisce la dipendenza dall'hardware un male necessario dei protocolli in virgola mobile e avverte che qualsiasi tolleranza nella verifica invita ad attacchi di grinding. Il 12 novembre i parametri di mainnet del ramo fp8-scheme non fissavano alcuna altezza di attivazione. Con la bozza attuale la serie RTX 40, la A100 e le schede AMD non sono ammesse, mentre un modello a bassa affidabilità di prlstats stima che circa l'82% dell'hashrate venga da schede consumer. Forzare il fork rischia una divisione della catena o una rivolta dei miner.

Un trilemma del formato numerico

Pearl affronta un trilemma: un formato numerico può essere utile per l'AI (FP8 o FP4), verificabile in modo esatto al bit su hardware diversi e neutrale rispetto all'hardware, ma al massimo due di queste tre cose. Lo schema INT7 è verificabile e neutrale ma diverge dalla direzione del silicio AI. FP8 con una whitelist è utile e verificabile ma lega il consenso a un solo fornitore. FP8 con una tolleranza è utile e neutrale ma apre la porta agli attacchi di grinding. L'aritmetica in virgola mobile dei tensor core in genere non segue uno standard numerico e non è riproducibile tra generazioni; su Hopper, per esempio, l'accumulo FP8 conserva solo 13 bit frazionari e tronca. Dal punto di vista di un progettista di chip, un "profilo Hopper" è un'aritmetica completamente determinata che un ASIC potrebbe implementare bit per bit, forse in modo più efficiente della GPU. Ma farlo sarebbe un'attività in zona grigia, contraria all'intento dichiarato del protocollo, con un rischio di governance molto alto analisi.

Governance

La governance è concentrata. Il laboratorio ha scritto ogni proposta di protocollo; i fork si attivano ad altezze fissate nel codice senza segnalazione dei miner; non esiste una seconda implementazione del client; e a inizio ottobre non era stato pubblicato alcun audit di terze parti del nodo o dei circuiti a conoscenza zero. Il quotidiano israeliano TheMarker ha riferito che il laboratorio ha raccolto circa 120–130 milioni di dollari in un round riservato (circa 55 milioni di dollari di equity più un fondo per infrastruttura GPU di circa 75 milioni di dollari); Pearl non lo ha confermato attraverso canali ufficiali. “Nessuna allocazione del token agli investitori” e “una società con azionisti venture” possono essere vere entrambe, ma la tensione resta: i ricavi dell’AI possono andare all’equity mentre la diluizione ricade sui detentori del token.

Nel complesso, Pearl è finora l’esperimento più pulito di “calcolo come consenso”: un meccanismo nuovo, verificabile on-chain, con un’unità che corrisponde uno a uno all’hardware AI. Ma l’utilità non è arrivata, e il suo budget di sicurezza, l’hashrate e l’entusiasmo dei miner dipendono tutti da un token la cui domanda principale deve ancora formarsi e la cui offerta crescerà di circa il 78% in 12 mesi. Vale la pena seguire tre segnali indicati da Blockworks: se l’aggiornamento va a buon fine senza scosse, se la quota di blocchi con lavoro di AI cresce in modo sostanziale, e se un qualche fornitore di inferenza trasferisce ai clienti il reddito del mining.

CAPITOLO 05

CATENE A LAVORO UTILE

05 Catene a lavoro utile: né le matrici né la conoscenza zero fermano gli ASIC

Kaspa: una matrice non è né utile né a prova di ASIC

Kaspa è il caso da manuale. Il suo kHeavyHash inserisce tra due hash un prodotto matrice-vettore 64 × 64 su valori a 4 bit, e il prodotto viene scartato. Circa 16 mesi dopo il suo fair launch del novembre 2021, IceRiver ha annunciato i primi ASIC per Kaspa il 14 marzo 2023, e a fine 2023 il mining con GPU era finito. Confrontare una RTX 4090 a 3,8 GH/s e 300 W con il KS7 di Bitmain a 77 J/TH dà un divario di efficienza di circa **1.000 volte** stima. Dopo l’abbuffata è arrivata la compressione: il 13 novembre la rete girava a 358,9 PH/s, circa il 64% sotto l’1 EH/s dell’agosto 2024, e il miglior KS7 rende solo **0,86 dollari al giorno** a 0,10 dollari per kWh.

Aleo: nemmeno la conoscenza zero ferma l’hardware specializzato

Il puzzle per i prover di Aleo è progettato per imitare il lavoro di generazione di una prova SNARK; non prova le transazioni degli utenti, e il progetto accoglie apertamente gli ASIC ZK. Circa cinque mesi dopo la sua mainnet del settembre 2024 si consegnavano già macchine dedicate. Nel terzo trimestre del 2025 i prover hanno ricevuto 31,2 milioni di ALEO (circa 7,8 milioni di dollari), mentre la rete ha incassato solo circa 43.000–44.000 dollari di commissioni: **un multiplo del sussidio di circa 175–180**. Il prezzo è sceso da circa 9 dollari nelle prime contrattazioni a 0,22 dollari a fine di quel trimestre, e ARC-46 richiede ora uno stake da 100.000 a 2,5 milioni di ALEO per soluzione, trasformando il mining in una soglia di stake.

Qubic e Bittensor

Qubic promuove una “proof of work utile” che addestra la sua rete neurale Aigarth, ma il prodotto più visibile del suo calcolo nel 2025–2026 sono stati i blocchi di altre catene. Prima ha minato Monero, portando la sua quota da

meno del 2% a metà maggio 2025 a più del 25% a fine luglio e dichiarando il 51% ad agosto, il che ha causato una riorganizzazione di sei blocchi (un'altra fonte dice 18). Il 1° aprile 2026 è passata a minare Dogecoin con ASIC, usando i proventi per riacquisti e burn. Non c'è alcuna evidenza indipendente che il suo addestramento di AI sia utile, e il suo valore di mercato è sceso da circa 280–300 milioni di dollari a circa 50 milioni il 13 novembre. Ha anche messo in luce un'esternalità: una catena che paga calcolo generico crea un grande pool di hashrate coordinato e noleggiabile, che può essere rivolto contro le catene adatte alle CPU.

Bittensor non è proof of work ma un punteggio soggettivo pesato sullo stake: i validatori assegnano punteggi ai miner, e i pesi vengono tagliati alla mediana pesata sullo stake. È l'unica rete di questo gruppo con ricavi esterni su scala. L'indice SubConnect, che conta solo i clienti esterni verificabili, indica circa **70 milioni di dollari l'anno** a inizio novembre 2026 (circa 35 milioni il 1° settembre), dal 78% all'82% dei quali da calcolo e infrastruttura; ma ai prezzi di inizio novembre le sue emissioni superano i 300 milioni di dollari l'anno, quindi i ricavi ne coprono solo dal 9% al 23%. Le stime dei ricavi variano molto a seconda della base, da circa 17 milioni di dollari (Artemis) a 172 milioni annualizzati dai totali autodichiarati. Ad aprile 67 subnet su 128 bruciavano più del 99% dei loro incentivi, e quando il team di punta Covenant se n'è andato per questioni di governance, TAO è sceso del 24% in poche ore.

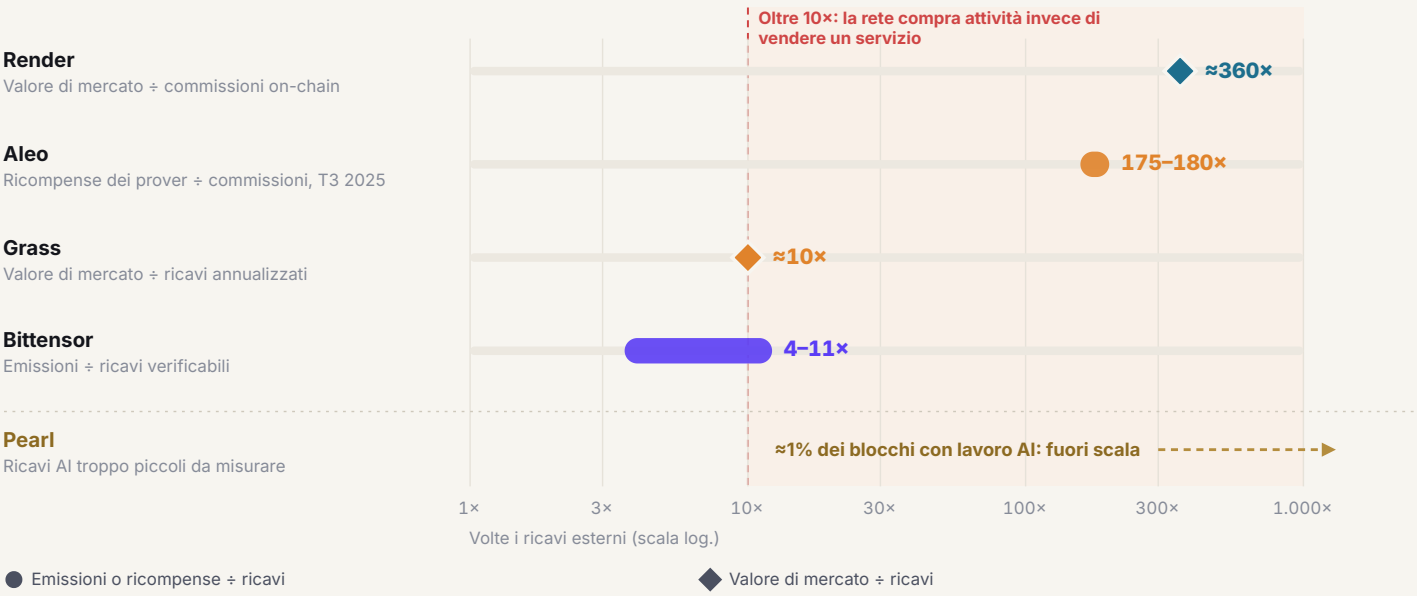
La storia dà la stessa risposta

Le catene di numeri primi di Primecoin erano rapide da verificare, ma nessuno ne aveva bisogno. Gridcoin si affidava a siti di statistiche di terze parti per attestare i contributi BOINC, e dei ricercatori hanno falsificato blocchi senza eseguire alcun calcolo. Le prove di archiviazione di Filecoin funzionano tecnicamente, ma nel terzo trimestre del 2025 l'utilizzo era solo del 36%, e il 99,5% dei suoi 793.000 dollari di commissioni erano penali, con commissioni base di appena 4.100 dollari.

Progetto	Lavoro o consenso	Valore di mercato (data)	Percorso hardware	Domanda pagata contro sussidio	Giudizio
Pearl	Moltiplicazione di matrici rumorose (MAC INT7 più certificato ZK)	Circa 500 milioni di dollari (13 novembre)	GPU; la bozza FP8 metterebbe in whitelist Hopper e Blackwell	Circa l'1% dei blocchi con lavoro di AI; reddito dei miner per lo più sussidio	Il meccanismo più credibile; domanda non dimostrata
Kaspa	kHeavyHash (un prodotto matrice-vettore 64 × 64 a 4 bit tra due hash)	Circa 840 milioni di dollari (13 novembre)	ASIC circa 16 mesi dopo il lancio; ora solo ASIC	Non dichiara alcuna utilità; ricavi dei miner circa 18 milioni di dollari l'anno stima	Una matrice non è né utile né a prova di ASIC
Aleo	Puzzle SNARK sintetico più proof of stake	Circa 118 milioni di dollari (fine T3 2025)	Macchine dedicate dopo circa cinque mesi	Ricompense dei prover circa 175–180 volte le commissioni	Lavoro sintetico, trainato dal sussidio
Qubic	"uPoW" (dichiara di addestrare Aigarth)	Circa 50 milioni di dollari (13 novembre)	CPU e GPU; ASIC per Dogecoin da aprile 2026	Nessuna evidenza indipendente di utilità; calcolo usato per mining esterno e riacquisti	Aggregazione di calcolo avvolta in una storia di AI
Bittensor	Consenso Yuma (punteggio pesato sullo stake)	Circa 2,55 miliardi di dollari (11 novembre)	GPU generiche; nessun percorso ASIC	Circa 70 milioni di dollari l'anno di ricavi verificabili contro più di 300 milioni di emissioni	L'unica con ricavi esterni su scala; comunque un sussidio di 4–11 volte
Filecoin (riferimento)	Prove di replica e di spazio-tempo	Circa 1,5 miliardi di dollari (fine T3 2025)	—	Utilizzo 36%; commissioni base trimestrali 4.100 dollari	Utile, ma a corto di domanda

FIG. 04 **Il multiplo del sussidio: quante volte i ricavi esterni paga una rete**

Emissioni o valore di mercato divisi per i ricavi pagati dall'esterno della rete, scala logaritmica



Fonte: Blockworks (Aleo T3 2025), DefiLlama (commissioni Render), CoinGecko, hoge.gg (Bittensor), Dealroom (Grass), Blockworks 0xResearch (Pearl); calcoli dell'autore.

La lezione per l'hardware: un lavoro fisso e regolare, pagato in un token liquido, ottiene gli ASIC entro circa 3–16 mesi, spesso da sfidanti prima che dagli operatori affermati. Le affermazioni di "utilità" hanno retto solo dove un cliente esterno paga per il risultato. E i principali motori dell'hardware rimasto inutilizzabile sono un sussidio alto rispetto ai ricavi e il controllo degli sviluppatori sulla funzione di lavoro.

CAPITOLO 06

IL PANORAMA PIÙ AMPIO

06 Il resto di AI × crypto: domanda reale ma piccola, e il valore che passa all'equity

Mercati decentralizzati del calcolo

Il calcolo decentralizzato ha clienti paganti, ma la domanda è piccola, concentrata e per lo più rivendita di inferenza su modelli aperti. Le commissioni verificabili on-chain sono piccole: DefiLlama indica circa 2,35 milioni di dollari di commissioni per Render nell'ultimo anno e circa 680.000 dollari per Livepeer, e l'intero settore DePIN ha guadagnato solo circa 72 milioni di dollari di ricavi on-chain nel 2025. I numeri autodichiarati sono più grandi ma definiti in modo vago: Aethir dichiara "più di 127,8 milioni di dollari" di ricavi per l'esercizio 2025 senza dire se si tratti di fatturato lordo o della quota del protocollo. Valori dei token e ricavi si sono scollati in entrambe le direzioni: il valore di mercato di Render è circa 360 volte le sue commissioni on-chain, mentre il valore di mercato di Aethir, circa 87,7 milioni di dollari, è inferiore ai suoi ricavi autodichiarati stima.

L'affermazione di costare dal 70% all'85% in meno regge solo contro i prezzi on-demand degli hyperscaler. Il 3 novembre il prezzo medio ponderato di una H100 su Akash era di circa **2,73 dollari per GPU-ora** (intervallo 2,04–3,16 dollari), sotto i circa 6,88 dollari di AWS e gli 11,06 dollari di Google Cloud, ma sopra i circa 1,73 dollari di Vast.ai, gli 1,99 dollari del community cloud di RunPod e i circa 2,44 dollari di una prenotazione annuale (prezzi delle altre piattaforme rilevati il 12 settembre; Akash, che li pubblica, è parte interessata). La domanda è per lo più inferenza standard distribuita tramite OpenRouter, circa 1,7 miliardi di token al giorno su AkashML e più di 4 miliardi su io.net, in concorrenza diretta con ogni neocloud. I dollari più consistenti del "calcolo decentralizzato" si stanno ricentralizzando: Axe Compute, che detiene il token di Aethir come asset di tesoreria, ha firmato un contratto quinquennale da 1,5 miliardi di dollari per noleggiare il proprio cluster di 9.200 B300, a fronte di ricavi nel primo trimestre 2026 di 35.000 dollari.

Addestramento e inferenza decentralizzati

L'addestramento via internet è stato dimostrato tra 8 e 40 miliardi di parametri. Pluralis-8B è stato addestrato per 40 giorni su 330 nodi, per lo più GPU consumer, e ha raggiunto il 63% dell'efficienza di un riferimento centralizzato su H100, anche se il suo articolo ammette che l'esecuzione non era decentralizzata e ha richiesto nodi di Pluralis stessa e un'autorità di ammissione. Psyche di Nous Research ha completato il pre-addestramento Consilience da 40 miliardi di parametri. Alla scala di frontiera il lavoro si è ricentralizzato: INTELLECT-3 di Prime Intellect, un modello mixture-of-experts da 106 miliardi di parametri, è stato addestrato per due mesi su 512 H200 collegate da InfiniBand a 400 Gbps. I vincoli sono chiari: i nodi hanno bisogno di almeno 200 Mbit/s di banda e di una compressione delle comunicazioni di circa 100 volte, e fiducia e verifica dipendono ancora dai ruoli privilegiati dell'operatore.

Il denaro è andato a società con equity che vendono prodotti centralizzati. Prime Intellect aveva più di 100 milioni di dollari di ricavi annualizzati prima di raccogliere 130 milioni di dollari a una valutazione di 1 miliardo a luglio; Venice, un'app di inferenza privata, dichiara più di 70 milioni di dollari di ricavi annualizzati e utili dal primo trimestre 2026 (dati della società), ha raccolto 65 milioni di dollari a una valutazione di 1 miliardo e gira su GPU in leasing. Sul lato dei token, la rete di addestramento principale di Gensyn è ancora "da definire" e il suo token vale circa

17,2 milioni di dollari; il token di Ritual, un layer-1 nativo AI, è sceso del 72% dal massimo dopo la mainnet (solo titolo).

AI verificabile

La produzione ha scelto gli ambienti di esecuzione affidabili (TEE). Il confidential computing di Hopper aggiunge meno del 7% di overhead alle tipiche query LLM e quasi nulla per i modelli grandi; l'overhead in produzione di Phala è di circa lo 0,5–5%, ma a fine 2025 pagavano 398 dei suoi 10.004 utenti. La radice di fiducia di un TEE è il produttore del chip: TEE.fail, reso noto nell'ottobre 2025, ha usato un interposer sul bus DDR5 dal costo inferiore a 1.000 dollari per estrarre le chiavi di attestazione da macchine Intel TDX e AMD SEV-SNP completamente aggiornate, il che a sua volta mina il confidential computing delle GPU NVIDIA, ed entrambi i produttori affermano che gli attacchi fisici sono fuori dal loro modello di minaccia. È esattamente lo scenario DePIN, in cui gli operatori dei nodi hanno accesso fisico. Lo zkML oggi può provare un LLM completo: DeepProve di Lagrange prova GPT-2 a circa 174 token al minuto e Gemma 3 a circa 86, ordini di grandezza lontano dall'inferenza di produzione e senza alcun uso pagato dichiarato.

Pagamenti tra agenti e dati

x402 è l'esempio ammonitore migliore. TRM Labs conta circa 198,9 milioni di transazioni e 52,7 milioni di dollari regolati da facilitatori noti dal maggio 2025, per il 99,6% in USDC; tolti gli autopagamenti, i flussi a lotti e i venditori con meno di dieci acquirenti, la quota avviata davvero da agenti è solo dallo 0,6% al 7,5%, **circa 5.000–11.000 dollari al mese**. Artemis giudica che più del 95% dell'attività sia segnalazione di protocollo e non compravendita. Nel frattempo i binari vengono standardizzati dagli operatori affermati: la x402 Foundation è stata costituita sotto la Linux Foundation il 14 luglio, con Visa, Mastercard, Stripe, Google, AWS, Cloudflare e Coinbase come membri principali, e Tempo, che ha scritto con Stripe il Machine Payments Protocol, ha lanciato la sua mainnet a marzo con una valutazione di 5 miliardi di dollari. Le piattaforme di token per agenti seguono ancora la logica dei launchpad: Virtuals autodichiara 479 milioni di dollari di "aGDP", e il fondatore di ElizaOS ha dichiarato morto il suo token dopo la transazione di una class action. È più probabile che il valore venga catturato dagli emittenti di stablecoin e dagli operatori affermati dei pagamenti; il ruolo distintivo della crypto qui potrebbe essere l'identità e la responsabilità degli agenti.

I dati sono uno dei pochi sottosettori con acquirenti reali fuori catena. Grass dichiara 17 milioni di dollari di ricavi dati nel 2025 e altri 17 milioni nel primo semestre 2026, punta a 75 milioni per l'anno e si dice in utile; GRASS vale circa 328 milioni di dollari, circa dieci volte i ricavi annualizzati del primo semestre stima. Il suo vantaggio competitivo è la scala di raccolta della banda residenziale distribuita, non la crypto in sé. DAO di dati e marketplace di dati non hanno una domanda pagata visibile: Vana vale circa 21,5 milioni di dollari e Ocean circa 27,6 milioni, e Ocean è in causa con Fetch.ai. La proof of personhood sta ottenendo integrazioni reali man mano che gli agenti si moltiplicano: World dice di aver verificato quasi 18 milioni di persone, e Zoom, Docusign e Tinder supportano parti di World ID.

Il capitale ha votato

Settore	Domanda pagata verificabile	Segnali di sussidio e di valutazione	Giudizio
Catene a lavoro utile	Bittensor circa 70 milioni di dollari l'anno di ricavi esterni verificabili; Pearl circa l'1% dei blocchi con lavoro di AI	Aleo 175–180 volte; Bittensor 4–11 volte; Pearl per lo più sussidio	I meccanismi hanno valore; l'utilità richiede clienti paganti
Mercati decentralizzati del calcolo	Commissioni on-chain circa 0,68–2,35 milioni di dollari l'anno (Livepeer, Render); autodichiarati 4–128 milioni	Valore di mercato di Render circa 360 volte le commissioni	Reale ma piccolo, per lo più rivendita di inferenza standard

Settore	Domanda pagata verificabile	Segnali di sussidio e di valutazione	Giudizio
Addestramento e inferenza decentralizzati	Società con equity: Prime Intellect oltre 100 milioni di dollari l'anno, Venice oltre 70 milioni (infrastruttura centralizzata)	Mainnet di addestramento di Gensyn non attiva; token di Ritual −72% dal massimo	Il progresso tecnico è reale; i profitti vanno all'equity
AI verificabile	TEE in produzione (398 utenti paganti su Phala); nessuno zkML pagato dichiarato	—	TEE utili; zkML agli inizi
Pagamenti tra agenti	Flussi avviati da agenti circa 5.000–11.000 dollari al mese	Circa 200 milioni di transazioni, per lo più segnalazione e speculazione	Prima i binari, la domanda dopo
Reti di dati	Grass 17 milioni di dollari nel primo semestre 2026 (autodichiarati)	GRASS circa 10 volte i ricavi annualizzati	Uno dei pochi settori con acquirenti reali

Messi fianco a fianco i sei settori, la conclusione è coerente: **il sussidio è carburante per l'avvio, non un modello di business.** Il 13 novembre la categoria AI di CoinGecko valeva in totale circa 18,6 miliardi di dollari, ma circa il 56% è costituito da layer-1 generici come NEAR e ICP, più progetti a livello di consenso come TAO, TAO in staking (forse contato due volte con TAO) e Pearl; i token AI dei livelli protocollo e applicazione arrivano in totale a circa 8 miliardi di dollari o meno stima. È più o meno la valutazione complessiva di quattro società con equity: Tempo (5 miliardi di dollari), Nous (1,5 miliardi secondo quanto riportato), Prime Intellect (1 miliardo) e Venice (1 miliardo). Il venture capital crypto è fermo a circa 4,5–5 miliardi di dollari a trimestre, con l'aumento che viene dall'infrastruttura legata all'AI, la cui quota sugli importi dichiarati è salita dall'8,1% al 18,2% nel terzo trimestre del 2026; e l'84,7% dei 118 token lanciati nel 2025 scambia sotto il prezzo di lancio. In una frase: **la crypto fa distribuzione e regolamento, i profitti vanno all'equity**, almeno nel 2026.

CAPITOLO 07

ECONOMIA DEGLI ASIC

07 La questione degli ASIC: una formula, tre metri di misura

Una formula: perché l'efficienza smette di pagare oltre k = 5

Sia R il ricavo giornaliero per unità di calcolo (reddito dei miner della rete diviso per l'hashrate della rete), c il costo giornaliero di un'unità di calcolo sull'hardware marginale (su Pearl, una GPU), e k quante volte un ASIC è più economico di quell'hardware marginale (c_ASIC = c/k). Finché le GPU restano i miner marginali, la libertà di entrata e di uscita spinge R verso c, quindi un ASIC guadagna R − c/k per unità di calcolo:

Tetto di profitto di un ASIC ≈ R × (1 − 1/k)

Sull'intera rete, il tetto dell'utile lordo annuo di una flotta di ASIC è all'incirca i ricavi dei miner della rete × (1 − 1/k), e solo finché gli ASIC non sostituiscono le GPU come miner marginali.

La formula ha tre conseguenze. Primo, (1 − 1/k) è concava: a k = 2 un ASIC trattiene il 50% dei ricavi, a k = 5 l'80%, a k = 10 il 90%, a k = 100 il 99%. **Oltre k = 5–10, più efficienza non aggiunge quasi alcun profitto**; decidono la grandezza di R e del piatto. Secondo, il riferimento si azzera: quando gli ASIC sono i miner marginali, k va misurato contro altri ASIC. Gli ASIC di Kaspa avevano un vantaggio di efficienza di circa 1.000 volte sulle GPU,

che ha comprato solo una breve manna; oggi il miglior KS7 rende 0,86 dollari al giorno. Terzo, su una catena adatta alle GPU come Pearl, il riferimento spesso non è il prezzo d'acquisto di una GPU ma **l'elettricità di una GPU inattiva**: una RTX 5090 di proprietà a 525 W e 0,10 dollari per kWh costa circa 1,26 dollari al giorno di energia e rende circa 5–6 dollari stima. Contro GPU che sono costi sommersi, anche un ASIC che usa un quinto dell'energia ha un k effettivo di appena 2,5 circa se il suo ammortamento è pari alla sua elettricità, e un tetto di profitto di circa il 60% di R (un'ipotesi illustrativa).

FIG. 05 **Perché l'efficienza smette di pagare: un ASIC tiene al massimo 1 − 1/k dei ricavi**

Quota dei ricavi che un ASIC può tenere rispetto al miner marginale, in base al suo vantaggio di costo k



Fonte: Modello dell'autore: quando l'hardware marginale fa il prezzo, i ricavi per unità di calcolo tendono al suo costo; dati di Kaspa e RTX 5090 dal rapporto.

Tre metri di misura: il piatto, la scala di rientro e le regole

Il primo metro di misura è **il piatto**. Annualizzando circa 2.893 BTC di ricompense di blocco nella prima settimana di novembre (circa 222 milioni di dollari), i miner di bitcoin guadagnano circa 11,5 miliardi di dollari l'anno stima. I miner di Pearl guadagnano circa 0,46–0,49 miliardi di dollari (ai prezzi attuali); le ricompense dei prover di Aleo, annualizzate dal terzo trimestre del 2025, circa 31 milioni di dollari; e i miner di Kaspa, con una ricompensa di blocco di 1,945 KAS e dieci blocchi al secondo, circa 18 milioni di dollari stima. **Un intero anno di ricavi dei miner di Kaspa vale meno della fascia bassa del budget di progetto di un singolo chip a 5/4 nm (circa 30 milioni di dollari).**

Il secondo metro di misura è **l'hashrate necessario per rientrare dell'NRE**:

h* = NRE ÷ [R × (1 − 1/k) × T × p]

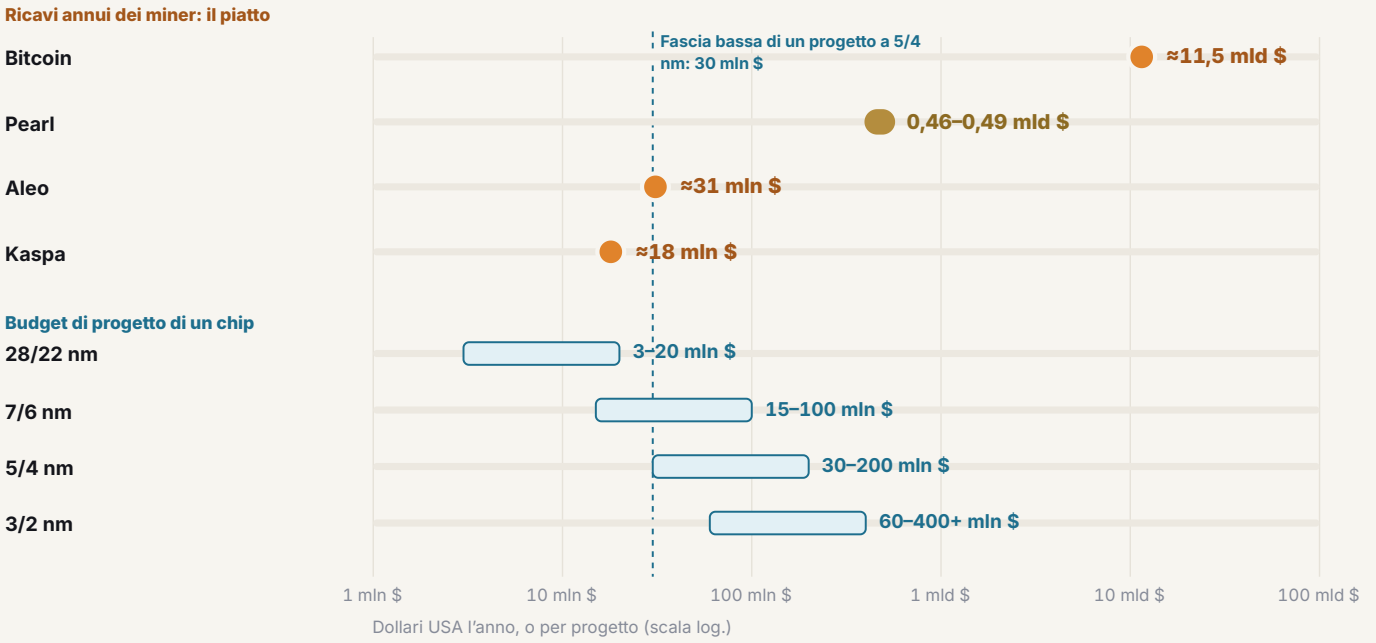
dove T è la finestra utile in giorni e p la probabilità che le regole restino invariate in quella finestra. Per Pearl, prendiamo R = 14 dollari per PH/s al giorno, k = 5, T = 365 giorni e un NRE di 20 milioni di dollari (un chip ad array ripetitivo N7/N6; nella fascia bassa degli intervalli di 5–30 milioni di dollari per i nodi di fascia media secondo Silicon Analysts e di 15–100 milioni di dollari per gli ASIC di calcolo a 7/6 nm secondo AnySilicon, perché un chip di mining usa poca IP e una struttura ripetuta; ipotesi mia). Con p = 1 servono circa **4,9 EH/s** in funzione per un anno intero, circa il 5% della rete; con p = 0,5, circa **9,8 EH/s**, circa l'11%; e se R torna al livello estivo di circa 10 dollari, anche con p = 1 servono circa 6,9 EH/s stima. La scala non è fuori portata. Il problema sono p e T.

Il terzo metro di misura è **il rischio di regole**, e spesso è decisivo. Il fork anti-ASIC di Monero del 2018 ha reso inutilizzabile l'X3 di Bitmain e per breve tempo è costato alla rete quasi metà del suo hashrate. Il fork di Sia dello stesso anno ha messo fuori uso i miner di Bitmain e di Innosilicon, mentre il prezzo dell'usato del miner Obelisk, legato agli sviluppatori, è salito da 2.499 a più di 5.000 dollari. Poco più di un mese dopo il Merge di Ethereum, meno del 20% degli ex miner di Ethereum aveva trovato nuovo lavoro. **Chi controlla la funzione di lavoro controlla il destino di chi investe in hardware**, e Sia mostra che la "resistenza agli ASIC" può diventare il fossato difensivo degli stessi sviluppatori. Sopravvive l'hardware con un secondo mercato: dopo il Merge le GPU sono passate all'AI e al cloud, mentre gli ASIC a singolo algoritmo si sono affollati su Ethereum Classic. Anche la finestra si sta restringendo: il primo ASIC di Kaspa ha richiesto circa 16 mesi, quello di Aleo circa cinque, e il primo a raggiungere la frontiera dell'efficienza è spesso uno sfidante piuttosto che un operatore affermato.

Rete	Ricavi annui dei miner (circa)	Hardware marginale o di riferimento	Stabilità delle regole	k disponibile per un nuovo entrante	Giudizio
Bitcoin	11,5 miliardi di dollari	Bitmain S23 XP Hyd (8,9 J/TH, 25,5 dollari per TH)	Molto alta	≤1	Un piatto grande, ma k appartiene agli operatori affermati
Pearl	0,46–0,49 miliardi di dollari (ai prezzi attuali)	GPU di proprietà o inattive (costo marginale ≈ elettricità) e GPU da data center	Molto bassa: quattro fork in due mesi; whitelist FP8 in sospeso	Forse 3–10 volte in efficienza analisi ; meno contro GPU a costo sommerso	Il piatto è abbastanza grande; il rischio di regole pone il veto
Aleo	Circa 31 milioni di dollari (T3 2025, annualizzati)	IceRiver AE2 e altre macchine dedicate	Media: soglia di stake ARC-46	≈1	Un piatto piccolo, e la soglia sale
Kaspa	Circa 18 milioni di dollari	KS7 (77 J/TH)	Alta	≈1	Ricavi annui sotto la fascia bassa del budget di un progetto a 5/4 nm

FIG. 06 **La maggior parte dei piatti è più piccola del budget di un solo chip**

Ricavi annui dei miner di quattro reti contro il costo di progettazione di un chip per nodo di processo, scala log.



Fonte: Hashrate Index (bitcoin, novembre 2026), prlstats e Kryptex (Pearl), Blockworks (Aleo, T3 2025), minerstat (Kaspa), AnySilicon (budget di progetto, 2026); annualizzazione dell'autore.

Il quadrante "piatto grande, regole stabili e k > 1" oggi è vuoto: bitcoin ha il piatto e le regole stabili, ma il suo k appartiene agli operatori affermati; il piatto di Pearl è abbastanza grande ma le sue regole sono le meno stabili; Kaspa e Aleo hanno regole abbastanza stabili e piatti troppo piccoli.

CAPITOLO 08

08 CHIP PER CHIP
Cinque tipi di chip, cinque risposte

PoUW stile Pearl: nessun chip INT7 dedicato; puntare su un miner che sappia servire inferenza

Gli argomenti a favore di un ASIC sono reali. Oggi solo circa l'1% dei blocchi contiene lavoro di AI, quindi l'argomento economico del whitepaper, secondo cui un ASIC dedicato perde la capacità di fare lavoro utile e quindi guadagna meno, è al suo punto più debole proprio adesso. Le parti più costose di una GPU, HBM e packaging CoWoS, valgono circa il 63% del costo di produzione di una H100, mentre il die logico vale solo circa il 9%. Un chip che non fa altro che moltiplicazioni-accumulo intere a 7 bit, tiene gli operandi nella SRAM on-chip e lavora a bassa tensione potrebbe plausibilmente raggiungere da 3 a 10 volte l'energia per operazione di una GPU sullo stesso nodo e un vantaggio maggiore nel costo per operazione [analisi; ancora nessuna misura]. Gli argomenti contrari pesano di più. I tempi di consegna delle fonderie sui nodi di punta possono superare le 50 settimane, e un team nuovo ha bisogno di circa 18–24 mesi dall'avvio ai volumi stima; quando un chip arriva, Pearl potrebbe benissimo essere passata a FP8 e poi a FP4. E poiché PIP-3 definisce l'aritmetica ammessa attraverso una whitelist di dispositivi, un chip INT7 smetterebbe semplicemente di funzionare una volta attivata.

La mia conclusione quindi è: **non fare ora il tape-out di un chip Pearl INT7 dedicato.** Ci sono due scommesse migliori. Una è un acceleratore pensato per l'inferenza che sappia anche minare: costruito attorno a GEMM INT8/FP8 generale e operandi residenti in SRAM, in modo che, se l'accesso al mining viene negato, il valore finisca in un mercato reale dell'inferenza. L'altra è entrare nel dibattito sul protocollo e spingere per un'aritmetica canonica neutrale rispetto ai fornitori, che un ASIC possa implementare in modo esatto, insieme a un impegno alla stabilità dell'algoritmo, che è la precondizione perché esista il quadrante "piatto grande, regole stabili, k > 1". Fino ad allora, un'architettura si può validare su FPGA o su wafer multiprogetto (MPW) a nodo maturo, mantenendo il rischio di NRE nell'ordine del milione di dollari. Il capitolo 10 traduce tutto questo in un budget a tappe.

Miner in stile Bitcoin: k appartiene agli operatori affermati, e l'energia è l'asset scarso

Il rischio di regole di Bitcoin è il più basso, ma la concorrenza spinge il k di un nuovo entrante sotto 1. L'S23 XP Hyd di Bitmain raggiunge 8,9 J/TH, si vende a 25,5 dollari per TH e viene consegnato da novembre; l'efficienza dei modelli di punta è scesa da 17,5 J/TH dell'S21 nel 2023 a 8,9 J/TH, circa il 49% in tre anni. Un nuovo entrante che consegnasse volumi nel 2027–2028 avrebbe bisogno di circa 8 J/TH o meglio per reggere stima. Gli sfidanti se ne vanno: Canaan ha riportato ricavi nel secondo trimestre 2026 di 31,9 milioni di dollari con una perdita lorda di 29,3 milioni e ha indicato solo 11–15 milioni per il terzo trimestre; le vendite a terzi delle macchine proprie di Bitdeer nel terzo trimestre sono scese a 0,3 milioni di dollari, con un margine lordo di gruppo del 3,4%; Auradine ha cambiato nome in Velaura AI e si è rivolta ai chip AI; e Core Scientific ha annullato un ordine di circa 15 EH/s di macchine a 3 nm dall'unità Proto di Block, registrando una perdita di 41,9 milioni di dollari.

Anche la domanda si restringe. Il 9 novembre l'hashprice era di circa 31,89 dollari per PH/s al giorno, la curva forward fino ad aprile 2027 era in media di circa 32,7 dollari, e le commissioni erano solo lo 0,72% delle ricompense di blocco. A 0,045 dollari per kWh, l'S23 XP Hyd rientra in circa 3,1 anni stima, vicino al prossimo halving previsto intorno al 2028. Più importante è il costo opportunità dell'energia: i contratti di leasing per AI e calcolo ad alte prestazioni rendono oggi circa 1,67–1,9 milioni di dollari per megawatt contrattato all'anno, che con un reddito operativo netto dall'80% all'85% è circa da 1,5 a 2,9 volte l'utile lordo del mining per megawatt (circa 0,56–0,9 milioni di dollari) [stima; le due basi in megawatt sono diverse]. Core Scientific smetterà di minare nel quarto trimestre; Bitfarms, ora Keel, ha chiuso tutti i suoi siti di mining negli USA; IREN in un trimestre ha guadagnato 200 milioni di dollari dal cloud AI e solo 28 milioni dal mining. **Energia e capacità di data center sono gli asset; le macchine da mining sono materiale di consumo.**

I conti dell'NRE sono altrettanto implacabili. L'efficienza di frontiera richiede oggi processi della classe 3 nm (secondo fonti secondarie MicroBT usa Samsung SF3, e i chip di Auradine e di Proto sono a 3 nm). Le maschere costano circa 10–20 milioni di dollari e l'NRE completo 30–100 milioni o più; a 20–25 dollari per TH e con un margine lordo del 30%, ogni 1 EH/s venduto rende circa 6–7,5 milioni di dollari di utile lordo, quindi coprire una spesa annua di R&S della scala di Bitdeer, circa 110–160 milioni di dollari, richiede di vendere a terzi circa 20–25 EH/s l'anno, circa il 2% della rete stima. La conclusione: **un team nuovo non dovrebbe entrare nel mercato dei miner bitcoin venduti a terzi.** Tre casi hanno ancora senso: l'integrazione verticale con energia propria a basso costo; formati di nicchia come i miner domestici e per il riscaldamento (anche Canaan lancia prodotti per il riscaldamento domestico); e partecipare alle roadmap degli operatori affermati tramite IP e servizi di progettazione.

Chip per l'addestramento AI: non da team piccolo

Per i chip di addestramento k è il più difficile da guadagnare. A parità di generazione e di precisione, Ironwood di Google supera la B300 in efficienza FP8 a livello di chip solo di circa 1,1–1,8 volte (la potenza di Ironwood è una stima di un giornalista) stima; le 30–80 volte spesso citate dall'articolo sulla TPU v1 erano misurate contro una K80, che non aveva tensor core. La vera leva è il formato numerico: passare da FP8 a FP4 dà di solito da 2 a 3 volte il throughput sullo stesso silicio, e lì NVIDIA si muove più in fretta di tutti: dichiara 50 PF di inferenza NVFP4 per Rubin, in volumi nel secondo semestre 2026. Le soglie di costo e di approvvigionamento sono ancora più alte:

i budget di progetto a 3/2 nm vanno da 60 a 400 milioni di dollari o più; la capacità CoWoS nel 2026 è di circa 106.000–130.000 wafer al mese, e si riporta che TSMC stia valutando aumenti di prezzo fino al 20%; a un prezzo contrattuale di 15,6 dollari per GB per la HBM3E, la sola HBM in un acceleratore da 288 GB costa circa 4.500 dollari stima. Questo mercato appartiene al co-sviluppo con gli hyperscaler in stile Broadcom (capitolo 11). La conclusione: **i team piccoli non dovrebbero costruire chip di addestramento autonomi**; meglio costruire IP, chiplet o servizi di progettazione.

Chip per l'inferenza AI: scambiare generalità con k, e ripaga

L'inferenza è il campo in cui gli ASIC hanno più spazio, perché k si può comprare rinunciando alla generalità. Etched dichiara che il suo server a otto chip esegue Llama-70B a più di 500.000 token al secondo, circa 20 volte otto H100; Taalas cabla i pesi del modello in una mask ROM e dichiara una potenza 10 volte inferiore e un costo di produzione 20 volte inferiore (entrambi dichiarato dal fornitore). Le architetture incentrate sulla SRAM evitano HBM e CoWoS, e i loro nodi sono più gestibili: Taalas usa TSMC 6 nm, e Groq è passata da 14 nm a Samsung 4 nm. La distribuzione crypto non aiuta i chip di inferenza, perché le reti di calcolo decentralizzato girano quasi interamente su NVIDIA; le vie d'ingresso sono l'economia degli acquisti dei neocloud e il grande mercato interno cinese dell'inferenza, dove i data center finanziati dallo Stato devono usare chip AI nazionali, le importazioni di H200 vengono strozzate e Pechino vuole che le grandi aziende di internet facciano girare l'inferenza su chip nazionali. La conclusione: **vale la pena costruirli**, purché ci sia un cliente chiaro e il grado di vincolo a modelli e architettura sia una scelta deliberata. È anche il miglior secondo mercato per un chip di mining PoUW.

Acceleratori ZK e FHE: coprirsi sul ZK, posizionarsi presto sull'FHE

Le GPU hanno già raggiunto la prova ZK in tempo reale: SP1 Hypercube di Succinct ha provato il 99,7% dei blocchi di Ethereum in meno di 12 secondi su 16 RTX 5090. Cysic dichiara che il suo chip ZK C1 è da 10 a 100 volte più efficiente dei prover su GPU e che sarà consegnato nel 2026, ma la sua domanda viene in gran parte dalle ricompense del suo stesso token dichiarato dal fornitore; non ho trovato evidenze pubbliche che gli ASIC per il puzzle di Aleo siano usati per vere prove ZK. La prova ZK è intera e deterministica, simile nella forma a ciò che fanno i chip di mining, ma i sistemi di prova cambiano in fretta (SP1, RISC Zero, Halo2, Plonky2 e altri). La conclusione: **attendere e coprirsi**, con motori MSM, NTT e di hash programmabili, prima su FPGA, poi come ASIC.

L'FHE è la primitiva più vincolata all'hardware e con il maggior potenziale. Heracles di Intel, su Intel 3 con 197 mm², 176 W e 48 GB di HBM3, esegue operazioni FHE da 1.074 a 5.547 volte più velocemente di uno Xeon a 24 core, ma Intel non ha piani commerciali e il responsabile del progetto se n'è andato nel dicembre 2025. Niobium sta sviluppando su Samsung 8LPU con SemiFive come partner di progettazione (un contratto di sviluppo di circa 6,86 milioni di dollari) e ha raccolto più di 23 milioni di dollari; Optalysys, con un approccio fotonico, ha raccolto 31 milioni di dollari e dice che un chip commerciale completo è ancora lontano due o tre anni; e il progetto FPGA open source di Zama raggiunge circa 13.000 bootstrap programmabili al secondo a circa 200 W, il riferimento pubblico che qualsiasi ASIC deve battere nettamente. La conclusione: **posizionarsi presto e validare con asset leggeri**, per una finestra intorno al 2027–2029. È più probabile che la domanda si ancori all'AI cifrata per imprese e governi che alle criptovalute stesse, e l'FHE richiede HBM, quindi la sua struttura dei costi somiglia a quella di un chip di addestramento.

Tipo di chip	Da dove viene k	Domanda e piatto	NRE e soglie	Rischio di regole o di vincolo	Secondo mercato	Giudizio
Chip di mining PoUW stile Pearl	Eliminare HBM, CoWoS e unità generiche: forse 3–10 volte in J/op analisi	Reddito dei miner circa 0,46–0,49 miliardi di dollari l'anno, volatile	Array ripetitivo N7/N6 circa 10–30 milioni di dollari analisi	Molto alto: il protocollo decide il formato numerico e la whitelist dei dispositivi	Inferenza AI, se costruito come GEMM INT8/FP8 generale	Nessun chip INT7 dedicato; costruire pensando all'inferenza e al mining, o attendere un'aritmetica canonica
Chip di mining Bitcoin SHA-256	k ≤ 1 contro Bitmain	Circa 11,5 miliardi di dollari l'anno, mercato delle vendite a terzi in mano a due aziende	Maschere classe 3 nm 10–20 milioni di dollari; NRE completo 30–100 milioni o più	Basso (algoritmo stabile); alto rischio di concorrenza e di halving	Quasi nessuno (singolo algoritmo)	I nuovi entranti stanno fuori dal mercato delle vendite a terzi; solo energia propria, formati di nicchia o IP
Chip per l'addestramento AI	Array sistolici della stessa generazione solo 1,1–1,8 volte; la leva è il formato	Il più grande (Broadcom circa 58 miliardi di dollari di ricavi da semiconduttori AI nell'esercizio 2026)	Progetto a 3/2 nm 60–400 milioni di dollari o più; CoWoS e HBM	La cadenza annuale di NVIDIA	—	I team piccoli costruiscono IP, chiplet e servizi di progettazione, non chip autonomi
Chip per l'inferenza AI	Rinunciare alla generalità: 10–20 volte dichiarati dichiarato dal fornitore	Grande e frammentato; il mercato interno della Cina continentale è grande di per sé	Architetture SRAM a N6/N7–N4 gestibili	Vincolo a modelli e architettura	È il mercato principale	Vale la pena costruirlo: firmare il cliente, poi scegliere il vincolo
Acceleratore ZK	10–100 volte le GPU, dichiarati dichiarato dal fornitore	Piccolo, trainato soprattutto dalle ricompense in token	N7/N6 poi N5; prima FPGA	I sistemi di prova cambiano in fretta	Limitato fuori dal ZK	Attendere e coprirsi: motori MSM e NTT programmabili
Acceleratore FHE	1.074–5.547 volte uno Xeon (Heracles)	Non ancora formato; ancorato a imprese e governi	Richiede HBM; struttura dei costi vicina ai chip di addestramento	Standard ed ecosistema non consolidati	Calcolo cifrato, AI privata	Posizionarsi presto, validare con asset leggeri; finestra 2027–2029

FIG. 07 **Cinque tipi di chip, cinque risposte**

Da dove viene il vantaggio di costo k, su quale mercato può vendere il silicio se le regole cambiano, e il verdetto



Chip di mining PoUW stile Pearl

NON ORA

k: niente HBM né unità generali: 3–10× per op (dedotto)

Secondo mercato: inferenza AI, se GEMM INT8/FP8 generale

Puntare su un chip per l'inferenza che sa minare, non su un chip INT7 dedicato



Chip di mining Bitcoin SHA-256

STARE FUORI

k: ≤1 contro gli operatori affermati

Secondo mercato: quasi nessuno

Solo con energia propria, un formato di nicchia, o come IP



Chip per l'addestramento AI

IP, NON CHIP

k: 1,1–1,8× alla stessa precisione

Secondo mercato: —

Decidono packaging e HBM; costruire IP, chiplet, servizi



Chip per l'inferenza AI

COSTRUIRE

k: rinunciare alla generalità: 10–20× dichiarati

Secondo mercato: è il mercato principale

Firmare il cliente, poi scegliere il vincolo



Acceleratore di prove ZK

ATTENDERE E COPRIRSI

k: 10–100× le GPU, dichiarati

Secondo mercato: limitato fuori dal ZK

Motori MSM e NTT programmabili; prima FPGA



Acceleratore FHE

POSIZIONARSI PRESTO

k: 1.074–5.547× uno Xeon (Heracles)

Secondo mercato: calcolo cifrato, AI privata

Una finestra 2027–2029; validare con asset leggeri

Fonte: Sintesi dell'autore dei capitoli 7 e 8; dati dei fornitori etichettati come nel rapporto.

CAPITOLO 09



NODI DI PROCESSO

09 Nodi di processo: lasciare che scelga il costo più grande sull'intera vita

La scala dei costi: i wafer all'avanguardia costano dieci volte quelli maturi

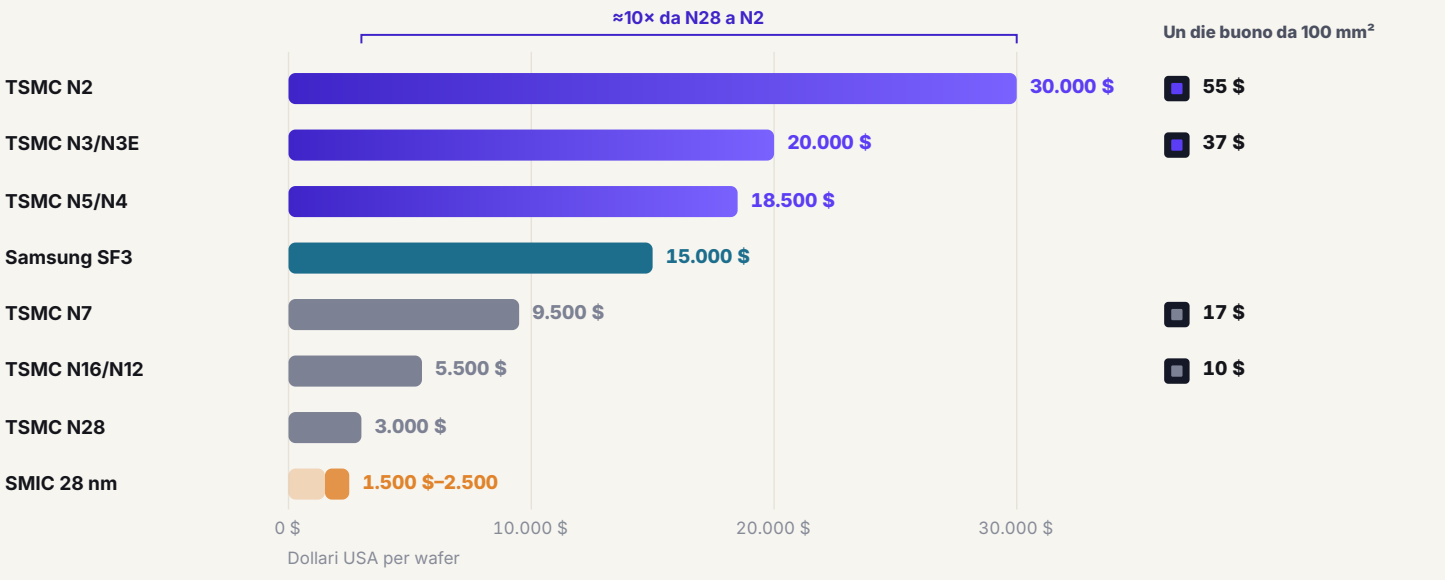
Partiamo dal registro dei conti. Un wafer all'avanguardia costa oggi circa dieci volte uno maturo: TSMC N2 circa 30.000 dollari, N3 circa 20.000, N7 circa 9.500 e N28 circa 3.000; Samsung SF3 circa 15.000, circa il 25% sotto N3; e si riporta che il costo di produzione della classe 7 nm di SMIC sia dal 40% al 50% superiore a quello di TSMC. Si riporta che TSMC aumenterà i prezzi di un ulteriore 3–6% da gennaio 2027, soprattutto per i 2 nm e i 3 nm. I costi delle maschere e di progettazione differiscono ancora di più, e le fonti non concordano, quindi la tabella li riporta entrambi.

Nodo	Prezzo del wafer (stima)	Set di maschere (valore di modello / intervallo indicativo)	Budget di progetto (AnySilicon)	MPW (per progetto)
TSMC N2	Circa 30.000 dollari	—	3/2 nm: 60–400 milioni di dollari o più	—

Nodo	Prezzo del wafer (stima)	Set di maschere (valore di modello / intervallo indicativo)	Budget di progetto (AnySilicon)	MPW (per progetto)
TSMC N3/N3E	Circa 20.000 dollari (18.000–27.000)	15 milioni di dollari / 10–20 milioni	Come sopra	—
Samsung SF3	Circa 15.000 dollari (13.000–17.000)	12 milioni di dollari	—	—
TSMC N5/N4	Circa 18.500 dollari (16.000–21.300, bassa affidabilità)	6,5 milioni di dollari / 10–20 milioni	5/4 nm: 30–200 milioni di dollari	—
TSMC N7	Circa 9.500 dollari (8.100–10.900)	3,5 milioni di dollari / 5–10 milioni	7/6 nm: 15–100 milioni di dollari	Commerciale 100.000–500.000 dollari
TSMC N16/N12	Circa 5.500 dollari	1,8 milioni di dollari	16/12 nm: 6–40 milioni di dollari	Accademico circa 35.000 dollari per mm²
TSMC N28	Circa 3.000 dollari	1 milione di dollari / 1–3 milioni	28/22 nm: 3–20 milioni di dollari	Commerciale 50.000–150.000 dollari; accademico circa 12.000 dollari per mm²
SMIC 28 nm	1.500–2.500 dollari	—	—	—
SMIC classe 7 nm (N+2/N+3)	40–50% sopra TSMC	—	—	—

FIG. 08 **Un wafer all'avanguardia costa dieci volte uno maturo**

Prezzo stimato per wafer da 300 mm per nodo, e costo di un die buono da 100 mm² con resa dell'85%



Fonte: Prezzi dei wafer di Silicon Analysts (stime, N5 bassa affidabilità), Wccftech (SMIC); i costi dei die sono calcoli dell'autore con una resa ipotizzata.

Poi la fisica. Le misure classiche di Horowitz mostrano che un'istruzione programmabile costa circa 70 pJ mentre l'aritmetica in sé costa pochi picojoule: **l'aritmetica è economica e l'overhead è caro**, ed è il motivo per cui esistono gli array sistolici e i chip dedicati. Dimezzare la dimensione caratteristica dimezza grosso modo l'energia per operazione, ma la tensione di alimentazione si è fermata intorno a 0,9–1,1 V. Un wafer N2 costa dal 50% al

67% più di uno N3, quindi se la densità non cresce di più di circa 1,5 volte il costo per transistor sale invece di scendere stima: i nodi di punta devono ripagarsi con efficienza e prestazioni. Per un die da 100 mm² con resa dell'85%, il costo di ogni die buono è di circa 55 dollari a N2, 37 a N3, 29–34 a N5, 17 a N7 e 10 a N16 [stima; la resa è un'ipotesi].

Scegliere un nodo per tipo di chip e di team

La mia regola è: **lasciare che sia la voce più grande nel costo del chip sull'intera vita a scegliere il nodo.** Per i chip di calcolo che girano a pieno carico giorno e notte (mining, ZK) l'elettricità domina il costo sull'intera vita, quindi una volta stabile l'algoritmo dovrebbero passare a N5/N4 e N3/SF3; finché l'algoritmo non è consolidato o i soldi scarseggiano, conviene validare prima su MPW a nodo maturo. Per i chip di addestramento il collo di bottiglia sono packaging e memoria (HBM e CoWoS valgono circa il 63% della distinta base di una H100), quindi il nodo è secondario. Per i chip di inferenza la prima decisione è l'architettura della memoria: i progetti incentrati sulla SRAM possono restare tra N6/N7 e N4, mentre i progetti con HBM ereditano la struttura dei costi dei chip di addestramento. L'FHE è vincolato dalla banda di memoria e non può fare a meno dell'HBM.

Tipo di chip	Start-up o team piccolo	Azienda già in scala	Percorso Cina continentale	Punti chiave
Chip di mining PoUW a moltiplicazione di matrici (residente in SRAM, senza HBM)	Prototipo su MPW a 28 nm o 16/12 nm; primi volumi su N7/N6; N5/N4 una volta stabili algoritmo e ricavi; die da 50–150 mm² in package flip-chip standard	N3/N3E o Samsung SF3	SMIC N+1/N+2 è praticabile (non servono HBM né CoWoS), con circa il 40–50% in più per die buono	Domina l'energia; le regole del protocollo sono il rischio maggiore
Chip di mining SHA-256	Non inseguire la frontiera; i formati di nicchia possono usare nodi maturi	Classe 3 nm; gli operatori affermati passano da TSMC 5 nm a 2 nm e da Samsung SF3 a 2 nm (fonti secondarie)	Difficile eguagliare l'efficienza dei nodi di punta	Il prezzo dell'energia e il ciclo degli halving decidono i margini
Chip per l'addestramento AI	Nessun chip autonomo: IP, chiplet, servizi di progettazione	N3P poi N2/A16; 2–8 chiplet; assicurarsi presto CoWoS e HBM	SMIC N+2/N+3 con HBM nazionale (il modello Huawei); FP8 per chip circa 0,2–0,5 volte quello di NVIDIA, compensato dalla scala del sistema	Packaging e HBM sono il collo di bottiglia
Chip per l'inferenza AI	Incentrato sulla SRAM o cablato, N6/N7 o N4/N5/SF4; solo INT8, FP8 e FP4	N3/N2 con HBM (contesti lunghi, mixture of experts), o SRAM multi-chiplet	SMIC N+2 con LPDDR o GDDR; la memoria di classe HBM è il punto di strozzatura	Il grado di vincolo architettturale è la decisione centrale
Acceleratore ZK	Partire su FPGA o MPW a 16/12 nm; poi N7/N6 e N5	N5/N3	SMIC N+1/N+2 è praticabile	Intero e deterministico; facile da verificare
Acceleratore FHE	FPGA (contro il riferimento open source) o MPW a 16/12 nm; i volumi richiedono HBM	N5/N3 con HBM	Limitato dall'accesso all'HBM	È la banda di memoria, non l'aritmetica, a fissare la velocità

Avanzato o maturo, quattro tipi di fonderia e chiplet

Opzione	Punti di forza principali	Punti deboli principali	Ideale per
Nodi di punta (N3/N2, SF3/SF2, Intel 18A)	Ogni nodo completo dimezza grosso modo l'energia per operazione; SRAM densa	Wafer 3–10 volte, maschere 5–15 volte e budget di progetto 10–20 volte quelli dei nodi maturi; tempi di consegna oltre le 50 settimane; il costo per transistor di N2 potrebbe salire	Chip di calcolo con algoritmi stabili che girano giorno e notte; chip di addestramento
Nodi maturi e di fascia media (N7/N6, N16/12, N28)	Maschere nell'ordine del milione di dollari, MPW economici, iterazioni rapide, molte fonderie (TSMC, UMC, GlobalFoundries, SMIC)	Efficienza non competitiva a pieno carico giorno e notte	Prototipi, mercati di nicchia, inferenza edge
TSMC	La resa e l'ecosistema migliori	I prezzi più alti; capacità N2/N3 scarsa; lunghe code per CoWoS	Team che inseguono la frontiera dell'efficienza con impegni di volume
Samsung Foundry	Wafer circa il 20–25% più economici (SF3 circa 15.000 dollari contro circa 20.000 per N3); usata dalla seconda generazione di Groq e da altri	Il business della fonderia è in perdita e punta al pareggio entro fine 2027	Chip di calcolo sensibili al costo
Intel Foundry	Chip complessi come Heracles dimostrati su Intel 3	Nessun prezzo pubblico per clienti esterni né dati sui clienti per 18A; il business della fonderia è in perdita	Valutare ecosistema e supporto caso per caso
Fonderie della Cina continentale (SMIC e altre)	Vicinanza a un grande mercato interno; wafer a 28 nm a 1.500–2.500 dollari	Capacità della classe 7 nm circa 45.000–60.000 wafer al mese, in gran parte assorbita da Huawei; rese circa un terzo di quelle di TSMC; costi più alti del 40–50%; solo DUV	Chip di inferenza, di mining e ZK per il mercato interno
Chiplet	Superano i limiti di reticolo e di resa; standard per i sistemi con HBM (MI355X, Ironwood, Trainium3)	Costo di packaging più alto e dipendenza da CoWoS; il test dei die noti come buoni aggiunge costi	Chip di addestramento e grandi chip di inferenza; non necessari per piccoli chip di mining o ZK

CAPITOLO 10

UN BUDGET A TAPPE

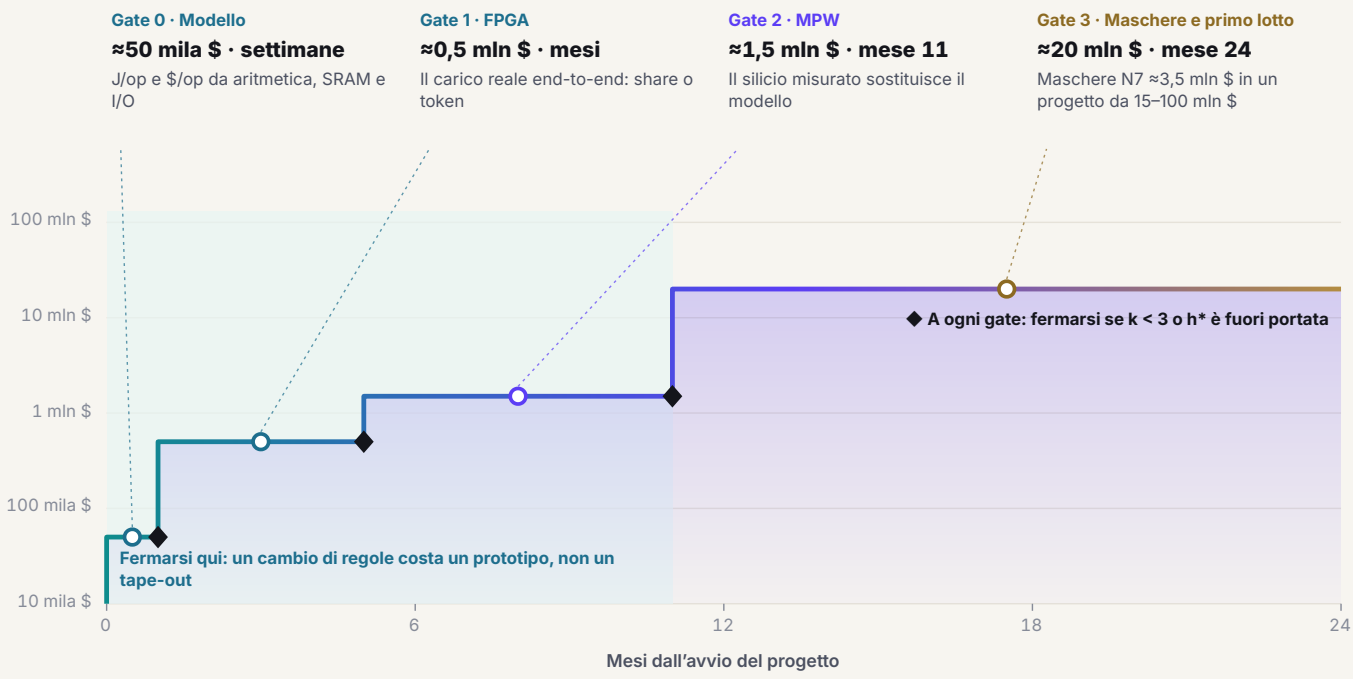
10 Dal modello al silicio: un budget a tappe per un primo chip

Comprare prove prima delle maschere

Un chip full-custom è una serie di scommesse, ognuna più grande della precedente, e la disciplina sta nel comprare prove prima di comprare maschere. Uso quattro gate analisi. Il Gate 0 è un modello: uno studio di architettura che stima J/op e $/op a partire da aritmetica, SRAM e I/O, contro l'hardware di riferimento del capitolo 7; costa settimane e decine di migliaia di dollari. Il Gate 1 è un prototipo su FPGA che esegue il carico reale end-to-end, share di mining o token di inferenza, e dimostra il datapath e l'integrazione con il protocollo; costa qualche mese e qualche centinaio di migliaia di dollari. Il Gate 2 è un wafer multiprogetto: uno slot commerciale costa circa 50.000–150.000 dollari a N28 e 100.000–500.000 dollari a N7, e il silicio misurato sostituisce le stime energetiche del modello. Il Gate 3 è il set completo di maschere e il primo lotto di produzione: maschere da circa 1 milione di dollari a N28, 3,5 milioni a N7 e 15 milioni a N3, dentro un budget di progetto di 15–100 milioni di dollari a 7/6 nm. Ogni gate ha bisogno di un numero che fermi il progetto, come un k sotto 3 rispetto all'hardware marginale, o un hashrate di rientro h* che la finestra non può raggiungere.

FIG. 09 **Prove prima delle maschere: un budget a tappe per un primo chip**

Spesa cumulata in dollari USA di un progetto N7/N6 attraverso quattro gate, scala log.; ogni gate ha un numero che ferma il progetto



Fonte: Prezzi di wafer, MPW e maschere come nel capitolo 9 (Silicon Analysts, AnySilicon, valori di modello); tempi e spese dei gate sono ipotesi illustrative dell'autore.

La scala è ripida. In un progetto N7/N6 i primi tre gate costano insieme circa 1,5 milioni di dollari e circa undici mesi; il quarto costa più di dieci volte tanto stima. Ecco perché l'ordine conta più del nodo: se il protocollo cambia la sua aritmetica, o l'acquirente si tira indietro, un team che si è fermato al Gate 2 perde il prezzo di un prototipo, non quello di un tape-out.

Quanto costa ogni chip, e che cosa deve essere vero

Progetto	Gate 2 (prototipo)	Gate 3 (nodo e budget di progetto)	Un die buono da 100 mm²	Che cosa deve essere vero per rientrare
Acceleratore pensato per l'inferenza che sa minare	MPW a 16/12 nm o N7, 0,1–0,5 milioni di dollari	N6/N7, circa 15–30 milioni di dollari stima	Circa 17 dollari a N7, prima di packaging e memoria	Un cliente di inferenza firmato copre la maggior parte dell'NRE; il reddito del mining è un extra, non lo scenario di base
Miner Pearl INT7 dedicato (oggi sconsigliato)	MPW a 28 nm, 0,05–0,15 milioni di dollari	N7/N6, circa 20 milioni di dollari	Circa 17 dollari a N7	Circa 4,9 EH/s per un anno intero con p = 1, o 9,8 EH/s con p = 0,5 (capitolo 7)
Acceleratore ZK	FPGA, poi MPW a 16/12 nm	N7/N6, poi N5	Circa 10 dollari a N16, 17 a N7	I sistemi di prova convergono su kernel MSM e NTT che un progetto fisso può servire
Acceleratore FHE	FPGA contro il riferimento open source	N5/N3 con HBM, 30 milioni di dollari o più	Dominato da HBM e packaging	Un acquirente di AI cifrata su volumi, che ancora non esiste

Ne seguono due considerazioni commerciali. Primo, l'efficienza più economica è architetturale: tenere gli operandi in SRAM e fare a meno di HBM e CoWoS elimina la maggior parte del costo di una GPU (circa il 63% della distinta base di una H100) prima ancora di scegliere un nodo. Secondo, il tempo è un costo: un team nuovo ha bisogno di circa 18–24 mesi dall'avvio ai volumi e i tempi di consegna all'avanguardia superano le 50 settimane, quindi un protocollo che cambia la sua aritmetica ogni pochi mesi può rendere obsoleto un chip prima che il Gate 3 si ripaghi.

Tre regole per un team piccolo

Nel complesso, questi numeri danno tre regole analisi. Primo, **restare al Gate 1 o 2 finché non ci sono un acquirente ben identificato o una versione di protocollo congelata**; né il prezzo del token né l'hashrate sostituiscono l'uno o l'altra. Secondo, **comprare k prima con l'architettura e poi con il nodo**: un array residente in SRAM su N7 può battere un componente generico su N3 per un singolo carico di lavoro, a una frazione del costo delle maschere. Terzo, **mettere per iscritto l'hashrate o il volume di rientro, con p, prima del tape-out**, e prezzare i pagamenti a milestone rispetto a una versione di protocollo fissata, così che a sostenere il rischio di regole sia il cliente e non la società di progettazione.

CAPITOLO 11

11 OPPORTUNITÀ
Dove sono le opportunità, e che cosa farei

Chip custom: la domanda viene in larghissima parte dall'AI

Nel 2026 la domanda di chip custom viene in larghissima parte dall'AI. I ricavi da semiconduttori AI di Broadcom nel terzo trimestre del suo esercizio 2026 sono stati di 16,7 miliardi di dollari, in crescita del 221% su base annua, circa 58 miliardi per l'anno, per il 73% da XPU per sei clienti (da una trascrizione automatica della conference call sui risultati). I ricavi del secondo trimestre di Global Unichip sono cresciuti del 128% su base annua, con i ricavi da produzione chiavi in mano in crescita del 188% e un margine lordo del 21,5%. La ripartizione per nodo dei ricavi del primo trimestre di Alchip era del 21% a 3/2 nm, del 63% a 7/5 nm e del 16% a 12 nm e oltre. I nuovi ordini di VeriSilicon per il 2026 hanno raggiunto 8,24 miliardi di RMB (circa 1,16 miliardi di dollari) a fine aprile, il 91,37% dei quali per il calcolo AI. Il mercato ha tre livelli: in cima, co-sviluppi pluriennali da miliardi; nel mezzo, progetti da 7 nm a 2 nm con NRE da pochi milioni a oltre cento milioni di dollari (5–30 milioni di dollari ai nodi di fascia media, 30–100 milioni o più all'avanguardia); e una coda lunga di ASIC a nodo maturo con NRE da 0,5–5 milioni di dollari. I clienti crypto e del mining sono una nicchia piccola e ciclica, ma sono il "primo cliente" più rapido e meno vincolato a permessi che oggi paga per il calcolo.

Le capacità di un team di chip di mining che si trasferiscono meglio, il funzionamento a bassa tensione vicino alla soglia, gli array di calcolo massicciamente replicati, l'alimentazione e il raffreddamento di classe kilowatt e i tape-out rapidi sui nodi di punta, sono esattamente ciò di cui hanno bisogno i chip di inferenza e gli acceleratori ZK e FHE. I mercati dei capitali lo prezzano: Velaura AI (ex Auradine) ha raccolto 110 milioni di dollari ad agosto a una valutazione superiore a 1 miliardo. Il controesempio è l'unità di chip per l'edge AI di Canaan, che ha avuto ricavi di 0,9 milioni di dollari nell'esercizio 2024 consumando il 15% delle spese operative, ed è stata chiusa nel giugno 2025. La differenza sta in un canale di vendita indipendente, non nel riutilizzo dei canali del mining.

Sviluppo AI: le ore sono una commodity; il valore sta nelle giunture

Nella fascia bassa i servizi di sviluppo AI sono una commodity: le tariffe tipiche su Clutch sono di 25–49 dollari l'ora, e di 50–99 negli Stati Uniti, in Canada e in mercati simili; un team di sviluppo blockchain costa da circa 15.000 a più di 90.000 dollari al mese. Vendere ore non dà alcun fossato difensivo. Il valore sta nelle giunture tra i campi: kernel e compilatori per chip custom, mining incorporato nei servizi di inferenza (il plug-in di Pearl per vLLM è un esempio), inferenza verificabile che combina TEE e controlli a campione ZK, e pagamenti e identità per gli agenti.

Sistemi AI × crypto: firmare l'acquirente di calcolo prima del token

Per chi progetta una catena nativa AI o una rete di calcolo, la cassetta degli attrezzi della verifica nel 2026 ha preso forma. In ordine di costo per unità di lavoro, dal più basso al più alto: campionamento statistico (la proof of sampling di Hyperbolic), delega con arbitro (Verde di Gensyn, che richiede un solo verificatore onesto ma un'esecuzione esatta al bit), TEE (velocità quasi nativa, ma fiducia nel produttore del chip), ridondanza completa e zkML generale. Il progetto pratico è a livelli: TEE o campionamento per la bassa latenza, giochi di contestazione o controlli a campione ZK per la finalità; un articolo del dicembre 2025 ha simulato il confidential computing della H100 con prove di frode e controlli a campione ZK casuali al 99% del throughput centralizzato e a un costo marginale

di 0,07 dollari per query. La filosofia dell'hardware decide sicurezza e distribuzione: una PoUW nativa per GPU è protetta all'inizio da GPU cloud noleggiabili, quindi il costo di un attacco è legato ai prezzi di noleggio, e Qubic mostra che una proof of work su hardware generico può essere catturata in blocco da incentivi esterni.

Costi e tempi non sono la barriera. Pearl è passata dall'articolo (aprile 2025) alla mainnet in circa 12 mesi, con un fork di btcd, un nodo completo, un wallet, un light client e un miner già al lancio. La maggior parte degli audit DeFi costa 25.000–100.000 dollari, gli audit enterprise o multi-chain 150.000–250.000 dollari o più, i circuiti ZK dall'80% al 120% in più, e i progetti con valore reale a rischio mettono a budget 150.000–500.000 dollari l'anno per la sicurezza. Le vere barriere sono distribuzione e liquidità. Alcuni fondatori riferiscono che Binance ha chiesto dallo 0% al 3,5%, e presumibilmente l'8%, dell'offerta del token per una quotazione (Binance nega di trarre profitto dalle quotazioni); tutti e sei i progetti esaminati da Cointelegraph che avevano usato accordi di market making "prestito di token più opzione call" sono scesi, e il market maker di Movement ha venduto circa 38 milioni di dollari di token il giorno della quotazione. Le alternative sono la liquidità on-chain, come nel caso di Hyperliquid, che ha superato i 10 miliardi di dollari di valore di mercato senza quotazione su Binance e i cui mercati perpetui HIP-3 hanno superato a giugno 3,2 miliardi di dollari di open interest, e le vendite pubbliche attraverso grandi exchange, come quando Monad ha raccolto 269 milioni di dollari tramite Coinbase. Un valore completamente diluito alto con un flottante basso è stato nel 2025–2026 il predittore più chiaro di un calo dopo il lancio.

La progettazione del token fa parte della progettazione del sistema. L'emissione dovrebbe seguire la domanda pagata, non anticiparla: le reti di questo rapporto con i multipli del sussidio più sani legano le ricompense a un lavoro che un acquirente esterno paga, e quelle con i peggiori pagano hashrate o stake senza nulla dietro. Una regola pratica è pubblicare, prima del lancio, il multiplo del sussidio a cui la rete punta nel primo, secondo e terzo anno, e tagliare l'emissione quando non viene raggiunto analisi.

Direzione	Chi compra, e disponibilità a pagare	Finestra	Capacità chiave	Rischio principale
Acceleratore pensato per l'inferenza che sa minare	Neocloud, implementazioni private, reti PoUW	2027–2028 (in funzione dei fork FP8 e FP4)	GEMM INT8, FP8 e FP4; architettura SRAM; integrazione con il protocollo	Ammissione al protocollo, vincolo architetturale
Chip di inferenza SRAM o cablato	Neocloud, data center nella Cina continentale, edge	Ora	Compilatori e adattamento dei modelli; progettazione N6–N4	I modelli cambiano troppo in fretta
Chip di mining (di nicchia o integrato verticalmente)	Miner con energia propria; miner domestici e per il riscaldamento	Ora, entro i limiti di halving e hashprice	Array a bassa tensione, efficienza, catena di fornitura per volumi	Guerre di prezzo degli operatori affermati, prezzi dell'energia
Accelerazione delle prove ZK	Reti di prover, rollup	Attendere che i sistemi di prova convergano	MSM e NTT programmabili	La domanda dipende dalle ricompense in token
Accelerazione FHE	AI cifrata per imprese e governi; catene per la privacy	2027–2029	NTT, sistemi con HBM, ingegneria crittografica	Domanda non ancora formata; cicli di vendita lunghi
Rafforzamento dei TEE e inferenza verificabile	Reti di inferenza riservata, DePIN	Ora	Cifratura della memoria con integrità, radici di attestazione indipendenti, controlli a campione ZK	Dipendenza dagli ecosistemi dei fornitori
Catena nativa AI o rete di calcolo	Acquirenti di inferenza, miner, sviluppatori	Ora	Progettazione della verifica, progettazione dell'emissione, distribuzione	Calo dopo il lancio, liquidità del token

Tre percorsi che prenderei, ognuno con un primo passo

Se dovessi scegliere, prenderei al massimo tre percorsi, ognuno con un primo passo piccolo che lo metta alla prova prima di impegnare denaro analisi.

- **Percorso A: un acceleratore pensato per l'inferenza che sa anche minare.** Primo passo: uno studio di architettura di quattro settimane che modelli J/op e $/op di un array GEMM INT8/FP8 residente in SRAM contro RTX 5090 e H100, sia sul lavoro di Pearl sia sull'inferenza dei transformer, seguito da una validazione su FPGA o MPW a nodo maturo con un NRE nell'ordine del milione di dollari.
- **Percorso B: servizi di silicio custom con la crypto come primo cliente e l'AI come secondo mercato.** Primo passo: trasformare il modello di NRE e di pareggio di questo rapporto in uno studio di fattibilità a pagamento per progetti di chip di calcolo, con pagamenti a milestone e collaudo rispetto a una versione di protocollo fissata.
- **Percorso C: una rete nativa AI che firma il suo acquirente di calcolo prima del token.** Primo passo: una nota di progetto sulla verifica a livelli (TEE più campionamento più controlli a campione ZK) e su un'aritmetica canonica neutrale rispetto ai fornitori, scritta con un acquirente di inferenza ben identificato prima di qualsiasi progettazione del token.

CAPITOLO 12

12 SEGNALI DA SEGUIRE
Che cosa seguire, da novembre 2026 al 2027

Quando	Che cosa seguire	Perché conta
In qualsiasi momento	L'altezza di attivazione di FP8 su Pearl	Decide se l'hardware intero e le schede non NVIDIA restano ammessi, e se la catena si divide
Ogni mese	La quota di blocchi di Pearl con lavoro di AI	Il miglior singolo test per capire se l'"utile" sta diventando vero
Ogni mese	Se i fornitori di inferenza trasferiscono ai clienti il reddito del mining	La domanda pagata che incontra l'emissione: il modello di cui Pearl ha bisogno
Ogni trimestre	I ricavi verificabili di Bittensor rispetto alle sue emissioni	Se la più grande rete a lavoro utile sta colmando il suo divario di sussidio
Ogni trimestre	La quota di flussi x402 avviati davvero da agenti	Se i pagamenti tra agenti passano dalla segnalazione al commercio
Gennaio 2027	L'aumento di prezzo del 3–6% riportato per TSMC, soprattutto per i 2 nm e i 3 nm	Il costo dei tape-out sui nodi di punta
Primo trimestre del 2027	FP4 in Pearl, come riportato dall'AMA dei fondatori	Un ulteriore spostamento del lavoro verso i formati più recenti di NVIDIA
Fine 2027	L'obiettivo di pareggio di Samsung Foundry	Se una seconda fonte più economica resta in salute
Aprile 2028	Il prossimo halving di Bitcoin	Dimezza il piatto dei miner SHA-256

CAPITOLO 13

13 DOMANDE DA PORRE
Le domande che non avete fatto ma avreste dovuto fare

Le domande che seguono sono quelle che porrei prima di impegnare denaro o silicio in qualsiasi progetto AI × crypto.

Chi controlla la funzione di lavoro? Se un piccolo team può cambiare l'aritmetica a un'altezza fissata nel codice, la vita del vostro hardware è una sua decisione. Chiedete un impegno scritto alla stabilità, oppure progettate hardware con un secondo mercato.

Qual è il multiplo del sussidio? Dividete le emissioni (o il valore di mercato) per i ricavi pagati dall'esterno della rete. Un multiplo superiore a 10 significa che la rete compra attività, non vende un servizio.

Quanto costa il prossimo gate, e quale numero vi ferma? Un wafer multiprogetto costa decine o centinaia di migliaia di dollari; un set di maschere costa milioni. Decidete il numero di arresto prima di spendere.

Qual è l'hashrate di rientro, con p? Dividete l'NRE per $R \times (1 - 1/k) \times T \times p$. Se il risultato è una quota grande della rete, il chip è una scommessa sulle regole, non sul silicio.

Qual è il secondo mercato del chip? Se il protocollo fa un fork o il token crolla, su quale altro mercato può vendere il silicio? Le GPU sono sopravvissute al Merge di Ethereum perché ne avevano uno; gli ASIC a singolo algoritmo no.

Il valore finisce nel token o nell'equity? Nel 2026 le aziende AI × crypto in utile sono per lo più società con equity e distribuzione crypto; decidete quale delle due state costruendo prima di raccogliere capitali.

Chi è il primo cliente pagante, con nome e cognome? Ogni caso duraturo di questo rapporto, da Grass alle subnet di calcolo di Bittensor, è cominciato con un acquirente ben identificato esterno alla rete.

Sono ambiti che vorrei esplorare insieme ad altri, con scambi e collaborazioni tecniche: chip custom per il mining, per l'addestramento e l'inferenza AI e acceleratori; sviluppo AI; e sistemi AI × crypto dalla catena al nodo fino al chip. Potete scrivermi a i@tanshuai.com.

Limiti di questo rapporto

Questa è una ricerca a tavolino basata su documenti pubblici fino al 13 novembre 2026. Per questa edizione non ho svolto interviste dirette né misure indipendenti. Diversi numeri chiave sono dichiarazioni di aziende o fornitori, come i ricavi di Aethir, i ricavi di Grass, le prestazioni di Etched e Taalas e l'efficienza di Cysic; sono etichettati come tali. Le metriche di catena di Pearl vengono da un nodo completo indipendente e dai pool e cambiano di ora in ora; ho usato le letture intorno alle 07:13 UTC del 13 novembre e riporto gli intervalli della stessa giornata. I costi di wafer, maschere e progettazione sono stime di terze parti e valori di modello che non concordano tra loro; ri-

porto degli intervalli. I miei calcoli di pareggio degli ASIC e il budget a tappe usano ipotesi illustrative (k, T, p, NRE e costi dei gate) che il lettore dovrebbe sostituire con le proprie.

L'autore

Tan Shuai è un tecnologo e fondatore con più di vent'anni di lavoro tra chip, hardware e software di sistema, e 20 brevetti concessi. Ha alle spalle molti anni di ricerca e sviluppo e di gestione di chip fabless, compresa la progettazione full-custom di datapath ASIC per chip di mining di criptovalute e per chip multimediali per fotocamere, che è la lente di questo rapporto: energia per operazione, costo per wafer, che cosa un protocollo consente di fissare nel silicio, e chi paga. Il suo lavoro lo ha portato nella Silicon Valley, a New York, a Chicago, a Taiwan, nella Greater Bay Area (Shenzhen e Hong Kong), in Vietnam e in Malesia. Oggi lavora come ricercatore e sviluppatore di AI.

Discussione tecnica: i@tanshuai.com · tanshuai.com

Note e avvertenze

Questo rapporto è una ricerca a tavolino indipendente basata su documenti pubblici disponibili fino al 13 novembre 2026. È fornito solo a scopo informativo e di studio e non costituisce consulenza in materia di investimenti o di ingegneria. Considera solo tecnologia e business. Prezzi, metriche di catena e regole dei protocolli cambiano; verificate sulla fonte primaria tutto ciò su cui fate affidamento. I dati segnati dichiarato dal fornitore, stima o analisi vanno letti come tali. I prezzi dei cripto-asset sono molto volatili.

© 2026 Tan Shuai. Potete condividere questo rapporto senza modifiche, citandone l'autore.

Fonti

Tutte consultate tra il 10 e il 13 novembre 2026. Raggruppate per capitolo.

Pearl: protocollo, codice e dati della catena

- Komargodski e Weinstein, "Proofs of Useful Work from Arbitrary Matrix Multiplication", arXiv 2504.09971: https://arxiv.org/abs/2504.09971
- Pearl Research Labs, whitepaper INT: https://pearlresearch.ai/research/int-whitepaper
- Codice sorgente di Pearl: blocco genesi: https://github.com/pearl-research-labs/pearl/blob/main/node/chaincfg/genesis.go ; parametri di consenso e altezze dei fork: https://github.com/pearl-research-labs/pearl/blob/main/node/chaincfg/params.go ; contabilizzazione dell'hashrate: https://github.com/pearl-research-labs/pearl/blob/main/zk-pow/src/api/sanity_checks.rs ; guida all'aggiornamento per il fork del seed con salt: https://github.com/pearl-research-labs/pearl/blob/main/docs/salted-seed-fork-upgrade-guide.md ; parametri del ramo fp8-scheme: https://github.com/pearl-research-labs/pearl/blob/fp8-scheme/node/chaincfg/params.go
- Pearl Improvement Proposals: https://github.com/pearl-research-labs/pips ; PIP-3, mining in virgola mobile: https://github.com/pearl-research-labs/pips/blob/main/PIPS/pip-0003.md
- prlstats (nodo completo indipendente): rete: https://prlstats.com/api/network ; serie: https://prlstats.com/api/series ; home page e modello del reddito dei miner: https://prlstats.com/api/home ; offerta minata: https://prlstats.com/api/mined ; pool: https://prlstats.com/api/pools ; detentori: https://prlstats.com/api/holders ; ipotesi su hardware e overhead: https://prlstats.com/api/assumptions
- Kryptex Pool: rete 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 ; storico dei prezzi: https://api.coingecko.com/api/v3/coins/pearl-2/market_chart?vs_currency=usd&days=365&interval=daily ; ricerca del ticker "Perle": https://api.coingecko.com/api/v3/search?query=perle
- SimeonOnSecurity, "Mining Pearl (PRL) in 2026": https://simeononsecurity.com/articles/pearl-prl-mining-guide-2026/
- Hashrate Index, "Pearl (PRL): Inside the AI-Compute Cryptocurrency", 2 giugno 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, partnership con Pearl Research Labs: https://together.ai/blog/together-ai-partners-with-pearl-research-labs ; pagina del modello: https://together.ai/models/gemma-4-31b-it-pearl
- Basu, censimento delle istanze di mining di Pearl, arXiv 2606.04819: https://arxiv.org/html/2606.04819v2
- Blockworks, newsletter 0xResearch, 28 ottobre 2026: https://blockworks.com/newsletter/0xresearch/issue/post_daf0d74a-2bf4-41e4-86d3-2a1275475a24
- Datawallet, "Pearl Explained": https://www.datawallet.com/crypto/pearl-explained
- Coingabbar sull'articolo di TheMarker relativo al finanziamento di Pearl Research Labs: https://www.coingabbar.com/en/pearl-research-labs-funding-130m-stealth-round-themarker

Catene a lavoro utile e storia

- Grothe et al., analisi di sicurezza di Gridcoin, USENIX WOOT 2017: https://www.usenix.org/system/files/conference/woot17/woot17-paper-grothe.pdf
- Falsificazione della proof of learning, arXiv 2208.03567: https://arxiv.org/abs/2208.03567
- hoge.gg, verifica e fiducia nell'inferenza decentralizzata: https://hoge.gg/decentralized-inference-2026-verification-trust/ ; ricavi e tesoreria di Bittensor: https://hoge.gg/bittensor-quantum-wallet-revenue-tao-treasury-2026/
- Kaspa: codice della matrice di kHeavyHash: https://github.com/kaspanet/rusty-kaspa/blob/master/consensus/pow/src/matrix.rs ; lancio: https://coinranking.com/coin/V8GxkwWow+kaspa-kas/about ; annuncio di IceRiver KS1/KS2: https://www.iceriver.io/2023/03/14/1677644580/ ; storia dal mining con GPU agli ASIC: https://captainmining.com/blogs/crypto-news/the-evolution-of-kaspa-mining-from-gpus-to-asic-powerhouses ; hashrate delle GPU: https://coincalculators.io/algorithms/kheavyhash ; KS7: https://www.asicminervalue.com/miners/bitmain/antminer-ks7-40th ; rete: https://minerstat.com/coin/kas ; 1 EH/s nel 2024: https://ourcryptotalk.com/news/kaspa-reaches-1-ehs-in-hashrate ; Crescendo: https://ourcryptotalk.com/news/kaspa-crescendo-live
- Aleo: FAQ sulla mainnet: https://aleo.org/post/aleo-mainnet-faq/ ; prova decentralizzata: https://aleo.org/post/decentralized-proving-advantages/ ; ASIC: https://whattomine.com/coins/432-aleo-zksnark/asics ; Blockworks, "State of Aleo Q3 2025": https://app.blockworks.com/report/state-of-aleo-q3-2025 ; ARC-46: https://chaincatcher.com/en/article/2144326
- Qubic: The Block sulla riorganizzazione di 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 ; mining di Dogecoin: https://hackernoon.com/qubic-is-now-mining-dogecoin-on-its-ai-compute-infrastructure ; CoinGecko: https://www.coingecko.com/en/coins/qubic

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

Settori AI × crypto

- DefiLlama: Render: https://defillama.com/protocol/render-network-bme ; Livepeer: https://defillama.com/protocol/livepeer
- Cointelegraph sui ricavi DePIN secondo Messari: https://cointelegraph.com/news/depin-revenues-grow-despite-token-losses
- Aethir, bilancio del 2025: https://aethir.com/blog-posts/aethirs-2025-wrap-up-decentralized-gpu-cloud-milestones
- CoinGecko, categoria AI: https://www.coingecko.com/en/categories/artificial-intelligence
- Akash, prezzi di noleggio delle H100: https://akash.network/the-bid/h100-rental-price-2026-cost-per-hour/ ; rapporto T1 2026: https://akash.network/blog/akash-network-q1-2026-report/
- io.net, accordo da 8 milioni di dollari e token giornalieri: 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, contratto di 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 su INTELLECT-3: https://www.implicator.ai/prime-intellects-intellect-3-open-source-ambition-meets-centralized-reality/
- SiliconANGLE, finanziamento di 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, finanziamento di Nous: https://runtimewire.com/article/nous-research-75-million-round-1-5-billion-valuation
- ownyourmind, Gensyn: https://ownyourmind.ai/projects/gensyn/ ; Phemex, Ritual (solo titolo): https://phemex.com/news/article/ritual-launches-ai-layer-1-mainnet-and-ritual-token-price-dives-72-from-high-101259
- Overhead del confidential computing, arXiv 2409.03992: https://arxiv.org/abs/2409.03992v2 ; rapporto 2025 di 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, misurare i pagamenti degli agenti AI: https://www.trmlabs.com/trm-tech-blog/whos-actually-paying-measuring-ai-agent-payments-onchain ; Forkast sulla 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 di Tempo: https://www.theblock.co/amp/post/394131/tempo-mainnet-goes-live-with-machine-payments-protocol-for-agents ; BlockEden sull'aGDP di Virtuals: https://blockeden.xyz/blog/2026/04/21/agdp-agent-gdp-virtuals-protocol-ai-blockchain-valuation-primitive-tvl-displacement/ ; The Block su ElizaOS: https://theblock.co/post/410774/eliza-labs-native-token-dead
- Dealroom, ricavi di Grass: https://dealroom.co/news/157794-grass-backed-by-multicoin-projects-75m-in-2026-data-revenue/ ; CryptoSlate, Fetch.ai e Ocean: https://cryptoslate.com/cryptos-flagship-ai-pact-fracture-fetch-sues-ocean-over-263m-fet-community-sales/ ; Fortune su World: https://fortune.com/2026/06/01/internet-according-to-sam-altman-seeing-orb/
- Galaxy, venture capital crypto: https://www.galaxy.com/insights/research/crypto-blockchain-venture-capital-q3 ; Crypto Briefing (solo titolo): https://cryptobriefing.com/crypto-vc-investment-q3-2026-galaxy/ ; Blockhead su Tiger Research: https://www.blockhead.co/2026/10/14/bitcoin-rose-43-in-q3-crypto-venture-funding-did-not-follow/ ; ForkLog sui lanci di token del 2025: https://forklog.com/en/85-of-tokens-launched-in-2025-fell-below-initial-prices/

Economia degli ASIC e tipi di chip

- Hashrate Index, rassegna del 9 novembre 2026: https://hashrateindex.com/blog/hashrate-index-roundup-november-9-2026/
- AnySilicon, costo di progettazione degli ASIC per nodo nel 2026: https://anysilicon.com/news/how-much-does-asic-design-cost-in-2026-real-price-ranges-by-process-node/ ; Silicon Analysts, guida ai costi di tape-out per aziende fabless: https://siliconanalysts.com/analysis/fabless-startup-tapeout-cost-guide ; dati di mercato: https://siliconanalysts.com/market-data
- MillionMiner, Bitmain S23 XP Hyd: https://millionminer.com/news/bitmain-wdms-2026-s23-xp-hyd-launch
- Khattak e Mikaitis, accumulo FP8 nei tensor core, arXiv 2512.07004: https://arxiv.org/html/2512.07004v2
- Canaan, risultati del secondo trimestre 2026: https://finviz.com/news/389337/canaan-inc-reports-unaudited-second-quarter-2026-financial-results ; Bitdeer, risultati del terzo trimestre 2026: https://greenstocknews.com/news/nasdaq/btdr/bitdeer-reports-unaudited-third-quarter-2026-results-record-revenue-of-us-320-7-million-adjusted-ebitda-up-173-to-us-69-6-million
- TheEnergyMag: Velaura AI: https://theenergymag.com/news/market-news/velaura-ai-series-a-bitcoin-pivot ; Core Scientific e Block: https://theenergymag.com/news/2026-07-29/core-scientific-exit-block-bitcoin-deal-ai ; economia dei contratti di leasing per l'AI: https://theenergymag.com/news/2026-08-06/miner-weekly-ai-pumpability-out-fuel
- Blockworks, Cipher e Fluidstack: https://blockworks.co/news/cipher-mining-inks-3b-ai-hosting-deal-with-fluidstack ; Cointelegraph, Core Scientific: https://cointelegraph.com/news/core-scientific-to-quit-bitcoin-mining-by-year-end-as-ai-business-surges ; ForkLog, Bitfarms: https://forklog.com/en/former-bitfarms-shuts-down-all-mining-sites-in-the-us/ ; IREN, risultati del T1 dell'esercizio 2027: https://globenewswire.com/news-release/2026/11/05/3397020/0/en/iren-reports-q1-fy27-results.html
- Spark, concentrazione del mercato degli ASIC per bitcoin: https://www.spark.money/research/bitcoin-asic-manufacturer-market-concentration

- Google, documentazione della 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 ; articolo sulla 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, prezzi di 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, confronto tra acceleratori per l'inferenza AI: https://themenonlab.blog/blog/ai-inference-accelerators-compared ; Introl, acceleratori AI oltre le GPU: https://introl.com/blog/ai-accelerators-beyond-gpus-tpu-trainium-gaudi-cerebras-2025
- StratNews (Reuters), chip AI nazionali nei data center finanziati dallo Stato: https://stratnewsglobal.com/team-sng/china-orders-state-funded-data-centres-to-use-domestic-ai-chips/ ; TNW, importazioni di H200: https://thenextweb.com/news/nvidia-h200-chips-hong-kong-china-import-approval
- Succinct, prova in tempo reale su 16 GPU: https://blog.succinct.xyz/real-time-proving-16-gpus/ ; prodotti ASIC ZK di Cysic: https://cysic-docs.pages.dev/hardware-products/zk-asic-products
- Tom's Hardware, Intel Heracles: https://www.tomshardware.com/tech-industry/cyber-security/intels-heracles-chip-computes-fully-encrypted-data-without-decrypting-it-chip-is-1-074-to-5-547-times-faster-than-a-24-core-intel-xeon-in-fhe-math-operations ; IEEE Spectrum: https://spectrum.ieee.org/fhe-intel ; Evertiq, Niobium e SemiFive: https://evertiq.com/design/2026-02-20-semifive-secures-design-win-with-niobium-for-fhe-accelerator ; The Quantum Insider, finanziamento di 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 su FPGA: https://www.zama.org/post/announcing-hpu-on-fpga-the-first-open-source-hardware-accelerator-for-fhe

Nodi di processo

- Silicon Analysts, prezzi dei wafer: https://siliconanalysts.com/data/wafer-pricing
- Tom's Hardware, indiscrezioni sui prezzi di 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, prezzi di SMIC: https://wccftech.com/smic-5nm-chip-prices-up-to-50-percent-more-expensive/amp/
- EUROPRACTICE, calendario MPW di 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, roadmap di 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, la catena di fornitura dei chip AI in Cina: 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

Opportunità ed economia dei lanci

- Webull (Benzinga), conference call sui risultati di 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, l'unità AI di Canaan: https://www.theblock.co/post/359250/crypto-miner-canaan-to-shutter-non-core-ai-semiconductor-business-unit-amid-strategic-realignment
- Clutch, tariffe dello sviluppo AI: https://clutch.co/developers/artificial-intelligence/pricing ; PixelPlex, costo dello sviluppo 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 ; verifica a livelli, arXiv 2512.20176: https://arxiv.org/abs/2512.20176v1
- Sherlock, riferimento sui prezzi degli audit per il 2026: https://sherlock.xyz/post/smart-contract-audit-pricing-a-market-reference-for-2026
- Blockworks, criteri di quotazione di Binance: https://blockworks.com/news/binance-listing-standards-scrutiny ; Cointelegraph, accordi con i market maker: 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

PROOF OF USEFUL WORK · CRYPTO × SILICIO

Il calcolo come consenso

AI × crypto × chip nel 2026, a partire dalla proof of useful work di Pearl

Ogni numero chiave di questo rapporto riporta la sua fonte e la sua data. Dove l'autore ricava un numero, questo è segnato come stima; dove lo riporta solo il venditore, è segnato come dichiarato dal fornitore. Il rapporto considera solo tecnologia e business.

Tan Shuai

Tecnologo e fondatore · Ricercatore e sviluppatore di AI

tanshuai.com · i@tanshuai.com

EDIZIONE PUBBLICA · Dati aggiornati al 13 novembre 2026

