Nota di ricerca · Novembre 2026

Il calcolo come consenso

AI × crypto × chip nel 2026, a partire da Pearl

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

Tan Shuai
tanshuai.com · i@tanshuai.com



Il teorema regge; l'incentivo no

01 Pearl funziona come ingegneria, non ancora come lavoro utile

Circa 90 EH/s di moltiplicazione di matrici intere proteggono la catena, ma solo circa l'1% dei blocchi contiene lavoro di AI reale.

02 La domanda pagata sta nell'inferenza, nei dati e nella privacy

Bittensor circa 70 milioni di dollari l'anno di ricavi verificabili, Grass 17 milioni nel primo semestre 2026, TEE in produzione; pagamenti tra agenti circa 5.000–11.000 dollari al mese.

03 Un ASIC ripaga solo con un piatto grande, regole stabili e un secondo mercato

Il suo tetto di profitto è R × (1 − 1/k). Oggi nessuna rete offre tutte e tre le cose analisi .

≈90 EH/s

Pearl il 13 novembre 2026: circa 122.000 H100 di throughput intero

≈1%

dei blocchi di Pearl contiene lavoro di AI reale (Blockworks)

4–11×

le emissioni di Bittensor rispetto ai suoi ricavi verificabili

AI × crypto in sei settori: dove la domanda pagata è reale

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

> Il sussidio è carburante per l'avvio, non un modello di business. La crypto fa distribuzione e regolamento; i profitti vanno all'equity, almeno nel 2026.

Una moltiplicazione-accumulo è un hash

Dal teorema alla mainnet

Komargodski e Weinstein (aprile 2025): una proof of useful work per qualsiasi prodotto di matrici con overhead 1 + o(1). La mainnet di Pearl è partita il 27 aprile 2026.

Come funziona

I miner moltiplicano matrici rumorose, trasformano con un hash l'accumulatore di ogni tile in un biglietto della lotteria e provano il tile vincente con una prova Plonky2 sotto i 60 KB.

L'unità

1 H è una moltiplicazione-accumulo INT7 × INT7 → INT32 accreditata: il budget di sicurezza di una blockchain prezzato in throughput tensoriale.

PEARL AL 13 NOVEMBRE 2026

89,77 EH/s

hashrate di rete (nodo completo indipendente)

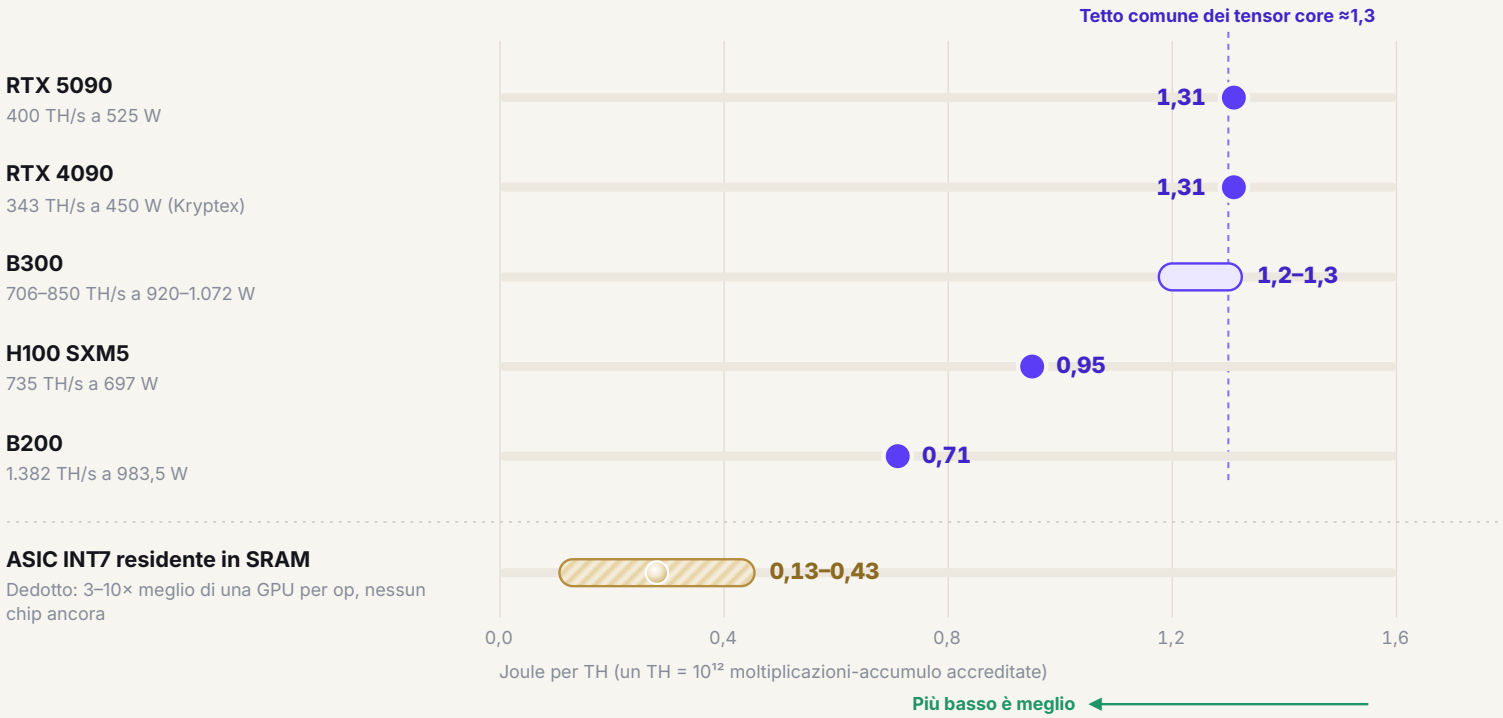
≈500 mln $

valore di mercato; completamente diluito circa 2,8 miliardi di dollari

+78%

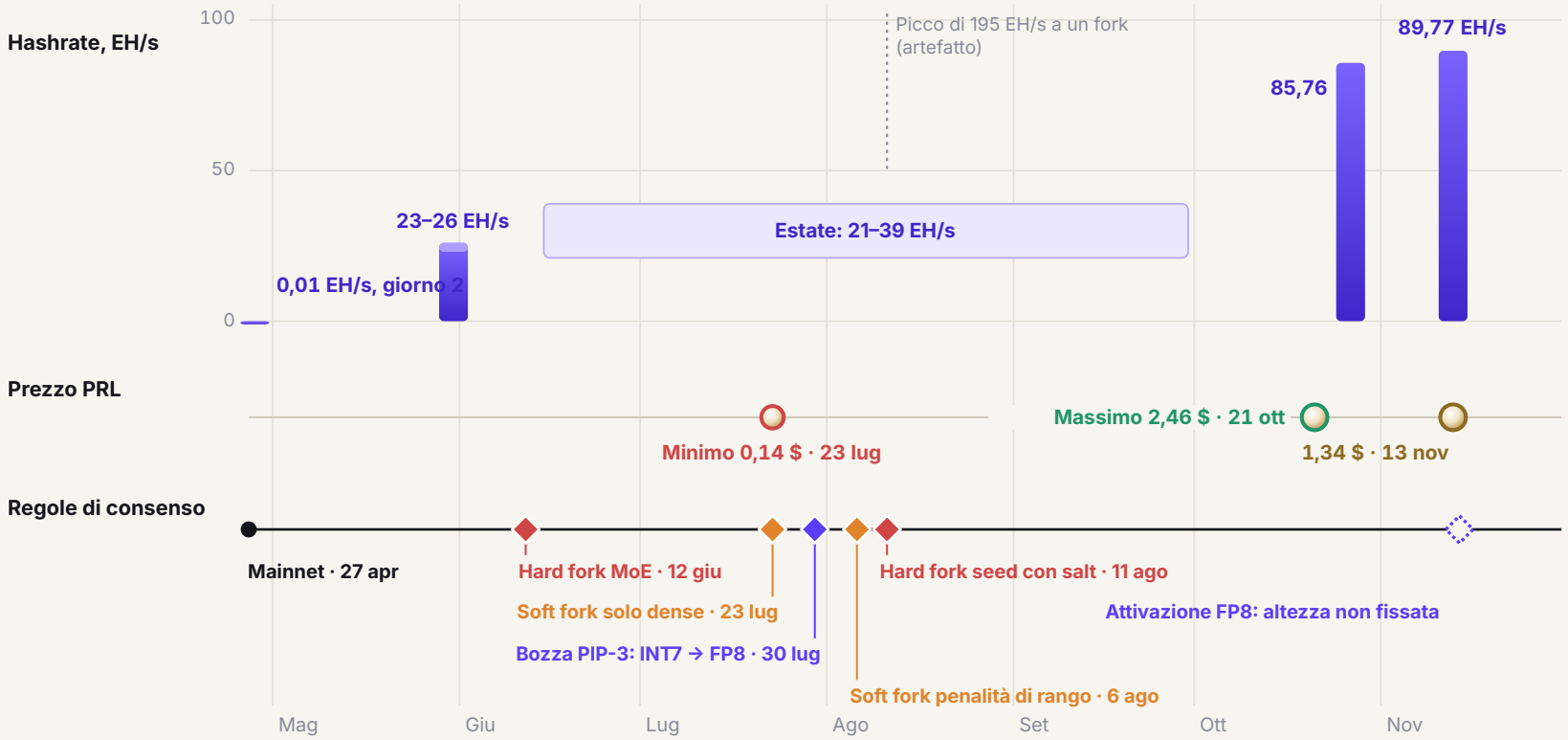
nuova offerta nei prossimi 12 mesi

Sul lavoro intero di Pearl le GPU hanno un tetto energetico comune



> Schede da gaming e B300 si incontrano a circa 1,3 J/TH; anche la B200 arriva solo a 0,71. Un chip che moltiplica soltanto interi a 7 bit dalla SRAM potrebbe plausibilmente fare 3–10 volte meglio **analisi**.

L'hashrate segue il prezzo; le regole continuano a cambiare



Circa 120 milioni di monete, il 32% dell'offerta attuale, sono state minate nei primi sei giorni. Una RTX 5090 oggi rende circa 5–6 dollari al giorno; noleggiare schede solo per minare fa perdere denaro.

Il teorema regge; l'incentivo no

QUANTO COSTA IL LAVORO UTILE

Lato inferenza: overhead ridotto

Pearl dichiara il 5,08% di tempo in più per Llama 70B su quattro H200; un test indipendente ha misurato il 9,2% di throughput in meno su una GPU Hopper.

QUANTO RENDE IL LAVORO UTILE

Lato mining: troppo poco hashrate

Una scheda che serve inferenza mina a circa il 27% di un miner dedicato in prefill, al 5,4% in chat, circa all'1% quando domina la decodifica.

Quindi i miner razionali moltiplicano matrici casuali. L'utilità è un problema commerciale fuori catena: qualcuno deve pagare l'inferenza.

Quattro fork in due mesi, e poi FP8

12 giugno
Hard fork mixture-of-experts (blocco 71.935)

23 luglio
Soft fork solo dense (blocco 91.630)

6 agosto
Soft fork con penalità di rango: alcune configurazioni perdono il 50–87,5% dell'hashrate effettivo

11 agosto
Hard fork del seed con salt: i vecchi miner producono solo share non valide

Non fissata
PIP-3: da INT7 a FP8, ammesse solo Hopper e Blackwell

2 su 3
utile per l'AI, verificabile bit per bit, neutrale rispetto all'hardware: un formato numerico ne ottiene al massimo due

13 bit
conservati dall'accumulo FP8 di Hopper, con troncamento: non riproducibile tra generazioni

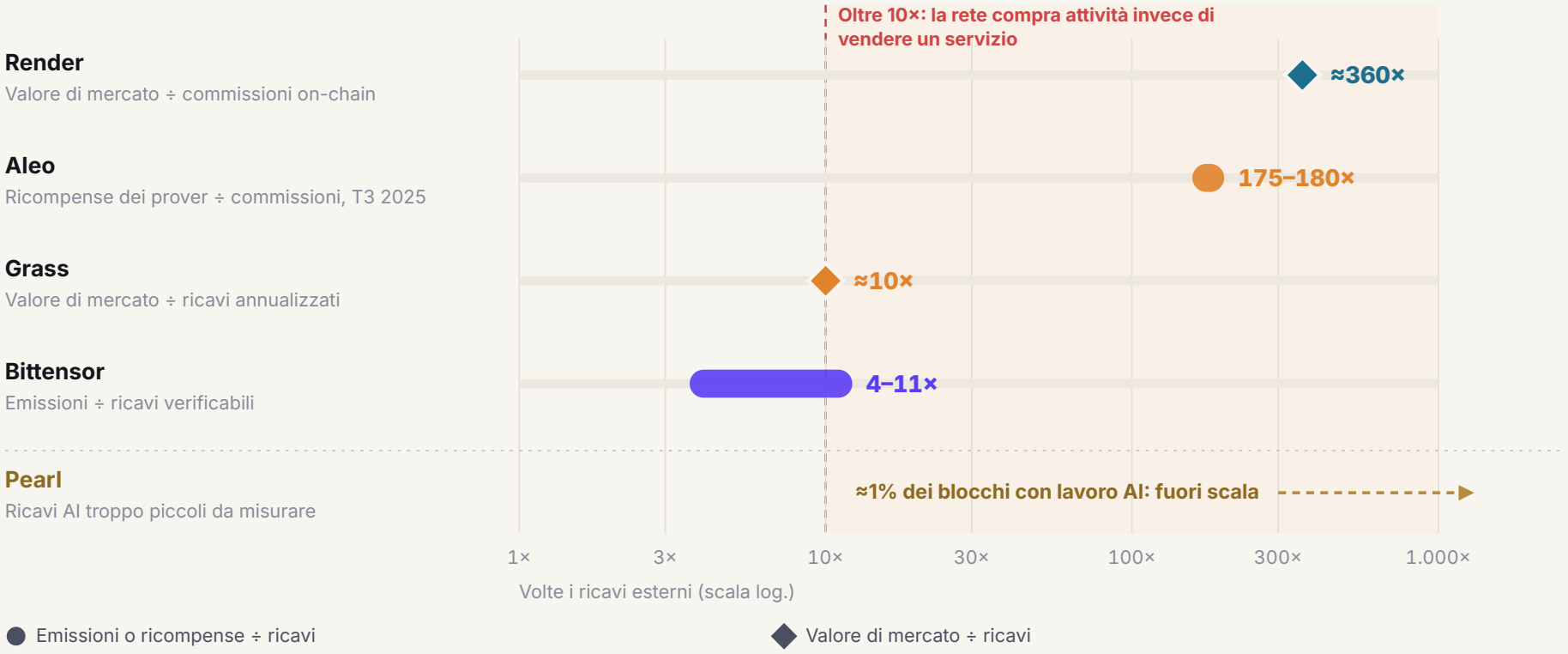
≈82%
dell'hashrate su schede consumer che PIP-3 non ammetterebbe (modello a bassa affidabilità)

Chi controlla la funzione di lavoro controlla il destino di chi investe in hardware analisi **.**

Né le matrici né la conoscenza zero fermano gli ASIC

Rete	Che cosa è successo	Che cosa insegna
Kaspa	Un prodotto di matrici 64 × 64 tra due hash; ASIC dopo 16 mesi, circa 1.000× l'efficienza delle GPU; il migliore oggi rende 0,86 dollari al giorno	Una matrice non è né utile né a prova di ASIC
Aleo	Un puzzle SNARK sintetico; macchine dedicate dopo cinque mesi; ricompense dei prover 175–180× le commissioni	Il lavoro sintetico vive di sussidio
Qubic	Dichiara addestramento AI; il suo calcolo ha minato Monero (51% dichiarato) e ora Dogecoin	Una PoW generica può essere catturata dall'esterno
Bittensor	Punteggio pesato sullo stake; circa 70 milioni di dollari l'anno di ricavi verificabili contro più di 300 milioni di emissioni	L'unica con ricavi esterni su scala
Monero, Sia, Ethereum	I fork hanno reso inutilizzabili gli ASIC; dopo il Merge meno del 20% dei miner ha trovato nuovo lavoro	Sopravvive l'hardware con un secondo mercato

Il multiplo del sussidio è il controllo di salute più onesto



> Oltre circa 10 volte i ricavi esterni, una rete compra attività invece di vendere un servizio. I ricavi AI di Pearl sono troppo piccoli per stare sulla scala.

Domanda reale ma piccola, e il valore che passa all'equity

- **Mercati del calcolo**

  Commissioni on-chain di Render circa 2,35 milioni di dollari l'anno; la H100 di Akash a 2,73 dollari l'ora batte AWS e Google ma non Vast.ai o RunPod.

- **Addestramento**

  8–40 miliardi di parametri addestrati via internet; INTELLECT-3 è tornato a 512 H200 su InfiniBand.

- **AI verificabile**

  I TEE aggiungono meno del 7% e girano in produzione; TEE.fail ha violato l'attestazione con un interposer da 1.000 dollari; lo zkML prova GPT-2 a 174 token al minuto.

- **Agenti e dati**

  Flussi x402 avviati da agenti circa 5.000–11.000 dollari al mese; Grass 17 milioni di dollari di ricavi dati nel primo semestre 2026.

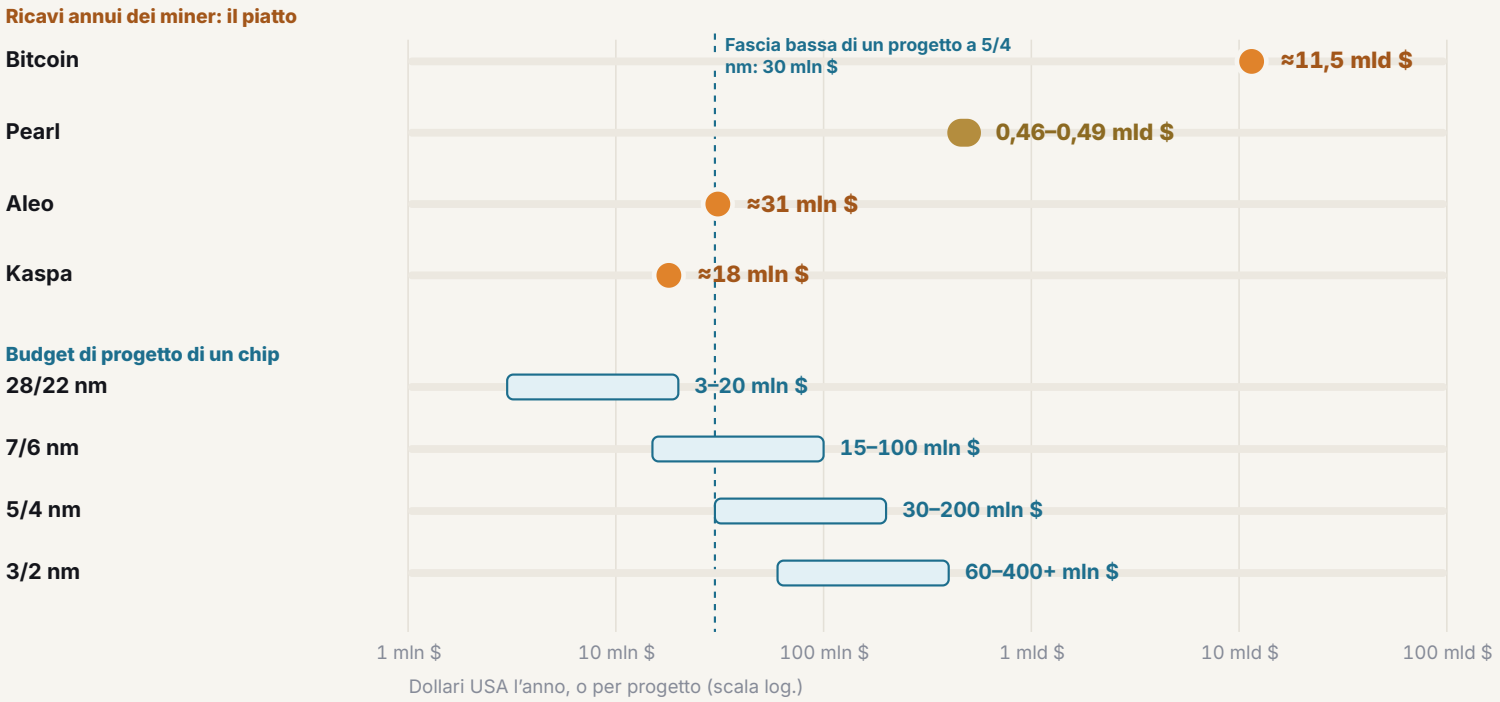
> **Quattro società con equity su infrastruttura centralizzata valgono circa 8,5 miliardi di dollari, più o meno quanto tutti i token AI dei livelli protocollo e applicazione.**

Perché l'efficienza smette di pagare oltre k = 5



Tetto di profitto ≈ R × (1 − 1/k). Quando gli ASIC diventano i miner marginali, k si azzera contro altri ASIC; contro GPU inattive che pagano solo l'energia, k è circa 2,5 **stima**.

La maggior parte dei piatti è più piccola del budget di un solo chip



Hashrate di rientro h* = NRE ÷ [R × (1 − 1/k) × T × p]. Per Pearl con 20 milioni di dollari di NRE e k = 5: circa 4,9 EH/s per un anno se le regole reggono, 9,8 EH/s se reggono con probabilità pari **stima**.

Cinque tipi di chip, cinque risposte

 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

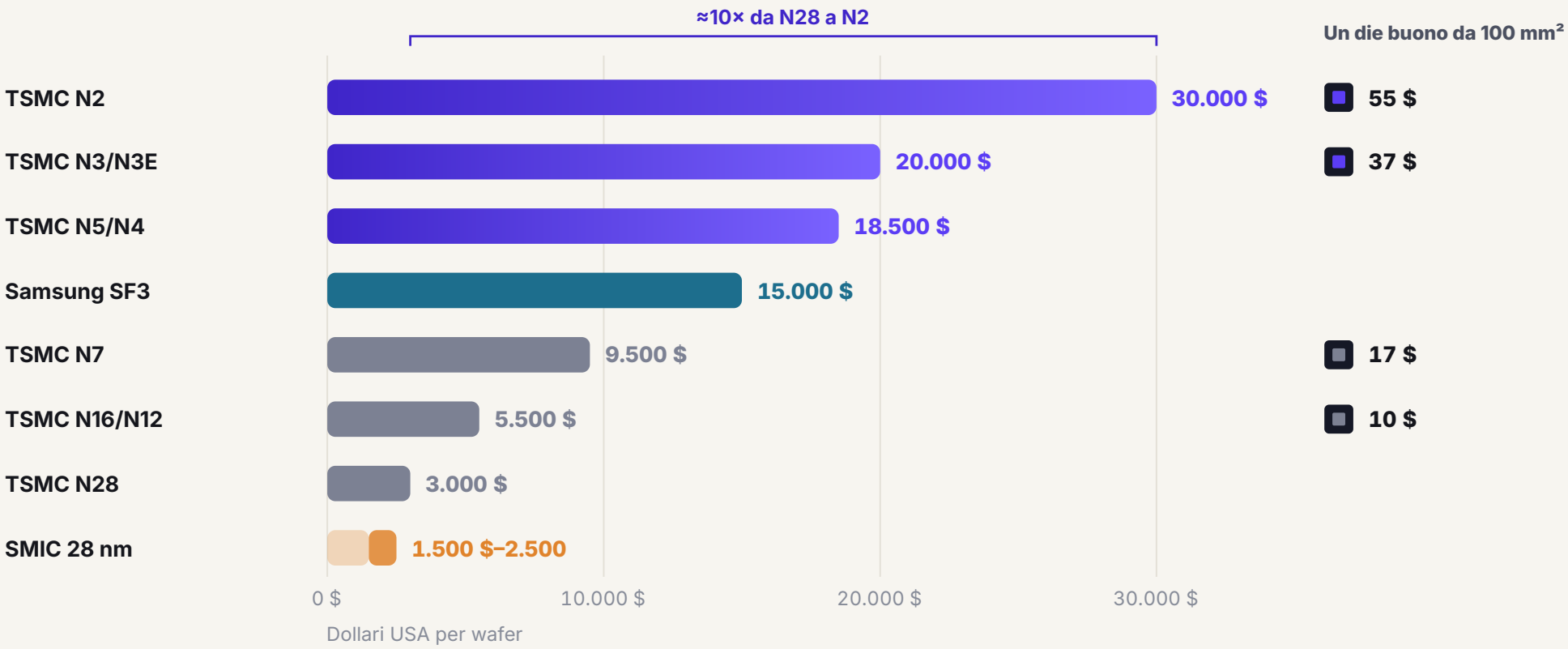
k: 1.074–5.547× uno Xeon (Heracles)

Secondo mercato: calcolo cifrato, AI privata

Una finestra 2027–2029; validare con asset leggeri

> Oggi nessuna rete offre insieme un piatto grande, regole stabili e k sopra 1: il k di bitcoin appartiene agli operatori affermati, le regole di Pearl sono le meno stabili, i piatti di Kaspa e Aleo sono troppo piccoli.

Un wafer all’avanguardia costa dieci volte uno maturo



> Lasciate scegliere il costo più grande sull’intera vita: l’energia per i chip di calcolo sempre accesi, packaging e HBM per i chip di addestramento (63% del costo di una H100), la scelta SRAM o HBM per i chip di inferenza.

Un nodo per ogni chip e per ogni team

Chip	Team piccolo	Azienda già in scala	Percorso Cina continentale
Chip di mining PoUW	MPW a 28 o 16/12 nm; primi volumi N7/N6	N3/N3E o Samsung SF3	SMIC N+1/N+2 funziona, circa il 40–50% in più per die
Chip di mining SHA-256	Non inseguire la frontiera	Classe 3 nm	Difficile eguagliare l'efficienza dei nodi di punta
Chip per l'addestramento AI	IP, chiplet, servizi di progettazione	N3P poi N2/A16 con CoWoS e HBM	SMIC N+2/N+3 con HBM nazionale
Chip per l'inferenza AI	Incentrato sulla SRAM a N6/N7 o N4/N5/SF4	N3/N2 con HBM, o chiplet SRAM	SMIC N+2 con LPDDR o GDDR
Acceleratori ZK e FHE	Prima FPGA o MPW a 16/12 nm	N5/N3; FHE con HBM	ZK praticabile; FHE limitato dall'HBM

UN BUDGET A TAPPE

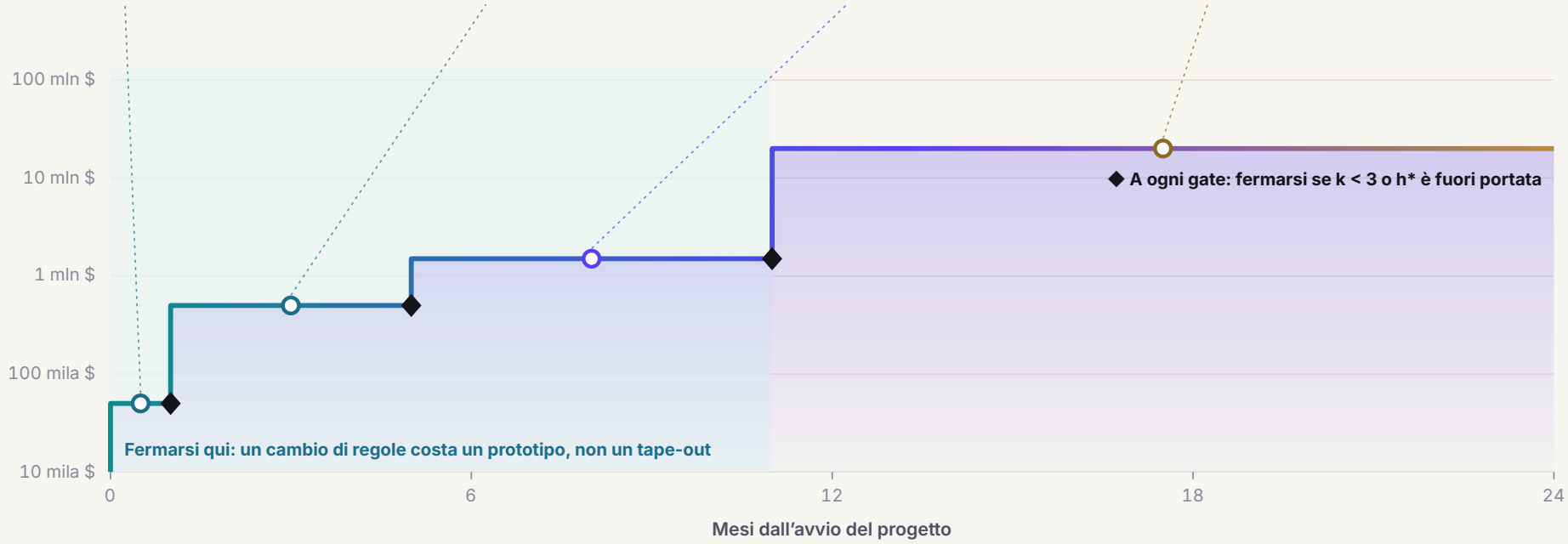
Prove prima delle maschere

Gate 0 · Modello
≈50 mila $ · settimane
J/op e $/op da aritmetica, SRAM e I/O

Gate 1 · FPGA
≈0,5 mln $ · mesi
Il carico reale end-to-end: share o token

Gate 2 · MPW
≈1,5 mln $ · mese 11
Il silicio misurato sostituisce il modello

Gate 3 · Maschere e primo lotto
≈20 mln $ · mese 24
Maschere N7 ≈3,5 mln $ in un progetto da 15–100 mln $



> Un modello, un FPGA e un MPW arrivano al silicio misurato con circa 1,5 milioni di dollari in undici mesi; il set di maschere e il primo lotto costano più di dieci volte tanto. Fermarsi prima delle maschere finché l'acquirente o la versione del protocollo non sono fissati.

OPPORTUNITÀ

La crypto come primo cliente, l'AI come secondo mercato

Chip custom

- La domanda è in larghissima parte AI: Broadcom circa 58 miliardi di dollari di semiconduttori AI nell'esercizio 2026
- Le competenze dei chip di mining si trasferiscono: array a bassa tensione, alimentazione da kilowatt, tape-out rapidi
- Si vince con un canale di vendita indipendente, non con i canali del mining

Sviluppo AI

- Le ore sono una commodity: 25–49 dollari l'ora su Clutch
- Il valore sta nelle giunture: kernel per chip custom, mining dentro i servizi di inferenza
- Inferenza verificabile: TEE più controlli a campione ZK

Sistemi AI × crypto

- Verifica a livelli: TEE o campionamento, poi contestazioni o controlli a campione ZK
- Le barriere sono distribuzione e liquidità, non il costo
- Firmare l'acquirente di calcolo prima del token

Tre percorsi, ognuno con un primo passo piccolo

Percorso	Primo passo
A. Un acceleratore pensato per l'inferenza che sa anche minare	Uno studio di quattro settimane su J/op e $/op di un array INT8/FP8 residente in SRAM contro RTX 5090 e H100, poi FPGA o MPW su nodo maturo
B. Servizi di silicio custom: prima la crypto, poi l'AI	Trasformare il modello di NRE e di pareggio di questo rapporto in studi di fattibilità a pagamento, con milestone e una versione di protocollo fissata
C. Una rete nativa AI con un acquirente prima del token	Una nota di progetto sulla verifica a livelli e su un'aritmetica canonica neutrale rispetto ai fornitori, scritta con un acquirente di inferenza ben identificato

Al massimo tre percorsi, ognuno messo alla prova prima di impegnare denaro analisi.

Che cosa seguire e quali domande porre

SEGNALI DA SEGUIRE

In qualsiasi momento	L'altezza di attivazione di FP8 su Pearl, e se la catena si divide
Ogni mese	La quota di blocchi di Pearl con lavoro di AI
Ogni trimestre	Ricavi verificabili di Bittensor contro emissioni; flussi x402 avviati da agenti
Gennaio 2027	L'aumento di prezzo del 3–6% riportato per TSMC sui 2 nm e 3 nm
T1 2027	FP4 in Pearl, come riportato dall'AMA dei fondatori
Aprile 2028	Il prossimo halving di Bitcoin

DOMANDE DA PORRE

01 Chi controlla la funzione di lavoro?

02 Qual è il multiplo del sussidio?

03 Quanto costa il prossimo gate, e che cosa vi ferma?

04 Qual è l'hashrate di rientro, con p?

05 Qual è il secondo mercato del chip?

06 Chi è il primo cliente pagante, con nome e cognome?

Sono ambiti che possiamo esplorare insieme, con scambi e collaborazioni tecniche: chip custom, sviluppo AI e sistemi AI × crypto.

i@tanshuai.com · tanshuai.com