Note de recherche · Novembre 2026

Le calcul comme consensus

IA × crypto × puces en 2026, en partant de Pearl

Pearl · Travail utile · DePIN et agents · Économie des ASIC · Nœuds de gravure · Budgets par étapes

Tan Shuai
tanshuai.com · i@tanshuai.com



LE VERDICT

Le théorème tient, l'incitation non

01 Pearl fonctionne techniquement, mais pas encore comme travail utile

Environ 90 EH/s de multiplication matricielle entière sécurisent la chaîne, mais seuls environ 1 % des blocs portent un vrai travail d'IA.

02 La demande payante est dans l'inférence, les données et la confidentialité

Bittensor environ 70 millions de dollars par an de revenus vérifiables, Grass 17 millions au S1 2026, TEE en production ; paiements d'agents environ 5 000–11 000 dollars par mois.

03 Un ASIC ne paie qu'avec une grande cagnotte, des règles stables et un second débouché

Son plafond de profit est R × (1 − 1/k). Aujourd'hui, aucun réseau n'offre les trois analyse .

≈90 EH/s

Pearl le 13 novembre 2026 : environ 122 000 H100 de débit entier

≈1 %

des blocs de Pearl portent un vrai travail d'IA (Blockworks)

4–11×

les émissions de Bittensor face à ses revenus vérifiables

L'IA × crypto en six secteurs : où la demande payante est réelle

Chaînes à travail utile — 01

RÉEL MAIS FAIBLE

≈70 M$ par an

revenus externes vérifiables de Bittensor

Signal : émissions 4–11× les revenus ; Pearl ≈1 % de blocs IA

Les mécanismes marchent ; l'utilité veut des payeurs

Calcul décentralisé — 02

RÉEL MAIS FAIBLE

2,35 M$ par an

frais on-chain de Render, les plus élevés vérifiables

Signal : valeur de marché ≈360× les frais

Surtout de la revente d'inférence banalisée

Entraînement et inférence — 03

PROFIT EN ACTIONS

100 M$+ par an

Prime Intellect, société par actions sur GPU centralisés

Signal : mainnet d'entraînement de Gensyn non lancé

Progrès réels ; le profit va aux actions

IA vérifiable — 04

PRÉCOCE

398

utilisateurs payants du cloud TEE de Phala, fin 2025

Signal : aucun zkML payant déclaré

TEE en production ; zkML précoce

Paiements d'agents — 05

AVANT LA DEMANDE

5–11 k$ par mois

paiements x402 réellement initiés par des agents

Signal : ≈200 M de transactions, surtout de la signalisation

Les rails d'abord ; la demande n'a pas suivi

Réseaux de données — 06

VRAIS ACHETEURS

17 M$

revenus de données de Grass, S1 2026 (autodéclarés)

Signal : valeur de marché ≈10× les revenus annualisés

L'un des rares secteurs à vrais acheteurs

> La subvention est un carburant pour un démarrage à froid, pas un modèle économique. La crypto assure distribution et règlement ; les actions captent les profits, du moins en 2026.

Une multiplication-accumulation vaut un hash

Du théorème au mainnet

Komargodski et Weinstein (avril 2025) : une preuve de travail utile pour tout produit matriciel à un surcoût de 1 + o(1). Le mainnet de Pearl a démarré le 27 avril 2026.

Le fonctionnement

Les mineurs multiplient des matrices bruitées, hachent l'accumulateur de chaque tuile en ticket de loterie et prouvent la tuile gagnante par une preuve Plonky2 de moins de 60 Ko.

L'unité

1 H est une multiplication-accumulation INT7 × INT7 → INT32 créditée : le budget de sécurité d'une blockchain exprimé en débit tensoriel.

PEARL LE 13 NOVEMBRE 2026

89,77 EH/s

hashrate du réseau (nœud complet indépendant)

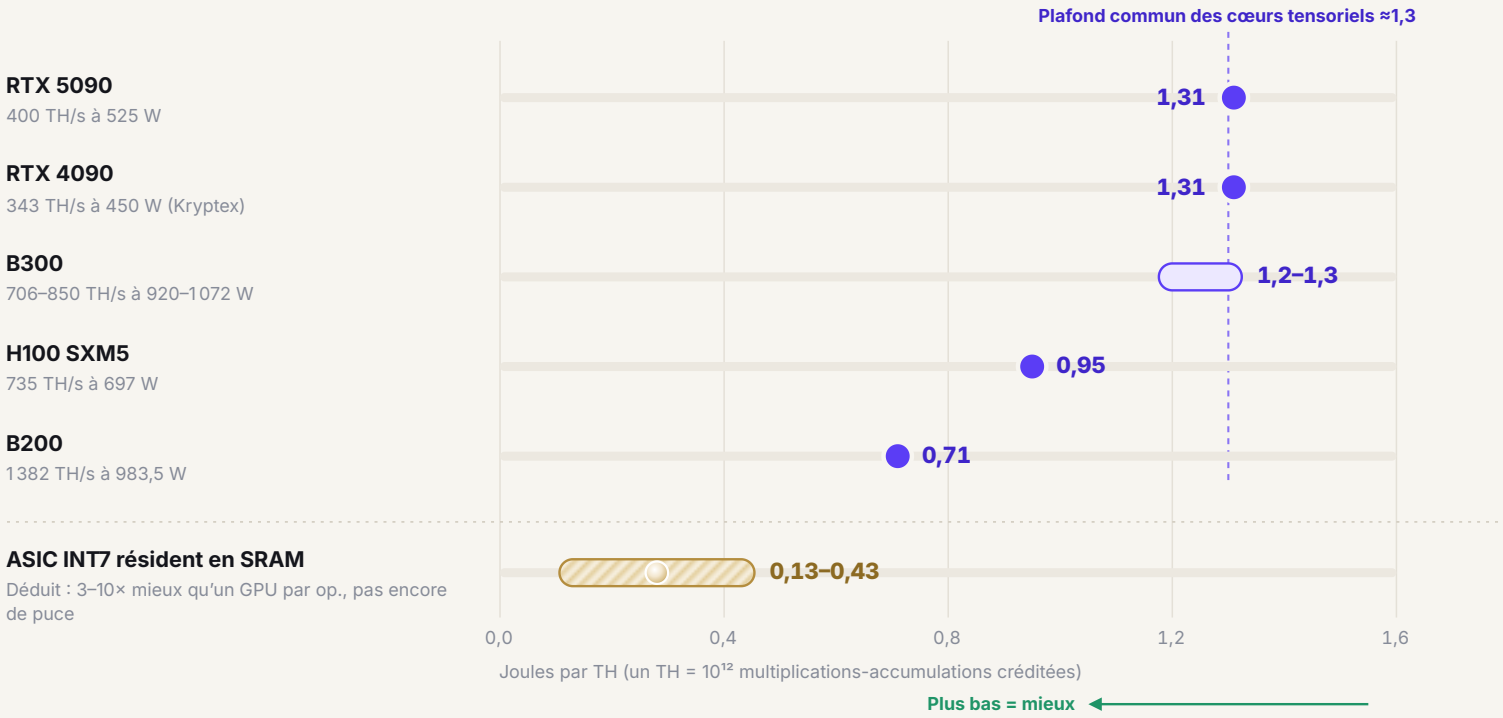
≈500 M$

valeur de marché ; entièrement diluée environ 2,8 milliards de dollars

+78 %

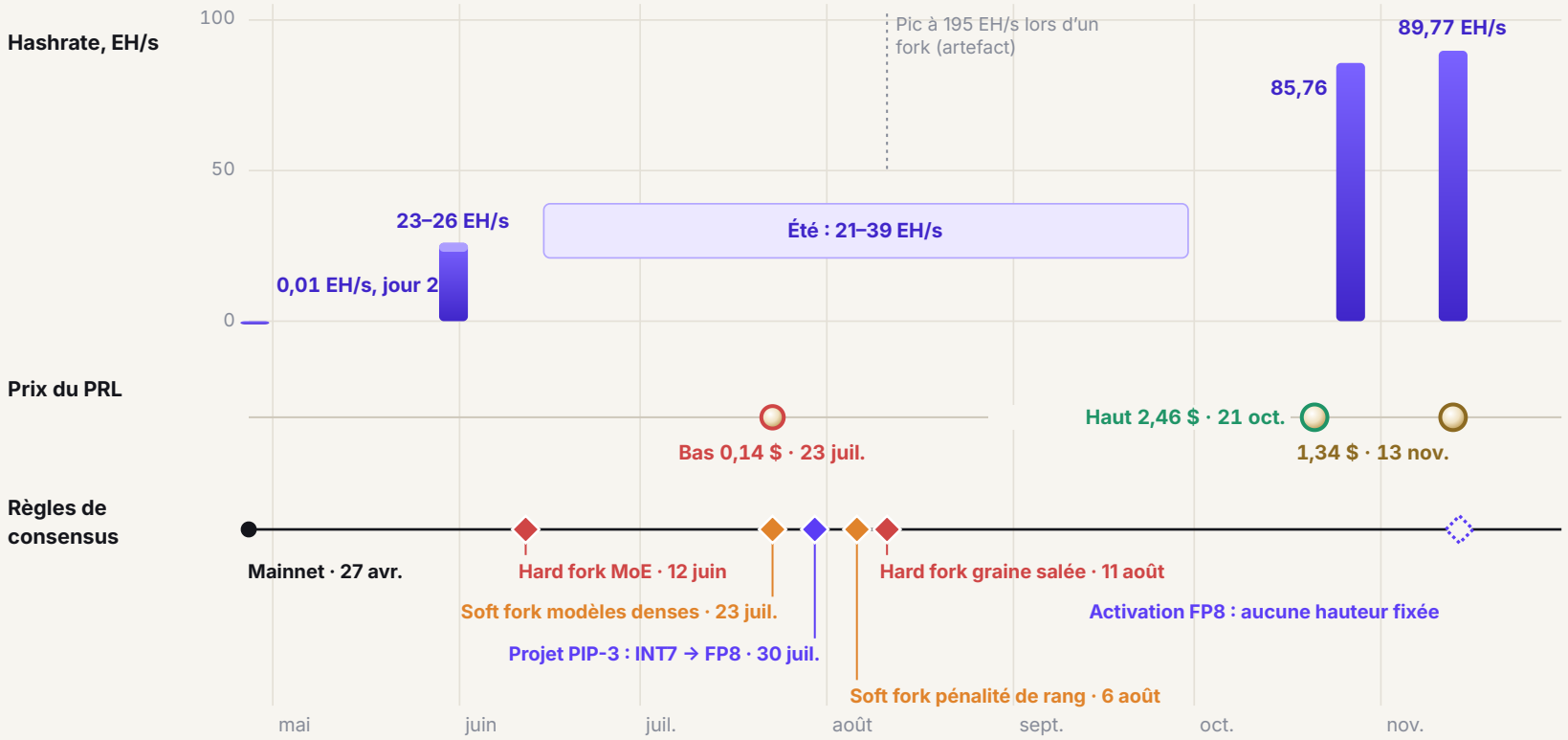
d'offre nouvelle sur les 12 prochains mois

Sur le travail entier de Pearl, les GPU partagent un plafond énergétique



> Cartes de jeu et B300 se rejoignent vers 1,3 J/TH ; même le B200 n'atteint que 0,71. Une puce qui ne ferait que multiplier des entiers de 7 bits depuis la SRAM pourrait plausiblement faire 3 à 10 fois mieux analyse.

Le hashrate suit le prix ; les règles changent sans cesse



Environ 120 millions de pièces, 32 % de l'offre actuelle, ont été minées les six premiers jours. Une RTX 5090 rapporte environ 5–6 dollars par jour ; louer des cartes pour miner fait perdre de l'argent.

Le théorème tient, l’incitation non

CE QUE COÛTE LE TRAVAIL UTILE

Côté inférence : surcoût faible

Pearl indique 5,08 % de temps en plus pour Llama 70B sur quatre H200 ; un test indépendant a mesuré 9,2 % de débit en moins sur un GPU Hopper.

CE QUE RAPPORTE LE TRAVAIL UTILE

Côté minage : trop peu de hashrate

Une carte qui sert de l’inférence mine à environ 27 % d’un mineur dédié en préremplissage, 5,4 % en conversation, environ 1 % quand le décodage domine.

Les mineurs rationnels multiplient donc des matrices aléatoires. L’utilité est un problème commercial hors chaîne : quelqu’un doit payer l’inférence.

Quatre forks en deux mois, et bientôt le FP8

12 juin
Hard fork mixture-of-experts (bloc 71 935)

23 juillet
Soft fork réservé aux modèles denses (bloc 91 630)

6 août
Soft fork de pénalité de rang : certaines configurations perdent 50–87,5 % de hashrate effectif

11 août
Hard fork à graine salée : les anciens mineurs ne produisent que des parts invalides

Non fixé
PIP-3 : de l'INT7 au FP8, seuls Hopper et Blackwell admis

2 sur 3

utile pour l'IA, vérifiable au bit près, neutre vis-à-vis du matériel : un format en obtient au plus deux

13 bits

conservés par l'accumulation FP8 de Hopper, avec troncature : non reproductible d'une génération à l'autre

≈82 %

du hashrate sur des cartes grand public que PIP-3 n'admettrait pas (modèle à faible confiance)

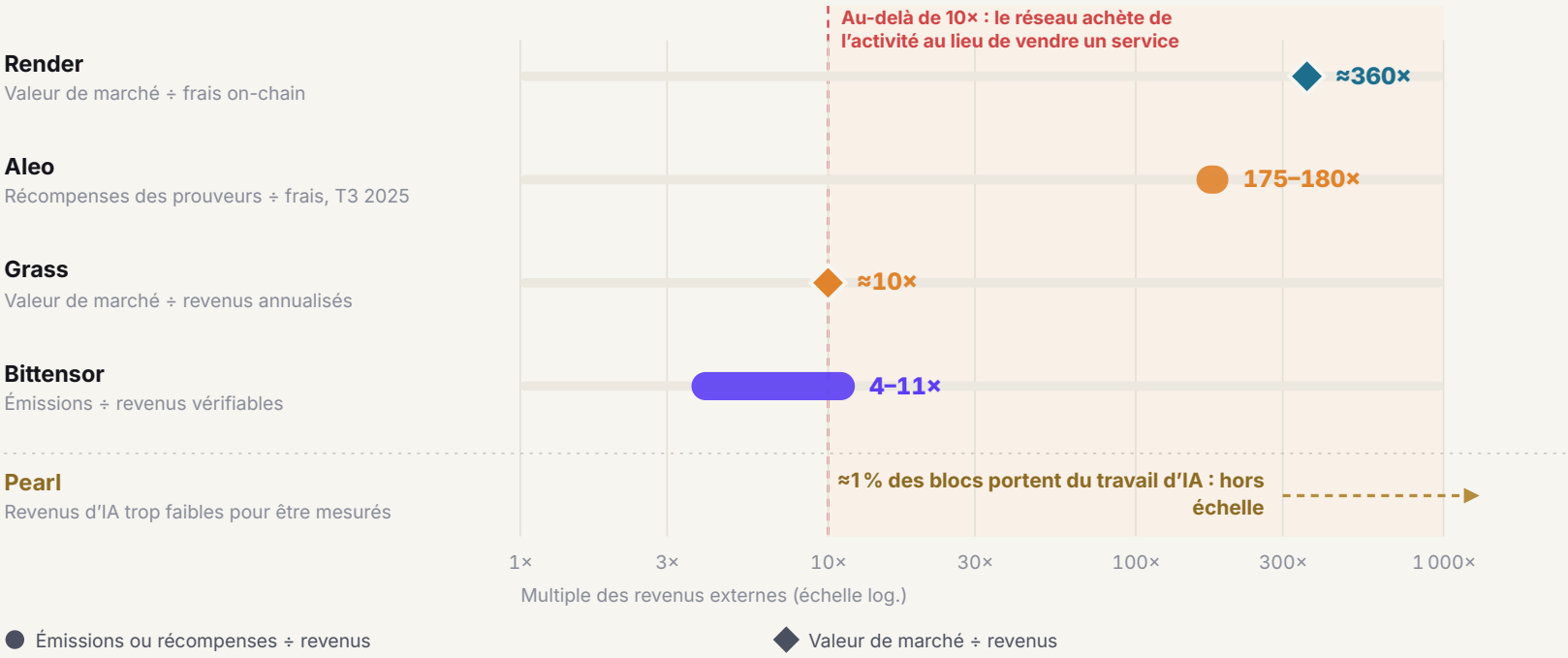
Qui contrôle la fonction de travail contrôle le sort des investisseurs en matériel analyse **.**

Ni les matrices ni la divulgation nulle n'arrêtent les ASIC

Réseau	Ce qui s'est passé	Ce que cela enseigne
Kaspa	Un produit matriciel 64 × 64 entre deux hachages ; des ASIC après 16 mois, environ 1 000× l'efficacité d'un GPU ; le meilleur rapporte 0,86 dollar par jour	Une matrice n'est ni utile ni à l'abri des ASIC
Aleo	Un puzzle SNARK synthétique ; des machines dédiées après cinq mois ; récompenses des prouveurs 175–180× les frais	Le travail synthétique vit de la subvention
Qubic	Prétend entraîner une IA ; son calcul a miné du Monero (51 % revendiqués), puis du Dogecoin	Une PoW généraliste peut être captée de l'extérieur
Bittensor	Notation pondérée par la mise ; environ 70 millions de dollars par an de revenus vérifiables face à plus de 300 millions d'émissions	Le seul doté de revenus externes à grande échelle
Monero, Sia, Ethereum	Des forks ont mis des ASIC hors service ; après le Merge, moins de 20 % des mineurs ont trouvé un nouveau travail	Le matériel doté d'un second débouché survit

SUBVENTION

Le multiple de subvention est le bilan de santé le plus honnête



Au-delà d'environ 10 fois les revenus externes, un réseau achète de l'activité au lieu de vendre un service. Les revenus d'IA de Pearl sont trop faibles pour figurer sur l'échelle.

Une demande réelle mais faible, et une valeur qui passe aux actions

Marchés de calcul

Frais on-chain de Render environ 2,35 millions de dollars par an ; le H100 d'Akash à 2,73 dollars de l'heure bat AWS et Google, mais pas Vast.ai ni RunPod.

Entraînement

8–40 milliards de paramètres entraînés via Internet ; INTELLECT-3 est revenu à 512 H200 sur InfiniBand.

IA vérifiable

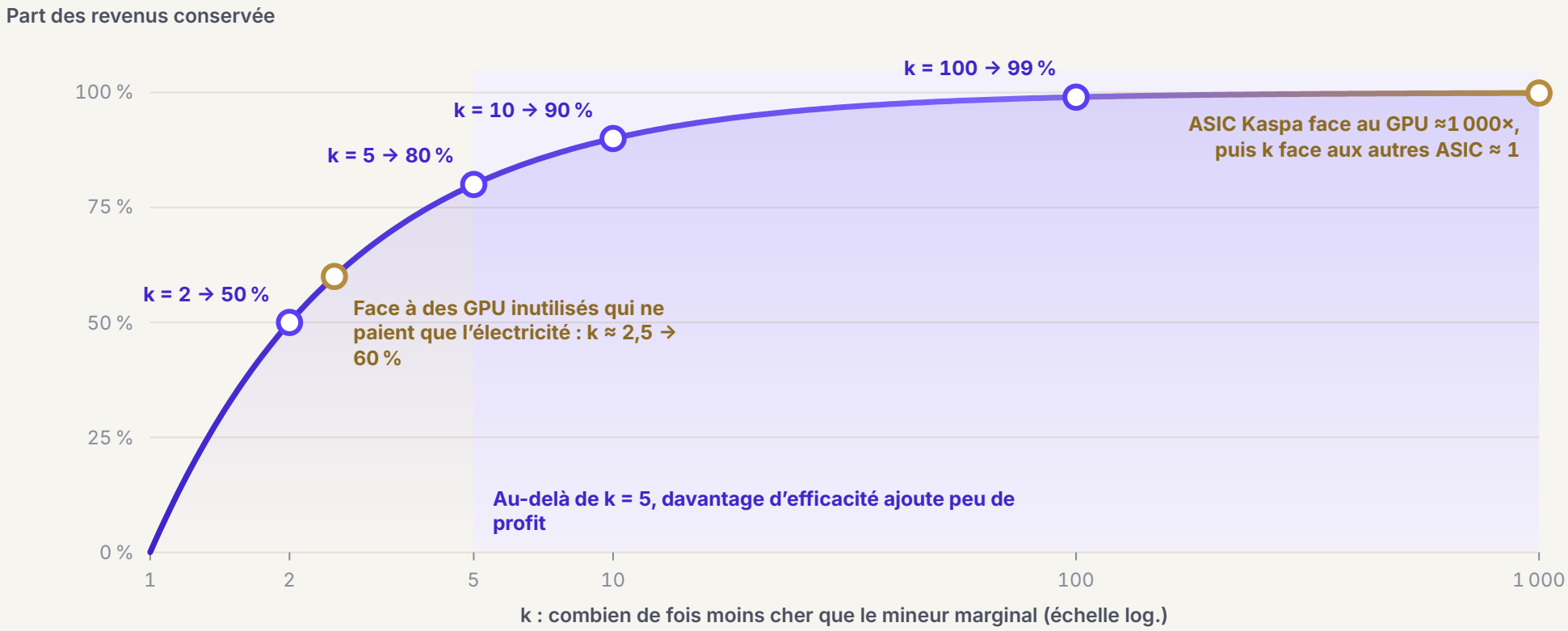
Les TEE ajoutent moins de 7 % et tournent en production ; TEE.fail a cassé l'attestation avec un interposeur à 1 000 dollars ; le zkML prouve GPT-2 à 174 tokens par minute.

Agents et données

Flux x402 initiés par des agents environ 5 000–11 000 dollars par mois ; Grass 17 millions de dollars de revenus de données au S1 2026.

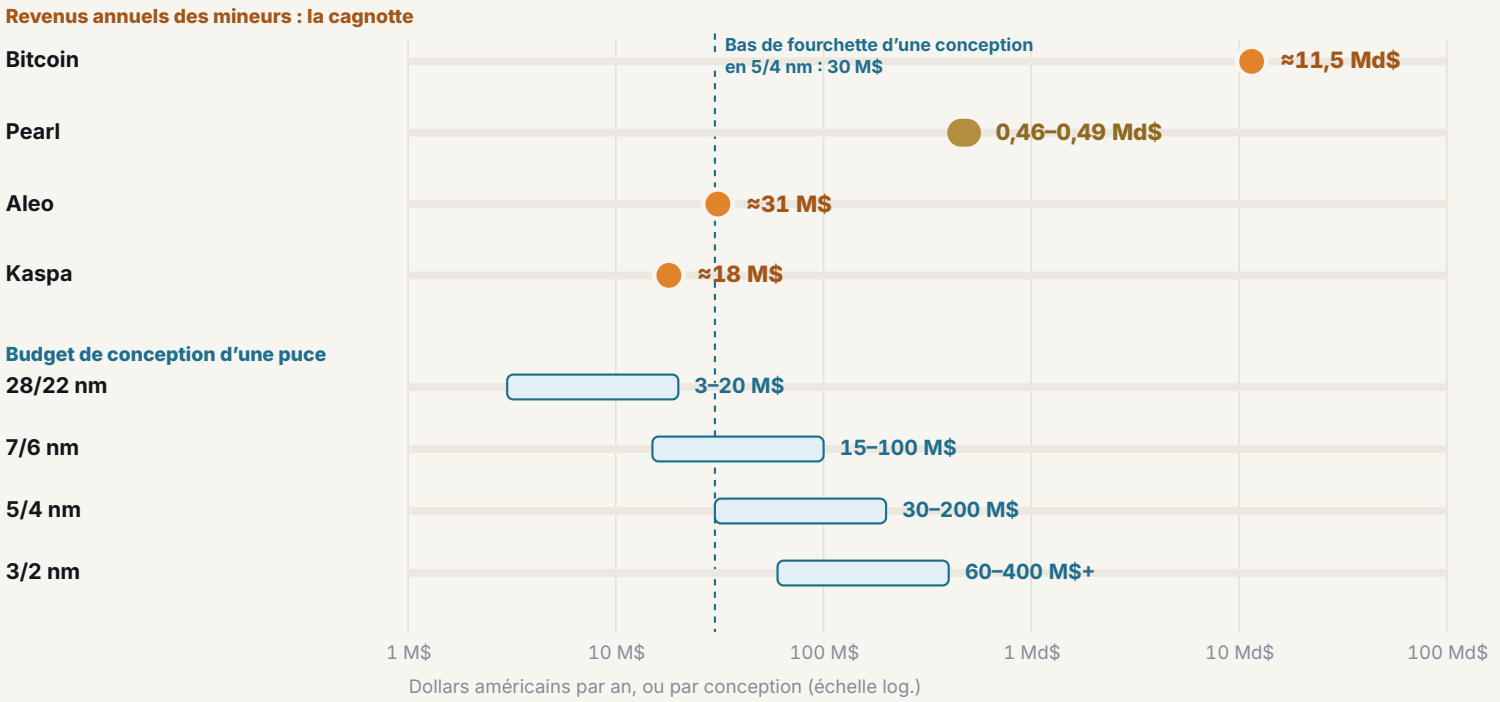
Quatre sociétés par actions sur infrastructure centralisée valent environ 8,5 milliards de dollars, à peu près autant que tous les tokens d'IA des couches protocole et application.

Pourquoi l'efficacité cesse de payer au-delà de k = 5



Plafond de profit ≈ R × (1 − 1/k). Dès que les ASIC deviennent les mineurs marginaux, k se réinitialise face aux autres ASIC ; face à des GPU inutilisés qui ne paient que l'électricité, k vaut environ 2,5 estimation.

La plupart des cagnottes valent moins qu'une conception de puce



Hashrate d'amortissement h* = NRE ÷ [R × (1 − 1/k) × T × p]. Pour Pearl, avec 20 millions de dollars de NRE et k = 5 : environ 4,9 EH/s pendant un an si les règles tiennent, 9,8 EH/s si elles n'ont qu'une chance sur deux de tenir **estimation**.

Cinq types de puces, cinq réponses

Puce de minage PoUW à la Pearl



PAS MAINTENANT

k : sans HBM ni unités généralistes : 3–10× par op. (déduit)

Second débouché : inférence d'IA, si GEMM INT8/FP8 généraliste

Parier sur une puce d'inférence capable de miner, pas sur un composant INT7 dédié

Puce de minage bitcoin SHA-256



S'ABSTENIR

k : ≤1 face aux acteurs établis

Second débouché : presque aucun

Seulement avec électricité propre, forme de niche, ou en IP

Puce d'entraînement d'IA



DE L'IP, PAS DES PUCES

k : 1,1–1,8× à précision égale

Second débouché : —

Packaging et HBM décident ; faire IP, chiplets, services

Puce d'inférence d'IA



CONCEVOIR

k : renoncer à la généralité : 10–20× revendiqué

Second débouché : c'est le marché principal

Signer le client, puis choisir le verrouillage

Accélérateur de preuve ZK



ATTENDRE ET SE COUVRIR

k : 10–100× les GPU revendiqué

Second débouché : limité hors du ZK

Moteurs MSM et NTT programmables ; FPGA d'abord

Accélérateur FHE



SE POSITIONNER TÔT

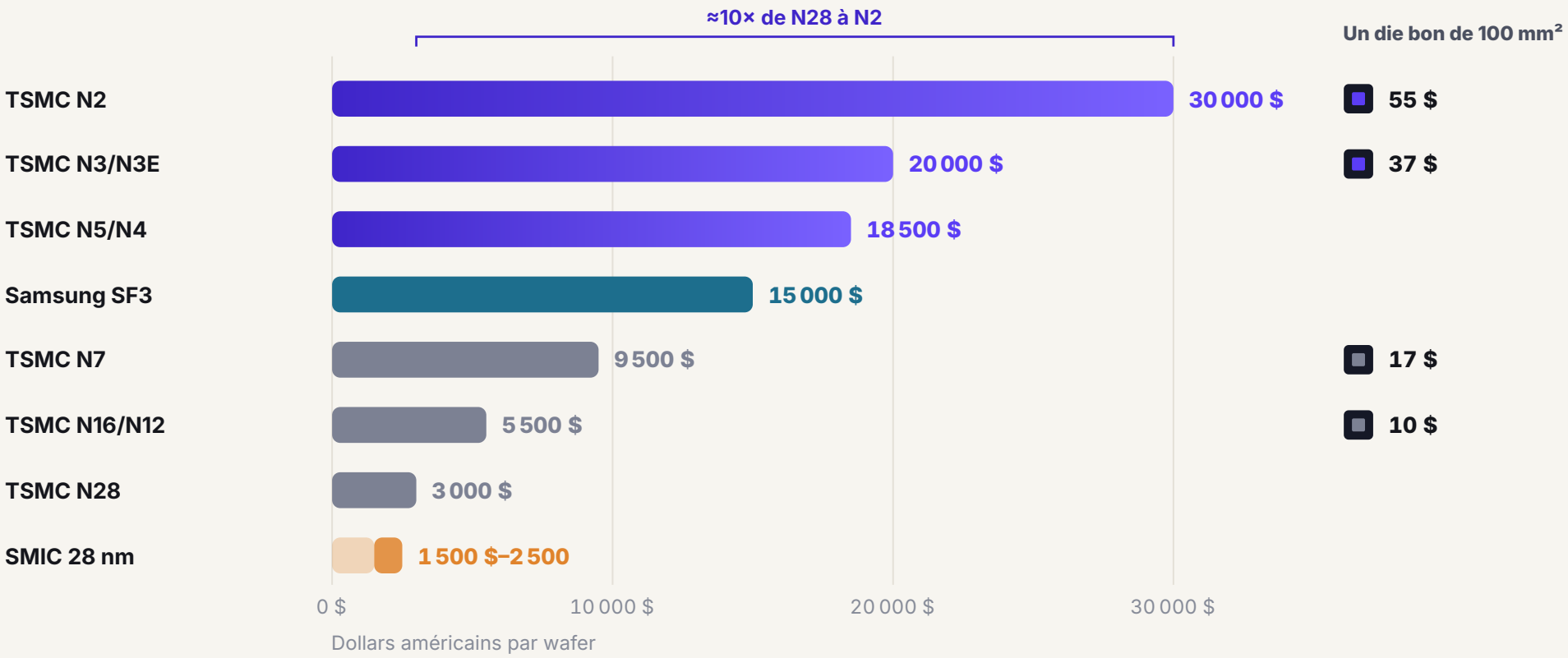
k : 1 074–5 547× un Xeon (Heracles)

Second débouché : calcul chiffré, IA privée

Une fenêtre 2027–2029 ; valider avec peu d'actifs

> Aucun réseau n'offre aujourd'hui à la fois une grande cagnotte, des règles stables et un k supérieur à 1 : le k du bitcoin appartient aux acteurs établis, les règles de Pearl sont les moins stables, les cagnottes de Kaspa et d'Aleo trop petites.

Un wafer de pointe coûte dix fois un wafer mature



> Laissez le coût le plus lourd sur la durée de vie choisir : l'énergie pour les puces de calcul en continu, packaging et HBM pour l'entraînement (63 % du coût d'un H100), le choix SRAM ou HBM pour l'inférence.

Un nœud pour chaque puce et chaque équipe

Puce	Petite équipe	Entreprise à l'échelle	Voie de la Chine continentale
Puce de minage PoUW	MPW 28 ou 16/12 nm ; premier volume N7/N6	N3/N3E ou Samsung SF3	SMIC N+1/N+2 praticable, environ 40–50 % de plus par die
Puce de minage SHA-256	Ne pas courir après la frontière	Classe 3 nm	Difficile d'égaler l'efficacité de pointe
Puce d'entraînement d'IA	IP, chiplets, services de conception	N3P puis N2/A16 avec CoWoS et HBM	SMIC N+2/N+3 avec HBM nationale
Puce d'inférence d'IA	Centrée SRAM en N6/N7 ou N4/N5/SF4	N3/N2 avec HBM, ou chiplets SRAM	SMIC N+2 avec LPDDR ou GDDR
Accélérateurs ZK et FHE	FPGA ou MPW 16/12 nm d'abord	N5/N3 ; FHE avec HBM	ZK praticable ; FHE limité par la HBM

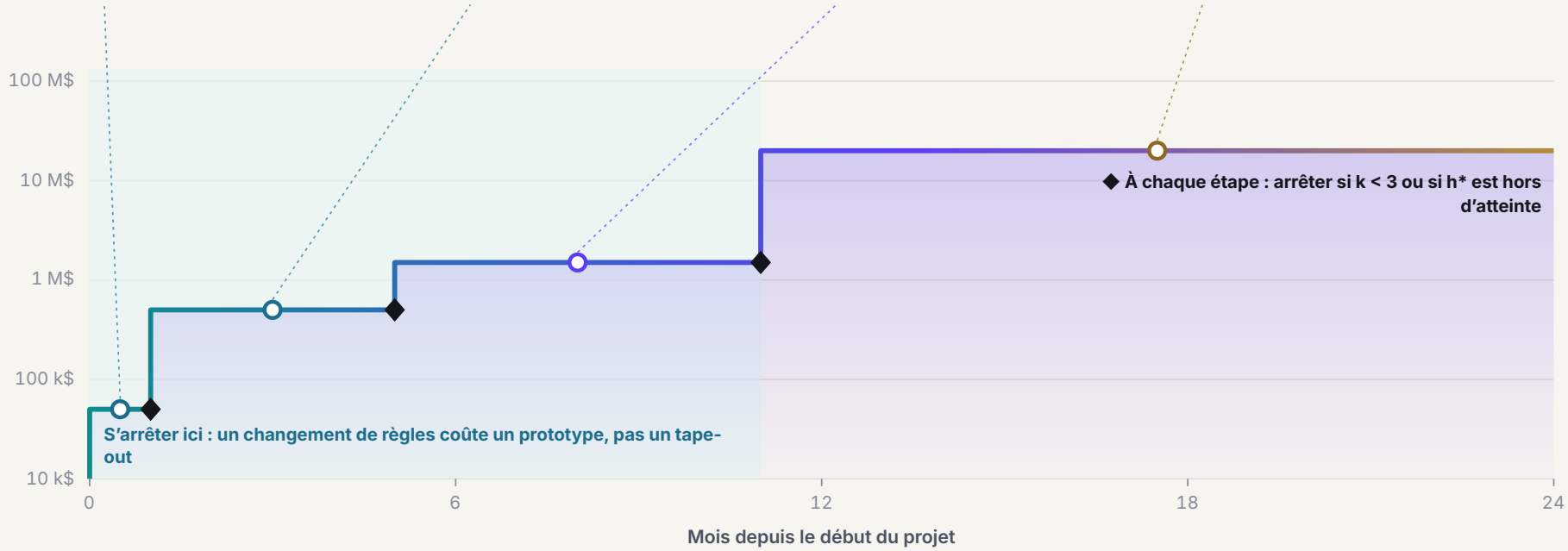
Acheter des preuves avant les masques

Étape 0 · Modèle
≈50 k$ · semaines
J/op et $/op d'après l'arithmétique, la SRAM et les E/S

Étape 1 · FPGA
≈0,5 M$ · mois
La vraie charge de bout en bout : parts ou tokens

Étape 2 · MPW
≈1,5 M$ · mois 11
Le silicium mesuré remplace le modèle

Étape 3 · Masques et premier lot
≈20 M$ · mois 24
Masques N7 ≈3,5 M$ dans une conception de 15–100 M$



Un modèle, un FPGA et un MPW permettent d'obtenir du silicium mesuré pour environ 1,5 million de dollars en onze mois ; le jeu de masques et le premier lot coûtent plus de dix fois plus. S'arrêter avant les masques tant que l'acheteur ou la version du protocole n'est pas fixé.

OPPORTUNITÉS

La crypto comme premier client, l'IA comme second débouché

Puces sur mesure

- La demande est massivement IA : Broadcom environ 58 milliards de dollars de semi-conducteurs d'IA sur l'exercice 2026
- Les compétences du minage se transfèrent : réseaux basse tension, alimentation kilowatt, tape-outs rapides
- Gagner avec un accès au marché indépendant, pas avec les canaux du minage

Développement d'IA

- Les heures sont une marchandise : 25–49 dollars de l'heure sur Clutch
- La valeur est aux jointures : noyaux pour puces sur mesure, minage intégré au service d'inférence
- Inférence vérifiable : TEE plus vérifications ponctuelles ZK

Systèmes IA × crypto

- Vérification en couches : TEE ou échantillonnage, puis contestations ou vérifications ZK
- Les barrières sont la distribution et la liquidité, pas le coût
- Signer l'acheteur de calcul avant le token

Trois voies, chacune avec un petit premier pas

Voie	Premier pas
A. Un accélérateur d'abord pour l'inférence, capable aussi de miner	Une étude de quatre semaines du J/op et du $/op d'un réseau INT8/FP8 résident en SRAM face à la RTX 5090 et au H100, puis FPGA ou MPW en nœud mature
B. Services de silicium sur mesure : la crypto d'abord, l'IA ensuite	Transformer le modèle de NRE et de point mort de ce rapport en études de faisabilité payantes, avec jalons et version de protocole figée
C. Un réseau natif IA avec un acheteur avant un token	Une note de conception sur la vérification en couches et une arithmétique canonique neutre, rédigée avec un acheteur d'inférence nommé

Au plus trois voies, chacune testée avant d'engager de l'argent analyse **.**

À SURVEILLER

Ce qu'il faut surveiller, et les questions à poser

JALONS

À tout moment	La hauteur d'activation FP8 de Pearl, et une éventuelle scission de la chaîne
Chaque mois	La part des blocs de Pearl portant du travail d'IA
Chaque trimestre	Revenus vérifiables de Bittensor face aux émissions ; flux x402 initiés par des agents
Janvier 2027	Hausse de prix de 3–6 % de TSMC pour le 2 nm et le 3 nm, selon la presse
T1 2027	Le FP4 dans Pearl, selon l'AMA des fondateurs tel que rapporté
Avril 2028	Le prochain halving du bitcoin

QUESTIONS À POSER

01 Qui contrôle la fonction de travail ?

02 Quel est le multiple de subvention ?

03 Que coûte la prochaine étape, et qu'est-ce qui vous arrête ?

04 Quel est le hashrate d'amortissement, avec p ?

05 Quel est le second débouché de la puce ?

06 Qui est le premier client payant, nommément ?

Ce sont des domaines que nous pouvons explorer ensemble, par l'échange technique et la collaboration technique : puces sur mesure, développement d'IA et systèmes IA × crypto.

i@tanshuai.com · tanshuai.com