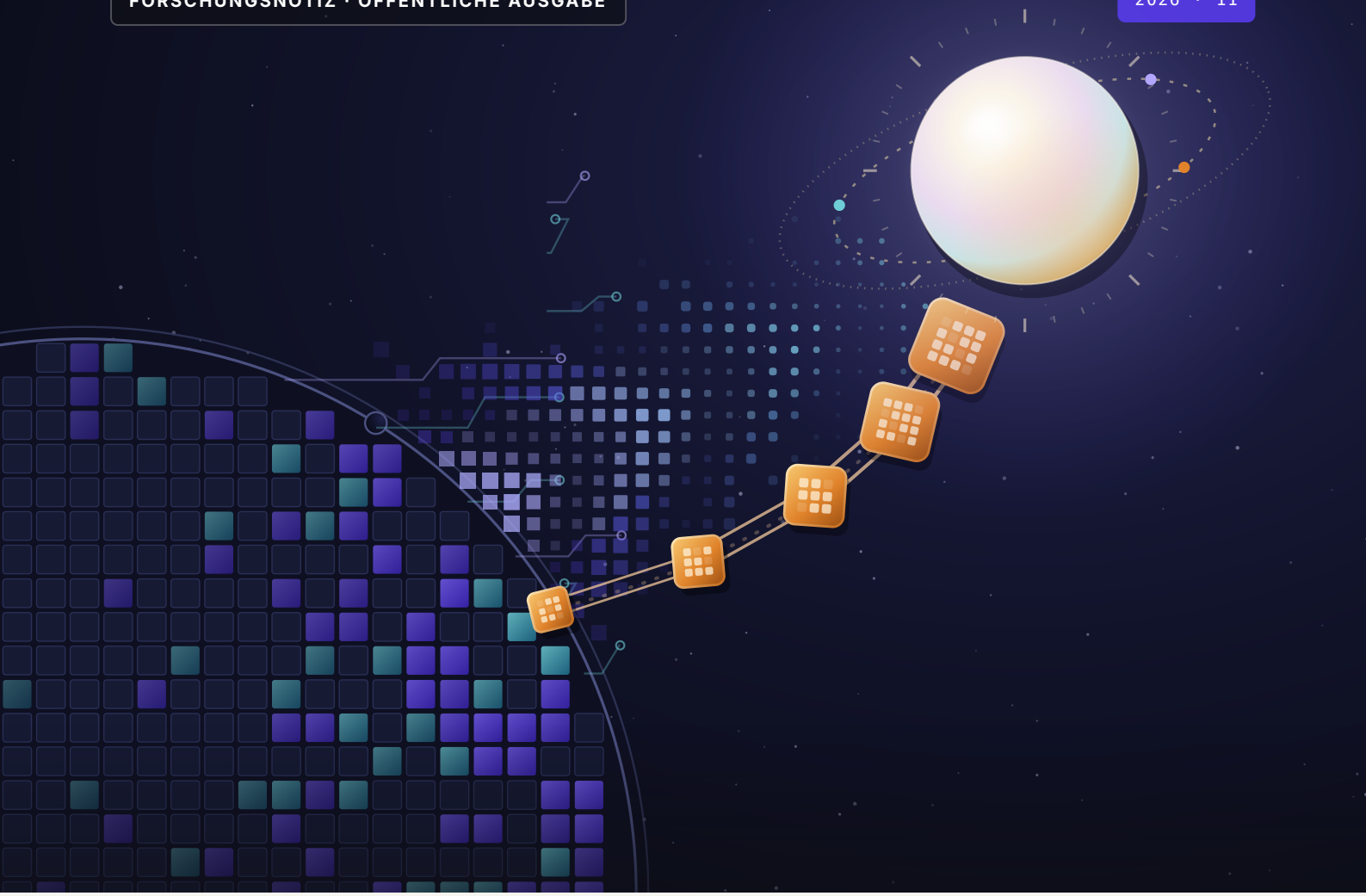
FORSCHUNGSNOTIZ · ÖFFENTLICHE AUSGABE

2026 · 11



PROOF OF USEFUL WORK · KRYPTO × SILIZIUM

Rechenleistung als Konsens

KI × Krypto × Chips 2026, ausgehend von Pearls Proof of Useful Work

Pearl · Nützliche Arbeit · DePIN und Agenten · ASIC-Ökonomie · Prozessknoten · Gestufte Budgets

DAS FAZIT

Pearl beweist, dass Matrixmultiplikation eine Blockchain absichern kann, aber noch nicht, dass die Arbeit nützlich ist. Im gesamten Feld KI × Krypto liegt die bezahlte Nachfrage bei Inferenz, Daten und Privatsphäre, und ein ASIC rechnet sich nur, wo der Topf groß ist, die Regeln stabil sind und der Chip einen Zweitmarkt hat.

Tan Shuai

Technologe und Gründer · KI-Forscher und -Entwickler

tanshuai.com · i@tanshuai.com

Datenstand: 13. November 2026

Beträge in US-Dollar

INHALT

Inhalt



◆ **Zusammenfassung** — Das Theorem hält, der Anreiz nicht — 04

01 **Umfang und Methode** — Umfang, Methode und vier Prüfsteine — 07

02 **Pearl · der Mechanismus** — Pearl: Matrixmultiplikation als Konsens — 08

03 **Pearl · die Ökonomie** — Pearl: Größe, Emission und die Nützlichkeitslücke — 10

04 **Pearl · die Regeln** — Pearl: Regelrisiko, Numerik und Governance — 12

05 **Ketten mit nützlicher Arbeit** — Ketten mit nützlicher Arbeit: Weder Matrizen noch Zero Knowledge stoppen ASICs — 13

06 **Das weitere Feld** — Der Rest von KI × Krypto: reale, aber kleine Nachfrage, und Wert wandert ins Eigenkapital — 16

07 **ASIC-Ökonomie** — Die ASIC-Frage: eine Formel, drei Maßstäbe — 19

08 **Chip für Chip** — Fünf Chiptypen, fünf Antworten — 22

09 **Prozessknoten** — Prozessknoten: Die größten Lebenszeitkosten sollen entscheiden — 26

10 **Ein gestuftes Budget** — Vom Modell zum Silizium: ein gestuftes Budget für einen ersten Chip — 29

11 **Chancen** — Wo die Chancen liegen, und was ich tun würde — 31

12 **Wegmarken** — Worauf zu achten ist, November 2026 bis 2027 — 34

13 **Fragen, die man stellen sollte** — Fragen, die Sie nicht gestellt haben, aber hätten stellen sollen — 35

◆ **Über den Autor** — Tan Shuai — 37

◆ **Quellen** — 38

ABBILDUNGEN

Abbildungen



ABB. 01	KI × Krypto in sechs Sektoren: wo bezahlte Nachfrage real ist	06
ABB. 02	Bei Pearls Ganzzahlarbeit teilen GPUs eine Energiegrenze	09
ABB. 03	Pearls erste sechs Monate: Hashrate folgt dem Preis, Regeln ändern sich ständig	11
ABB. 04	Das Subventionsvielfache: wie viele externe Einnahmen ein Netz zahlt	15
ABB. 05	Warum Effizienz sich nicht mehr auszahlt: ein ASIC behält höchstens 1 − 1/k	20
ABB. 06	Die meisten Töpfe sind kleiner als das Designbudget eines Chips	21
ABB. 07	Fünf Chiptypen, fünf Antworten	25
ABB. 08	Ein Wafer am Leading Edge kostet zehnmal so viel wie ein ausgereifter	27
ABB. 09	Belege vor Masken kaufen: ein gestuftes Budget für einen ersten Chip	30

Markierungen im Text

Schätzung — Rechnung oder Schlussfolgerung des Autors aus zitierten Daten

Herstellerangabe — vom Anbieter des Produkts veröffentlicht, nicht unabhängig geprüft

Analyse — Deutung der Belege durch den Autor

Dies ist eine Übersetzung der englischen Ausgabe; bei Abweichungen ist die englische Ausgabe maßgeblich.

Datenstand: 13. November 2026 · Beträge in US-Dollar

ZUSAMMENFASSUNG

Das Theorem hält, der Anreiz nicht

DAS FAZIT

Pearl zeigt, dass „Rechenleistung als Konsens" als Technik funktioniert, aber noch nicht, dass sie nützlich ist.

≈90 EH/s
Pearls Netzwerk am 13. November 2026, wobei ein H eine INT7-Multiply-Accumulate-Operation ist

≈1 %
der Pearl-Blöcke enthalten echte KI-Arbeit (Blockworks, 28. Oktober 2026)

≈70 Mio. $
jährliche verifizierbare externe Einnahmen bei Bittensor, gegenüber mehr als 300 Millionen US-Dollar an Emissionen

175–180×
Aleos Prover-Belohnungen im Verhältnis zu seinen Gebühren im dritten Quartal 2025

R × (1 – 1/k)
die Gewinnobergrenze eines ASIC, der k-mal günstiger ist als der Grenzminer
Schätzung

≈1,5 Mio. $
bis zu gemessenem Silizium auf N7 über Modell, FPGA und MPW, vor einem Maskensatz für ≈3,5 Mio. $ Schätzung

Pearl

Am 13. November 2026 lief Pearls Netz mit rund 90 EH/s. In seinem Konsenscode ist ein „Hash" ein ganzzahliges Multiply-Accumulate, sodass das Netz dem Ganzzahldurchsatz von etwa 122.000 H100 entspricht. Doch nur etwa 1 % der Blöcke enthält echte KI-Arbeit, und die einzige öffentliche Inferenzpartnerschaft wird in der Serverless-API ihres Anbieters nicht mehr angeboten. Die Ursache ist ökonomisch, nicht kryptografisch: Eine GPU, die mint, während sie Inferenz bedient, erzielt nur etwa 1 % bis 27 % der Hashrate eines reinen Miners, also multiplizieren rationale Miner Zufallsmatrizen.

Bezahlte Nachfrage

Im gesamten Feld KI × Krypto zeigt sich bezahlte Nachfrage, die sich verifizieren lässt, an drei Stellen: beim Verkauf von Inferenz, von Daten und von Privatsphäre. Bittensors verifizierbare externe Einnahmen liegen bei etwa 70 Millionen US-Dollar pro Jahr, rund vier Fünftel davon aus Subnetzen für Rechenleistung und Infrastruktur. Grass meldet 17 Millionen US-Dollar Datenerlöse im ersten Halbjahr 2026. Vertrauliche Inferenz in Trusted Execution Environments ist in Produktion. Dagegen kommen tatsächlich von KI-Agenten ausgelöste Zahlungen auf etwa 5.000–11.000 US-Dollar im Monat, und für zkML und dezentrales Training ist bisher keine zahlende Nachfrage offengelegt.

Subvention und Wert

Das Subventionsvielfache ist der ehrlichste Gesundheitscheck. Aleos Prover-Belohnungen betrugen etwa das 175- bis 180-Fache seiner Gebühren; Bittensors Emissionen liegen beim etwa 4- bis 11-Fachen seiner verifizierbaren Einnahmen; Renders Marktwert beträgt etwa das 360-Fache seiner On-Chain-Gebühren; und Pearls Miner-Einkommen von rund 460 Millionen US-Dollar pro Jahr ist größtenteils Blocksubvention, nahezu unverbunden mit bezahlter KI-Nachfrage. Subvention ist Treibstoff für einen Kaltstart, kein Geschäftsmodell. Zugleich wandert

Wert von Token zu Eigenkapital: Vier Unternehmen mit Eigenkapital, die zentrale Infrastruktur und kryptonativen Vertrieb nutzen, werden mit etwa 8,5 Milliarden US-Dollar bewertet, ungefähr so hoch wie die KI-Token der Protokoll- und Anwendungsebene zusammen.

ASICs

Für Chips gilt: Die Gewinnobergrenze eines ASIC ist ungefähr der Erlös je Recheneinheit mal (1 – 1/k), wobei k angibt, wie viel günstiger er ist als die Grenzhardware. Weil dieser Term konkav ist, bringt Effizienz jenseits von k = 5–10 fast keinen zusätzlichen Gewinn; über das Ergebnis entscheiden die Größe des Topfs, wie lange die Regeln Bestand haben und ob der Chip einen Zweitmarkt hat. Meine Antworten nach Chiptyp: jetzt keinen dedizierten INT7-Chip für Pearl zum Tape-out bringen, sondern auf einen inferenzorientierten Beschleuniger setzen, der auch minen kann; als Neueinsteiger dem freien Markt für Bitcoin-Miner fernbleiben; als kleines Team keinen eigenständigen Trainingschip bauen; Inferenzchips bauen, wenn ein Kunde unterschrieben hat; bei ZK-Beschleunigern absichern und sich bei FHE früh positionieren.

Knoten und Budgets

Der Prozessknoten sollte nach dem größten Posten der Lebenszeitkosten des Chips gewählt werden: Energie bei Chips, die rund um die Uhr laufen, Packaging und Speicher bei Trainingschips, die Entscheidung SRAM oder HBM bei Inferenzchips. Und ein erster Chip sollte in Stufen eingekauft werden, ein Modell, ein FPGA, ein Multi-Project-Wafer und erst dann ein Maskensatz, damit eine Änderung des Protokolls den Preis eines Prototyps kostet und nicht den eines Tape-outs. Die Chancen liegen dort, wo sich die Felder kreuzen: Krypto als erster Kunde, KI als Zweitmarkt.

Frage	Urteil	Kernzahlen (Quellen in den Kapiteln)
① **Funktioniert Pearl, und ist seine Arbeit nützlich?**	Es funktioniert: eine Blockchain, die in großem Maßstab durch Matrixmultiplikation gesichert ist. Noch nicht nützlich: Fast das gesamte Mining läuft auf Zufallsmatrizen	Rund 90 EH/s am 13. November 2026; 1 H = 1 INT7-Multiply-Accumulate; etwa 1 % der Blöcke mit KI-Arbeit
② **Wo ist die echte bezahlte Nachfrage in KI × Krypto?**	Inferenz, Daten und Privatsphäre. Agentenzahlungen, zkML und dezentrales Training liegen noch vor der Nachfrage	Bittensor etwa 70 Millionen US-Dollar verifizierbare Einnahmen pro Jahr; Grass 17 Millionen US-Dollar im 1. Halbjahr 2026; Agentenzahlungen etwa 5.000–11.000 US-Dollar im Monat
③ **Lohnt sich der Bau eines ASIC?**	Nur wo der Topf groß ist, die Regeln stabil sind und der Chip einen Zweitmarkt hat. Heute bietet kein Netz alle drei	Gewinnobergrenze ≈ R × (1 – 1/k); Bitcoin etwa 11,5 Milliarden US-Dollar Miner-Einnahmen pro Jahr, Pearl etwa 0,46–0,49 Milliarden, Kaspa etwa 18 Millionen
④ **Welche Chips sollte ein kleines Team bauen?**	Inferenzorientierte Beschleuniger, die auch minen können; Inferenzchips für einen unterschriebenen Kunden; frühe FHE-Arbeit. Keine Bitcoin-Miner für den freien Markt und keine Trainingschips	Kaspas bester ASIC verdient 0,86 US-Dollar am Tag; Designbudgets bei 3/2 nm 60–400 Millionen US-Dollar oder mehr
⑤ **Welcher Prozessknoten?**	Die größten Lebenszeitkosten entscheiden: Energie, Packaging und Speicher oder die Speicherarchitektur	TSMC N2 etwa 30.000 US-Dollar je Wafer gegenüber etwa 3.000 bei N28; HBM und CoWoS etwa 63 % der Kosten einer H100
⑥ **Was kostet ein erster Chip, und wie sollte man ihn einkaufen?**	In vier Stufen: ein Modell, ein FPGA, ein Multi-Project-Wafer, dann Masken. Vor den Masken stoppen, wenn die Regeln oder der Käufer nicht feststehen	Ein kommerzieller MPW-Platz 50.000–500.000 US-Dollar; N7-Masken etwa 3,5 Millionen US-Dollar; ein Designbudget bei 7/6 nm 15–100 Millionen US-Dollar

ABB. 01

KI × Krypto in sechs Sektoren: wo bezahlte Nachfrage real ist

Verifizierbare bezahlte Nachfrage gegenüber Subventions- und Bewertungssignalen, November 2026

Ketten mit nützlicher Arbeit 01

REAL, ABER KLEIN

≈70 Mio. $/Jahr

verifizierbare externe Einnahmen von Bittensor

Signal: Emissionen 4–11× Einnahmen; Pearl ≈1 % KI-Blöcke

Die Mechanismen funktionieren; Nutzen braucht Zahler

Dezentrale Rechenleistung 02

REAL, ABER KLEIN

2,35 Mio. $/Jahr

Renders On-Chain-Gebühren, die größten verifizierbaren

Signal: Marktwert ≈360× Gebühren

Meist Weiterverkauf von Standard-Inferenz

Training und Inferenz 03

GEWINN IM EIGENKAPITAL

100+ Mio. $/Jahr

Prime Intellect, ein Unternehmen mit Eigenkapital auf zentralen GPUs

Signal: Gensyns Trainings-Mainnet nicht live

Fortschritt ist real; Gewinn geht ans Eigenkapital

Verifizierbare KI 04

FRÜH

398

zahlende Nutzer von Phalas TEE-Cloud, Ende 2025

Signal: kein offengelegtes bezahltes zkML

TEEs in Produktion; zkML früh

Agentenzahlungen 05

VOR DER NACHFRAGE

5–11 Tsd. $/Monat

tatsächlich von Agenten ausgelöste x402-Zahlungen

Signal: ≈200 Mio. Transaktionen, meist Signalisierung

Die Schienen kamen zuerst; die Nachfrage nicht

Datennetzwerke 06

ECHTE KÄUFER

17 Mio. $

Datenerlöse von Grass, 1. Halbjahr 2026 (Eigenangabe)

Signal: Marktwert ≈10× annualisierte Einnahmen

Einer der wenigen Sektoren mit echten Käufern

Quelle: DefiLlama, SubConnect (über hoge.gg), TRM Labs, Phala, Dealroom, SiliconANGLE, CoinGecko, Blockworks; Synthese des Autors.

01 UMFANG UND METHODE
Umfang, Methode und vier Prüfsteine

Warum dieser Bericht, und warum jetzt

Im April 2026 brachte eine neue Blockchain, Pearl, eine lange gehegte Idee in Produktion: Miner sollen um Blöcke konkurrieren, indem sie die Matrixmultiplikationen ausführen, auf denen KI läuft, statt zu hashen. Binnen sechs Monaten erreichte ihr Netz rund 90 EH/s und ihr Token einen Marktwert von etwa 500 Millionen US-Dollar. Zur gleichen Zeit sammelten KI-Rechenmarktplätze, dezentrale Trainingsnetze, Zahlungsschienen für Agenten und Datennetzwerke Milliarden ein, und Bitcoin-Miner begannen, ihre Standorte mit Stromanschluss in KI-Rechenzentren umzuwandeln. Ich wollte wissen, welche davon echte Kunden haben, ob kundenspezifisches Silizium dort irgendwo sinnvoll ist und welchen Prozessknoten und welche Foundry ein Chipteam tatsächlich nutzen kann.

Vier Prüfsteine für jedes Projekt aus KI × Krypto

Eine Sicherheitsstudie zu Gridcoin aus dem Jahr 2017 argumentierte, dass nützliche Berechnung eine Blockchain nur dann absichern kann, wenn sie schwer zu lösen, leicht zu verifizieren und leicht zu belohnen ist. Ich füge einen vierten Prüfstein hinzu. Zusammen bestimmen sie den Großteil dessen, was folgt.

01 **Schwer zu lösen.** Ehrliche Berechnung muss der billigste Weg sein, einen gültigen Beweis zu erzeugen; es darf keine Abkürzung geben.

02 **Leicht zu verifizieren.** Die Prüfung der Arbeit muss weit weniger kosten als die Arbeit selbst, sonst zahlt das Netz doppelt.

03 **Leicht zu belohnen.** Das Protokoll muss die Belohnung allein aus dem Beweis berechnen können, ohne einem Dritten zu vertrauen.

04 **Jemand bezahlt für das Ergebnis.** Solange kein Kunde außerhalb des Netzes für das Resultat bezahlt, ist die „nützliche“ Arbeit eine Subvention mit einer Erzählung.

Hash-Rätsel bestehen die ersten drei Prüfsteine und scheitern am vierten. Echte KI-Aufgaben bestehen den vierten, scheitern aber am zweiten: Eine KI-Berechnung in Zero Knowledge zu beweisen, kostet das 100- bis 10.000-Fache der Berechnung selbst, und für Proof-of-Learning-Verfahren wurde gezeigt, dass sie sich billig fälschen lassen. Pearls verrauschte Matrixmultiplikation ist die erste glaubwürdige Brücke, die den Verifikationstest besteht und dabei dieselbe Form wie KI-Rechenlast behält.

Wie ich gearbeitet habe

Dies ist Schreibtischrecherche, so durchgeführt, wie ich einen Chip oder einen Geschäftsplan prüfen würde. Ich habe Pearls Code, seine Protokollvorschläge und sein Whitepaper gelesen; Chain-Daten eines unabhängigen Full Node; Unternehmensberichte und Quartalsmitteilungen; Fachartikel aus Kryptografie, Rechnerarchitektur und maschinellem Lernen; Preisdaten von Foundrys und Anbietern; sowie Fach- und Technologiepresse. Jede Zahl trägt ihr Datum. Wo ich eine Zahl selbst herleite, sage ich das und markiere sie mit Schätzung. Zahlen, die nur der Ver-

käufer angibt, sind mit Herstellerangabe markiert; meine Deutung der Belege mit Analyse. Beträge sind in US-Dollar. Stand ist der 13. November 2026.

KAPITEL 02

02

PEARL · DER MECHANISMUS

Pearl: Matrixmultiplikation als Konsens

Die Idee, vom Theorem zum Mainnet

Im April 2025 veröffentlichten Ilan Komargodski und Omri Weinstein „Proofs of Useful Work from Arbitrary Matrix Multiplication". Sie konstruieren einen Proof of Useful Work für jedes Matrixprodukt mit einem Overhead von nur dem 1 + o(1)-Fachen einer naiven Multiplikation; seine Sicherheit beruht auf der Schwierigkeit, Stapel zufälliger linearer Gleichungen niedrigen Rangs zu lösen, was eine Vermutung ist, kein Beweis. Das Paper benennt die eigentliche Schwierigkeit des Feldes: Ein Proof of Useful Work mit konstantem Overhead-Faktor ist trivial, aber nutzlos; schwierig ist es, ehrliche Berechnung als billigsten Weg zum Gewinn zu erhalten, wenn Miner ihre Eingaben selbst wählen.

Pearls Mainnet startete am 27. April 2026 um 09:00 UTC. Miner berechnen ein verrauschtes Produkt (A + E)(B + F), wobei das Rauschen niedrigen Rangs aus einem BLAKE3-Commitment abgeleitet wird, und entfernen dann das Rauschen, um A·B zurückzugewinnen. Der Akkumulatorzustand jeder Ausgabekachel wird zu einem Los gehasht; eine Gewinnerkachel wird geöffnet und mit einem rekursiven Plonky2-Zero-Knowledge-Beweis von unter 60 KB bewiesen, innerhalb eines Blockzertifikats mit einer Obergrenze von 65 KB. Die Verifikation ist ein kleiner Beweis konstanter Größe, unabhängig von der Matrixgröße: so billig zu prüfen wie ein Hash-Rätsel, während die Berechnung selbst nützlich sein kann.

Ein Multiply-Accumulate ist ein „Hash"

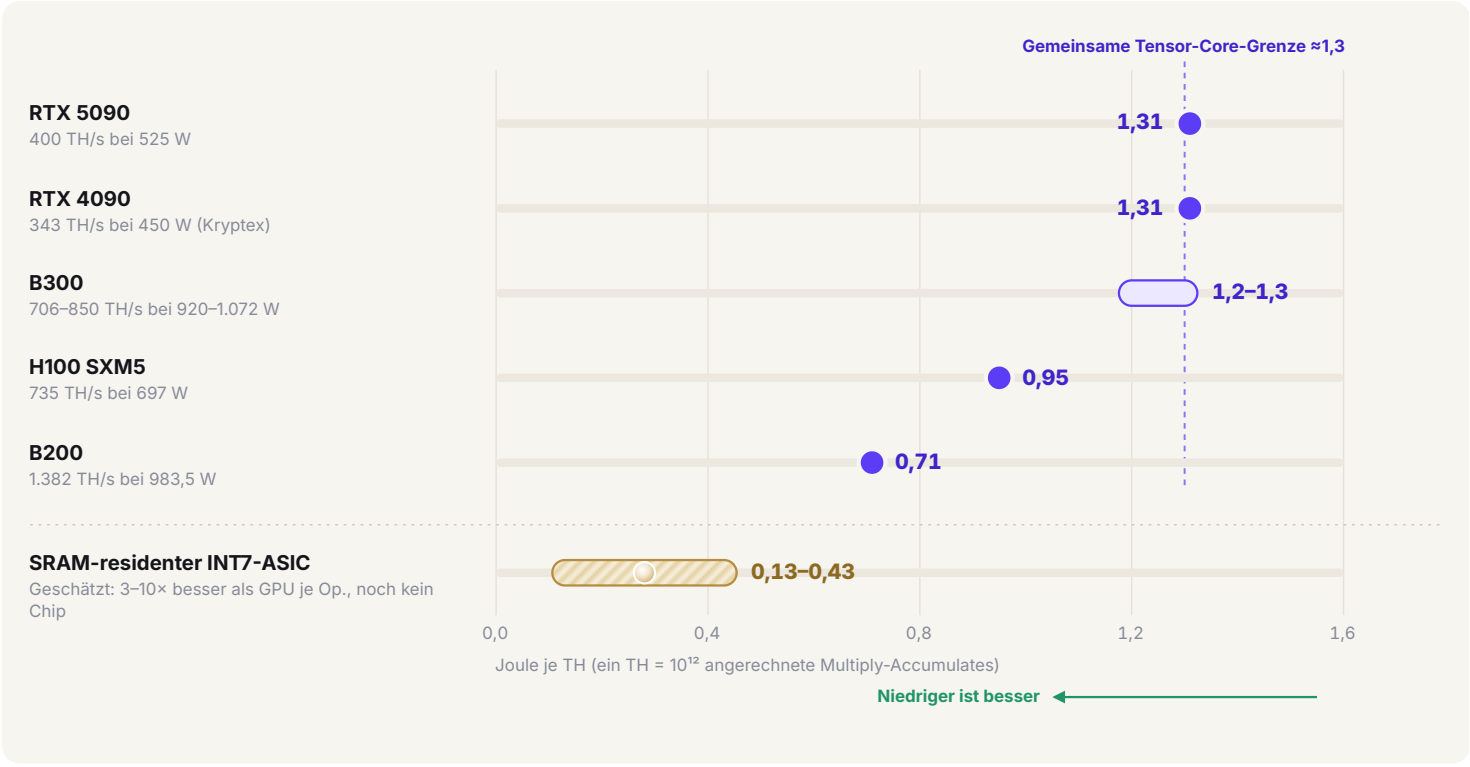
Für einen Chipdesigner ist die wichtigste Tatsache die Einheit. Pearls Konsenscode skaliert das Ziel jedes Versuchs mit der Kachelgröße mal der Länge des Skalarprodukts, also ist **1 H genau ein angerechnetes INT7 × INT7 → INT32-Multiply-Accumulate (MAC)**. Seit Block 96.251 verlangt eine Rang-Strafe einen Rauschrang von mindestens 128; beim Mindestrang ist 1 H weiterhin gleich 1 MAC. Pearls „EH/s" bedeutet daher buchstäblich 10^18 ganzzahlige Multiply-Accumulates pro Sekunde: **Erstmals ist das Sicherheitsbudget einer Blockchain im Tensordurchsatz von KI-Beschleunigern bepreist.** Gemessene GPU-Hashraten liegen bei 60 % bis 95 % der jeweiligen dichten INT8-Multiply-Accumulate-Spitzenleistung einer Karte, was zu dieser Lesart passt.

Hardware	Pearl-Hashrate	Leistung	Joule je Billion MACs (J/TH)	Quelle (Datum)
RTX 5090	400 TH/s	525 W	1,31 Schätzung	Kryptex (12. November)
RTX 4090	etwa 305 TH/s (Spanne der Community 280–330); Kryptex nennt 343 TH/s	450 W (Kryptex)	1,31 (Kryptex-Zahlen) Schätzung	prlstats (3. Oktober); Kryptex
H100 SXM5	735 TH/s	697 W	0,95 Schätzung	PeakMiner, über prlstats (19. September)
B200	1,382 PH/s	983,5 W	0,71 Schätzung	PeakMiner, über prlstats

Hardware	Pearl-Hashrate	Leistung	Joule je Billion MACs (J/TH)	Quelle (Datum)
B300	706–850 TH/s	920–1.072 W	1,2–1,3 Schätzung	prlstats; SimeonOnSecurity

ABB. 02 **Bei Pearls Ganzzahlarbeit teilen GPUs eine Energiegrenze**

Joule je Billion INT7-Multiply-Accumulates (J/TH) bei gemessener Pearl-Hashrate; niedriger ist besser



Quelle: Kryptex (12. November 2026); PeakMiner- und SRBMiner-Benchmarks über prlstats; SimeonOnSecurity; die ASIC-Spanne ist eine Schlussfolgerung des Autors, keine Messung.

Zwei Details dieser Tabelle sind für Chipdesigner wichtig. Erstens konvergieren die Consumer-Karten RTX 4090 und 5090 und die Rechenzentrums-GPU B300 bei etwa 1,3 J/TH, und selbst die beste Karte, die B200, erreicht nur etwa 0,71 J/TH. Eine Rechenzentrums-GPU für Zehntausende Dollar verbraucht je ganzzahligem Multiply-Accumulate fast so viel Energie wie eine Gaming-Karte, was auf eine gemeinsame Grenze der Tensor-Core-Energie je INT8-MAC hindeutet, und genau diese Grenze könnte ein dedizierter Chip durchbrechen. Zweitens ist die B300 langsamer als die B200: NVIDIA hat den INT8-Durchsatz in Blackwell Ultra stark gekürzt, sodass die Ganzzahlarbeit auf diesem Chip einen langsameren Weg nehmen muss. Gängiges KI-Silizium bewegt sich weg von INT8 hin zu FP8 und FP4. Wenn das Zahlenformat, das ein Konsens wählt, von der Richtung der Hardware abweicht, ist das Protokoll gezwungen, seine Regeln zu ändern, und das ist die Wurzel der FP8-Fork und meines ASIC-Urteils später in diesem Bericht.

KAPITEL 03

03 PEARL · DIE ÖKONOMIE
Pearl: Größe, Emission und die Nützlichkeitslücke

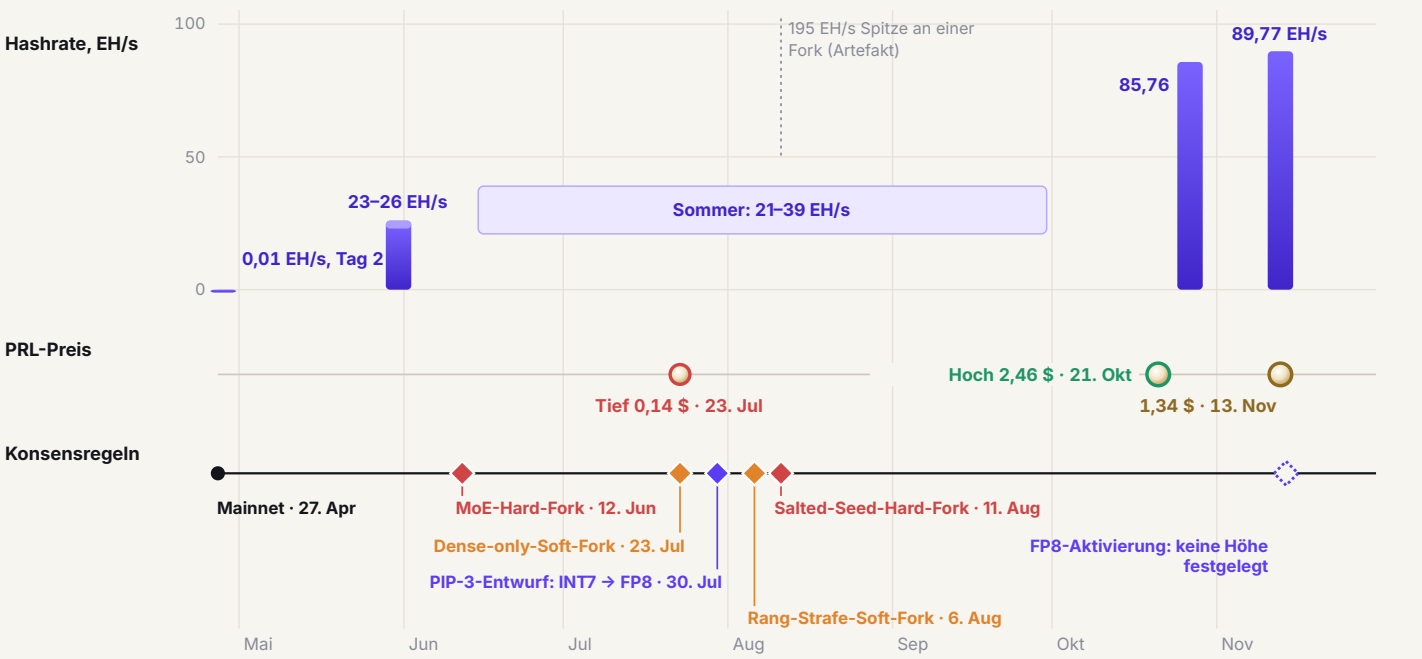
Von 0,01 auf 90 EH/s in sechs Monaten

Kennzahl	Wert (13. November 2026)	Grundlage und Quelle
Netzwerk-Hashrate	Rund 90 EH/s (Basiswert 89,77; Spanne am selben Tag 88,20–93,22; Tageshoch am 12. November von 110–123)	Unabhängiger Full Node prlstats; Kryptex
Difficulty und Blockhöhe	58.841.982 und 142.530	prlstats
Äquivalente Hardware	Etwa 122.000 H100 oder 225.000 RTX 5090	Bei 735 und 400 TH/s je Karte Schätzung
Preis, Marktwert und vollständig verwässerter Wert	1,32–1,34 US-Dollar; etwa 500 Millionen US-Dollar; etwa 2,8 Milliarden US-Dollar (CoinGecko-Rang 101)	CoinGecko
Allzeithoch und -tief	2,46 US-Dollar (21. Oktober) und 0,1399 US-Dollar (23. Juli)	CoinGecko
Geminte Coins	377,6 Millionen PRL, 17,98 % der Obergrenze von 2,1 Milliarden	prlstats
Blocksubvention und tägliche Emission	Etwa 2.173 PRL und etwa 0,97–1,03 Millionen PRL	prlstats; Kryptex
Neues Angebot in den nächsten 12 Monaten	Etwa 78 % des aktuellen Angebots	prlstats
Miner-Einkommen	Etwa 1,25 Millionen US-Dollar am Tag (Modell), etwa 460 Millionen US-Dollar pro Jahr	prlstats
Erlös je Einheit Hashrate, R	Etwa 14–15 US-Dollar je PH/s und Tag; eine RTX 5090 verdient etwa 5–6 US-Dollar am Tag	Schätzung; Kryptex
Blöcke mit echter KI-Arbeit	Etwa 1 % (Schätzung vom 28. Oktober)	Blockworks
Pool-Konzentration	Kryptex fand 46,4 % der Blöcke der letzten 24 Stunden; die beiden größten Pools etwa 62 %	prlstats
Konzentration der Halter	Gini-Koeffizient der Adressen 0,969; die 100 größten Adressen halten 53,0 %	prlstats

Pearls Hashrate ist nahezu eine nachlaufende Funktion seines Preises. Am Tag nach dem Start lief das Netz mit etwa 0,01 EH/s; bis Ende Mai erreichte es 23–26 EH/s. Über den Sommer fiel der Preis um etwa 86 % vom Mai-Hoch, und die Hashrate bewegte sich seitwärts bei 21–39 EH/s. Nachdem sich der Preis im September und Oktober etwa verzehnfacht hatte, erreichte die Hashrate am 27. Oktober 85,76 EH/s und im November rund 90 EH/s. Eine Spitze auf 195 EH/s am 11. August fiel mit einer Fork zusammen und ist als Schätzartefakt zu lesen. Der Energiebedarf wuchs mit: Eine Studie schätzte etwa 112 MW für das gesamte Netz bei 24 EH/s, und ein Modell von prlstats setzt ihn heute bei etwa 190–215 MW an.

ABB. 03 **Pearls erste sechs Monate: Hashrate folgt dem Preis, Regeln ändern sich ständig**

Netzwerk-Hashrate, drei Preismarken und jede Konsensänderung, 27. April bis 13. November 2026



Quelle: prlstats-Full-Node und -Zeitreihen; Kryptex; CoinGecko; Pearls params.go und PIP-3. Die Spitze am 11. August ist ein Schätzartefakt an einer Fork.

Die Emission verstärkt die Reflexivität

Die Angebotsobergrenze liegt bei 2,1 Milliarden Coins, und es gibt keine Halvings: Die Blocksubvention sinkt gleichmäßig nach E(t) = S·H / ((t + H)(t + H − 1)) mit H = 650.226. Die Emission wird jedoch in Blöcken gezählt, nicht in Zeit, und die anfängliche Difficulty startete zu niedrig und passte sich langsam an. **Etwa 44,06 Millionen Coins wurden am ersten Tag gemint und etwa 120 Millionen in den ersten sechs Tagen**, rund 32 % des heutigen Angebots Schätzung. Neuere Blöcke kommen mit etwa 171–183 Sekunden immer noch schneller als das Ziel von 194 Sekunden. Die Folge: Das Angebot wird in den nächsten 12 Monaten um etwa 78 % wachsen, eine Verwässerung, der sich Miner, Hardwarehersteller und Halter gleichermaßen stellen müssen.

Daraus folgt die Ökonomie der Miner. Das netzweite Miner-Einkommen liegt bei etwa 1,25 Millionen US-Dollar am Tag (Modell); die tägliche Emission laut Kryptex mal dem aktuellen Preis ergibt etwa 1,35 Millionen US-Dollar am Tag, also **Erlös je Einheit Hashrate R ≈ 14–15 US-Dollar je PH/s und Tag** Schätzung. Eine RTX 5090 verdient etwa 5–6 US-Dollar am Tag. Eine H100 oder H200 verdient etwa 11 US-Dollar, während die günstigste öffentliche On-Demand-Miete bei etwa 2,09 US-Dollar je Karte und Stunde liegt, rund 50 US-Dollar am Tag; Karten nur zum Minen zu mieten, bringt also hohe Verluste. R schwankt stark: Laut der Zeitreihe von prlstats lag es im August und September bei etwa 10 US-Dollar und Mitte Oktober kurzzeitig bei etwa 33 US-Dollar Schätzung. Binnen Wochen nach dem Start halbierte sich der geschätzte Tageserlös einer RTX 5090 von etwa 33,80 auf etwa 17,19 US-Dollar. Cloud-GPUs zum Minen zu mieten, lohnte sich nur in den ersten Wochen und während der Preisspitze im Oktober.

Das Theorem hält; der Anreiz nicht

Das eigentliche Problem ist die Nachfrage, nicht die Kryptografie. Pearls einzige öffentliche kommerzielle Integration ist ein Gemma-4-31B-it-pearl-Endpunkt von Together AI, gestartet am 15. Mai zu einem Preis mehr als 25 % unter dem üblichen, wobei der Rabatt aus dem erwarteten künftigen Wert der Emission finanziert wurde. Am 13.

November zeigte die Modellseite, dass er **in Togethers Serverless-API nicht verfügbar ist**. Im Juni kam eine Zählung von 8.012 Online-Mining-Instanzen, etwa 21 % des Netzes, zu dem Schluss, dass das Netz überhaupt keine nützliche KI-Berechnung erzeugte (der Autor nennt die Belege suggestiv, nicht schlüssig), und zeigte, dass Pools Shares aus rein zufälligen Matrizen akzeptierten. Am 28. Oktober schätzte Blockworks, dass **nur etwa 1 % der Blöcke echte KI-Arbeit enthält**; ohne sie, schrieb Blockworks, sei Pearl ein kleinerer, weniger sicherer, schneller inflationierender Bitcoin.

Der Grund lässt sich berechnen. Der Overhead auf der Inferenzseite ist tatsächlich gering: Pearl meldet eine Zunahme der End-to-End-Zeit um 5,08 % für Llama 70B auf vier H200, und ein unabhängiger Test maß einen Durchsatzverlust von 9,2 % auf einer Hopper-GPU. Die Mining-Seite bringt zu wenig ein. Das offizielle vLLM-Plug-in mint nur bei Prefill-Batches von mindestens 1.024 Token und nie bei Decode-Schritten. Eine Karte, die Inferenz bedient, erreicht etwa 27 % der Hashrate eines reinen Miners bei reinem Prefill-Verkehr, etwa 5,4 % bei einem Chat-Mix mit 80 % Decode und nur etwa 1 % bei Decode-dominiertem Verkehr. Ohne bezahlte Inferenznachfrage multipliziert ein rationaler Miner Zufallsmatrizen; das Protokoll verifiziert, dass die festgelegten Matrizen ehrlich multipliziert wurden, nicht, ob sie nützlich waren. **Das Theorem hält; der Anreiz nicht.** Pearls Nützlichkeit ist ein kommerzielles Problem außerhalb der Kette: Jemand muss für Inferenz bezahlen.

KAPITEL 04

04 PEARL · DIE REGELN
Pearl: Regelrisiko, Numerik und Governance

Vier Forks in zwei Monaten

Die andere Seite von Pearl ist, wie schnell sich seine Regeln ändern. Von Juni bis August 2026 setzte das Labor vier Konsensänderungen bei fest einprogrammierten Höhen durch: eine Mixture-of-Experts-Hard-Fork (Block 71.935, etwa 12. Juni), eine Dense-only-Soft-Fork (91.630, etwa 23. Juli), eine Rang-Strafe-Soft-Fork (96.251, etwa 6. August) und eine Salted-Seed-Hard-Fork (99.000, etwa 11. August). Nach der Salted-Seed-Fork konnten alte Miner nur noch ungültige Shares erzeugen; die Rang-Strafe senkte die effektive Hashrate von Konfigurationen mit hohem Rang um 50 % bis 87,5 %.

Der nächste Schritt ist PIP-3, noch ein Entwurf: Er verlegt das Mining von INT7 auf FP8 × FP8 → FP32 und verlangt bitgenaue Geräteprofile, **die nur zwei Architekturen unterstützen, Hopper und Blackwell**. Das Dokument nennt Hardwareabhängigkeit ein notwendiges Übel von Gleitkommaprotokollen und warnt, dass jede Toleranz bei der Verifikation Grinding-Angriffe einlädt. Am 12. November legten die Mainnet-Parameter des Zweigs fp8-scheme keine Aktivierungshöhe fest. Nach dem aktuellen Entwurf sind die RTX-40-Serie, die A100 und AMD-Karten nicht zugelassen, während ein Modell von prlstats mit geringer Sicherheit schätzt, dass etwa 82 % der Hashrate von Consumer-Karten stammen. Die Fork durchzudrücken, riskiert eine Spaltung der Kette oder eine Gegenreaktion der Miner.

Ein Trilemma der Zahlenformate

Pearl steht vor einem Trilemma: Ein Zahlenformat kann nützlich für KI (FP8 oder FP4), über Hardware hinweg bitgenau verifizierbar und hardwareneutral sein, aber höchstens zwei der drei Eigenschaften haben. Das INT7-Schema ist verifizierbar und neutral, weicht aber von der Richtung des KI-Siliziums ab. FP8 mit einer Whitelist ist nützlich und verifizierbar, bindet den Konsens aber an einen Anbieter. FP8 mit Toleranz ist nützlich und neutral, öffnet aber Grinding-Angriffen die Tür. Gleitkommaarithmetik in Tensor Cores folgt im Allgemeinen keinem nume-

rischen Standard und ist über Generationen nicht reproduzierbar; auf Hopper zum Beispiel behält die FP8-Akkumulation nur 13 Nachkommabits und schneidet ab. Aus Sicht eines Chipdesigners ist ein „Hopper-Profil" eine vollständig festgelegte Arithmetik, die ein ASIC Bit für Bit implementieren könnte, vielleicht effizienter als die GPU. Das wäre jedoch ein Geschäft in einer Grauzone, gegen die erklärte Absicht des Protokolls, mit sehr hohem Governance-Risiko Analyse.

Governance

Die Governance ist konzentriert. Das Labor hat jeden Protokollvorschlag geschrieben; Forks werden bei fest einprogrammierten Höhen ohne Signalisierung der Miner aktiviert; es gibt keine zweite Client-Implementierung; und bis Anfang Oktober war kein Audit des Nodes oder der Zero-Knowledge-Schaltungen durch Dritte veröffentlicht. Die israelische Zeitung TheMarker berichtete, das Labor habe in einer verdeckten Finanzierungsrunde etwa 120–130 Millionen US-Dollar eingesammelt (etwa 55 Millionen US-Dollar Eigenkapital plus einen GPU-Infrastrukturfonds von etwa 75 Millionen US-Dollar); Pearl hat das über offizielle Kanäle nicht bestätigt. „Keine Investorenzuteilung im Token" und „ein Unternehmen mit Risikokapitalgebern als Gesellschaftern" können beide zutreffen, doch die Spannung bleibt: KI-Einnahmen könnten ins Eigenkapital fließen, während die Verwässerung die Token-Halter trifft.

Insgesamt ist Pearl das bisher sauberste Experiment zu „Rechenleistung als Konsens": ein neuartiger Mechanismus, on-chain verifizierbar, mit einer Einheit, die sich eins zu eins auf KI-Hardware abbilden lässt. Doch die Nützlichkeit ist nicht angekommen, und Sicherheitsbudget, Hashrate und Begeisterung der Miner hängen an einem Token, dessen Hauptnachfrage sich erst noch bilden muss und dessen Angebot in 12 Monaten um etwa 78 % wachsen wird. Drei Signale von Blockworks lohnen die Beobachtung: ob das Upgrade reibungslos landet, ob der Anteil der Blöcke mit KI-Arbeit deutlich steigt und ob irgendein Inferenzanbieter Mining-Einnahmen an seine Kunden weitergibt.

KAPITEL 05

KETTEN MIT NÜTZLICHER ARBEIT

05 Ketten mit nützlicher Arbeit: Weder Matrizen noch Zero Knowledge stoppen ASICs

Kaspa: Eine Matrix ist weder nützlich noch ASIC-fest

Kaspa ist das Lehrbuchbeispiel. Sein kHeavyHash setzt ein 64 × 64-Matrix-Vektor-Produkt über 4-Bit-Werte zwischen zwei Hashes, und das Produkt wird verworfen. Etwa 16 Monate nach dem Fair Launch im November 2021 kündigte IceRiver am 14. März 2023 die ersten Kaspa-ASICs an, und bis Ende 2023 war das GPU-Mining vorbei. Der Vergleich einer RTX 4090 mit 3,8 GH/s und 300 W mit Bitmains KS7 bei 77 J/TH ergibt eine Effizienzlücke von etwa **1.000-fach** Schätzung. Nach dem Festmahl kam die Kompression: Am 13. November lief das Netz mit 358,9 PH/s, etwa 64 % unter seinem 1 EH/s vom August 2024, und der beste KS7 verdient bei 0,10 US-Dollar je kWh nur **0,86 US-Dollar am Tag**.

Aleo: Auch Zero Knowledge stoppt Spezialhardware nicht

Aleos Prover-Rätsel soll die Arbeit der Erzeugung eines SNARK-Beweises nachahmen; es beweist keine Transaktionen von Nutzern, und das Projekt begrüßt ZK-ASICs offen. Etwa fünf Monate nach dem Mainnet im September 2024 wurden dedizierte Maschinen ausgeliefert. Im dritten Quartal 2025 erhielten Prover 31,2 Millionen ALEO

(etwa 7,8 Millionen US-Dollar), während das Netz nur etwa 43.000–44.000 US-Dollar an Gebühren einnahm: **ein Subventionsvielfaches von etwa 175–180**. Der Preis fiel von etwa 9 US-Dollar im frühen Handel auf 0,22 US-Dollar bis zum Ende dieses Quartals, und ARC-46 verlangt jetzt einen Stake von 100.000 bis 2,5 Millionen ALEO je Lösung, was das Mining zu einer Stake-Schwelle macht.

Qubic und Bittensor

Qubic vermarktet „nützlichen Proof of Work“, der sein neuronales Netz Aigarth trainiert, doch der sichtbarste Ertrag seiner Rechenleistung 2025–2026 waren Blöcke anderer Ketten. Zuerst minte es Monero, steigerte seinen Anteil von unter 2 % Mitte Mai 2025 auf mehr als 25 % Ende Juli und beanspruchte im August 51 %, was eine Reorganisation von sechs Blöcken verursachte (eine andere Quelle nennt 18). Am 1. April 2026 wechselte es dazu, Dogecoin mit ASICs zu minen, und verwendete die Erlöse für Rückkäufe und Burns. Es gibt keine unabhängigen Belege, dass sein KI-Training nützlich ist, und sein Marktwert fiel von etwa 280–300 Millionen US-Dollar auf etwa 50 Millionen US-Dollar am 13. November. Es legte zudem eine Externalität offen: Eine Kette, die für Allzweck-Rechenleistung zahlt, schafft einen großen, mietbaren, koordinierten Hashrate-Pool, der sich gegen CPU-freundliche Ketten richten lässt.

Bittensor ist kein Proof of Work, sondern eine Stake-gewichtete subjektive Bewertung: Validatoren bewerten Miner, und die Gewichte werden am Stake-gewichteten Median gekappt. Es ist das einzige Netz dieser Gruppe mit externen Einnahmen in nennenswerter Größenordnung. Der SubConnect-Index, der nur verifizierbare externe Kunden zählt, zeigt Anfang November 2026 etwa **70 Millionen US-Dollar pro Jahr** (etwa 35 Millionen US-Dollar am 1. September), davon 78 % bis 82 % aus Rechenleistung und Infrastruktur; doch zu Preisen von Anfang November übersteigen seine Emissionen 300 Millionen US-Dollar pro Jahr, sodass die Einnahmen nur 9 % bis 23 % decken. Die Einnahmeschätzungen variieren je nach Grundlage stark, von etwa 17 Millionen US-Dollar (Artemis) bis zu annualisiert 172 Millionen US-Dollar aus selbst gemeldeten Summen. Im April verbrannten 67 von 128 Subnetzen mehr als 99 % ihrer Anreize, und als das führende Team Covenant wegen der Governance ausstieg, fiel TAO binnen Stunden um 24 %.

Die Geschichte gibt dieselbe Antwort

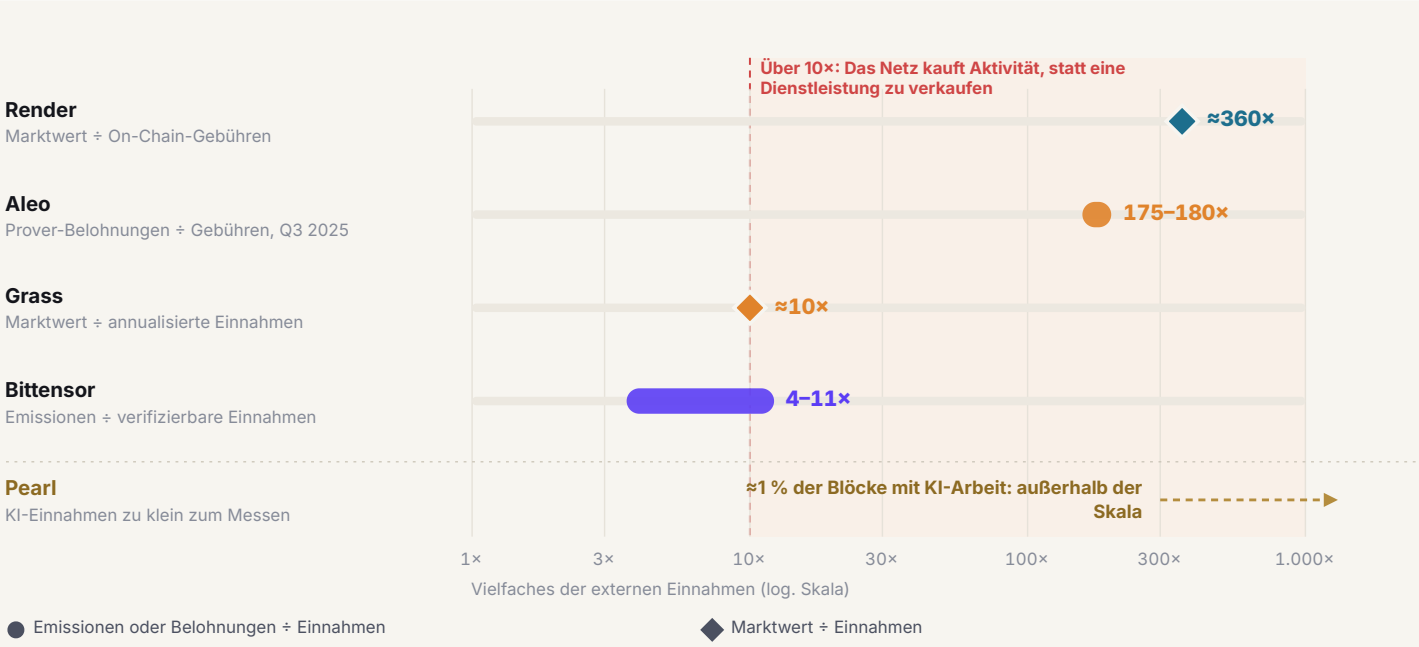
Primecoins Primzahlketten waren schnell zu verifizieren, aber niemand brauchte sie. Gridcoin stützte sich auf Statistikseiten Dritter, um BOINC-Beiträge zu bestätigen, und Forscher fälschten Blöcke, ohne irgendeine Berechnung auszuführen. Filecoins Speicherbeweise funktionieren technisch, doch im dritten Quartal 2025 lag die Auslastung bei nur 36 %, und 99,5 % seiner Gebühren von 793.000 US-Dollar waren Strafzahlungen, bei Basisgebühren von nur 4.100 US-Dollar.

Projekt	Arbeit oder Konsens	Marktwert (Datum)	Hardwarepfad	Bezahlte Nachfrage gegenüber Subvention	Urteil
Pearl	Verrauschte Matrixmultiplikation (INT7-MAC plus ZK-Zertifikat)	Etwa 500 Millionen US-Dollar (13. November)	GPUs; der FP8-Entwurf würde Hopper und Blackwell auf eine Whitelist setzen	Etwa 1 % der Blöcke mit KI-Arbeit; Miner-Einkommen größtenteils Subvention	Der glaubwürdigste Mechanismus; Nachfrage unbewiesen
Kaspa	kHeavyHash (ein 64 × 64-Matrix-Vektor-Produkt mit 4 Bit zwischen Hashes)	Etwa 840 Millionen US-Dollar (13. November)	ASICs etwa 16 Monate nach dem Start; heute nur noch ASICs	Beansprucht keinen Nutzen; Miner-Einnahmen etwa 18 Millionen US-Dollar pro Jahr Schätzung	Eine Matrix ist weder nützlich noch ASIC-fest
Aleo	Synthetisches SNARK-Rätsel plus Proof of Stake	Etwa 118 Millionen US-Dollar (Ende Q3 2025)	Dedizierte Maschinen nach etwa fünf Monaten	Prover-Belohnungen etwa das 175- bis 180-Fache der Gebühren	Synthetische Arbeit, angetrieben von Subvention

Projekt	Arbeit oder Konsens	Marktwert (Datum)	Hardwarepfad	Bezahlte Nachfrage gegenüber Subvention	Urteil
Qubic	„uPoW“ (behauptet, Aigarth zu trainieren)	Etwa 50 Millionen US-Dollar (13. November)	CPU und GPU; ASICs für Dogecoin seit April 2026	Keine unabhängigen Belege für Nutzen; Rechenleistung für externes Mining und Rückkäufe genutzt	Bündelung von Rechenleistung, verpackt in eine KI-Geschichte
Bittensor	Yuma-Konsens (Stake-gewichtete Bewertung)	Etwa 2,55 Milliarden US-Dollar (11. November)	Allzweck-GPUs; kein ASIC-Pfad	Etwa 70 Millionen US-Dollar verifizierbare Einnahmen pro Jahr gegenüber mehr als 300 Millionen US-Dollar Emissionen	Das einzige mit externen Einnahmen in Größenordnung; dennoch eine 4- bis 11-fache Subvention
Filecoin (Referenz)	Proofs of Replication und Proofs of Spacetime	Etwa 1,5 Milliarden US-Dollar (Ende Q3 2025)	—	Auslastung 36 %; vierteljährliche Basisgebühren 4.100 US-Dollar	Nützlich, aber zu wenig Nachfrage

ABB. 04 **Das Subventionsvielfache: wie viele externe Einnahmen ein Netz zahlt**

Emissionen oder Marktwert geteilt durch Einnahmen von außerhalb des Netzes, logarithmische Skala



Quelle: Blockworks (Aleo Q3 2025), DefiLlama (Render-Gebühren), CoinGecko, hoge.gg (Bittensor), Dealroom (Grass), Blockworks 0xResearch (Pearl); Rechnung des Autors.

Die Lehre für Hardware: Feste, regelmäßige Arbeit, die in einem liquiden Token bezahlt wird, bekommt binnen etwa 3 bis 16 Monaten ASICs, oft von Herausforderern vor den Etablierten. Behauptungen von „Nützlichkeit“ hielten nur dort, wo ein externer Kunde für das Ergebnis bezahlt. Und die Hauptursachen für gestrandete Hardware sind eine hohe Subvention im Verhältnis zu den Einnahmen und die Kontrolle der Entwickler über die Arbeitsfunktion.

KAPITEL 06

DAS WEITERE FELD

06 Der Rest von KI × Krypto: reale, aber kleine Nachfrage, und Wert wandert ins Eigenkapital

Dezentrale Rechenmärkte

Dezentrale Rechenleistung hat zahlende Kunden, doch die Nachfrage ist klein, konzentriert und überwiegend Weiterverkauf von Inferenz mit offenen Modellen. On-Chain verifizierbare Gebühren sind gering: DefiLlama zeigt für Render über das vergangene Jahr etwa 2,35 Millionen US-Dollar an Gebühren und für Livepeer etwa 680.000 US-Dollar, und der gesamte DePIN-Sektor erzielte 2025 nur etwa 72 Millionen US-Dollar On-Chain-Einnahmen. Selbst gemeldete Zahlen sind größer, aber locker definiert: Aethir meldet für das Geschäftsjahr 2025 „mehr als 127,8 Millionen US-Dollar“ Umsatz, ohne zu sagen, ob das Bruttoabrechnung oder Protokollanteil ist. Tokenwerte und Einnahmen haben sich in beide Richtungen entkoppelt: Renders Marktwert beträgt etwa das 360-Fache seiner On-Chain-Gebühren, während Aethirs Marktwert von etwa 87,7 Millionen US-Dollar unter seinem selbst gemeldeten Umsatz liegt Schätzung.

Die Behauptung, 70 % bis 85 % günstiger zu sein, gilt nur gegenüber den On-Demand-Preisen der Hyperscaler. Am 3. November lag der gewichtete Durchschnittspreis einer H100 auf Akash bei etwa **2,73 US-Dollar je GPU-Stunde** (Spanne 2,04–3,16 US-Dollar), unter etwa 6,88 US-Dollar bei AWS und 11,06 US-Dollar bei Google Cloud, aber über etwa 1,73 US-Dollar bei Vast.ai, 1,99 US-Dollar in RunPods Community Cloud und etwa 2,44 US-Dollar für eine einjährige Reservierung (Preise anderer Plattformen abgelesen am 12. September; Akash, der Herausgeber, ist eine interessierte Partei). Die Nachfrage ist überwiegend Standard-Inferenz, vertrieben über OpenRouter, etwa 1,7 Milliarden Token am Tag auf AkashML und mehr als 4 Milliarden auf io.net, in direkter Konkurrenz zu jeder Neocloud. Die größten Dollarbeträge der „dezentralen Rechenleistung“ rezentralisieren sich: Axe Compute, das Aethirs Token als Treasury-Vermögenswert hält, unterzeichnete einen Fünfjahresvertrag über 1,5 Milliarden US-Dollar zur Vermietung seines eigenen Clusters aus 9.200 B300, bei einem Umsatz von 35.000 US-Dollar im ersten Quartal 2026.

Dezentrales Training und dezentrale Inferenz

Training über das Internet wurde bei 8 bis 40 Milliarden Parametern demonstriert. Pluralis-8B trainierte 40 Tage lang auf 330 Knoten, überwiegend Consumer-GPUs, und erreichte 63 % der Effizienz einer zentralen H100-Basislinie, wobei das zugehörige Paper einräumt, dass der Lauf nicht dezentral war und Pluralis’ eigene Knoten sowie eine Zulassungsinstanz brauchte. Psyche von Nous Research schloss den Pretraining-Lauf Consilience mit 40 Milliarden Parametern ab. Im Frontier-Maßstab rezentralisierte sich die Arbeit: Prime Intellects INTELLECT-3, ein Mixture-of-Experts-Modell mit 106 Milliarden Parametern, wurde zwei Monate lang auf 512 H200 trainiert, verbunden über InfiniBand mit 400 Gbps. Die Einschränkungen sind klar: Knoten brauchen mindestens 200 Mbit/s Bandbreite und etwa 100-fache Kommunikationskompression, und Vertrauen und Verifikation hängen weiterhin von privilegierten Rollen des Betreibers ab.

Das Geld ging an Unternehmen mit Eigenkapital, die zentrale Produkte verkaufen. Prime Intellect hatte mehr als 100 Millionen US-Dollar annualisierten Umsatz, bevor es im Juli 130 Millionen US-Dollar bei einer Bewertung von 1 Milliarde US-Dollar einsammelte; Venice, eine App für private Inferenz, meldet mehr als 70 Millionen US-Dollar annualisierten Umsatz und Gewinne seit dem ersten Quartal 2026 (Unternehmensangaben), sammelte 65 Millio-

nen US-Dollar bei einer Bewertung von 1 Milliarde US-Dollar ein und läuft auf gemieteten GPUs. Auf der Token-Seite ist Gensyns zentrales Trainingsnetz noch „offen“, und sein Token ist etwa 17,2 Millionen US-Dollar wert; der Token von Ritual, einem KI-nativen Layer-1, fiel nach dem Mainnet um 72 % gegenüber seinem Hoch (nur Schlagzeile).

Verifizierbare KI

Die Produktion hat sich für Trusted Execution Environments (TEEs) entschieden. Hoppers Confidential Computing verursacht bei typischen LLM-Anfragen unter 7 % Overhead und bei großen Modellen fast keinen; Phalas Overhead in Produktion liegt bei etwa 0,5 % bis 5 %, doch Ende 2025 zahlten 398 seiner 10.004 Nutzer. Der Vertrauensanker eines TEE ist der Chiphersteller: TEE.fail, offengelegt im Oktober 2025, nutzte einen DDR5-Bus-Interposer für unter 1.000 US-Dollar, um Attestierungsschlüssel aus vollständig aktualisierten Intel-TDX- und AMD-SEV-SNP-Maschinen zu extrahieren, was wiederum NVIDIAs GPU-Confidential-Computing untergräbt, und beide Hersteller sagen, physische Angriffe lägen außerhalb ihres Bedrohungsmodells. Genau das ist die DePIN-Situation, in der Knotenbetreiber physischen Zugang haben. zkML kann inzwischen ein vollständiges LLM beweisen: Lagranges DeepProve beweist GPT-2 mit etwa 174 Token pro Minute und Gemma 3 mit etwa 86, Größenordnungen von produktiver Inferenz entfernt und ohne offengelegte bezahlte Nutzung.

Agentenzahlungen und Daten

x402 ist das beste warnende Beispiel. TRM Labs zählt seit Mai 2025 etwa 198,9 Millionen Transaktionen und 52,7 Millionen US-Dollar, die von bekannten Facilitatoren abgewickelt wurden, zu 99,6 % in USDC; nach Abzug von Selbstzahlungen, Batch-Flüssen und Verkäufern mit weniger als zehn Käufern liegt der tatsächlich von Agenten ausgelöste Anteil bei nur 0,6 % bis 7,5 %, **etwa 5.000–11.000 US-Dollar im Monat**. Artemis urteilt, dass mehr als 95 % der Aktivität Protokollsignalisierung sind und kein Kaufen und Verkaufen. Zugleich standardisieren Etablierte die Schienen: Die x402 Foundation wurde am 14. Juli unter dem Dach der Linux Foundation gegründet, mit Visa, Mastercard, Stripe, Google, AWS, Cloudflare und Coinbase als Kernmitgliedern, und Tempo, das gemeinsam mit Stripe das Machine Payments Protocol verfasst hat, startete sein Mainnet im März bei einer Bewertung von 5 Milliarden US-Dollar. Plattformen für Agenten-Token folgen weiter der Launchpad-Logik: Virtuals meldet selbst 479 Millionen US-Dollar „aGDP“, und der Gründer von ElizaOS erklärte dessen Token nach einem Vergleich in einer Sammelklage für tot. Wert dürfte eher von Stablecoin-Emittenten und etablierten Zahlungsanbietern abgeschöpft werden; die besondere Rolle von Krypto könnte hier Identität und Verantwortlichkeit für Agenten sein.

Daten sind einer der wenigen Teilsektoren mit echten Käufern außerhalb der Kette. Grass meldet 17 Millionen US-Dollar Datenerlöse für 2025 und weitere 17 Millionen US-Dollar im ersten Halbjahr 2026, peilt 75 Millionen US-Dollar für das Jahr an und sagt, es sei profitabel; GRASS ist etwa 328 Millionen US-Dollar wert, etwa das Zehnfache seines annualisierten Umsatzes aus dem ersten Halbjahr Schätzung. Sein Burggraben ist das Sammelvolumen verteilter privater Bandbreite, nicht Krypto selbst. Daten-DAOs und Datenmarktplätzen fehlt sichtbare bezahlte Nachfrage: Vana ist etwa 21,5 Millionen US-Dollar wert und Ocean etwa 27,6 Millionen US-Dollar, und Ocean liegt im Rechtsstreit mit Fetch.ai. Proof of Personhood gewinnt echte Integrationen, während sich Agenten vervielfachen: World sagt, es habe fast 18 Millionen Menschen verifiziert, und Zoom, Docusign und Tinder unterstützen Teile von World ID.

Das Kapital hat abgestimmt

Sektor	Verifizierbare bezahlte Nachfrage	Subventions- und Bewertungssignale	Urteil
Ketten mit nützlicher Arbeit	Bittensor etwa 70 Millionen US-Dollar verifizierbare externe Einnahmen pro Jahr; Pearl etwa 1 % der Blöcke mit KI-Arbeit	Aleo das 175- bis 180-Fache; Bittensor das 4- bis 11-Fache; Pearl größtenteils Subvention	Die Mechanismen haben Wert; Nützlichkeit braucht zahlende Kunden
Dezentrale Rechenmärkte	On-Chain-Gebühren etwa 0,68–2,35 Millionen US-Dollar pro Jahr (Livepeer, Render); selbst gemeldet 4–128 Millionen US-Dollar	Renders Marktwert etwa das 360-Fache der Gebühren	Real, aber klein, meist Weiterverkauf von Standard-Inferenz
Dezentrales Training und dezentrale Inferenz	Unternehmen mit Eigenkapital: Prime Intellect über 100 Millionen US-Dollar pro Jahr, Venice über 70 Millionen US-Dollar (zentrale Infrastruktur)	Gensyns Trainings-Mainnet nicht live; Ritual-Token −72 % gegenüber seinem Hoch	Der technische Fortschritt ist real; Gewinne gehen ans Eigenkapital
Verifizierbare KI	TEEs in Produktion (398 zahlende Nutzer bei Phala); kein offengelegtes bezahltes zkML	—	TEEs nützlich; zkML früh
Agentenzahlungen	Von Agenten ausgelöste Flüsse etwa 5.000–11.000 US-Dollar im Monat	Etwa 200 Millionen Transaktionen, meist Signalisierung und Spekulation	Erst die Schienen, später die Nachfrage
Datennetzwerke	Grass 17 Millionen US-Dollar im 1. Halbjahr 2026 (Eigenangabe)	GRASS etwa das 10-Fache des annualisierten Umsatzes	Einer der wenigen Sektoren mit echten Käufern

Stellt man die sechs Sektoren nebeneinander, ist das Ergebnis eindeutig: **Subvention ist Treibstoff für einen Kaltstart, kein Geschäftsmodell.** Am 13. November war die KI-Kategorie von CoinGecko insgesamt etwa 18,6 Milliarden US-Dollar wert, doch etwa 56 % davon entfallen auf Allzweck-Layer-1 wie NEAR und ICP sowie auf Projekte auf Konsensebene wie TAO, gestaktes TAO (womöglich doppelt mit TAO gezählt) und Pearl; KI-Token der Protokoll- und Anwendungsebene kommen zusammen auf etwa 8 Milliarden US-Dollar oder weniger Schätzung. Das entspricht etwa der gemeinsamen Bewertung von vier Unternehmen mit Eigenkapital: Tempo (5 Milliarden US-Dollar), Nous (berichtet 1,5 Milliarden US-Dollar), Prime Intellect (1 Milliarde US-Dollar) und Venice (1 Milliarde US-Dollar). Risikokapital für Krypto verharrt bei etwa 4,5–5 Milliarden US-Dollar pro Quartal, wobei der Zuwachs aus KI-bezogener Infrastruktur kommt, deren Anteil an offengelegten Beträgen im dritten Quartal 2026 von 8,1 % auf 18,2 % stieg; und 84,7 % der 118 im Jahr 2025 gestarteten Token notieren unter ihrem Startpreis. In einem Satz: **Krypto leistet Vertrieb und Abwicklung, die Gewinne gehen ans Eigenkapital**, zumindest 2026.

KAPITEL 07

07

ASIC-ÖKONOMIE

Die ASIC-Frage: eine Formel, drei Maßstäbe

Eine Formel: Warum Effizienz sich ab k = 5 nicht mehr auszahlt

Sei R der tägliche Erlös je Recheneinheit (netzweites Miner-Einkommen geteilt durch netzweite Hashrate), c die täglichen Kosten einer Recheneinheit auf der Grenzhardware (bei Pearl eine GPU) und k, wie viel günstiger ein ASIC ist als diese Grenzhardware (c_ASIC = c/k). Solange GPUs die Grenzminer bleiben, drücken freier Markteintritt und -austritt R in Richtung c, sodass ein ASIC je Recheneinheit R − c/k verdient:

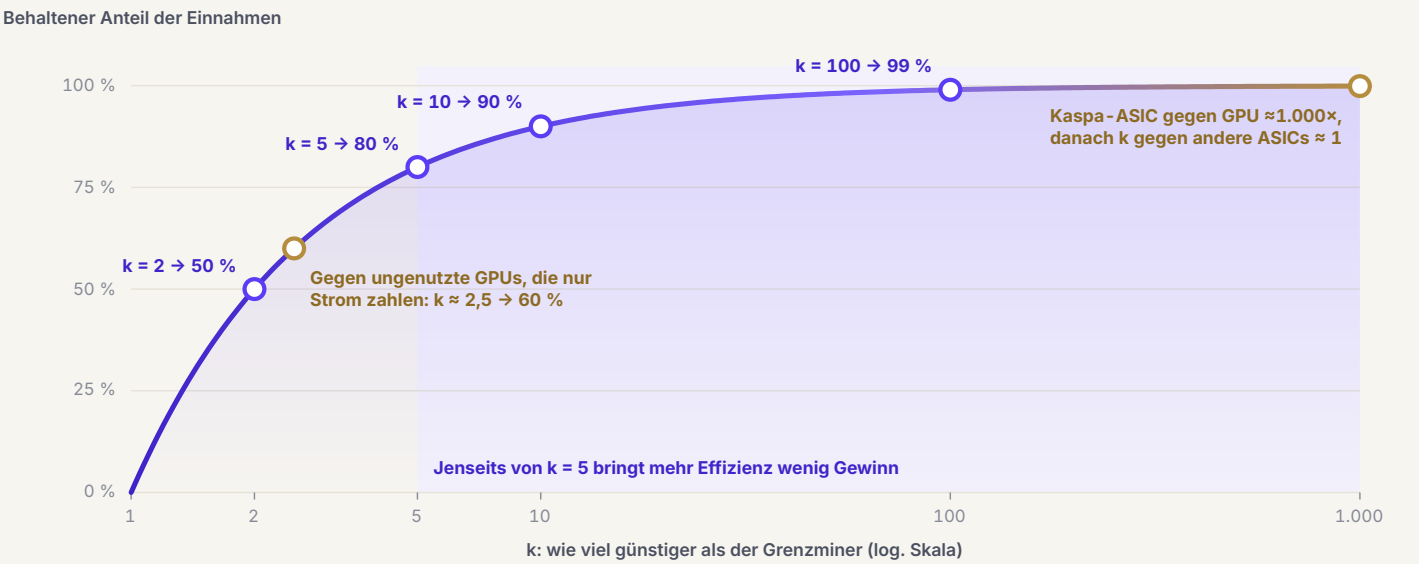
Gewinnobergrenze eines ASIC ≈ R × (1 − 1/k)

Über das gesamte Netz liegt die jährliche Bruttogewinnobergrenze einer ASIC-Flotte ungefähr bei netzweiten Miner-Einnahmen × (1 − 1/k), und das nur, bis ASICs die GPUs als Grenzminer ablösen.

Die Formel hat drei Folgen. Erstens ist (1 − 1/k) konkav: Bei k = 2 behält ein ASIC 50 % der Einnahmen, bei k = 5 80 %, bei k = 10 90 %, bei k = 100 99 %. **Jenseits von k = 5–10 bringt mehr Effizienz fast keinen zusätzlichen Gewinn**; die Größe von R und des Topfs entscheidet. Zweitens setzt sich der Bezugspunkt zurück: Sobald ASICs die Grenzminer sind, muss k gegenüber anderen ASICs gemessen werden. Kaspas ASICs hatten gegenüber GPUs einen etwa 1.000-fachen Effizienzvorteil, der nur einen kurzen Geldsegen brachte; heute verdient der beste KS7 0,86 US-Dollar am Tag. Drittens ist auf einer GPU-freundlichen Kette wie Pearl der Bezugspunkt oft nicht der Kaufpreis einer GPU, sondern **der Strom einer ungenutzten GPU**: Eine eigene RTX 5090 mit 525 W kostet bei 0,10 US-Dollar je kWh etwa 1,26 US-Dollar Strom am Tag und verdient etwa 5–6 US-Dollar Schätzung. Gegenüber GPUs, die versunkene Kosten sind, hat selbst ein ASIC, der ein Fünftel der Energie verbraucht, nur ein effektives k von etwa 2,5, wenn seine Abschreibung seinen Stromkosten entspricht, und eine Gewinnobergrenze von etwa 60 % von R (eine illustrative Annahme).

ABB. 05 **Warum Effizienz sich nicht mehr auszahlt: ein ASIC behält höchstens 1 – 1/k**

Anteil der Einnahmen, den ein ASIC gegenüber dem Grenzminer behalten kann, nach Kostenvorteil k



Quelle: Modell des Autors: Wenn die Grenzhardware den Preis setzt, strebt der Erlös je Recheneinheit gegen ihre Kosten; Kaspa- und RTX-5090-Zahlen aus dem Bericht.

Drei Maßstäbe: der Topf, die Amortisationsgröße und die Regeln

Der erste Maßstab ist **der Topf**. Annualisiert man die rund 2.893 BTC an Blockbelohnungen der ersten Novemberwoche (etwa 222 Millionen US-Dollar), verdienen Bitcoin-Miner etwa 11,5 Milliarden US-Dollar pro Jahr Schätzung. Pearls Miner verdienen etwa 0,46–0,49 Milliarden US-Dollar (zu aktuellen Preisen); Aleos Prover-Belohnungen, annualisiert aus dem dritten Quartal 2025, etwa 31 Millionen US-Dollar; und Kaspas Miner, bei einer Blockbelohnung von 1,945 KAS und zehn Blöcken pro Sekunde, etwa 18 Millionen US-Dollar Schätzung. **Ein ganzes Jahr Miner-Einnahmen von Kaspa ist weniger als die Untergrenze des Designbudgets eines einzigen 5/4-nm-Chips (etwa 30 Millionen US-Dollar).**

Der zweite Maßstab ist **die Hashrate, die nötig ist, um die NRE zu amortisieren**:

$$h^* = \mathrm{NRE} \div [R \times (1 - 1/k) \times T \times p]$$

wobei T das nutzbare Zeitfenster in Tagen und p die Wahrscheinlichkeit ist, dass die Regeln in diesem Fenster unverändert bleiben. Für Pearl nehme man R = 14 US-Dollar je PH/s und Tag, k = 5, T = 365 Tage und eine NRE von 20 Millionen US-Dollar an (ein N7/N6-Chip mit repetitivem Array; am unteren Ende der Spannen von 5–30 Millionen US-Dollar für mittlere Knoten laut Silicon Analysts und 15–100 Millionen US-Dollar für Rechen-ASICs bei 7/6 nm laut AnySilicon, weil ein Mining-Chip wenig IP und eine wiederholte Struktur nutzt; meine Annahme). Mit p = 1 braucht es etwa **4,9 EH/s**, die ein volles Jahr laufen, etwa 5 % des Netzes; mit p = 0,5 etwa **9,8 EH/s**, etwa 11 %; und fällt R auf das Sommerniveau von etwa 10 US-Dollar zurück, braucht selbst p = 1 etwa 6,9 EH/s Schätzung. Die Größenordnung ist nicht unerreichbar. Das Problem sind p und T.

Der dritte Maßstab ist **das Regelrisiko**, und es ist oft entscheidend. Moneros Anti-ASIC-Fork von 2018 ließ Bitmains X3 stranden und kostete das Netz kurzzeitig fast die Hälfte seiner Hashrate. Sias Fork im selben Jahr machte die Miner von Bitmain und Innosilicon zu Elektroschrott, während der Gebrauchtpreis des Obelisk-Miners, der mit den Entwicklern verbunden war, von 2.499 auf mehr als 5.000 US-Dollar stieg. Gut einen Monat nach dem Merge von Ethereum hatten weniger als 20 % der früheren Ethereum-Miner neue Arbeit gefunden. **Wer die Arbeitsfunktion kontrolliert, bestimmt das Schicksal der Hardware-Investoren**, und Sia zeigt, dass „ASIC-Resistenz“ zum eigenen Burggraben der Entwickler werden kann. Hardware mit Zweitmarkt überlebt: Nach dem Merge

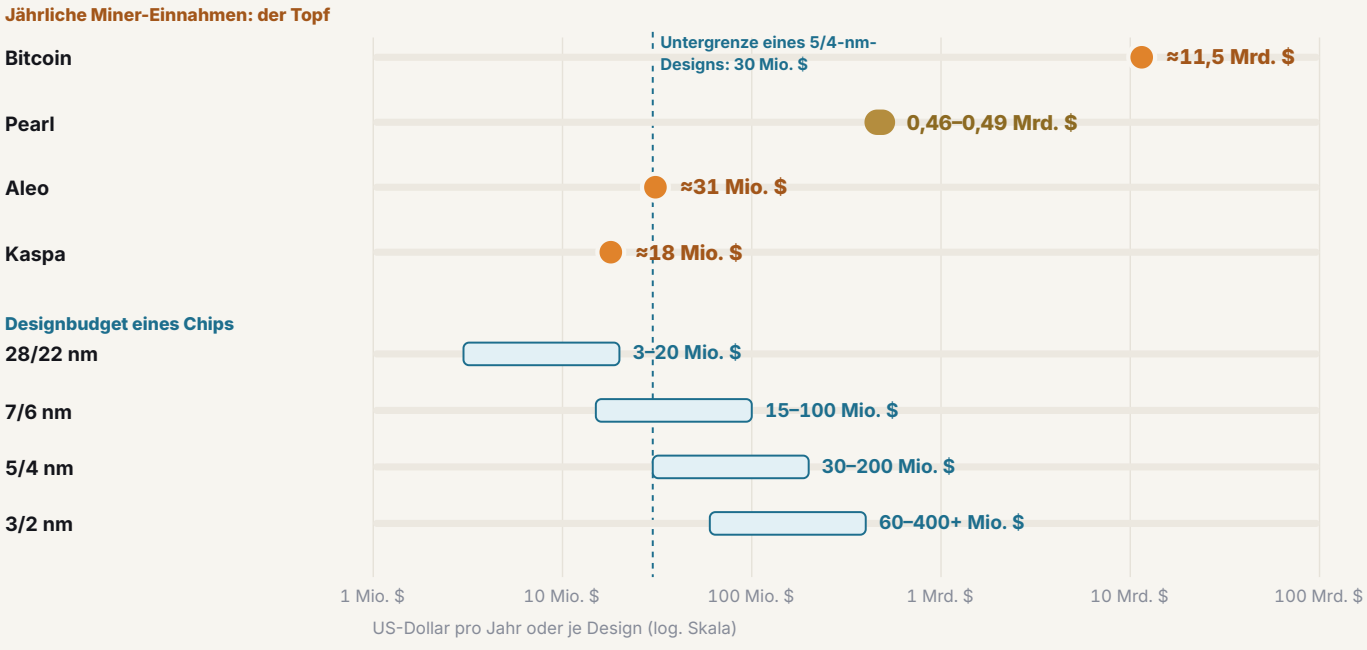
wanderten GPUs zu KI und in die Cloud, während Ein-Algorithmus-ASICs sich in Ethereum Classic drängten. Auch das Zeitfenster schrumpft: Kaspas erster ASIC brauchte etwa 16 Monate, Aleos etwa fünf, und wer als Erster die Effizienzgrenze erreicht, ist oft ein Herausforderer und kein Etablierter.

Netz	Jährliche Miner-Einnahmen (etwa)	Grenz- oder Referenzhardware	Regelstabilität	Für Neueinsteiger erreichbares k	Urteil
Bitcoin	11,5 Milliarden US-Dollar	Bitmain S23 XP Hyd (8,9 J/TH, 25,5 US-Dollar je TH)	Sehr hoch	≤1	Ein großer Topf, aber k gehört den Etablierten
Pearl	0,46–0,49 Milliarden US-Dollar (zu aktuellen Preisen)	Eigene oder ungenutzte GPUs (Grenzkosten ≈ Strom) und Rechenzentrums-GPUs	Sehr niedrig: vier Forks in zwei Monaten; FP8-Whitelist ausstehend	Vielleicht das 3- bis 10-Fache an Effizienz Analyse; weniger gegenüber GPUs mit versunkenen Kosten	Der Topf ist groß genug; das Regelrisiko legt ein Veto ein
Aleo	Etwa 31 Millionen US-Dollar (Q3 2025, annualisiert)	IceRiver AE2 und andere dedizierte Maschinen	Mittel: Stake-Schwelle ARC-46	≈1	Ein kleiner Topf, und die Schwelle steigt
Kaspa	Etwa 18 Millionen US-Dollar	KS7 (77 J/TH)	Hoch	≈1	Jahreseinnahmen unter der Untergrenze eines einzigen 5/4-nm-Designbudgets

ABB. 06

Die meisten Töpfe sind kleiner als das Designbudget eines Chips

Jährliche Miner-Einnahmen von vier Netzen gegenüber den Designkosten eines Chips je Prozessknoten, log. Skala



Quelle: Hashrate Index (Bitcoin, November 2026), prlstats und Kryptex (Pearl), Blockworks (Aleo, Q3 2025), minerstat (Kaspa), AnySilicon (Designbudgets, 2026); annualisiert vom Autor.

Der Quadrant „großer Topf, stabile Regeln und k > 1“ ist heute leer: Bitcoin hat den Topf und die stabilen Regeln, aber sein k gehört den Etablierten; Pearls Topf ist groß genug, aber seine Regeln sind am wenigsten stabil; Kaspa und Aleo haben hinreichend stabile Regeln und zu kleine Töpfe.

KAPITEL 08

08

CHIP FÜR CHIP

Fünf Chiptypen, fünf Antworten

PoUW à la Pearl: kein dedizierter INT7-Chip; auf einen Miner setzen, der Inferenz bedienen kann

Das Argument für einen ASIC ist real. Nur etwa 1 % der Blöcke enthält heute KI-Arbeit, also ist das ökonomische Argument des Whitepapers, ein dedizierter ASIC verliere die Fähigkeit zu nützlicher Arbeit und verdiene deshalb weniger, gerade jetzt am schwächsten. Die teuersten Teile einer GPU, HBM und CoWoS-Packaging, machen etwa 63 % der Herstellungskosten einer H100 aus, das Logik-Die nur etwa 9 %. Ein Chip, der nichts anderes tut als 7-Bit-Ganzzahl-Multiply-Accumulates, seine Operanden im On-Chip-SRAM hält und bei niedriger Spannung läuft, könnte plausibel das 3- bis 10-Fache der Energie je Operation einer GPU auf gleichem Knoten erreichen und einen größeren Vorteil bei den Kosten je Operation [Analyse; noch keine Messungen]. Das Gegenargument wiegt schwerer. Die Lieferzeiten der Foundrys bei führenden Knoten können 50 Wochen übersteigen, und ein neues Team braucht etwa 18 bis 24 Monate vom Start bis zur Serie Schätzung; bis ein Chip ankommt, ist Pearl womöglich auf FP8 und dann FP4 umgestiegen. Und weil PIP-3 die zugelassene Arithmetik über eine Geräte-Whitelist definiert, würde ein INT7-Chip mit der Aktivierung schlicht aufhören zu funktionieren.

Mein Schluss lautet daher: **Jetzt keinen dedizierten INT7-Chip für Pearl zum Tape-out bringen.** Es gibt zwei bessere Wetten. Die eine ist ein inferenzorientierter Beschleuniger, der auch minen kann: aufgebaut um allgemeine INT8/FP8-GEMM und SRAM-residente Operanden, damit der Wert in einem echten Inferenzmarkt landet, falls der Zugang zum Mining verwehrt wird. Die andere ist, sich an der Protokolldebatte zu beteiligen und auf eine herstellerneutrale kanonische Arithmetik zu drängen, die ein ASIC exakt implementieren kann, zusammen mit einer Zusage zur Stabilität des Algorithmus, was die Voraussetzung dafür ist, dass der Quadrant „großer Topf, stabile Regeln, k > 1“ überhaupt existiert. Bis dahin lässt sich eine Architektur auf FPGAs oder Multi-Project-Wafern (MPW) auf ausgereiften Knoten validieren, sodass das NRE-Risiko im Bereich von einer Million US-Dollar bleibt. Kapitel 10 macht daraus ein gestuftes Budget.

Miner nach Bitcoin-Art: Die Etablierten besitzen k, und Strom ist das knappe Gut

Bitcoins Regelrisiko ist das niedrigste, doch der Wettbewerb drückt das k eines Neueinsteigers unter 1. Bitmains S23 XP Hyd erreicht 8,9 J/TH, kostet 25,5 US-Dollar je TH und wird ab November ausgeliefert; die Effizienz der Flaggschiffe fiel von 17,5 J/TH beim S21 im Jahr 2023 auf 8,9 J/TH, um etwa 49 % in drei Jahren. Ein Neueinsteiger, der 2027–2028 in Serie liefert, bräuchte etwa 8 J/TH oder besser, um zu bestehen Schätzung. Herausforderer ziehen sich zurück: Canaan meldete für das zweite Quartal 2026 einen Umsatz von 31,9 Millionen US-Dollar bei einem Bruttoverlust von 29,3 Millionen US-Dollar und stellte für das dritte Quartal nur 11–15 Millionen US-Dollar in Aussicht; Bitdeers Verkäufe eigener Rigs an externe Käufer fielen im dritten Quartal auf 0,3 Millionen US-Dollar, bei einer Bruttomarge des Konzerns von 3,4 %; Auradine benannte sich in Velaura AI um und wandte sich KI-Chips zu; und Core Scientific stornierte eine Bestellung von etwa 15 EH/s an 3-nm-Maschinen bei Blocks Einheit Proto und verbuchte einen Verlust von 41,9 Millionen US-Dollar.

Auch die Nachfrage schrumpft. Am 9. November lag der Hashprice bei etwa 31,89 US-Dollar je PH/s und Tag, die Terminkurve bis April 2027 im Mittel bei etwa 32,7 US-Dollar, und Gebühren machten nur 0,72 % der Blockbelohnungen aus. Bei 0,045 US-Dollar je kWh amortisiert sich der S23 XP Hyd in etwa 3,1 Jahren Schätzung, nahe am nächsten Halving, das um 2028 erwartet wird. Wichtiger sind die Opportunitätskosten des Stroms: Leasingverträge für KI und Hochleistungsrechnen bringen inzwischen etwa 1,67–1,9 Millionen US-Dollar je vertraglich gebundenem Megawatt und Jahr, was bei 80 % bis 85 % Nettobetriebsergebnis etwa dem 1,5- bis 2,9-Fachen des Bruttogewinns aus Mining je Megawatt entspricht (etwa 0,56–0,9 Millionen US-Dollar) [Schätzung; die beiden Megawatt-Grundlagen unterscheiden sich]. Core Scientific wird im vierten Quartal das Mining einstellen; Bitfarms, inzwischen Keel, hat alle seine US-Mining-Standorte geschlossen; IREN verdiente in einem Quartal 200 Millionen US-Dollar mit KI-Cloud und nur 28 Millionen US-Dollar mit Mining. **Strom und Rechenzentrumskapazität sind die Vermögenswerte; Mining-Maschinen sind Verbrauchsgüter.**

Die NRE-Rechnung ist ebenso unerbittlich. Effizienz an der Spitze verlangt inzwischen Prozesse der 3-nm-Klasse (laut Sekundärquellen nutzt MicroBT Samsung SF3, und die Chips von Auradine und Proto sind 3 nm). Masken kosten etwa 10–20 Millionen US-Dollar und die vollständige NRE 30–100 Millionen US-Dollar oder mehr; bei 20–25 US-Dollar je TH und 30 % Bruttomarge bringt jedes verkaufte 1 EH/s etwa 6–7,5 Millionen US-Dollar Bruttogewinn, sodass die Deckung jährlicher F&E-Ausgaben in der Größenordnung von Bitdeer, etwa 110–160 Millionen US-Dollar, den Verkauf von etwa 20–25 EH/s pro Jahr an externe Käufer erfordert, etwa 2 % des Netzes Schätzung. Der Schluss: **Ein neues Team sollte nicht in den freien Markt für Bitcoin-Miner einsteigen.** Drei Fälle ergeben noch Sinn: vertikale Integration mit eigenem günstigem Strom; Nischenformen wie Heim- und Heizungsminer (auch Canaan bringt Produkte zum Heizen für zu Hause auf den Markt); und die Beteiligung an den Roadmaps der Etablierten über IP und Designdienste.

KI-Trainingschips: nicht als kleines Team

Bei Trainingschips ist k am schwersten zu verdienen. Bei gleicher Generation und Präzision liegt Googles Ironwood bei der FP8-Effizienz auf Chipebene nur etwa 1,1- bis 1,8-fach vor der B300 (Ironwoods Leistungsaufnahme ist die Schätzung eines Reporters) Schätzung; das oft zitierte 30- bis 80-Fache aus dem Paper zur TPU v1 wurde gegenüber einer K80 gemessen, die keine Tensor Cores hatte. Der eigentliche Hebel ist das Zahlenformat: Der Wechsel von FP8 auf FP4 bringt auf demselben Silizium meist das 2- bis 3-Fache an Durchsatz, und NVIDIA bewegt sich dort am schnellsten: Es beansprucht für Rubin, in Serie ab dem zweiten Halbjahr 2026, 50 PF NVFP4-Inferenz. Die Kosten- und Versorgungsschwellen sind noch höher: Designbudgets bei 3/2 nm liegen bei 60–400 Millionen US-Dollar oder mehr; die CoWoS-Kapazität beträgt 2026 etwa 106.000–130.000 Wafer im Monat, und TSMC erwägt Berichten zufolge Preiserhöhungen von bis zu 20 %; bei einem Vertragspreis von 15,6 US-Dollar je GB für HBM3E kostet allein das HBM in einem Beschleuniger mit 288 GB etwa 4.500 US-Dollar Schätzung. Dieser Markt gehört der gemeinsamen Entwicklung mit Hyperscalern nach Art von Broadcom (Kapitel 11). Der Schluss: **Kleine Teams sollten keine eigenständigen Trainingschips bauen**; stattdessen IP, Chiplets oder Designdienste.

KI-Inferenzchips: Allgemeinheit gegen k tauschen, und es zahlt sich aus

Inferenz ist der Bereich, in dem ASICs den meisten Spielraum haben, weil sich k durch den Verzicht auf Allgemeinheit erkaufen lässt. Etched behauptet, sein Server mit acht Chips führe Llama-70B mit mehr als 500.000 Token pro Sekunde aus, etwa das 20-Fache von acht H100; Taalas verdrahtet Modellgewichte fest in Masken-ROM und beansprucht eine 10-mal geringere Leistungsaufnahme und 20-mal geringere Herstellungskosten (beides Herstellerangabe). SRAM-zentrierte Architekturen vermeiden HBM und CoWoS, und ihre Knoten sind besser beherrschbar: Taalas nutzt TSMC 6 nm, und Groq wechselte von 14 nm zu Samsung 4 nm. Krypto-Vertrieb hilft Inferenzchips nicht, da dezentrale Rechennetze fast ausschließlich auf NVIDIA laufen; der Zugang führt über die Beschaffungsökonomie der Neoclouds und den großen heimischen Inferenzmarkt in China, wo staatlich finanzierte Rechenzentren heimische KI-Chips nutzen müssen, H200-Importe gedrosselt werden und Peking will, dass große Internetunternehmen Inferenz auf heimischen Chips betreiben. Der Schluss: **lohnt sich zu bauen**, solange es ei-

nen klaren Kunden gibt und der Grad der Bindung an Modelle und Architektur eine bewusste Entscheidung ist. Es ist auch der beste Zweitmarkt für einen PoUW-Mining-Chip.

ZK- und FHE-Beschleuniger: bei ZK absichern, bei FHE früh positionieren

GPUs haben beim Echtzeit-ZK-Proving bereits aufgeholt: Succincts SP1 Hypercube bewies 99,7 % der Ethereum-Blöcke in unter 12 Sekunden auf 16 RTX 5090. Cysic behauptet, sein ZK-Chip C1 sei 10- bis 100-mal so effizient wie GPU-Prover und werde 2026 ausgeliefert, doch seine Nachfrage stammt größtenteils aus den eigenen Token-Belohnungen Herstellerangabe; ich habe keine öffentlichen Belege gefunden, dass ASICs für das Aleo-Rätsel für echtes ZK-Proving eingesetzt werden. ZK-Proving ist ganzzahlig und deterministisch, in seiner Form dem nahe, was Mining-Chips tun, doch Beweissysteme ändern sich schnell (SP1, RISC Zero, Halo2, Plonky2 und weitere). Der Schluss: **abwarten und absichern**, mit programmierbaren MSM-, NTT- und Hash-Engines, zuerst auf FPGAs, dann als ASICs.

FHE ist das am stärksten hardwaregebundene Primitiv mit dem größten Potenzial. Intels Heracles, auf Intel 3 mit 197 mm², 176 W und 48 GB HBM3, führt FHE-Operationen 1.074- bis 5.547-mal schneller aus als ein Xeon mit 24 Kernen, doch Intel hat keine kommerziellen Pläne, und der Leiter des Projekts ist im Dezember 2025 gegangen. Niobium baut auf Samsung 8LPU mit SemiFive als Designpartner (ein Entwicklungsvertrag über etwa 6,86 Millionen US-Dollar) und hat mehr als 23 Millionen US-Dollar eingesammelt; Optalysys mit einem photonischen Ansatz sammelte 31 Millionen US-Dollar ein und sagt, ein vollständiger kommerzieller Chip sei noch zwei bis drei Jahre entfernt; und Zamas quelloffenes FPGA-Design erreicht etwa 13.000 programmierbare Bootstraps pro Sekunde bei etwa 200 W, die öffentliche Basislinie, die jeder ASIC deutlich schlagen muss. Der Schluss: **früh positionieren und mit wenig Kapital validieren**, für ein Fenster um 2027–2029. Die Nachfrage dürfte sich eher in verschlüsselter KI für Unternehmen und Behörden verankern als in Kryptowährungen selbst, und FHE braucht HBM, sodass seine Kostenstruktur der eines Trainingschips ähnelt.

Chiptyp	Woher k kommt	Nachfrage und Topf	NRE und Schwellen	Regel- oder Bindungsrisiko	Zweitmarkt	Urteil
PoUW-Mining-Chip à la Pearl	Verzicht auf HBM, CoWoS und Allzweckeinheiten: vielleicht das 3- bis 10-Fache bei J/Op. Analyse	Miner-Einkommen etwa 0,46–0,49 Milliarden US-Dollar pro Jahr, volatil	Repetitives Array auf N7/N6 etwa 10–30 Millionen US-Dollar Analyse	Sehr hoch: Das Protokoll bestimmt Zahlenformat und Geräte-Whitelist	KI-Inferenz, wenn als allgemeine INT8/FP8-GEMM gebaut	Kein dedizierter INT7-Chip; inferenzorientiert und minebar bauen oder auf kanonische Arithmetik warten
Bitcoin-SHA-256-Mining-Chip	k ≤ 1 gegenüber Bitmain	Etwa 11,5 Milliarden US-Dollar pro Jahr, freier Markt in der Hand zweier Firmen	Masken der 3-nm-Klasse 10–20 Millionen US-Dollar; vollständige NRE 30–100 Millionen US-Dollar oder mehr	Niedrig (stabiler Algorithmus); hoher Wettbewerb und Halving-Risiko	Fast keiner (ein einziger Algorithmus)	Neueinsteiger bleiben dem freien Markt fern; nur eigener Strom, Nischenformen oder IP
KI-Trainingschip	Systolische Arrays gleicher Generation nur das 1,1- bis 1,8-Fache; das Format ist der Hebel	Am größten (Broadcom etwa 58 Milliarden US-Dollar Umsatz mit KI-Halbleitern im Geschäftsjahr 2026)	Design bei 3/2 nm 60–400 Millionen US-Dollar oder mehr; CoWoS und HBM	NVIDIAs Jahresrhythmus	—	Kleine Teams bauen IP, Chiplets und Designdienste, keine eigenständigen Chips

Chiptyp	Woher k kommt	Nachfrage und Topf	NRE und Schwellen	Regel- oder Bindungsrisiko	Zweitmarkt	Urteil
KI-Inferenzchip	Verzicht auf Allgemeinheit: das 10- bis 20-Fache behauptet Herstellerangabe	Groß und fragmentiert; der heimische Markt Festlandchinas ist für sich allein groß	SRAM-Architekturen auf N6/N7–N4 sind beherrschbar	Bindung an Modelle und Architektur	Er ist der Hauptmarkt	Lohnt sich: erst den Kunden gewinnen, dann die Bindung wählen
ZK-Beschleuniger	Das 10- bis 100-Fache von GPUs behauptet Herstellerangabe	Klein, überwiegend von Token-Belohnungen getrieben	N7/N6, dann N5; zuerst FPGA	Beweissysteme ändern sich schnell	Begrenzt außerhalb von ZK	Abwarten und absichern: programmierbare MSM- und NTT-Engines
FHE-Beschleuniger	Das 1.074- bis 5.547-Fache eines Xeon (Heracles)	Noch nicht entstanden; verankert bei Unternehmen und Behörden	Braucht HBM; Kostenstruktur nahe an Trainingschips	Standards und Ökosystem ungeklärt	Verschlüsseltes Rechnen, private KI	Früh positionieren, mit wenig Kapital validieren; Fenster 2027–2029

ABB. 07

Fünf Chiptypen, fünf Antworten

Woher der Kostenvorteil k kommt, wohin das Silizium verkaufen kann, wenn sich die Regeln ändern, und das Urteil



PoUW-Mining-Chip à la Pearl

NICHT JETZT

k: ohne HBM und Allzweckeinheiten: 3–10× je Op. (geschätzt)

Zweitmarkt: KI-Inferenz, bei allgemeinem INT8/FP8-GEMM

Auf einen Inferenz-Chip setzen, der minen kann, nicht auf ein reines INT7-Teil



Bitcoin-SHA-256-Mining-Chip

FERNBLEIBEN

k: ≤1 gegenüber den Etablierten

Zweitmarkt: fast keiner

Nur mit eigenem Strom, Nischenform oder als IP



KI-Trainingschip

IP STATT CHIPS

k: 1,1–1,8× bei gleicher Präzision

Zweitmarkt: —

Packaging und HBM entscheiden; IP, Chiplets, Dienste bauen



KI-Inferenzchip

BAUEN

k: Allgemeinheit aufgeben: 10–20× behauptet

Zweitmarkt: er ist der Hauptmarkt

Erst den Kunden gewinnen, dann die Festlegung wählen



ZK-Proving-Beschleuniger

ABWARTEN, ABSICHERN

k: 10–100× GPUs behauptet

Zweitmarkt: begrenzt außerhalb von ZK

Programmierbare MSM- und NTT-Engines; zuerst FPGA



FHE-Beschleuniger

FRÜH POSITIONIEREN

k: 1.074–5.547× ein Xeon (Heracles)

Zweitmarkt: verschlüsseltes Rechnen, private KI

Ein Fenster 2027–2029; mit wenig Kapital validieren

Quelle: Synthese des Autors aus Kapitel 7 und 8; Herstellerzahlen wie im Bericht gekennzeichnet.

KAPITEL 09

PROZESSKNOTEN

09 Prozessknoten: Die größten Lebenszeitkosten sollen entscheiden

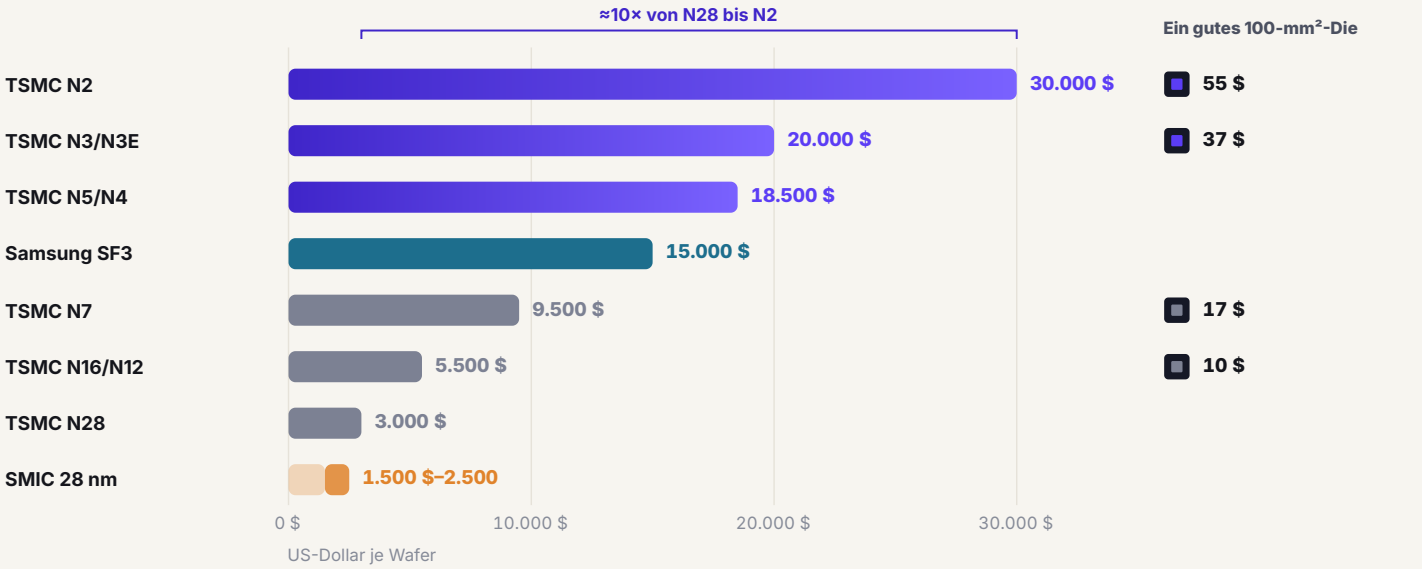
Die Kostenleiter: Wafer am Leading Edge kosten zehnmal so viel wie ausgereifte

Zuerst die Bilanz. Ein Wafer am Leading Edge kostet inzwischen etwa zehnmal so viel wie ein ausgereifter: TSMC N2 etwa 30.000 US-Dollar, N3 etwa 20.000 US-Dollar, N7 etwa 9.500 US-Dollar und N28 etwa 3.000 US-Dollar; Samsung SF3 etwa 15.000 US-Dollar, etwa 25 % unter N3; und SMICs Fertigungskosten in der 7-nm-Klasse liegen Berichten zufolge 40 % bis 50 % über denen von TSMC. TSMC erhöht die Preise Berichten zufolge ab Januar 2027 um weitere 3 % bis 6 %, am stärksten für 2 nm und 3 nm. Masken- und Designkosten unterscheiden sich noch stärker, und die Quellen widersprechen sich, daher zeigt die Tabelle beide.

Knoten	Waferpreis (Schätzung)	Maskensatz (Modellwert / Richtspanne)	Designbudget (AnySilicon)	MPW (je Projekt)
TSMC N2	Etwa 30.000 US-Dollar	—	3/2 nm: 60–400 Millionen US-Dollar oder mehr	—
TSMC N3/N3E	Etwa 20.000 US-Dollar (18.000–27.000)	15 Millionen US-Dollar / 10–20 Millionen US-Dollar	Wie oben	—
Samsung SF3	Etwa 15.000 US-Dollar (13.000–17.000)	12 Millionen US-Dollar	—	—
TSMC N5/N4	Etwa 18.500 US-Dollar (16.000–21.300, geringe Sicherheit)	6,5 Millionen US-Dollar / 10–20 Millionen US-Dollar	5/4 nm: 30–200 Millionen US-Dollar	—
TSMC N7	Etwa 9.500 US-Dollar (8.100–10.900)	3,5 Millionen US-Dollar / 5–10 Millionen US-Dollar	7/6 nm: 15–100 Millionen US-Dollar	Kommerziell 100.000–500.000 US-Dollar
TSMC N16/N12	Etwa 5.500 US-Dollar	1,8 Millionen US-Dollar	16/12 nm: 6–40 Millionen US-Dollar	Akademisch etwa 35.000 US-Dollar je mm²
TSMC N28	Etwa 3.000 US-Dollar	1 Million US-Dollar / 1–3 Millionen US-Dollar	28/22 nm: 3–20 Millionen US-Dollar	Kommerziell 50.000–150.000 US-Dollar; akademisch etwa 12.000 US-Dollar je mm²
SMIC 28 nm	1.500–2.500 US-Dollar	—	—	—
SMIC 7-nm-Klasse (N+2/N+3)	40–50 % über TSMC	—	—	—

ABB. 08 **Ein Wafer am Leading Edge kostet zehnmal so viel wie ein ausgereifter**

Geschätzter Preis je 300-mm-Wafer nach Knoten und Kosten eines guten 100-mm²-Dies bei 85 % Ausbeute



Quelle: Silicon Analysts Waferpreise (Schätzungen, N5 geringe Sicherheit), Wccftech (SMIC); Die-Kosten sind Rechnung des Autors mit angenommener Ausbeute.

Dann die Physik. Horowitz' klassische Messungen zeigen, dass eine programmierbare Instruktion etwa 70 pJ kostet, während die Arithmetik selbst nur wenige Pikojoule kostet: **Arithmetik ist billig, Overhead ist teuer**, und das ist der Grund, warum es systolische Arrays und dedizierte Chips gibt. Eine Halbierung der Strukturgröße halbiert ungefähr die Energie je Operation, doch die Versorgungsspannung stagniert bei etwa 0,9–1,1 V. Ein N2-Wafer kostet 50 % bis 67 % mehr als N3; solange die Dichte also nicht um mehr als etwa das 1,5-Fache steigt, steigen die Kosten je Transistor, statt zu sinken Schätzung: Führende Knoten müssen sich über Effizienz und Leistung bezahlt machen. Für ein 100-mm²-Die bei 85 % Ausbeute kostet jedes gute Die etwa 55 US-Dollar bei N2, 37 US-Dollar bei N3, 29–34 US-Dollar bei N5, 17 US-Dollar bei N7 und 10 US-Dollar bei N16 [Schätzung; die Ausbeute ist eine Annahme].

Die Wahl des Knotens nach Chiptyp und Team

Meine Regel lautet: **Der größte Posten der Lebenszeitkosten des Chips soll den Knoten bestimmen.** Bei Rechenchips, die rund um die Uhr unter Volllast laufen (Mining, ZK), dominiert der Strom die Lebenszeitkosten; sobald der Algorithmus stabil ist, sollten sie daher auf N5/N4 und N3/SF3 wechseln; solange der Algorithmus ungeklärt oder das Geld knapp ist, zuerst auf MPW mit ausgereiftem Knoten validieren. Bei Trainingschips ist der Engpass Packaging und Speicher (HBM und CoWoS machen etwa 63 % der Stückliste einer H100 aus), daher ist der Knoten zweitrangig. Bei Inferenzchips ist die erste Entscheidung die Speicherarchitektur: SRAM-zentrierte Designs können bei N6/N7 bis N4 bleiben, während HBM-Designs die Kostenstruktur von Trainingschips erben. FHE ist durch die Speicherbandbreite gebunden und kommt um HBM nicht herum.

Chiptyp	Start-up oder kleines Team	Etabliertes Unternehmen	Weg in Festlandchina	Kernpunkte
PoUW-Mining-Chip für Matrix-multiplikation (SRAM-resident, ohne HBM)	Prototyp auf MPW mit 28 nm oder 16/12 nm; erste Serie auf N7/N6; N5/N4, sobald Algorithmus und Einnahmen stabil sind; Dies von 50–150 mm² in Standard-Flip-Chip-Gehäusen	N3/N3E oder Samsung SF3	SMIC N+1/N+2 ist machbar (kein HBM oder CoWoS nötig), bei etwa 40–50 % mehr je gutem Die	Energie dominiert; Protokollregeln sind das größte Risiko
SHA-256-Mining-Chip	Nicht der Spitze nachjagen; Nischenformen können ausgereifte Knoten nutzen	3-nm-Klasse; Etablierte wechseln von TSMC 5 nm auf 2 nm und von Samsung SF3 auf 2 nm (Sekundärquellen)	Effizienz führender Knoten schwer erreichbar	Strompreis und Halving-Zyklus bestimmen die Margen
KI-Trainingschip	Kein eigenständiger Chip: IP, Chiplets, Designdienste	N3P, dann N2/A16; 2–8 Chiplets; CoWoS und HBM früh sichern	SMIC N+2/N+3 mit heimischem HBM (das Huawei-Modell); FP8 je Chip etwa das 0,2- bis 0,5-Fache von NVIDIA, ausgeglichen durch Systemgröße	Packaging und HBM sind der Engpass
KI-Inferenzchip	SRAM-zentriert oder fest verdrahtet, N6/N7 oder N4/N5/SF4; nur INT8, FP8 und FP4	N3/N2 mit HBM (langer Kontext, Mixture of Experts) oder SRAM mit mehreren Chiplets	SMIC N+2 mit LPDDR oder GDDR; Speicher der HBM-Klasse ist der Engpass	Der Grad der architektonischen Bindung ist die zentrale Entscheidung
ZK-Beschleuniger	Start auf FPGA oder MPW mit 16/12 nm; dann N7/N6 und N5	N5/N3	SMIC N+1/N+2 ist machbar	Ganzzahlig und deterministisch; leicht zu verifizieren
FHE-Beschleuniger	FPGA (gegen die quelloffene Basislinie) oder MPW mit 16/12 nm; Serie braucht HBM	N5/N3 mit HBM	Begrenzt durch Zugang zu HBM	Die Speicherbandbreite, nicht die Arithmetik, bestimmt die Geschwindigkeit

Fortschrittlich oder ausgereift, vier Arten von Foundry, und Chiplets

Option	Hauptstärken	Hauptschwächen	Am besten für
Führende Knoten (N3/N2, SF3/SF2, Intel 18A)	Jeder volle Knoten halbiert ungefähr die Energie je Operation; dichtes SRAM	Wafer das 3- bis 10-Fache, Masken das 5- bis 15-Fache und Designbudgets das 10- bis 20-Fache ausgereifter Knoten; Lieferzeiten über 50 Wochen; bei N2 können die Kosten je Transistor steigen	Rechenchips mit stabilen Algorithmen im Dauerbetrieb; Trainingschips
Ausgereifte und mittlere Knoten (N7/N6, N16/12, N28)	Masken im Millionen-Dollar-Bereich, günstige MPW, schnelle Iteration, viele Foundrys (TSMC, UMC, GlobalFoundries, SMIC)	Nicht wettbewerbsfähige Effizienz bei Volllast rund um die Uhr	Prototypen, Nischenmärkte, Edge-Inferenz
TSMC	Die beste Ausbeute und das beste Ökosystem	Die höchsten Preise; N2/N3-Kapazität knapp; lange Warteschlangen für CoWoS	Teams, die der Effizienzgrenze nachjagen und Volumenzusagen machen

Option	Hauptstärken	Hauptschwächen	Am besten für
Samsung Foundry	Wafer etwa 20–25 % günstiger (SF3 etwa 15.000 US-Dollar gegenüber etwa 20.000 US-Dollar für N3); genutzt u. a. von Groqs zweiter Generation	Das Foundry-Geschäft schreibt Verluste und peilt den Break-even bis Ende 2027 an	Kostensensible Rechenchips
Intel Foundry	Komplexe Chips wie Heracles auf Intel 3 erprobt	Keine öffentlichen externen Preise oder Kundendaten für 18A; das Foundry-Geschäft schreibt Verluste	Ökosystem und Unterstützung im Einzelfall prüfen
Foundrys in Festlandchina (SMIC und andere)	Nähe zu einem großen heimischen Markt; 28-nm-Wafer für 1.500–2.500 US-Dollar	Kapazität der 7-nm-Klasse etwa 45.000–60.000 Wafer im Monat, großteils von Huawei absorbiert; Ausbeute etwa ein Drittel der von TSMC; Kosten 40–50 % höher; nur DUV	Inferenz-, Mining- und ZK-Chips für den heimischen Markt
Chiplets	Überwinden Grenzen von Belichtungsfeld und Ausbeute; Standard für HBM-Systeme (MI355X, Ironwood, Trainium3)	Höhere Packaging-Kosten und Abhängigkeit von CoWoS; Tests auf Known Good Die erhöhen die Kosten	Trainings- und große Inferenzchips; für kleine Mining- oder ZK-Chips nicht nötig

KAPITEL 10

EIN GESTUFTES BUDGET

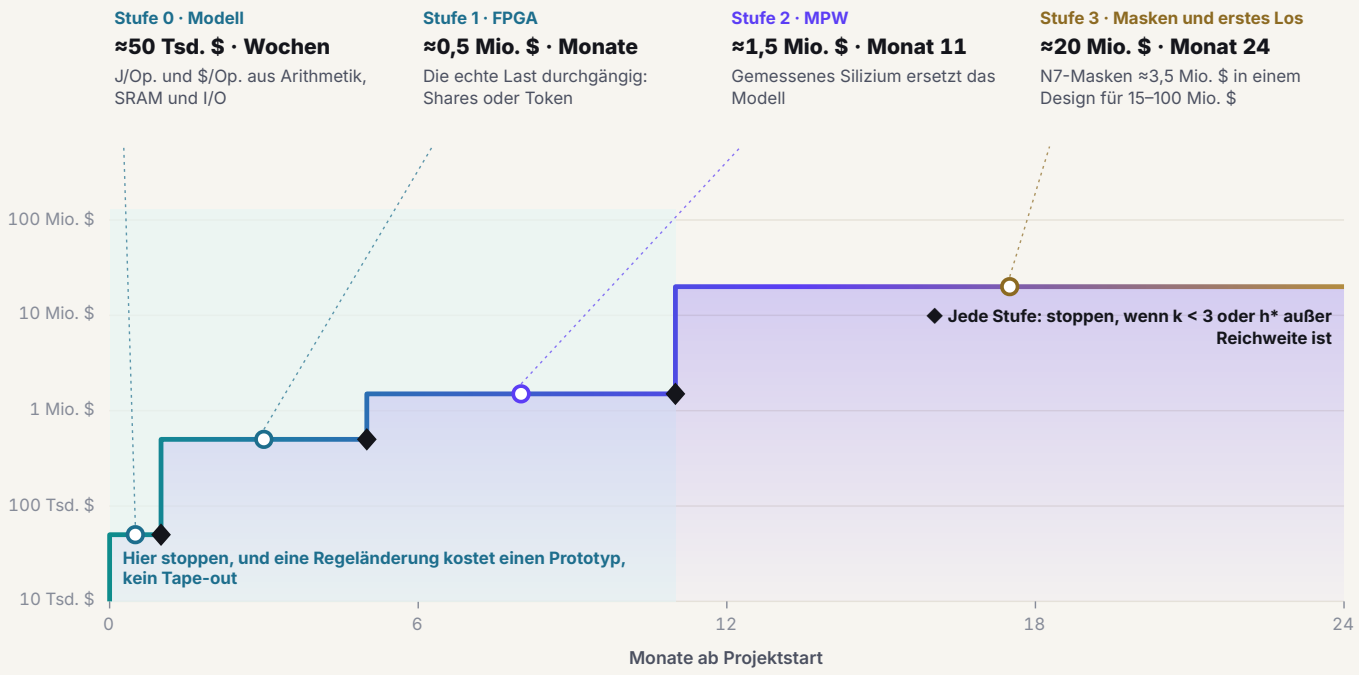
10 Vom Modell zum Silizium: ein gestuftes Budget für einen ersten Chip

Belege vor Masken kaufen

Ein Full-Custom-Chip ist eine Folge von Wetten, jede größer als die vorige, und die Disziplin besteht darin, Belege zu kaufen, bevor man Masken kauft. Ich nutze vier Stufen Analyse. Stufe 0 ist ein Modell: eine Architekturstudie, die J/Op. und $/Op. aus Arithmetik, SRAM und I/O schätzt, gegenüber der Referenzhardware aus Kapitel 7; sie kostet Wochen und einige Zehntausend US-Dollar. Stufe 1 ist ein FPGA-Prototyp, der die echte Last durchgängig ausführt, Mining-Shares oder Inferenz-Token, und Datenpfad sowie Protokollintegration nachweist; sie kostet einige Monate und einige Hunderttausend US-Dollar. Stufe 2 ist ein Multi-Project-Wafer: Ein kommerzieller Platz kostet etwa 50.000–150.000 US-Dollar bei N28 und 100.000–500.000 US-Dollar bei N7, und gemessenes Silizium ersetzt die Energieschätzungen des Modells. Stufe 3 ist der vollständige Maskensatz mit dem ersten Produktionslos: Masken für etwa 1 Million US-Dollar bei N28, 3,5 Millionen US-Dollar bei N7 und 15 Millionen US-Dollar bei N3, innerhalb eines Designbudgets von 15–100 Millionen US-Dollar bei 7/6 nm. Jede Stufe braucht eine Zahl, die das Projekt stoppt, etwa k unter 3 gegenüber der Grenzhardware oder eine Amortisations-Hashrate h*, die das Zeitfenster nicht erreichen kann.

ABB. 09 **Belege vor Masken kaufen: ein gestuftes Budget für einen ersten Chip**

Kumulierte Ausgaben in US-Dollar eines N7/N6-Projekts über vier Stufen, log. Skala; jede Stufe hat eine Zahl, die das Projekt stoppt



Quelle: Wafer-, MPW- und Maskenpreise wie in Kapitel 9 (Silicon Analysts, AnySilicon, Modellwerte); Zeitplan und Ausgaben der Stufen sind illustrative Annahmen des Autors.

Die Treppe ist steil. In einem N7/N6-Projekt kosten die ersten drei Stufen zusammen etwa 1,5 Millionen US-Dollar und etwa elf Monate; die vierte kostet mehr als zehnmal so viel Schätzung. Deshalb zählt die Reihenfolge mehr als der Knoten: Ändert das Protokoll seine Arithmetik oder springt der Käufer ab, verliert ein Team, das bei Stufe 2 aufgehört hat, den Preis eines Prototyps, nicht den eines Tape-outs.

Was jeder Chip kostet, und was zutreffen muss

Projekt	Stufe 2 (Prototyp)	Stufe 3 (Knoten und Designbudget)	Ein gutes 100-mm²-Die	Was zutreffen muss, damit es sich amortisiert
Inferenzorientierter Beschleuniger, der minen kann	MPW mit 16/12 nm oder N7, 0,1–0,5 Millionen US-Dollar	N6/N7, etwa 15–30 Millionen US-Dollar Schätzung	Etwa 17 US-Dollar bei N7, vor Packaging und Speicher	Ein unterschriebener Inferenzkunde trägt den Großteil der NRE; Mining-Einnahmen sind Zusatzertrag, nicht der Basisfall
Dedizierter INT7-Miner für Pearl (derzeit nicht empfohlen)	MPW mit 28 nm, 0,05–0,15 Millionen US-Dollar	N7/N6, etwa 20 Millionen US-Dollar	Etwa 17 US-Dollar bei N7	Etwa 4,9 EH/s für ein volles Jahr bei p = 1 oder 9,8 EH/s bei p = 0,5 (Kapitel 7)
ZK-Beschleuniger	FPGA, dann MPW mit 16/12 nm	N7/N6, dann N5	Etwa 10 US-Dollar bei N16, 17 US-Dollar bei N7	Beweissysteme einigen sich auf MSM- und NTT-Kernels, die ein festes Design bedienen kann
FHE-Beschleuniger	FPGA gegen die quelloffene Basislinie	N5/N3 mit HBM, 30 Millionen US-Dollar oder mehr	Von HBM und Packaging dominiert	Ein Käufer für verschlüsselte KI in Serie, den es noch nicht gibt

Zwei kaufmännische Punkte folgen daraus. Erstens ist die billigste Effizienz architektonisch: Operanden im SRAM zu halten und auf HBM und CoWoS zu verzichten, beseitigt den Großteil der Kosten einer GPU (etwa 63 % der Stückliste einer H100), bevor überhaupt ein Knoten gewählt ist. Zweitens ist Zeit ein Kostenfaktor: Ein neues Team braucht etwa 18 bis 24 Monate vom Start bis zur Serie, und die Lieferzeiten am Leading Edge übersteigen 50 Wochen; ein Protokoll, das seine Arithmetik alle paar Monate ändert, kann einen Chip also veralten lassen, bevor sich Stufe 3 bezahlt macht.

Drei Regeln für ein kleines Team

Zusammengenommen ergeben diese Zahlen drei Regeln Analyse. Erstens: **Bei Stufe 1 oder 2 bleiben, bis es einen namentlich bekannten Käufer oder eine eingefrorene Protokollversion gibt**; weder der Token-Preis noch die Hashrate ersetzt eines davon. Zweitens: **k zuerst mit Architektur kaufen und erst dann mit dem Knoten**: Ein SRAM-residentes Array auf N7 kann für eine einzelne Last einen Allzweckbaustein auf N3 schlagen, zu einem Bruchteil der Maskenkosten. Drittens: **Die Amortisations-Hashrate oder -Stückzahl samt p vor dem Tape-out aufschreiben** und Meilensteinzahlungen gegen eine feste Protokollversion bepreisen, damit der Kunde das Regelrisiko trägt, nicht das Designhaus.

KAPITEL 11

11 CHANCEN
Wo die Chancen liegen, und was ich tun würde

Kundenspezifische Chips: Die Nachfrage kommt überwiegend von KI

2026 kommt die Nachfrage nach kundenspezifischen Chips überwiegend von KI. Broadcoms Umsatz mit KI-Halbleitern im dritten Quartal seines Geschäftsjahres 2026 betrug 16,7 Milliarden US-Dollar, 221 % mehr als im Vorjahr, etwa 58 Milliarden US-Dollar für das Jahr, 73 % davon aus XPUs für sechs Kunden (laut einem automatischen Transkript der Telefonkonferenz zu den Quartalszahlen). Der Umsatz von Global Unichip stieg im zweiten Quartal gegenüber dem Vorjahr um 128 %, der Umsatz mit schlüsselfertiger Produktion um 188 %, bei 21,5 % Bruttomarge. Alchips Umsatz im ersten Quartal verteilte sich nach Knoten zu 21 % auf 3/2 nm, zu 63 % auf 7/5 nm und zu 16 % auf 12 nm und darüber. Die Neuaufträge von VeriSilicon für 2026 erreichten bis Ende April 8,24 Milliarden RMB (etwa 1,16 Milliarden US-Dollar), 91,37 % davon für KI-Rechenleistung. Der Markt hat drei Stufen: oben mehrjährige gemeinsame Entwicklungen im Milliardenumfang; in der Mitte Projekte von 7 nm bis 2 nm mit NRE von einigen Millionen bis über hundert Millionen US-Dollar (5–30 Millionen US-Dollar bei mittleren Knoten, 30–100 Millionen US-Dollar oder mehr am Leading Edge); und ein langer Schwanz von ASICs auf ausgereiften Knoten mit NRE von 0,5–5 Millionen US-Dollar. Krypto- und Mining-Kunden sind eine kleine, zyklische Nische, aber sie sind heute der schnellste und am wenigsten an Erlaubnisse gebundene „erste Kunde“, der für Rechenleistung bezahlt.

Die Fähigkeiten eines Mining-Chip-Teams, die sich am besten übertragen lassen, also Niederspannungsbetrieb nahe der Schwellspannung, massiv replizierte Rechenarrays, Stromversorgung und Kühlung im Kilowattbereich sowie schnelle Tape-outs auf führenden Knoten, sind genau das, was Inferenzchips sowie ZK- und FHE-Beschleuniger brauchen. Die Kapitalmärkte bepreisen das: Velaura AI (vormals Auradine) sammelte im August 110 Millionen US-Dollar bei einer Bewertung von über 1 Milliarde US-Dollar ein. Das Gegenbeispiel ist Canaans Einheit für Edge-KI-Chips, die im Geschäftsjahr 2024 einen Umsatz von 0,9 Millionen US-Dollar erzielte, dabei 15 % der

Betriebsausgaben verbrauchte und im Juni 2025 geschlossen wurde. Der Unterschied liegt in einem eigenständigen Marktzugang, nicht in der Wiederverwendung von Mining-Kanälen.

KI-Entwicklung: Stunden sind Massenware; der Wert liegt an den Nahtstellen

Am unteren Ende sind KI-Entwicklungsdienste zur Massenware geworden: Typische Sätze auf Clutch liegen bei 25–49 US-Dollar pro Stunde und bei 50–99 US-Dollar in den Vereinigten Staaten, Kanada und ähnlichen Märkten; ein Blockchain-Entwicklungsteam kostet etwa 15.000 bis mehr als 90.000 US-Dollar im Monat. Der Verkauf von Stunden hat keinen Burggraben. Der Wert liegt an den Nahtstellen zwischen den Feldern: Kernels und Compiler für kundenspezifische Chips, in den Inferenzbetrieb eingebettetes Mining (Pearls vLLM-Plug-in ist ein Beispiel), verifizierbare Inferenz, die TEEs mit ZK-Stichproben kombiniert, sowie Zahlungen und Identität für Agenten.

KI × Krypto-Systeme: erst den Compute-Käufer gewinnen, dann den Token

Für alle, die eine KI-native Kette oder ein Rechennetz entwerfen, hat der Verifikationsbaukasten 2026 Gestalt angenommen. Geordnet nach Kosten je Arbeitseinheit, vom niedrigsten zum höchsten: statistische Stichproben (Hyperbolics Proof of Sampling), schiedsrichterliche Delegation (Gensyns Verde, die nur einen ehrlichen Verifizierer braucht, aber bitgenaue Ausführung verlangt), TEEs (nahezu native Geschwindigkeit, aber Vertrauen in den Chiphersteller), vollständige Redundanz und allgemeines zkML. Das praktikable Design ist gestuft: TEEs oder Stichproben für niedrige Latenz, Streitspiele oder ZK-Stichproben für die Finalität; ein Paper vom Dezember 2025 simulierte H100-Confidential-Computing mit Betrugsbeweisen und zufälligen ZK-Stichproben bei 99 % des zentralen Durchsatzes und Grenzkosten von 0,07 US-Dollar je Anfrage. Die Hardwarephilosophie bestimmt Sicherheit und Verteilung: GPU-nativer PoUW ist anfangs durch mietbare Cloud-GPUs geschützt, sodass die Kosten eines Angriffs an Mietpreise gekoppelt sind, und Qubic zeigt, dass Proof of Work auf Allzweckhardware von externen Anreizen im großen Stil vereinnahmt werden kann.

Kosten und Zeit sind nicht die Hürde. Pearl kam vom Paper (April 2025) in etwa 12 Monaten zum Mainnet, als Fork von btcd, mit Full Node, Wallet, Light Client und Miner zum Start. Die meisten DeFi-Audits kosten 25.000–100.000 US-Dollar, Audits für Unternehmen oder mehrere Ketten 150.000–250.000 US-Dollar oder mehr, ZK-Schaltungen 80 % bis 120 % mehr, und Projekte, bei denen echter Wert auf dem Spiel steht, budgetieren 150.000–500.000 US-Dollar pro Jahr für Sicherheit. Die eigentlichen Hürden sind Vertrieb und Liquidität. Gründer berichten, dass Binance für ein Listing 0 % bis 3,5 %, angeblich auch 8 %, des Token-Angebots verlangt habe (Binance bestreitet, an Listings zu verdienen); alle sechs von Cointelegraph untersuchten Projekte, die Market-Making-Verträge nach dem Muster „Token-Leihe plus Kaufoption“ nutzten, fielen, und der Market Maker von Movement verkaufte am Tag des Listings Token für etwa 38 Millionen US-Dollar. Die Alternativen sind On-Chain-Liquidität wie bei Hyperliquid, das ohne Binance-Listing 10 Milliarden US-Dollar Marktwert überschritt und dessen HIP-3-Perpetual-Märkte im Juni 3,2 Milliarden US-Dollar Open Interest überschritten, sowie öffentliche Verkäufe über große Börsen, etwa als Monad über Coinbase 269 Millionen US-Dollar einsammelte. Ein hoher vollständig verwässerter Wert bei geringem Streubesitz war 2025–2026 der deutlichste Prädiktor für einen Rückgang nach dem Start.

Token-Design ist Teil des Systemdesigns. Die Emission sollte der bezahlten Nachfrage folgen, statt ihr vorauszugehen: Die Netze in diesem Bericht mit den gesündesten Subventionsvielfachen koppeln Belohnungen an Arbeit, für die ein externer Käufer bezahlt, und die mit den schlechtesten zahlen für Hashrate oder Stake ohne Gegenwert. Eine praktische Regel ist, vor dem Start das Subventionsvielfache zu veröffentlichen, das das Netz im ersten, zweiten und dritten Jahr anstrebt, und die Emission zu kürzen, wenn es verfehlt wird Analyse.

Richtung	Wer kauft, und Zahlungsbereitschaft	Zeitfenster	Schlüsselfähigkeiten	Hauptrisiko
Inferenzorientierter Beschleuniger, der minen kann	Neoclouds, private Bereitstellungen, PoUW-Netze	2027–2028 (abhängig von den FP8- und FP4-Forks)	INT8-, FP8- und FP4-GEMM; SRAM-Architektur; Protokollintegration	Zulassung durch das Protokoll, architektonische Bindung
SRAM- oder fest verdrahteter Inferenzchip	Neoclouds, Rechenzentren in Festlandchina, Edge	Jetzt	Compiler und Modellanpassung; Design auf N6–N4	Modelle ändern sich zu schnell
Mining-Chip (Nische oder vertikal integriert)	Miner mit eigenem Strom; Heim- und Heizungsminer	Jetzt, begrenzt durch Halving und Hashprice	Niederspannungs-Arrays, Effizienz, Lieferkette für Serien	Preiskriege der Etablierten, Strompreise
Beschleunigung von ZK-Proving	Prover-Netze, Rollups	Abwarten, bis Beweissysteme konvergieren	Programmierbare MSM und NTT	Nachfrage hängt von Token-Belohnungen ab
Beschleunigung von FHE	Verschlüsselte KI für Unternehmen und Behörden; Privacy-Chains	2027–2029	NTT, HBM-Systeme, kryptografisches Engineering	Nachfrage noch nicht entstanden; lange Verkaufszyklen
TEE-Härtung und verifizierbare Inferenz	Netze für vertrauliche Inferenz, DePIN	Jetzt	Speicherverschlüsselung mit Integrität, unabhängige Attestierungsanker, ZK-Stichproben	Abhängigkeit von Hersteller-Ökosystemen
KI-native Kette oder Rechennetz	Inferenzkäufer, Miner, Entwickler	Jetzt	Verifikationsdesign, Emissionsdesign, Vertrieb	Rückgang nach dem Start, Token-Liquidität

Drei Wege, die ich gehen würde, jeder mit einem ersten Schritt

Müsste ich wählen, würde ich höchstens drei Wege gehen, jeden mit einem kleinen ersten Schritt, der ihn prüft, bevor Geld gebunden wird Analyse.

- **Weg A: ein inferenzorientierter Beschleuniger, der auch minen kann.** Erster Schritt: eine vierwöchige Architekturstudie, die J/Op. und $/Op. eines SRAM-residenten INT8/FP8-GEMM-Arrays gegenüber RTX 5090 und H100 modelliert, sowohl für Pearls Arbeit als auch für Transformer-Inferenz, gefolgt von einer Validierung auf FPGA oder MPW mit ausgereiftem Knoten bei einer NRE im Bereich von einer Million US-Dollar.
- **Weg B: Custom-Silicon-Dienste mit Krypto als erstem Kunden und KI als Zweitmarkt.** Erster Schritt: das NRE- und Break-even-Modell dieses Berichts in eine bezahlte Machbarkeitsstudie für Rechenchip-Projekte verwandeln, mit Zahlungen nach Meilensteinen und Abnahme gegen eine feste Protokollversion.
- **Weg C: ein KI-natives Netz, das seinen Compute-Käufer vor seinem Token gewinnt.** Erster Schritt: eine Designnotiz zu gestufter Verifikation (TEE plus Stichproben plus ZK-Stichproben) und zu einer herstellerneutralen kanonischen Arithmetik, geschrieben mit einem namentlich bekannten Inferenzkäufer, vor jedem Token-Design.

KAPITEL 12

12

WEGMARKEN

Worauf zu achten ist, November 2026 bis 2027

Wann	Worauf zu achten ist	Warum es wichtig ist
Jederzeit	Pearls FP8-Aktivierungshöhe	Entscheidet, ob Ganzzahlhardware und Nicht-NVIDIA-Karten zugelassen bleiben und ob sich die Kette spaltet
Monatlich	Pearls Anteil an Blöcken mit KI-Arbeit	Der beste einzelne Test dafür, ob „nützlich“ wahr wird
Monatlich	Ob Inferenzanbieter Mining-Einnahmen an ihre Kunden weitergeben	Bezahlte Nachfrage trifft auf Emission: das Modell, das Pearl braucht
Quartalsweise	Bittensors verifizierbare Einnahmen gegenüber seinen Emissionen	Ob das größte Netz mit nützlicher Arbeit seine Subventionslücke schließt
Quartalsweise	Der Anteil der x402-Flüsse, die tatsächlich von Agenten ausgelöst werden	Ob Agentenzahlungen von Signalisierung zu Handel übergehen
Januar 2027	Die berichtete Preiserhöhung von TSMC um 3 % bis 6 %, am stärksten für 2 nm und 3 nm	Die Kosten von Tape-outs auf führenden Knoten
Erstes Quartal 2027	FP4 in Pearl, laut Berichten aus dem AMA der Gründer	Eine weitere Verschiebung der Arbeit hin zu NVIDIAs neuesten Formaten
Ende 2027	Das Ziel von Samsung Foundry, den Break-even zu erreichen	Ob eine günstigere zweite Bezugsquelle gesund bleibt
April 2028	Bitcoins nächstes Halving	Halbiert den Topf für SHA-256-Miner

KAPITEL 13

FRAGEN, DIE MAN STELLEN SOLLTE

13 Fragen, die Sie nicht gestellt haben, aber hätten stellen sollen

Die folgenden Fragen würde ich stellen, bevor ich Geld oder Silizium in ein Projekt aus dem Feld KI × Krypto stecke.

Wer kontrolliert die Arbeitsfunktion? Wenn ein kleines Team die Arithmetik bei einer fest einprogrammierten Höhe ändern kann, entscheidet es über die Lebensdauer Ihrer Hardware. Verlangen Sie eine schriftliche Stabilitätszusage, oder entwerfen Sie Hardware mit Zweitmarkt.

Wie hoch ist das Subventionsvielfache? Teilen Sie die Emissionen (oder den Marktwert) durch die Einnahmen, die von außerhalb des Netzes gezahlt werden. Ein Vielfaches über 10 bedeutet, dass das Netz Aktivität kauft, statt eine Dienstleistung zu verkaufen.

Was kostet die nächste Stufe, und welche Zahl stoppt Sie? Ein Multi-Project-Wafer kostet Zehn- oder Hunderttausende US-Dollar; ein Maskensatz kostet Millionen. Legen Sie die Stopp-Zahl fest, bevor Sie Geld ausgeben.

Wie hoch ist die Amortisations-Hashrate, samt p? Teilen Sie die NRE durch R × (1 − 1/k) × T × p. Ist das Ergebnis ein großer Anteil des Netzes, ist der Chip eine Wette auf die Regeln, nicht auf das Silizium.

Was ist der Zweitmarkt des Chips? Wenn das Protokoll forkt oder der Token fällt, wohin kann das Silizium sonst verkauft werden? GPUs überlebten den Merge von Ethereum, weil sie einen hatten; Ein-Algorithmus-ASICs nicht.

Landet der Wert im Token oder im Eigenkapital? 2026 sind die profitablen Unternehmen im Feld KI × Krypto überwiegend Unternehmen mit Eigenkapital und Krypto-Vertrieb; entscheiden Sie, welches von beiden Sie aufbauen, bevor Sie Geld einsammeln.

Wer ist der erste zahlende Kunde, namentlich? Jeder dauerhafte Fall in diesem Bericht, von Grass bis zu Bittensors Rechen-Subnetzen, begann mit einem namentlich bekannten Käufer außerhalb des Netzes.

Das sind Felder, die ich gern gemeinsam mit anderen erkunden würde, im fachlichen Austausch und in technischer Zusammenarbeit: kundenspezifische Chips für Mining, KI-Training und -Inferenz sowie Beschleuniger; KI-Entwicklung; und KI × Krypto-Systeme von der Kette über den Knoten bis zum Chip. Sie erreichen mich unter i@tanshuai.com.

Grenzen dieses Berichts

Dies ist Schreibtischrecherche auf Grundlage öffentlicher Dokumente bis zum 13. November 2026. Für diese Ausgabe habe ich keine eigenen Interviews und keine unabhängigen Messungen durchgeführt. Mehrere zentrale Zahlen sind Angaben von Unternehmen oder Herstellern, etwa Aethirs Umsatz, der Umsatz von Grass, die Leistung von Etched und Taalas und die Effizienz von Cysic; sie sind entsprechend gekennzeichnet. Chain-Kennzahlen für Pearl stammen von einem unabhängigen Full Node und von Pools und ändern sich stündlich; ich habe die Werte um 07:13 UTC am 13. November verwendet und nenne Spannen desselben Tages. Wafer-, Masken- und Designkosten sind Schätzungen Dritter und Modellwerte, die einander widersprechen; ich zeige Spannen. Meine Break-even-Rechnung für ASICs und das gestufte Budget nutzen illustrative Annahmen (k, T, p, NRE und Kosten der Stufen), die Leserinnen und Leser durch eigene ersetzen sollten.

ÜBER DEN AUTOR

Über den Autor

Tan Shuai ist Technologe und Gründer mit mehr als zwanzig Jahren Arbeit in Chips, Hardware und Systemsoftware sowie 20 erteilten Patenten. Er hat viele Jahre Fabless-Chip-Entwicklung und -Management hinter sich, darunter Full-Custom-ASIC-Datenpfaddesign für Mining-Chips für Kryptowährungen und Multimedia-Kamerachips, und das ist die Perspektive dieses Berichts: Energie je Operation, Kosten je Wafer, was ein Protokoll im Silizium festzuschreiben erlaubt und wer bezahlt. Seine Arbeit führte ihn durch das Silicon Valley, New York, Chicago, Taiwan, die Greater Bay Area (Shenzhen und Hongkong), Vietnam und Malaysia. Heute arbeitet er als KI-Forscher und -Entwickler.

Fachlicher Austausch: i@tanshuai.com · tanshuai.com

Hinweise und Haftungsausschluss

Dieser Bericht ist unabhängige Schreibtischrecherche auf Grundlage öffentlicher Dokumente, die bis zum 13. November 2026 verfügbar waren. Er dient nur der Information und dem Lernen und ist keine Anlage- oder technische Beratung. Er betrachtet ausschließlich Technik und Wirtschaft. Preise, Chain-Kennzahlen und Protokollregeln ändern sich; prüfen Sie alles, worauf Sie sich stützen, anhand der Primärquelle. Mit Herstellerangabe, Schätzung oder Analyse markierte Zahlen sind als solche zu lesen. Preise von Krypto-Assets sind hochvolatil.

© 2026 Tan Shuai. Sie dürfen diesen Bericht unverändert und unter Nennung des Autors weitergeben.

Quellen

Alle abgerufen zwischen dem 10. und dem 13. November 2026. Nach Kapiteln gruppiert.

Pearl: Protokoll, Code und Chain-Daten

- Komargodski und Weinstein, „Proofs of Useful Work from Arbitrary Matrix Multiplication“, arXiv 2504.09971: https://arxiv.org/abs/2504.09971
- Pearl Research Labs, INT-Whitepaper: https://pearlresearch.ai/research/int-whitepaper
- Quellcode von Pearl: Genesis-Block: https://github.com/pearl-research-labs/pearl/blob/main/node/chaincfg/genesis.go ; Konsensparameter und Fork-Höhen: https://github.com/pearl-research-labs/pearl/blob/main/node/chaincfg/params.go ; Hashrate-Abrechnung: https://github.com/pearl-research-labs/pearl/blob/main/zk-pow/src/api/sanity_checks.rs ; Upgrade-Anleitung zur Salted-Seed-Fork: https://github.com/pearl-research-labs/pearl/blob/main/docs/salted-seed-fork-upgrade-guide.md ; Parameter des Zweigs 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, Gleitkomma-Mining: https://github.com/pearl-research-labs/pips/blob/main/PIPS/pip-0003.md
- prlstats (unabhängiger Full Node): Netzwerk: https://prlstats.com/api/network ; Zeitreihen: https://prlstats.com/api/series ; Startseite und Modell des Miner-Einkommens: https://prlstats.com/api/home ; gemintes Angebot: https://prlstats.com/api/mined ; Pools: https://prlstats.com/api/pools ; Halter: https://prlstats.com/api/holders ; Annahmen zu Hardware und Overhead: https://prlstats.com/api/assumptions
- Kryptex Pool: Pearl-Netzwerk: 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 ; Preisverlauf: https://api.coingecko.com/api/v3/coins/pearl-2/market_chart?vs_currency=usd&days=365&interval=daily ; Tickersuche „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. Juni 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, Partnerschaft mit Pearl Research Labs: https://together.ai/blog/together-ai-partners-with-pearl-research-labs ; Modellseite: https://together.ai/models/gemma-4-31b-it-pearl
- Basu, Zählung von Pearl-Mining-Instanzen, arXiv 2606.04819: https://arxiv.org/html/2606.04819v2
- Blockworks, Newsletter 0xResearch, 28. Oktober 2026: https://blockworks.com/newsletter/0xresearch/issue/post_daf0d74a-2bf4-41e4-86d3-2a1275475a24
- Datawallet, „Pearl Explained“: https://www.datawallet.com/crypto/pearl-explained
- Coingabbar zum Bericht von TheMarker über die Finanzierung von Pearl Research Labs: https://www.coingabbar.com/en/pearl-research-labs-funding-130m-stealth-round-themarker

Ketten mit nützlicher Arbeit und Geschichte

- Grothe et al., Sicherheitsanalyse von Gridcoin, USENIX WOOT 2017: https://www.usenix.org/system/files/conference/woot17/woot17-paper-grothe.pdf
- Fälschung von Proof of Learning, arXiv 2208.03567: https://arxiv.org/abs/2208.03567
- hoge.gg, Verifikation und Vertrauen bei dezentraler Inferenz: https://hoge.gg/decentralized-inference-2026-verification-trust/ ; Einnahmen und Treasury von Bittensor: https://hoge.gg/bittensor-quantum-wallet-revenue-tao-treasury-2026/
- Kaspa: Matrixcode von kHeavyHash: https://github.com/kaspanet/rusty-kaspa/blob/master/consensus/pow/src/matrix.rs ; Start: https://coinranking.com/coin/V8GxkwWow+kaspa-kas/about ; Ankündigung von IceRiver KS1/KS2: https://www.iceriver.io/2023/03/14/1677644580/ ; Geschichte von GPU zu ASIC: https://captainmining.com/blogs/crypto-news/the-evolution-of-kaspa-mining-from-gpus-to-asic-powerhouses ; GPU-Hashraten: https://coincalculators.io/algorithms/kheavyhash ; KS7: https://www.asicminervalue.com/miners/bitmain/antminer-ks7-40th ; Netzwerk: https://minerstat.com/coin/kas ; 1 EH/s im Jahr 2024: https://ourcryptotalk.com/news/kaspa-reaches-1-ehs-in-hashrate ; Crescendo: https://ourcryptotalk.com/news/kaspa-crescendo-live
- Aleo: Mainnet-FAQ: https://aleo.org/post/aleo-mainnet-faq/ ; dezentrales Proving: https://aleo.org/post/decentralized-proving-advantages/ ; ASICs: https://whattomine.com/coins/432-aleo-zksnark/asics ; Blockworks, „State of Aleo Q3 2025“: https://app.blockworks.com/report/state-of-aleo-q3-2025 ; ARC-46: https://chaincatcher.com/en/article/2144326
- Qubic: The Block zur Monero-Reorganisation: 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 ; Dogecoin-Mining: https://hackernoon.com/qubic-is-now-mining-dogecoin-on-its-ai-compute-infrastructure ; CoinGecko: https://www.coingecko.com/en/coins/qubic

- Dokumentation zum Konsens von 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 und Forks für ASIC-Resistenz: https://lbry.com/news/forking-for-asic-resistance ; Sia-Fork: https://hotfix.coindesk.com/markets/2018/11/01/sia-completes-hard-fork-as-bricked-miners-form-rival-blockchain ; Obelisk-Preise: https://www.cryptoglobe.com/latest/2018/11/obelisk-hardware-skyrockets-in-selling-price-after-sia-hard-fork/ ; Miner nach dem Merge: https://hotfix.coindesk.com/business/2022/11/11/ex-ethereum-miners-token-hop-to-stay-alive-after-the-merge

Sektoren von KI × Krypto

- DefiLlama: Render: https://defillama.com/protocol/render-network-bme ; Livepeer: https://defillama.com/protocol/livepeer
- Cointelegraph zu den DePIN-Einnahmen laut Messari: https://cointelegraph.com/news/depin-revenues-grow-despite-token-losses
- Aethir, Jahresrückblick 2025: https://aethir.com/blog-posts/aethirs-2025-wrap-up-decentralized-gpu-cloud-milestones
- CoinGecko, KI-Kategorie: https://www.coingecko.com/en/categories/artificial-intelligence
- Akash, Mietpreise für H100: https://akash.network/the-bid/h100-rental-price-2026-cost-per-hour/ ; Bericht Q1 2026: https://akash.network/blog/akash-network-q1-2026-report/
- io.net, Geschäft über 8 Millionen US-Dollar und tägliche Token: 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, Vertrag von 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 zu INTELLECT-3: https://www.implicator.ai/prime-intellects-intellect-3-open-source-ambition-meets-centralized-reality/
- SiliconANGLE, Finanzierung von 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, Finanzierung von Nous: https://runtimewire.com/article/nous-research-75-million-round-1-5-billion-valuation
- ownyourmind, Gensyn: https://ownyourmind.ai/projects/gensyn/ ; Phemex, Ritual (nur Schlagzeile): https://phemex.com/news/article/ritual-launches-ai-layer-1-mainnet-and-ritual-token-price-dives-72-from-high-101259
- Overhead von Confidential Computing, arXiv 2409.03992: https://arxiv.org/abs/2409.03992v2 ; Phala, Bericht 2025: 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, Messung von Zahlungen durch KI-Agenten: https://www.trmlabs.com/trm-tech-blog/whos-actually-paying-measuring-ai-agent-payments-onchain ; Forkast zur 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 von Tempo: https://www.theblock.co/amp/post/394131/tempo-mainnet-goes-live-with-machine-payments-protocol-for-agents ; BlockEden zum aGDP von Virtuals: https://blockeden.xyz/blog/2026/04/21/agdp-agent-gdp-virtuals-protocol-ai-blockchain-valuation-primitive-tvl-displacement/ ; The Block zu ElizaOS: https://theblock.co/post/410774/eliza-labs-native-token-dead
- Dealroom, Umsatz von Grass: https://dealroom.co/news/157794-grass-backed-by-multicoin-projects-75m-in-2026-data-revenue/ ; CryptoSlate, Fetch.ai und Ocean: https://cryptoslate.com/cryptos-flagship-ai-pact-fracture-fetch-sues-ocean-over-263m-fet-community-sales/ ; Fortune zu World: https://fortune.com/2026/06/01/internet-according-to-sam-altman-seeing-orb/
- Galaxy, Risikokapital für Krypto: https://www.galaxy.com/insights/research/crypto-blockchain-venture-capital-q3 ; Crypto Briefing (nur Schlagzeile): https://cryptobriefing.com/crypto-vc-investment-q3-2026-galaxy/ ; Blockhead zu Tiger Research: https://www.blockhead.co/2026/10/14/bitcoin-rose-43-in-q3-crypto-venture-funding-did-not-follow/ ; ForkLog zu Token-Starts 2025: https://forklog.com/en/85-of-tokens-launched-in-2025-fell-below-initial-prices/

ASIC-Ökonomie und Chiptypen

- Hashrate Index, Wochenüberblick vom 9. November 2026: https://hashrateindex.com/blog/hashrate-index-roundup-november-9-2026/
- AnySilicon, Kosten von ASIC-Designs nach Knoten 2026: https://anysilicon.com/news/how-much-does-asic-design-cost-in-2026-real-price-ranges-by-process-node/ ; Silicon Analysts, Leitfaden zu Tape-out-Kosten für Fabless-Firmen: https://siliconanalysts.com/analysis/fabless-startup-tapeout-cost-guide ; Marktdaten: https://siliconanalysts.com/market-data
- MillionMiner, Bitmain S23 XP Hyd: https://millionminer.com/news/bitmain-wdms-2026-s23-xp-hyd-launch
- Khattak und Mikaitis, FP8-Akkumulation in Tensor Cores, arXiv 2512.07004: https://arxiv.org/html/2512.07004v2
- Canaan, Ergebnisse des zweiten Quartals 2026: https://finviz.com/news/389337/canaan-inc-reports-unaudited-second-quarter-2026-financial-results ; Bitdeer, Ergebnisse des dritten Quartals 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 und Block: https://theenergymag.com/news/2026-07-29/core-scientific-exit-block-bitcoin-deal-ai ; Ökonomie von KI-Leasingverträgen: https://theenergymag.com/news/2026-08-06/miner-weekly-ai-pumpability-out-fuel
- Blockworks, Cipher und 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, Ergebnisse Q1 GJ27: https://globenewswire.com/news-release/2026/11/05/3397020/0/en/iren-reports-q1-fy27-results.html
- Spark, Marktkonzentration bei Bitcoin-ASICs: https://www.spark.money/research/bitcoin-asic-manufacturer-market-concentration

- Google, Dokumentation zu 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 ; Paper zur 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, CoWoS-Preise: 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, KI-Inferenzbeschleuniger im Vergleich: https://themenonlab.blog/blog/ai-inference-accelerators-compared ; Introl, KI-Beschleuniger jenseits von GPUs: https://introl.com/blog/ai-accelerators-beyond-gpus-tpu-trainium-gaudi-cerebras-2025
- StratNews (Reuters), heimische KI-Chips in staatlich finanzierten Rechenzentren: https://stratnewsglobal.com/team-sng/china-orders-state-funded-data-centres-to-use-domestic-ai-chips/ ; TNW, H200-Importe: https://thenextweb.com/news/nvidia-h200-chips-hong-kong-china-import-approval
- Succinct, Echtzeit-Proving auf 16 GPUs: https://blog.succinct.xyz/real-time-proving-16-gpus/ ; Cysic, ZK-ASIC-Produkte: 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 und SemiFive: https://evertiq.com/design/2026-02-20-semifive-secures-design-win-with-niobium-for-fhe-accelerator ; The Quantum Insider, Finanzierung von 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 auf FPGA: https://www.zama.org/post/announcing-hpu-on-fpga-the-first-open-source-hardware-accelerator-for-fhe

Prozessknoten

- Silicon Analysts, Waferpreise: https://siliconanalysts.com/data/wafer-pricing
- Tom's Hardware, Gerüchte zu Preisen für 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, SMIC-Preise: https://wccftech.com/smic-5nm-chip-prices-up-to-50-percent-more-expensive/amp/
- EUROPRACTICE, MPW-Zeitplan von 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 von 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, Chinas Lieferkette für KI-Chips: 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

Chancen und Ökonomie von Starts

- Webull (Benzinga), Telefonkonferenz von Broadcom zu den Quartalszahlen: 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, Canaans KI-Einheit: https://www.theblock.co/post/359250/crypto-miner-canaan-to-shutter-non-core-ai-semiconductor-business-unit-amid-strategic-realignment
- Clutch, Preise für KI-Entwicklung: https://clutch.co/developers/artificial-intelligence/pricing ; PixelPlex, Kosten der Blockchain-Entwicklung: 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 ; gestufte Verifikation, arXiv 2512.20176: https://arxiv.org/abs/2512.20176v1
- Sherlock, Preisreferenz für Audits 2026: https://sherlock.xyz/post/smart-contract-audit-pricing-a-market-reference-for-2026
- Blockworks, Listing-Standards von Binance: https://blockworks.com/news/binance-listing-standards-scrutiny ; Cointelegraph, Market-Maker-Verträge: 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 · KRYPTO × SILIZIUM

Rechenleistung als Konsens

KI × Krypto × Chips 2026, ausgehend von Pearls Proof of Useful Work

Jede zentrale Zahl in diesem Bericht trägt ihre Quelle und ihr Datum. Wo der Autor eine Zahl herleitet, ist sie als Schätzung markiert; wo nur der Verkäufer sie angibt, als Herstellerangabe. Der Bericht betrachtet ausschließlich Technik und Wirtschaft.

Tan Shuai

Technologe und Gründer · KI-Forscher und -Entwickler

tanshuai.com · i@tanshuai.com

ÖFFENTLICHE AUSGABE · Datenstand: 13. November 2026

