NOTA DE INVESTIGACIÓN · EDICIÓN PÚBLICA

2026 · 11



PRUEBA DE TRABAJO ÚTIL · CRIPTO × SILICIO

El cómputo como consenso

IA × cripto × chips en 2026, a partir de la prueba de trabajo útil de Pearl

Pearl · Trabajo útil · DePIN y agentes · Economía de los ASIC · Nodos de proceso · Presupuestos por etapas

EL VEREDICTO

Pearl demuestra que la multiplicación de matrices puede proteger una cadena de bloques, pero todavía no que el trabajo sea útil. En todo IA × cripto, la demanda de pago está en la inferencia, los datos y la privacidad, y un ASIC solo compensa donde el bote es grande, las reglas son estables y el chip tiene un segundo mercado.

Tan Shuai

Tecnólogo y fundador · Investigador y desarrollador de IA

tanshuai.com · i@tanshuai.com

Datos a 13 de noviembre de 2026

Importes en dólares estadounidenses

ÍNDICE

Índice



◆ **Resumen ejecutivo** — El teorema se sostiene, el incentivo no — 04

01 **Alcance y método** — Alcance, método y cuatro pruebas — 07

02 **Pearl · el mecanismo** — Pearl: la multiplicación de matrices como consenso — 08

03 **Pearl · la economía** — Pearl: escala, emisión y la brecha de utilidad — 10

04 **Pearl · las reglas** — Pearl: riesgo de reglas, numérica y gobernanza — 12

05 **Cadenas de trabajo útil** — Cadenas de trabajo útil: ni las matrices ni el conocimiento cero frenan a los ASIC — 13

06 **El panorama general** — El resto de IA × cripto: demanda real pero pequeña, y el valor se va al capital — 16

07 **Economía de los ASIC** — La cuestión del ASIC: una fórmula, tres varas de medir — 18

08 **Chip a chip** — Cinco tipos de chip, cinco respuestas — 21

09 **Nodos de proceso** — Nodos de proceso: que decida el mayor coste a lo largo de la vida útil — 25

10 **Un presupuesto por etapas** — Del modelo al silicio: un presupuesto por etapas para un primer chip — 29

11 **Oportunidades** — Dónde están las oportunidades, y qué haría yo — 31

12 **Indicadores** — Qué vigilar, de noviembre de 2026 a 2027 — 34

13 **Preguntas que hacer** — Preguntas que no hizo pero debería haber hecho — 35

◆ **Sobre el autor** — Tan Shuai — 37

◆ **Fuentes** — 38

FIGURAS

Figuras



FIG. 01	IA × cripto en seis sectores: dónde hay demanda de pago real	06
FIG. 02	En el trabajo entero de Pearl, las GPU comparten un techo energético	09
FIG. 03	Los seis primeros meses de Pearl: el hashrate sigue al precio y las reglas cambian	11
FIG. 04	El múltiplo de subsidio: cuántas veces sus ingresos externos paga una red	15
FIG. 05	Por qué la eficiencia deja de pagar: un ASIC retiene como mucho 1 − 1/k de los ingresos	19
FIG. 06	La mayoría de los botes son menores que el presupuesto de diseño de un chip	21
FIG. 07	Cinco tipos de chip, cinco respuestas	25
FIG. 08	Una oblea de vanguardia cuesta diez veces una madura	26
FIG. 09	Pruebas antes que máscaras: un presupuesto por etapas para un primer chip	29

Etiquetas en el texto

- **estimación** cálculo o inferencia del autor a partir de datos citados
- **dato del proveedor** publicado por la empresa que vende el producto, sin verificación independiente
- **análisis** interpretación de los datos por parte del autor

Esta es una traducción de la edición en inglés; si difieren, prevalece la edición en inglés.

Datos a 13 de noviembre de 2026 · Importes en dólares estadounidenses

RESUMEN EJECUTIVO

El teorema se sostiene, el incentivo no

EL VEREDICTO

Pearl demuestra que «el cómputo como consenso» funciona como ingeniería, pero todavía no que sea útil.

≈90 EH/s
la red de Pearl el 13 de noviembre de 2026, donde un H es una multiplicación-acumulación INT7

≈1 %
de los bloques de Pearl llevan trabajo real de IA (Blockworks, 28 de octubre de 2026)

≈70 M$
al año de ingresos externos verificables en Bittensor, frente a más de 300 millones de dólares de emisiones

175–180×
las recompensas de los probadores de Aleo frente a sus comisiones en el tercer trimestre de 2025

R × (1 – 1/k)
el techo de beneficio de un ASIC k veces más barato que el minero marginal
estimación

≈1,5 M$
para llegar a silicio medido en N7 con un modelo, una FPGA y una MPW, antes de un juego de máscaras de ≈3,5 M$
estimación

Pearl

El 13 de noviembre de 2026, la red de Pearl funcionaba a unos 90 EH/s. En su código de consenso, un «hash» es una multiplicación-acumulación entera, así que la red equivale al rendimiento entero de unas 122.000 H100. Sin embargo, solo cerca del 1 % de los bloques lleva trabajo real de IA, y la única alianza pública de inferencia ya no se ofrece en la API serverless de su proveedor. La causa es económica, no criptográfica: una GPU que mina mientras sirve inferencia obtiene solo entre el 1 % y el 27 % del hashrate de un minero dedicado, así que los mineros racionales multiplican matrices aleatorias.

Demanda de pago

En todo IA × cripto, la demanda de pago verificable aparece en tres lugares: vender inferencia, vender datos y vender privacidad. Los ingresos externos verificables de Bittensor rondan los 70 millones de dólares al año, cerca de cuatro quintas partes procedentes de subredes de cómputo e infraestructura. Grass declara 17 millones de dólares de ingresos por datos en el primer semestre de 2026. La inferencia confidencial en entornos de ejecución de confianza está en producción. En cambio, los pagos iniciados de verdad por agentes de IA suman unos 5000–11.000 dólares al mes, y el zkML y el entrenamiento descentralizado aún no tienen demanda de pago declarada.

Subsidio y valor

El múltiplo de subsidio es el chequeo de salud más honesto. Las recompensas de los probadores de Aleo fueron unas 175–180 veces sus comisiones; las emisiones de Bittensor son unas 4–11 veces sus ingresos verificables; el valor de mercado de Render es unas 360 veces sus comisiones on-chain; y los cerca de 460 millones de dólares anuales que ingresan los mineros de Pearl son sobre todo subsidio de bloque, casi desconectado de la demanda de IA de pago. El subsidio es combustible para el arranque en frío, no un modelo de negocio. Mientras tanto, el

valor se desplaza de los tokens al capital: cuatro empresas de capital que usan infraestructura centralizada y distribución nativa cripto están valoradas en unos 8,5 mil millones de dólares, más o menos lo mismo que los tokens de IA de capa de protocolo y de aplicación juntos.

ASIC

En chips, el techo de beneficio de un ASIC son aproximadamente los ingresos por unidad de cómputo multiplicados por (1 − 1/k), donde k indica cuántas veces es más barato que el hardware marginal. Como ese término es cóncavo, la eficiencia más allá de k = 5–10 casi no añade beneficio; lo que decide el resultado es el tamaño del bote, cuánto duran las reglas y si el chip tiene un segundo mercado. Mis respuestas por tipo de chip: no hacer ahora el tape-out de un chip INT7 dedicado a Pearl y apostar en cambio por un acelerador pensado para inferencia que también pueda minar; mantenerse fuera del mercado abierto de mineros de bitcoin como nuevo entrante; no construir un chip de entrenamiento independiente como equipo pequeño; construir chips de inferencia cuando haya un cliente firmado; cubrirse con los aceleradores ZK y posicionarse pronto en FHE.

Nodos y presupuestos

El nodo de proceso debe elegirse según la mayor partida del coste del chip a lo largo de su vida útil: la energía en los chips que funcionan día y noche, el encapsulado y la memoria en los de entrenamiento, la decisión entre SRAM y HBM en los de inferencia. Y un primer chip debe comprarse por etapas —un modelo, una FPGA, una oblea multiproyecto y solo después un juego de máscaras—, para que un cambio en el protocolo cueste el precio de un prototipo y no el de un tape-out. Las oportunidades están donde se cruzan los campos: lo cripto como primer cliente, la IA como segundo mercado.

Pregunta	Juicio	Cifras clave (fuentes en los capítulos)
① **¿Funciona Pearl, y es útil su trabajo?**	Funciona: una cadena de bloques protegida por multiplicación de matrices a gran escala. Todavía no es útil: casi toda la minería se hace sobre matrices aleatorias	Unos 90 EH/s el 13 de noviembre de 2026; 1 H = 1 multiplicación-acumulación INT7; cerca del 1 % de bloques con trabajo de IA
② **¿Dónde está la demanda de pago real en IA × cripto?**	En la inferencia, los datos y la privacidad. Los pagos de agentes, el zkML y el entrenamiento descentralizado siguen antes de la demanda	Bittensor, unos 70 millones de dólares al año de ingresos verificables; Grass, 17 millones en el S1 de 2026; pagos de agentes, unos 5000–11.000 dólares al mes
③ **¿Merece la pena construir un ASIC?**	Solo donde el bote es grande, las reglas son estables y el chip tiene un segundo mercado. Hoy ninguna red ofrece las tres cosas	Techo de beneficio ≈ R × (1 − 1/k); bitcoin, unos 11,5 mil millones de dólares al año de ingresos de los mineros; Pearl, unos 0,46–0,49 mil millones; Kaspa, unos 18 millones
④ **¿Qué chips debería construir un equipo pequeño?**	Aceleradores pensados para inferencia que también puedan minar; chips de inferencia para un cliente firmado; trabajo temprano en FHE. No mineros de bitcoin para el mercado abierto ni chips de entrenamiento	El mejor ASIC de Kaspa gana 0,86 dólares al día; presupuestos de diseño en 3/2 nm de 60–400 millones de dólares o más
⑤ **¿Qué nodo de proceso?**	Decide el mayor coste a lo largo de la vida útil: la energía, el encapsulado y la memoria, o la arquitectura de memoria	TSMC N2, unos 30.000 dólares por oblea frente a unos 3000 en N28; HBM y CoWoS, cerca del 63 % del coste de una H100
⑥ **¿Cuánto cuesta un primer chip, y cómo debe comprarse?**	En cuatro etapas: un modelo, una FPGA, una oblea multiproyecto y después las máscaras. Parar antes de las máscaras si las reglas o el comprador no están asentados	Una plaza comercial en una MPW, 50.000–500.000 dólares; máscaras N7, unos 3,5 millones de dólares; presupuesto de diseño en 7/6 nm, 15–100 millones de dólares

FIG. 01 **IA × cripto en seis sectores: dónde hay demanda de pago real**

Demanda de pago verificable frente a las señales de subsidio y valoración, noviembre de 2026

Cadenas de trabajo útil 01

REAL PERO PEQUEÑO

≈70 M$ al año

ingresos externos verificables de Bittensor

Señal: emisiones 4–11× los ingresos; Pearl ≈1 % de bloques con IA

Los mecanismos funcionan; la utilidad necesita pagadores

Cómputo descentralizado 02

REAL PERO PEQUEÑO

2,35 M$ al año

comisiones on-chain de Render, las mayores verificables

Señal: valor de mercado ≈360× las comisiones

Sobre todo reventa de inferencia genérica

Entrenamiento e inferencia 03

BENEFICIO EN EL CAPITAL

100+ M$ al año

Prime Intellect, una empresa de capital en GPU centralizadas

Señal: la mainnet de entrenamiento de Gensyn no está activa

El avance es real; el beneficio va al capital

IA verificable 04

INCIPIENTE

398

usuarios de pago de la nube TEE de Phala, finales de 2025

Señal: sin zkML de pago declarado

TEE en producción; zkML incipiente

Pagos de agentes 05

ANTES DE LA DEMANDA

5–11 mil $ al mes

pagos x402 iniciados de verdad por agentes

Señal: ≈200 M de transacciones, casi todo señalización

Primero llegaron los raíles; la demanda, no

Redes de datos 06

COMPRADORES REALES

17 M$

ingresos por datos de Grass, primer semestre de 2026 (autodeclarados)

Señal: valor de mercado ≈10× los ingresos anualizados

Uno de los pocos sectores con compradores reales

Fuente: DefiLlama, SubConnect (vía hoge.gg), TRM Labs, Phala, Dealroom, SiliconANGLE, CoinGecko, Blockworks; síntesis del autor.

CAPÍTULO 01



01 ALCANCE Y MÉTODO
Alcance, método y cuatro pruebas

Por qué este informe, y por qué ahora

En abril de 2026 una nueva cadena de bloques, Pearl, llevó a producción una idea antigua: que los mineros compitan por los bloques haciendo las multiplicaciones de matrices de las que vive la IA, en lugar de calcular hashes. En seis meses su red alcanzó unos 90 EH/s y su token un valor de mercado de unos 500 millones de dólares. Al mismo tiempo, los mercados de cómputo de IA, las redes de entrenamiento descentralizado, los raíles de pago para agentes y las redes de datos captaron miles de millones, y los mineros de bitcoin empezaron a convertir sus emplazamientos con energía en centros de datos de IA. Quería saber cuáles de ellos tienen clientes reales, si el silicio a medida tiene sentido en alguno y qué nodo de proceso y qué fundición puede usar de verdad un equipo de chips.

Cuatro pruebas para cualquier proyecto de IA × cripto

Un estudio de seguridad de 2017 sobre Gridcoin sostenía que el cómputo útil solo puede proteger una cadena de bloques si es difícil de resolver, fácil de verificar y fácil de recompensar. Yo añado una cuarta prueba. Juntas deciden la mayor parte de lo que sigue.

01 **Difícil de resolver.** El cómputo honesto debe ser la forma más barata de producir una prueba válida; no puede haber atajos.

02 **Fácil de verificar.** Comprobar el trabajo debe costar mucho menos que hacerlo, o la red paga dos veces.

03 **Fácil de recompensar.** El protocolo debe poder calcular la recompensa solo a partir de la prueba, sin confiar en un tercero.

04 **Alguien paga por el resultado.** Si ningún cliente de fuera de la red paga por el resultado, el trabajo «útil» es un subsidio con relato.

Los rompecabezas de hash superan las tres primeras pruebas y fallan la cuarta. Las tareas reales de IA superan la cuarta pero fallan la segunda: probar un cómputo de IA con conocimiento cero cuesta de 100 a 10.000 veces el propio cómputo, y se ha demostrado que los esquemas de prueba de aprendizaje pueden falsearse a bajo coste. La multiplicación de matrices con ruido de Pearl es el primer puente creíble que supera la prueba de verificación sin dejar de tener la misma forma que el cómputo de IA.

Cómo he trabajado

Es una investigación documental, hecha como revisaría un chip o un plan de negocio. He leído el código de Pearl, sus propuestas de protocolo y su whitepaper; datos de la cadena de un nodo completo independiente; documentación presentada por empresas y comunicados de resultados; artículos de criptografía, arquitectura de computadores y aprendizaje automático; datos de precios de fundiciones y fabricantes; y prensa especializada y tecnológica. Cada cifra lleva su fecha. Cuando derivo yo mismo una cifra, lo digo y la marco como estimación. Las ci-

fras que solo publica el vendedor se marcan como dato del proveedor; mi interpretación de los datos se marca como análisis. Los importes están en dólares estadounidenses. La situación es a 13 de noviembre de 2026.

CAPÍTULO 02

02

PEARL · EL MECANISMO

Pearl: la multiplicación de matrices como consenso

La idea, del teorema a la mainnet

En abril de 2025, Ilan Komargodski y Omri Weinstein publicaron «Proofs of Useful Work from Arbitrary Matrix Multiplication». Construyen una prueba de trabajo útil para cualquier producto de matrices con un sobrecoste de solo 1 + o(1) veces una multiplicación ingenua; su seguridad descansa en la dificultad de resolver lotes de ecuaciones lineales aleatorias de bajo rango, que es una conjetura, no una demostración. El artículo nombra la verdadera dificultad del campo: una prueba de trabajo útil con un sobrecoste de factor constante es trivial pero inútil; lo difícil es que el cómputo honesto siga siendo la forma más barata de ganar cuando los mineros eligen sus propias entradas.

La mainnet de Pearl arrancó el 27 de abril de 2026 a las 09:00 UTC. Los mineros calculan un producto con ruido (A + E)(B + F), donde el ruido de bajo rango se deriva de un compromiso BLAKE3, y después eliminan el ruido para recuperar A·B. El estado del acumulador de cada tesela de salida se convierte mediante un hash en un boleto de lotería; una tesela ganadora se abre y se prueba con una prueba recursiva de conocimiento cero Plonky2 de menos de 60 KB, dentro de un certificado de bloque limitado a 65 KB. La verificación es una prueba pequeña de tamaño constante, independiente del tamaño de la matriz: tan barata de comprobar como un rompecabezas de hash, mientras que el cómputo en sí puede ser útil.

Una multiplicación-acumulación es un «hash»

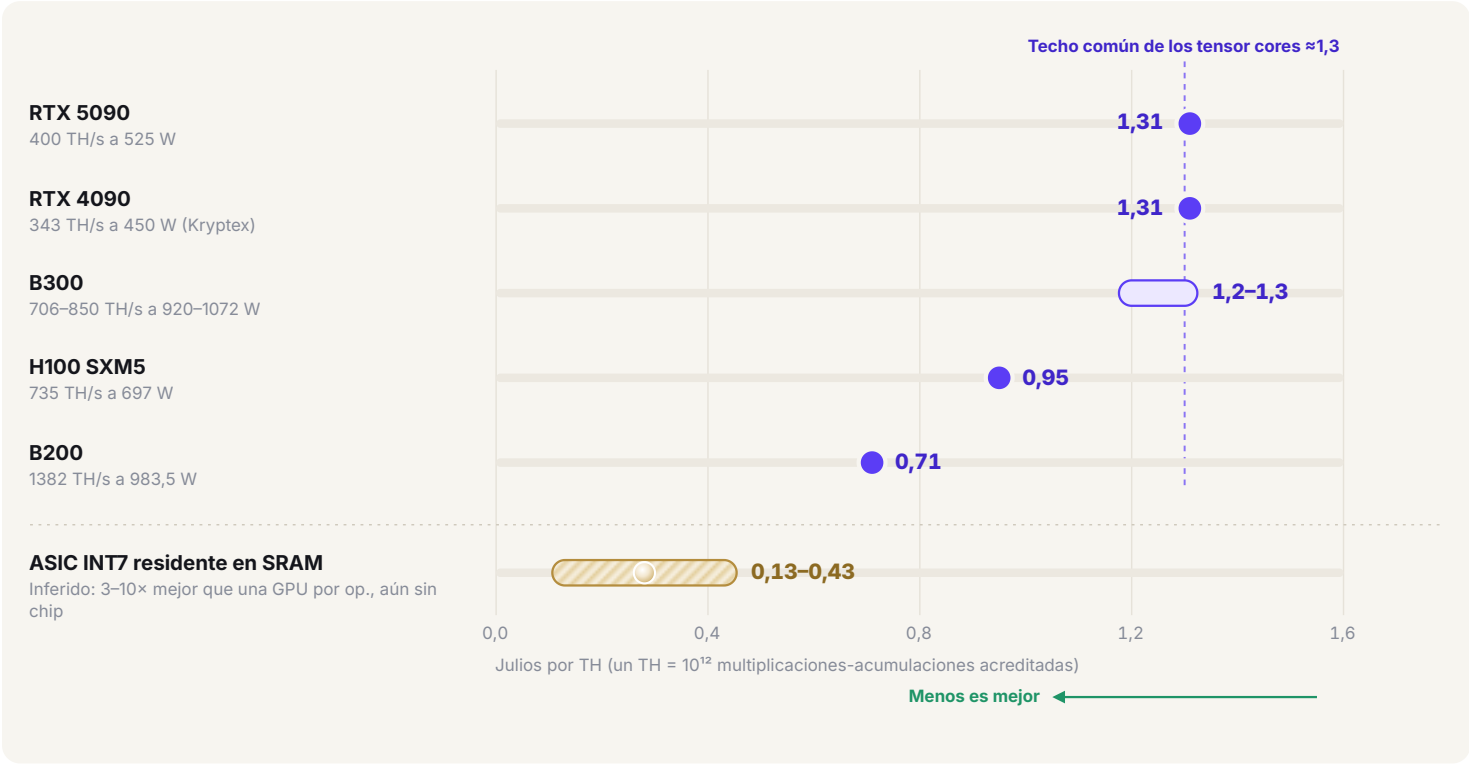
Para un diseñador de chips, el dato más importante es la unidad. El código de consenso de Pearl escala el objetivo de cada intento por el tamaño de la tesela multiplicado por la longitud del producto escalar, de modo que **1 H es exactamente una multiplicación-acumulación (MAC) INT7 × INT7 → INT32 acreditada**. Desde el bloque 96.251, una penalización de rango exige un rango de ruido de al menos 128; con el rango mínimo, 1 H sigue equivaliendo a 1 MAC. Por tanto, los «EH/s» de Pearl significan literalmente 10^18 multiplicaciones-acumulaciones enteras por segundo: **por primera vez, el presupuesto de seguridad de una cadena de bloques se valora en el rendimiento tensorial de los aceleradores de IA.** Los hashrates de GPU medidos se sitúan entre el 60 % y el 95 % del pico de multiplicación-acumulación INT8 densa de cada tarjeta, lo que encaja con esta lectura.

Hardware	Hashrate en Pearl	Potencia	Julios por billón de MAC (J/TH)	Fuente (fecha)
RTX 5090	400 TH/s	525 W	1,31 estimación	Kryptex (12 de noviembre)
RTX 4090	unos 305 TH/s (rango de la comunidad 280–330); Kryptex indica 343 TH/s	450 W (Kryptex)	1,31 (cifras de Kryptex) estimación	prlstats (3 de octubre); Kryptex
H100 SXM5	735 TH/s	697 W	0,95 estimación	PeakMiner, vía prlstats (19 de septiembre)
B200	1,382 PH/s	983,5 W	0,71 estimación	PeakMiner, vía prlstats

Hardware	Hashrate en Pearl	Potencia	Julios por billón de MAC (J/TH)	Fuente (fecha)
B300	706–850 TH/s	920–1072 W	1,2–1,3 estimación	prlstats; SimeonOnSecurity

FIG. 02 **En el trabajo entero de Pearl, las GPU comparten un techo energético**

Julios por billón de multiplicaciones-acumulaciones INT7 (J/TH) al hashrate medido en Pearl; menos es mejor



Fuente: Kryptex (12 de noviembre de 2026); pruebas de PeakMiner y SRBMiner vía prlstats; SimeonOnSecurity; el rango del ASIC es una inferencia del autor, no una medición.

Dos detalles de esta tabla importan a los diseñadores de chips. Primero, las tarjetas de consumo RTX 4090 y 5090 y la B300 de centro de datos convergen en torno a 1,3 J/TH, e incluso la mejor tarjeta, la B200, solo llega a unos 0,71 J/TH. Una GPU de centro de datos que cuesta decenas de miles de dólares gasta casi tanta energía por multiplicación-acumulación entera como una tarjeta gaming, lo que apunta a un techo común de energía de los tensor cores por MAC INT8, y ese techo es justo lo que un chip dedicado podría romper. Segundo, la B300 es más lenta que la B200: NVIDIA recortó mucho el rendimiento INT8 en Blackwell Ultra, así que el trabajo entero tiene que ir por una ruta más lenta en ese chip. El silicio de IA dominante se aleja de INT8 hacia FP8 y FP4. Cuando el formato numérico que elige un consenso diverge del rumbo del hardware, el protocolo se ve obligado a cambiar sus reglas, y esa es la raíz de la bifurcación a FP8 y de mi juicio sobre los ASIC más adelante en este informe.

CAPÍTULO 03

03 PEARL · LA ECONOMÍA
Pearl: escala, emisión y la brecha de utilidad

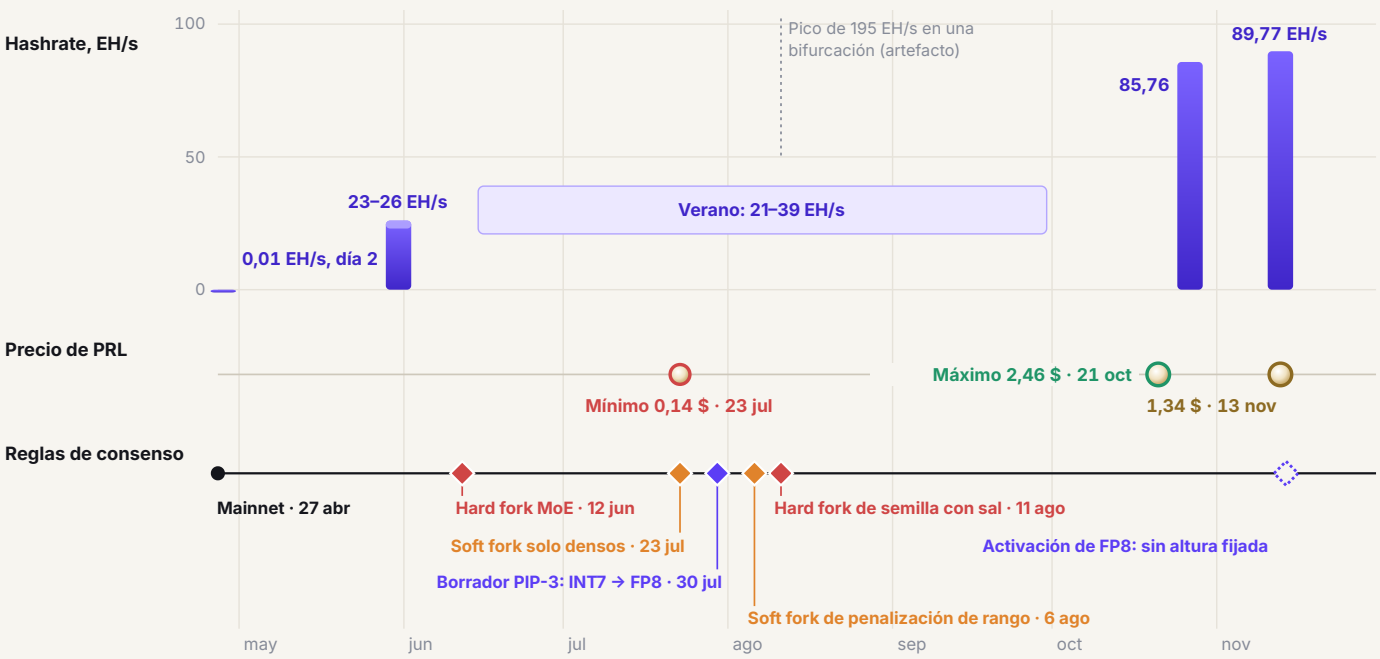
De 0,01 a 90 EH/s en seis meses

Métrica	Valor (13 de noviembre de 2026)	Base y fuente
Hashrate de la red	Unos 90 EH/s (referencia 89,77; rango del mismo día 88,20–93,22; máximo intradía del 12 de noviembre de 110–123)	Nodo completo independiente prlstats; Kryptex
Dificultad y altura de bloque	58.841.982 y 142.530	prlstats
Hardware equivalente	Unas 122.000 H100 o 225.000 RTX 5090	A 735 y 400 TH/s por tarjeta estimación
Precio, valor de mercado y valor totalmente diluido	1,32–1,34 $; unos 500 millones de dólares; unos 2,8 mil millones de dólares (puesto 101 en CoinGecko)	CoinGecko
Máximo y mínimo históricos	2,46 $ (21 de octubre) y 0,1399 $ (23 de julio)	CoinGecko
Monedas minadas	377,6 millones de PRL, el 17,98 % del límite de 2,1 mil millones	prlstats
Subsidio de bloque y emisión diaria	Unos 2173 PRL y unos 0,97–1,03 millones de PRL	prlstats; Kryptex
Nueva oferta en los próximos 12 meses	Cerca del 78 % de la oferta actual	prlstats
Ingresos de los mineros	Unos 1,25 millones de dólares al día (modelo), unos 460 millones al año	prlstats
Ingresos por unidad de hashrate, R	Unos 14–15 dólares por PH/s y día; una RTX 5090 gana unos 5–6 dólares al día	estimación; Kryptex
Bloques con trabajo real de IA	Cerca del 1 % (estimación del 28 de octubre)	Blockworks
Concentración de pools	Kryptex encontró el 46,4 % de los bloques de las últimas 24 horas; los dos primeros pools, cerca del 62 %	prlstats
Concentración de tenedores	Coeficiente de Gini por dirección 0,969; las 100 primeras direcciones tienen el 53,0 %	prlstats

El hashrate de Pearl es casi una función retardada de su precio. El día después del lanzamiento la red funcionaba a unos 0,01 EH/s; a finales de mayo llegó a 23–26 EH/s. Durante el verano el precio cayó cerca de un 86 % desde su máximo de mayo y el hashrate se movió de lado entre 21 y 39 EH/s. Tras multiplicarse el precio por diez en septiembre y octubre, el hashrate alcanzó 85,76 EH/s el 27 de octubre y unos 90 EH/s en noviembre. Un pico de 195 EH/s el 11 de agosto coincidió con una bifurcación y debe leerse como un artefacto de estimación. La energía creció con él: un estudio estimó unos 112 MW para toda la red a 24 EH/s, y un modelo de prlstats la sitúa hoy en unos 190–215 MW.

FIG. 03 **Los seis primeros meses de Pearl: el hashrate sigue al precio y las reglas cambian**

Hashrate de la red, tres marcas de precio y cada cambio de consenso, del 27 de abril al 13 de noviembre de 2026



Fuente: Nodo completo y series de prlstats; Kryptex; CoinGecko; params.go y PIP-3 de Pearl. El pico del 11 de agosto es un artefacto de estimación en una bifurcación.

La emisión amplifica la reflexividad

El límite de oferta es de 2,1 mil millones de monedas y no hay halvings: el subsidio de bloque baja de forma suave según E(t) = S·H / ((t + H)(t + H − 1)) con H = 650.226. Pero la emisión se cuenta en bloques, no en tiempo, y la dificultad inicial empezó demasiado baja y se ajustó despacio. **Se minaron unos 44,06 millones de monedas el primer día y unos 120 millones en los seis primeros días**, cerca del 32 % de la oferta actual estimación. Los bloques recientes, a unos 171–183 segundos, siguen llegando más rápido que el objetivo de 194 segundos. El resultado es que la oferta crecerá cerca de un 78 % en los próximos 12 meses, una dilución a la que tienen que hacer frente mineros, fabricantes de hardware y tenedores por igual.

De ahí se sigue la economía de los mineros. Los ingresos de los mineros en toda la red son de unos 1,25 millones de dólares al día (modelo); la emisión diaria de Kryptex por el precio actual da unos 1,35 millones de dólares al día, así que **los ingresos por unidad de hashrate son R ≈ 14–15 dólares por PH/s y día** estimación. Una RTX 5090 gana unos 5–6 dólares al día. Una H100 o H200 gana unos 11 dólares, mientras que el alquiler público bajo demanda más barato ronda los 2,09 dólares por tarjeta y hora, unos 50 dólares al día, así que alquilar tarjetas solo para minar pierde mucho dinero. R oscila mucho: según la serie de prlstats, rondaba los 10 dólares en agosto y septiembre y brevemente unos 33 dólares a mediados de octubre estimación. A las pocas semanas del lanzamiento, los ingresos diarios estimados de una RTX 5090 se redujeron a la mitad, de unos 33,80 a unos 17,19 dólares. Alquilar GPU en la nube para minar solo compensó en las primeras semanas y durante el pico de precio de octubre.

El teorema se sostiene; el incentivo no

El problema real es la demanda, no la criptografía. La única integración comercial pública de Pearl es un endpoint Gemma-4-31B-it-pearl de Together AI, lanzado el 15 de mayo a más de un 25 % por debajo del precio habitual, con el descuento financiado por el valor futuro esperado de la emisión. El 13 de noviembre, la página del modelo

indicaba que **no está disponible en la API serverless de Together**. En junio, un censo de 8012 instancias de minería en línea, cerca del 21 % de la red, concluyó que la red no estaba produciendo ningún cómputo útil de IA (el autor califica las pruebas de indicativas, no concluyentes), y mostró que los pools aceptaban shares construidas con matrices puramente aleatorias. El 28 de octubre, Blockworks estimó que **solo cerca del 1 % de los bloques lleva trabajo real de IA**; sin él, escribió, Pearl es un bitcoin más pequeño, menos seguro y con más inflación.

La razón puede calcularse. El sobrecoste en el lado de la inferencia es de verdad pequeño: Pearl declara un aumento del 5,08 % en el tiempo de extremo a extremo para Llama 70B en cuatro H200, y una prueba independiente midió una pérdida de rendimiento del 9,2 % en una GPU Hopper. El lado de la minería gana demasiado poco. El plug-in oficial de vLLM solo mina en lotes de prefill de al menos 1024 tokens y nunca en los pasos de decodificación. Una tarjeta que sirve inferencia obtiene cerca del 27 % del hashrate de un minero dedicado con tráfico de prefill puro, cerca del 5,4 % con una mezcla de chat que es un 80 % decodificación, y solo cerca del 1 % con tráfico dominado por la decodificación. Sin demanda de inferencia de pago, un minero racional multiplica matrices aleatorias; el protocolo verifica que las matrices comprometidas se multiplicaron honradamente, no que fueran útiles. **El teorema se sostiene; el incentivo no.** La utilidad de Pearl es un problema comercial fuera de la cadena: alguien tiene que pagar la inferencia.

CAPÍTULO 04

04 PEARL · LAS REGLAS
Pearl: riesgo de reglas, numérica y gobernanza

Cuatro bifurcaciones en dos meses

La otra cara de Pearl es la rapidez con que cambian sus reglas. Entre junio y agosto de 2026 el laboratorio impulsó cuatro cambios de consenso en alturas fijadas en el código: un hard fork de mezcla de expertos (bloque 71.935, hacia el 12 de junio), un soft fork solo para modelos densos (91.630, hacia el 23 de julio), un soft fork de penalización de rango (96.251, hacia el 6 de agosto) y un hard fork de semilla con sal (99.000, hacia el 11 de agosto). Tras la bifurcación de semilla con sal, los mineros antiguos solo podían producir shares inválidas; la penalización de rango recortó el hashrate efectivo de las configuraciones de rango alto entre un 50 % y un 87,5 %.

El siguiente paso es PIP-3, todavía un borrador: lleva la minería de INT7 a FP8 × FP8 → FP32 y exige perfiles de dispositivo exactos bit a bit, **con soporte solo para dos arquitecturas, Hopper y Blackwell**. El documento llama a la dependencia del hardware un mal necesario de los protocolos de coma flotante y advierte de que cualquier tolerancia en la verificación invita a ataques de grinding. El 12 de noviembre, los parámetros de mainnet de la rama fp8-scheme no fijaban ninguna altura de activación. Con el borrador actual no se admiten la serie RTX 40, la A100 ni las tarjetas de AMD, mientras que un modelo de baja confianza de prlstats estima que cerca del 82 % del hashrate procede de tarjetas de consumo. Forzar la bifurcación supone arriesgarse a una división de la cadena o a una rebelión de los mineros.

Un trilema del formato numérico

Pearl se enfrenta a un trilema: un formato numérico puede ser útil para la IA (FP8 o FP4), verificable bit a bit en distintos hardware y neutral respecto al hardware, pero como mucho dos de las tres cosas. El esquema INT7 es verificable y neutral, pero diverge del rumbo del silicio de IA. FP8 con lista blanca es útil y verificable, pero ata el consenso a un solo fabricante. FP8 con tolerancia es útil y neutral, pero abre la puerta a ataques de grinding. La aritmética de coma flotante de los tensor cores no suele seguir un estándar numérico y no es reproducible entre

generaciones; en Hopper, por ejemplo, la acumulación FP8 conserva solo 13 bits fraccionarios y trunca. Desde el punto de vista de un diseñador de chips, un «perfil Hopper» es una aritmética totalmente determinada que un ASIC podría implementar bit a bit, quizá con más eficiencia que la GPU. Pero hacerlo sería un negocio en zona gris, contra la intención declarada del protocolo y con un riesgo de gobernanza muy alto análisis.

Gobernanza

La gobernanza está concentrada. El laboratorio escribió todas las propuestas de protocolo; las bifurcaciones se activan en alturas fijadas en el código sin señalización de los mineros; no hay una segunda implementación del cliente; y a principios de octubre no se había publicado ninguna auditoría de terceros del nodo ni de los circuitos de conocimiento cero. El diario israelí TheMarker informó de que el laboratorio captó unos 120–130 millones de dólares en una ronda discreta (unos 55 millones de dólares de capital más un fondo de infraestructura de GPU de unos 75 millones); Pearl no lo ha confirmado por canales oficiales. «Sin asignación a inversores en el token» y «una empresa con accionistas de capital riesgo» pueden ser ciertas a la vez, pero la tensión persiste: los ingresos de IA pueden ir al capital mientras la dilución recae en los tenedores del token.

En conjunto, Pearl es hasta ahora el experimento más limpio de «el cómputo como consenso»: un mecanismo novedoso, verificable en la cadena, con una unidad que se corresponde uno a uno con el hardware de IA. Pero la utilidad no ha llegado, y su presupuesto de seguridad, su hashrate y el entusiasmo de sus mineros dependen de un token cuya demanda principal aún no se ha formado y cuya oferta crecerá cerca de un 78 % en 12 meses. Merece la pena seguir tres señales de Blockworks: si la actualización llega sin sobresaltos, si la proporción de bloques con trabajo de IA sube de forma apreciable y si algún proveedor de inferencia traslada los ingresos de la minería a sus clientes.

CAPÍTULO 05



CADENAS DE TRABAJO ÚTIL

05 Cadenas de trabajo útil: ni las matrices ni el conocimiento cero frenan a los ASIC

Kaspa: una matriz no es ni útil ni resistente a los ASIC

Kaspa es el caso de manual. Su kHeavyHash coloca un producto matriz-vector 64 × 64 sobre valores de 4 bits entre dos hashes, y el producto se desecha. Unos 16 meses después de su lanzamiento justo en noviembre de 2021, IceRiver anunció los primeros ASIC para Kaspa el 14 de marzo de 2023, y a finales de 2023 la minería con GPU había terminado. Comparar una RTX 4090 a 3,8 GH/s y 300 W con la KS7 de Bitmain a 77 J/TH da una diferencia de eficiencia de unas **1000 veces** estimación. Tras el festín llegó la compresión: el 13 de noviembre la red funcionaba a 358,9 PH/s, cerca de un 64 % por debajo de su 1 EH/s de agosto de 2024, y la mejor KS7 gana solo **0,86 dólares al día** a 0,10 dólares por kWh.

Aleo: el conocimiento cero tampoco frena al hardware especializado

El rompecabezas de los probadores de Aleo está diseñado para imitar el trabajo de generar una prueba SNARK; no prueba las transacciones de los usuarios, y el proyecto da la bienvenida abiertamente a los ASIC de ZK. Unos cinco meses después de su mainnet de septiembre de 2024 ya se vendían máquinas dedicadas. En el tercer trimestre de 2025, los probadores recibieron 31,2 millones de ALEO (unos 7,8 millones de dólares), mientras que la red solo recaudó unos 43.000–44.000 dólares en comisiones: **un múltiplo de subsidio de unas 175–180 veces**. El

precio cayó de unos 9 dólares en las primeras cotizaciones a 0,22 dólares a finales de ese trimestre, y ARC-46 exige ahora un stake de 100.000 a 2,5 millones de ALEO por solución, lo que convierte la minería en un umbral de stake.

Qubic y Bittensor

Qubic promociona una «prueba de trabajo útil» que entrena su red neuronal Aigarth, pero el resultado más visible de su cómputo en 2025–2026 fueron bloques de otras cadenas. Primero minó Monero: elevó su cuota de menos del 2 % a mediados de mayo de 2025 a más del 25 % a finales de julio y reivindicó el 51 % en agosto, lo que provocó una reorganización de seis bloques (otra fuente habla de 18). El 1 de abril de 2026 pasó a minar Dogecoin con ASIC, y destina lo obtenido a recompras y quemas. No hay pruebas independientes de que su entrenamiento de IA sea útil, y su valor de mercado cayó de unos 280–300 millones de dólares a unos 50 millones el 13 de noviembre. También dejó al descubierto una externalidad: una cadena que paga por cómputo de propósito general crea un gran fondo de hashrate coordinado y alquilable que puede volverse contra las cadenas pensadas para CPU.

Bittensor no es prueba de trabajo sino puntuación subjetiva ponderada por stake: los validadores puntúan a los mineros y los pesos se recortan en la mediana ponderada por stake. Es la única red de este grupo con ingresos externos a escala. El índice SubConnect, que solo cuenta clientes externos verificables, muestra unos **70 millones de dólares al año** a principios de noviembre de 2026 (unos 35 millones el 1 de septiembre), entre el 78 % y el 82 % procedentes de cómputo e infraestructura; pero a los precios de principios de noviembre sus emisiones superan los 300 millones de dólares al año, así que los ingresos cubren solo entre el 9 % y el 23 %. Las estimaciones de ingresos varían mucho según la base, desde unos 17 millones de dólares (Artemis) hasta 172 millones anualizados a partir de totales autodeclarados. En abril, 67 de 128 subredes quemaban más del 99 % de sus incentivos, y cuando el equipo líder Covenant se marchó por desacuerdos de gobernanza, TAO cayó un 24 % en pocas horas.

La historia da la misma respuesta

Las cadenas de números primos de Primecoin eran rápidas de verificar, pero nadie las necesitaba. Gridcoin dependía de sitios de estadísticas de terceros para certificar las contribuciones a BOINC, y unos investigadores falsificaron bloques sin hacer ningún cómputo. Las pruebas de almacenamiento de Filecoin funcionan técnicamente, pero en el tercer trimestre de 2025 la utilización fue solo del 36 %, y el 99,5 % de sus 793.000 dólares de comisiones fueron penalizaciones, con comisiones base de apenas 4100 dólares.

Proyecto	Trabajo o consenso	Valor de mercado (fecha)	Vía de hardware	Demanda de pago frente a subsidio	Juicio
Pearl	Multiplicación de matrices con ruido (MAC INT7 más certificado ZK)	Unos 500 millones de dólares (13 de noviembre)	GPU; el borrador FP8 pondría en lista blanca a Hopper y Blackwell	Cerca del 1 % de bloques con trabajo de IA; ingresos de los mineros, sobre todo subsidio	El mecanismo más creíble; demanda sin demostrar
Kaspa	kHeavyHash (un producto matriz-vector 64 × 64 de 4 bits entre hashes)	Unos 840 millones de dólares (13 de noviembre)	ASIC unos 16 meses después del lanzamiento; hoy todo ASIC	No afirma ser útil; ingresos de los mineros, unos 18 millones de dólares al año estimación	Una matriz no es ni útil ni resistente a los ASIC
Aleo	Rompecabezas SNARK sintético más prueba de participación	Unos 118 millones de dólares (final del T3 de 2025)	Máquinas dedicadas a los cinco meses aproximadamente	Recompensas de probadores de unas 175–180 veces las comisiones	Trabajo sintético, impulsado por el subsidio

Proyecto	Trabajo o consenso	Valor de mercado (fecha)	Vía de hardware	Demanda de pago frente a subsidio	Juicio
Qubic	«uPoW» (dice entrenar Aigarth)	Unos 50 millones de dólares (13 de noviembre)	CPU y GPU; ASIC para Dogecoin desde abril de 2026	Sin pruebas independientes de utilidad; cómputo usado para minería externa y recompras	Agregación de cómputo envuelta en un relato de IA
Bittensor	Consenso Yuma (puntuación ponderada por stake)	Unos 2,55 mil millones de dólares (11 de noviembre)	GPU de propósito general; sin vía ASIC	Unos 70 millones de dólares al año de ingresos verificables frente a más de 300 millones de emisiones	La única con ingresos externos a escala; aún un subsidio de 4–11 veces
Filecoin (referencia)	Pruebas de replicación y de espacio-tiempo	Unos 1,5 mil millones de dólares (final del T3 de 2025)	—	Utilización del 36 %; comisiones base trimestrales de 4100 dólares	Útil, pero falta demanda

FIG. 04 **El múltiplo de subsidio: cuántas veces sus ingresos externos paga una red**

Emisiones o valor de mercado divididos entre los ingresos pagados desde fuera de la red, escala logarítmica



Fuente: Blockworks (Aleo T3 2025), DefiLlama (comisiones de Render), CoinGecko, hoge.gg (Bittensor), Dealroom (Grass), Blockworks 0xResearch (Pearl); cálculos del autor.

La lección para el hardware: un trabajo fijo y regular pagado en un token líquido tiene ASIC en unos 3 a 16 meses, a menudo de aspirantes antes que de los líderes. Las afirmaciones de «utilidad» solo se sostuvieron donde un cliente externo paga por el resultado. Y los principales motores del hardware varado son un subsidio alto en relación con los ingresos y el control de los desarrolladores sobre la función de trabajo.

CAPÍTULO 06

EL PANORAMA GENERAL

06 El resto de IA × cripto: demanda real pero pequeña, y el valor se va al capital

Mercados de cómputo descentralizado

El cómputo descentralizado tiene clientes que pagan, pero la demanda es pequeña, concentrada y sobre todo reventa de inferencia de modelos abiertos. Las comisiones verificables en la cadena son pequeñas: DefiLlama muestra unos 2,35 millones de dólares de comisiones para Render en el último año y unos 680.000 dólares para Livepeer, y todo el sector DePIN ingresó solo unos 72 millones de dólares on-chain en 2025. Las cifras autodeclaradas son mayores pero de definición laxa: Aethir declara «más de 127,8 millones de dólares» de ingresos en el ejercicio 2025 sin aclarar si es facturación bruta o la parte del protocolo. El valor de los tokens y los ingresos se han separado en ambas direcciones: el valor de mercado de Render es unas 360 veces sus comisiones on-chain, mientras que el de Aethir, unos 87,7 millones de dólares, está por debajo de sus ingresos autodeclarados estimación.

La promesa de ser entre un 70 % y un 85 % más barato solo se cumple frente a los precios bajo demanda de los hiperescaladores. El 3 de noviembre, el precio medio ponderado de una H100 en Akash era de unos **2,73 dólares por GPU-hora** (rango 2,04–3,16 dólares), por debajo de unos 6,88 dólares en AWS y 11,06 en Google Cloud, pero por encima de unos 1,73 dólares en Vast.ai, 1,99 en la community cloud de RunPod y unos 2,44 con una reserva de un año (precios de las otras plataformas leídos el 12 de septiembre; Akash, que los publica, es parte interesada). La demanda es sobre todo inferencia genérica distribuida a través de OpenRouter, unos 1,7 mil millones de tokens al día en AkashML y más de 4 mil millones en io.net, en competencia directa con todas las neoclouds. Los mayores dólares del «cómputo descentralizado» se están recentralizando: Axe Compute, que tiene el token de Aethir como activo de tesorería, firmó un contrato a cinco años por 1,5 mil millones de dólares para alquilar su propio clúster de 9200 B300, frente a unos ingresos de 35.000 dólares en el primer trimestre de 2026.

Entrenamiento e inferencia descentralizados

El entrenamiento por internet se ha demostrado con 8000 a 40.000 millones de parámetros. Pluralis-8B entrenó durante 40 días en 330 nodos, sobre todo GPU de consumo, y alcanzó el 63 % de la eficiencia de una referencia centralizada con H100, aunque su artículo admite que la ejecución no fue descentralizada y necesitó nodos propios de Pluralis y una autoridad de admisión. Psyche, de Nous Research, completó la ejecución de preentrenamiento Consilience de 40.000 millones de parámetros. A escala de frontera el trabajo se recentralizó: INTELLECT-3 de Prime Intellect, un modelo de mezcla de expertos de 106.000 millones de parámetros, se entrenó durante dos meses en 512 H200 conectadas con InfiniBand a 400 Gbps. Las limitaciones son claras: los nodos necesitan al menos 200 Mbit/s de ancho de banda y una compresión de la comunicación de unas 100 veces, y la confianza y la verificación siguen dependiendo de los roles privilegiados del operador.

El dinero fue a empresas de capital que venden productos centralizados. Prime Intellect tenía más de 100 millones de dólares de ingresos anualizados antes de captar 130 millones con una valoración de 1 mil millones en julio; Venice, una aplicación de inferencia privada, declara más de 70 millones de dólares de ingresos anualizados y beneficios desde el primer trimestre de 2026 (cifras de la empresa), captó 65 millones con una valoración de 1 mil millones y funciona con GPU arrendadas. En el lado de los tokens, la red de entrenamiento central de Gensyn

sigue «por determinar» y su token vale unos 17,2 millones de dólares; el token de Ritual, una capa 1 nativa de IA, cayó un 72 % desde su máximo tras la mainnet (solo titular).

IA verificable

La producción eligió los entornos de ejecución de confianza (TEE). La computación confidencial de Hopper añade menos del 7 % de sobrecoste a las consultas típicas de LLM y casi nada en los modelos grandes; el sobrecoste en producción de Phala ronda entre el 0,5 % y el 5 %, pero a finales de 2025 pagaban 398 de sus 10.004 usuarios. La raíz de confianza de un TEE es el fabricante del chip: TEE.fail, divulgado en octubre de 2025, usó un interposer en el bus DDR5 de menos de 1000 dólares para extraer claves de atestación de máquinas Intel TDX y AMD SEV-SNP totalmente actualizadas, lo que a su vez socava la computación confidencial de las GPU de NVIDIA, y ambos fabricantes dicen que los ataques físicos quedan fuera de su modelo de amenazas. Ese es justo el escenario DePIN, donde los operadores de nodos tienen acceso físico. El zkML ya puede probar un LLM completo: DeepProve de Lagrange prueba GPT-2 a unos 174 tokens por minuto y Gemma 3 a unos 86, a órdenes de magnitud de la inferencia en producción y sin uso de pago declarado.

Pagos de agentes y datos

x402 es el mejor ejemplo aleccionador. TRM Labs cuenta unos 198,9 millones de transacciones y 52,7 millones de dólares liquidados por facilitadores conocidos desde mayo de 2025, el 99,6 % en USDC; tras eliminar los autopagos, los flujos por lotes y los vendedores con menos de diez compradores, la parte iniciada de verdad por agentes es solo del 0,6 % al 7,5 %, **unos 5000–11.000 dólares al mes**. Artemis estima que más del 95 % de la actividad es señalización del protocolo y no compraventa. Mientras tanto, los raíles los están estandarizando los actores establecidos: la x402 Foundation se creó dentro de la Linux Foundation el 14 de julio, con Visa, Mastercard, Stripe, Google, AWS, Cloudflare y Coinbase como miembros principales, y Tempo, que coescribió el Machine Payments Protocol con Stripe, lanzó su mainnet en marzo con una valoración de 5 mil millones de dólares. Las plataformas de tokens de agentes siguen funcionando con lógica de launchpad: Virtuals declara 479 millones de dólares de «aGDP», y el fundador de ElizaOS dio por muerto su token tras un acuerdo en una demanda colectiva. Es más probable que el valor lo capten los emisores de stablecoins y los actores establecidos de pagos; el papel distintivo de lo cripto aquí puede ser la identidad y la rendición de cuentas de los agentes.

Los datos son uno de los pocos subsectores con compradores reales fuera de la cadena. Grass declara 17 millones de dólares de ingresos por datos en 2025 y otros 17 millones en el primer semestre de 2026, aspira a 75 millones en el año y dice ser rentable; GRASS vale unos 328 millones de dólares, unas diez veces sus ingresos anualizados del primer semestre estimación. Su foso defensivo es la escala de captación del ancho de banda residencial distribuido, no lo cripto en sí. Las DAO de datos y los mercados de datos carecen de demanda de pago visible: Vana vale unos 21,5 millones de dólares y Ocean unos 27,6 millones, y Ocean está en litigio con Fetch.ai. La prueba de personalidad gana integraciones reales a medida que se multiplican los agentes: World dice haber verificado a casi 18 millones de personas, y Zoom, Docusign y Tinder admiten partes de World ID.

El capital ha votado

Sector	Demanda de pago verificable	Señales de subsidio y valoración	Juicio
Cadenas de trabajo útil	Bittensor, unos 70 millones de dólares al año de ingresos externos verificables; Pearl, cerca del 1 % de bloques con trabajo de IA	Aleo, 175–180 veces; Bittensor, 4–11 veces; Pearl, sobre todo subsidio	Los mecanismos tienen valor; la utilidad necesita clientes que paguen
Mercados de cómputo descentralizado	Comisiones on-chain de unos 0,68–2,35 millones de dólares al año (Livepeer, Render); autodeclarado, 4–128 millones	Valor de mercado de Render, unas 360 veces las comisiones	Real pero pequeño, sobre todo reventa de inferencia genérica

Sector	Demanda de pago verificable	Señales de subsidio y valoración	Juicio
Entrenamiento e inferencia descentralizados	Empresas de capital: Prime Intellect, más de 100 millones de dólares al año; Venice, más de 70 millones (infraestructura centralizada)	La mainnet de entrenamiento de Gensyn no está activa; el token de Ritual, −72 % desde su máximo	El avance técnico es real; los beneficios van al capital
IA verificable	TEE en producción (398 usuarios de pago en Phala); sin zkML de pago declarado	—	TEE útiles; zkML incipiente
Pagos de agentes	Flujos iniciados por agentes, unos 5000–11.000 dólares al mes	Unos 200 millones de transacciones, sobre todo señalización y especulación	Primero los raíles, después la demanda
Redes de datos	Grass, 17 millones de dólares en el S1 de 2026 (autodeclarado)	GRASS, unas 10 veces los ingresos anualizados	Uno de los pocos sectores con compradores reales

Si se ponen los seis sectores uno al lado del otro, la conclusión es coherente: **el subsidio es combustible para el arranque en frío, no un modelo de negocio.** El 13 de noviembre, la categoría de IA de CoinGecko valía en total unos 18,6 mil millones de dólares, pero cerca del 56 % corresponde a capas 1 de propósito general como NEAR e ICP, más proyectos de nivel de consenso como TAO, TAO en staking (quizá contado dos veces con TAO) y Pearl; los tokens de IA de capa de protocolo y de aplicación suman unos 8 mil millones de dólares o menos estimación. Es más o menos la valoración conjunta de cuatro empresas de capital: Tempo (5 mil millones de dólares), Nous (1,5 mil millones, según la prensa), Prime Intellect (1 mil millones) y Venice (1 mil millones). La financiación de capital riesgo cripto está estancada en unos 4,5–5 mil millones de dólares por trimestre, y el aumento procede de la infraestructura relacionada con la IA, cuya parte de los importes declarados subió del 8,1 % al 18,2 % en el tercer trimestre de 2026; y el 84,7 % de los 118 tokens lanzados en 2025 cotiza por debajo de su precio de lanzamiento. En una frase: **lo cripto aporta distribución y liquidación, el capital se lleva los beneficios**, al menos en 2026.

CAPÍTULO 07

07

ECONOMÍA DE LOS ASIC

La cuestión del ASIC: una fórmula, tres varas de medir

Una fórmula: por qué la eficiencia deja de pagar a partir de k = 5

Sea R el ingreso diario por unidad de cómputo (los ingresos de los mineros de la red divididos entre el hashrate de la red), c el coste diario de una unidad de cómputo en el hardware marginal (en Pearl, una GPU) y k cuántas veces es un ASIC más barato que ese hardware marginal (c_ASIC = c/k). Mientras las GPU sigan siendo los mineros marginales, la libre entrada y salida empuja R hacia c, así que un ASIC gana R − c/k por unidad de cómputo:

Techo de beneficio de un ASIC ≈ R × (1 − 1/k)

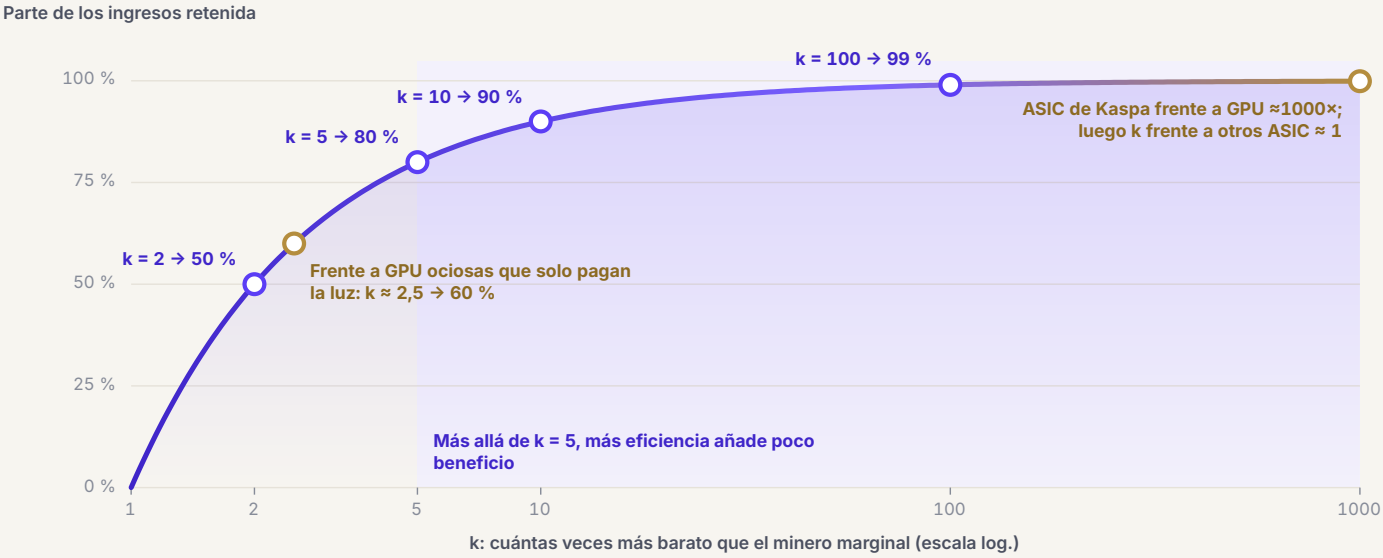
En el conjunto de la red, el techo de beneficio bruto anual de una flota de ASIC es aproximadamente los ingresos de los mineros de la red × (1 − 1/k), y solo hasta que los ASIC sustituyen a las GPU como mineros marginales.

La fórmula tiene tres consecuencias. Primera, (1 − 1/k) es cóncava: con k = 2 un ASIC retiene el 50 % de los ingresos, con k = 5 el 80 %, con k = 10 el 90 %, con k = 100 el 99 %. **Más allá de k = 5–10, más eficiencia casi no añade beneficio**; deciden el tamaño de R y del bote. Segunda, la referencia se reinicia: cuando los ASIC son los mineros marginales, k tiene que medirse frente a otros ASIC. Los ASIC de Kaspa tenían una ventaja de eficiencia

de unas 1000 veces sobre las GPU, que solo compró una ganancia breve; hoy la mejor KS7 gana 0,86 dólares al día. Tercera, en una cadena favorable a las GPU como Pearl, la referencia no suele ser el precio de compra de una GPU sino **la electricidad de una GPU ociosa**: una RTX 5090 propia a 525 W y 0,10 dólares por kWh cuesta unos 1,26 dólares al día de luz y gana unos 5–6 estimación. Frente a GPU que son costes hundidos, incluso un ASIC que use una quinta parte de la energía tiene una k efectiva de solo unos 2,5 si su amortización iguala a su electricidad, y un techo de beneficio de cerca del 60 % de R (un supuesto ilustrativo).

FIG. 05 **Por qué la eficiencia deja de pagar: un ASIC retiene como mucho 1 − 1/k de los ingresos**

Parte de los ingresos que un ASIC puede retener frente al minero marginal, según su ventaja de coste k



Fuente: Modelo del autor: cuando el hardware marginal fija el precio, los ingresos por unidad de cómputo tienden a su coste; cifras de Kaspa y de la RTX 5090 tomadas del informe.

Tres varas de medir: el bote, la escala de amortización y las reglas

La primera vara de medir es **el bote**. Si se anualizan las unas 2893 BTC de recompensas de bloque de la primera semana de noviembre (unos 222 millones de dólares), los mineros de bitcoin ingresan unos 11,5 mil millones de dólares al año estimación. Los mineros de Pearl ingresan unos 0,46–0,49 mil millones (a precios actuales); las recompensas de los probadores de Aleo, anualizadas a partir del tercer trimestre de 2025, unos 31 millones; y los mineros de Kaspa, con una recompensa de bloque de 1,945 KAS y diez bloques por segundo, unos 18 millones estimación. **Un año entero de ingresos de los mineros de Kaspa es menos que el extremo bajo del presupuesto de diseño de un solo chip de 5/4 nm (unos 30 millones de dólares).**

La segunda vara de medir es **el hashrate necesario para amortizar el NRE**:

h* = NRE ÷ [R × (1 − 1/k) × T × p]

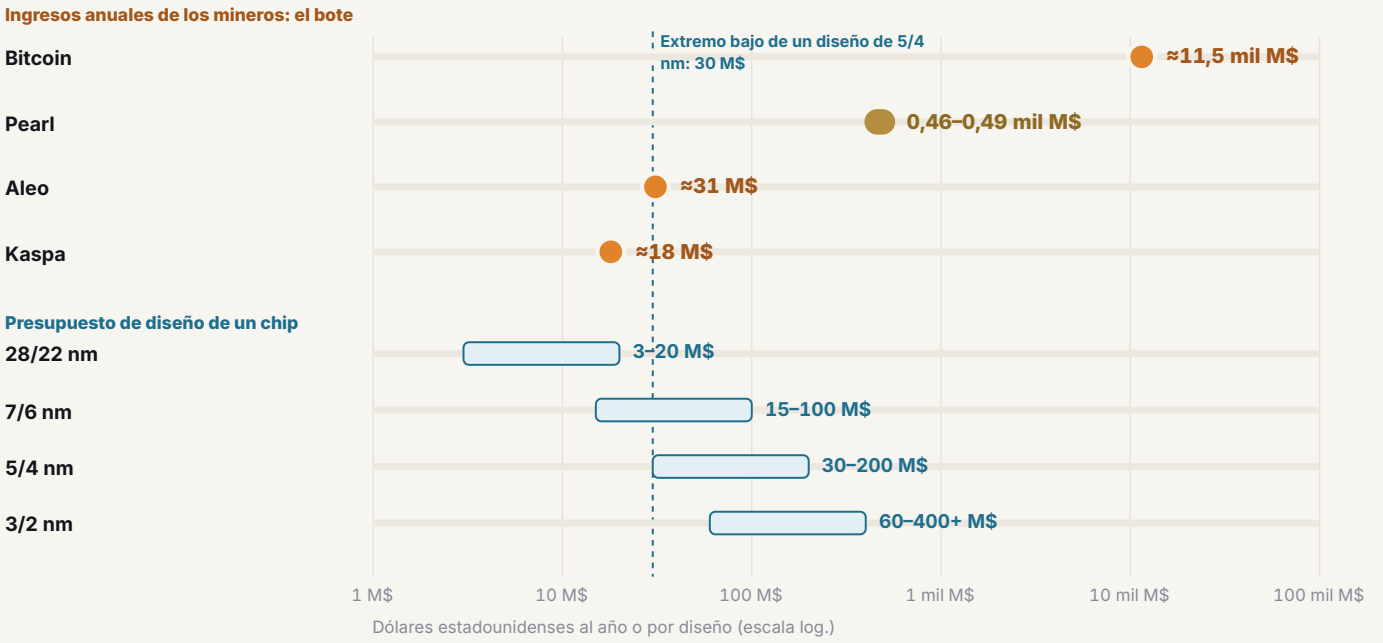
donde T es la ventana útil en días y p la probabilidad de que las reglas no cambien en esa ventana. Para Pearl, tomemos R = 14 dólares por PH/s y día, k = 5, T = 365 días y un NRE de 20 millones de dólares (un chip de matriz repetitiva en N7/N6; en el extremo bajo de los rangos de 5–30 millones de dólares para nodos intermedios de Silicon Analysts y de 15–100 millones para ASIC de cómputo en 7/6 nm de AnySilicon, porque un chip de minería usa poca IP y una estructura repetida; supuesto mío). Con p = 1, necesita unos **4,9 EH/s** funcionando un año entero, cerca del 5 % de la red; con p = 0,5, unos **9,8 EH/s**, cerca del 11 %; y si R vuelve al nivel del verano, unos 10 dólares, incluso con p = 1 hacen falta unos 6,9 EH/s estimación. La escala no es inalcanzable. El problema son p y T.

La tercera vara de medir es **el riesgo de reglas**, y a menudo es decisiva. La bifurcación anti-ASIC de Monero en 2018 dejó varada la X3 de Bitmain y le costó brevemente a la red casi la mitad de su hashrate. La bifurcación de Sia ese mismo año inutilizó los mineros de Bitmain y de Innosilicon, mientras que el precio de segunda mano del minero Obelisk, vinculado a los desarrolladores, subió de 2499 a más de 5000 dólares. Poco más de un mes después del Merge de Ethereum, menos del 20 % de los antiguos mineros de Ethereum había encontrado otro trabajo. **Quien controla la función de trabajo controla el destino de quienes invierten en hardware**, y Sia muestra que la «resistencia a los ASIC» puede convertirse en el foso defensivo de los propios desarrolladores. Sobrevive el hardware con un segundo mercado: tras el Merge, las GPU se fueron a la IA y a la nube, mientras que los ASIC de un solo algoritmo se amontonaron en Ethereum Classic. La ventana también se estrecha: el primer ASIC de Kaspa tardó unos 16 meses, el de Aleo unos cinco, y el primero en llegar a la frontera de eficiencia suele ser un aspirante y no un líder.

Red	Ingresos anuales de los mineros (aprox.)	Hardware marginal o de referencia	Estabilidad de las reglas	k al alcance de un nuevo entrante	Juicio
Bitcoin	11,5 mil millones de dólares	Bitmain S23 XP Hyd (8,9 J/TH, 25,5 dólares por TH)	Muy alta	≤1	Un bote grande, pero la k es de los líderes
Pearl	0,46–0,49 mil millones de dólares (a precios actuales)	GPU propias u ociosas (coste marginal ≈ electricidad) y GPU de centro de datos	Muy baja: cuatro bifurcaciones en dos meses; lista blanca FP8 pendiente	Quizá 3–10 veces en eficiencia análisis ; menos frente a GPU de coste hundido	El bote basta; el riesgo de reglas lo veta
Aleo	Unos 31 millones de dólares (T3 de 2025, anualizado)	IceRiver AE2 y otras máquinas dedicadas	Media: umbral de stake de ARC-46	≈1	Un bote pequeño, y el umbral sube
Kaspa	Unos 18 millones de dólares	KS7 (77 J/TH)	Alta	≈1	Ingresos anuales por debajo del extremo bajo del presupuesto de un diseño de 5/4 nm

FIG. 06 **La mayoría de los botes son menores que el presupuesto de diseño de un chip**

Ingresos anuales de los mineros de cuatro redes frente al coste de diseño de un chip por nodo, escala logarítmica



Fuente: Hashrate Index (bitcoin, noviembre de 2026), prlstats y Kryptex (Pearl), Blockworks (Aleo, T3 2025), minerstat (Kaspa), AnySilicon (presupuestos de diseño, 2026); anualizado por el autor.

El cuadrante de «bote grande, reglas estables y k > 1» está hoy vacío: bitcoin tiene el bote y las reglas estables, pero su k es de los líderes; el bote de Pearl basta, pero sus reglas son las menos estables; Kaspa y Aleo tienen reglas bastante estables y botes demasiado pequeños.

CAPÍTULO 08

08 CHIP A CHIP
Cinco tipos de chip, cinco respuestas

PoUW al estilo de Pearl: ningún chip INT7 dedicado; apostar por un minero que pueda servir inferencia

Los argumentos a favor de un ASIC son reales. Hoy solo cerca del 1 % de los bloques lleva trabajo de IA, así que el argumento económico del whitepaper —que un ASIC dedicado pierde la capacidad de hacer trabajo útil y por tanto gana menos— está ahora en su punto más débil. Las partes más caras de una GPU, la HBM y el encapsulado CoWoS, suponen cerca del 63 % del coste de fabricación de una H100, y el die lógico solo cerca del 9 %. Un chip que no haga más que multiplicaciones-acumulaciones de enteros de 7 bits, mantenga sus operandos en SRAM del propio chip y funcione a baja tensión podría plausiblemente lograr de 3 a 10 veces la energía por operación de una GPU del mismo nodo y una ventaja mayor en coste por operación [análisis; aún sin mediciones]. Los argumentos en contra pesan más. Los plazos de las fundiciones en nodos punteros pueden superar las 50 semanas, y un equipo nuevo necesita entre 18 y 24 meses desde el inicio hasta el volumen estimación; para cuando llegue un chip, es muy posible que Pearl haya pasado a FP8 y después a FP4. Y como PIP-3 define la

aritmética admitida mediante una lista blanca de dispositivos, un chip INT7 simplemente dejaría de funcionar en cuanto se active.

Así que mi conclusión es: **no hacer ahora el tape-out de un chip INT7 dedicado a Pearl.** Hay dos apuestas mejores. Una es un acelerador pensado para inferencia que también pueda minar: construido en torno a GEMM INT8/FP8 general y operandos residentes en SRAM, de modo que, si se niega el acceso a la minería, el valor acabe en un mercado de inferencia real. La otra es entrar en el debate del protocolo e impulsar una aritmética canónica neutral respecto al fabricante que un ASIC pueda implementar con exactitud, junto con un compromiso de estabilidad del algoritmo, que es la condición previa para que exista el cuadrante de «bote grande, reglas estables, k > 1». Hasta entonces, una arquitectura puede validarse en FPGA o en obleas multiproyecto (MPW) de nodo maduro, manteniendo el riesgo de NRE en el orden del millón de dólares. El capítulo 10 convierte esto en un presupuesto por etapas.

Mineros al estilo de bitcoin: la k es de los líderes, y la energía es el activo escaso

El riesgo de reglas de bitcoin es el más bajo, pero la competencia empuja la k de un nuevo entrante por debajo de 1. La S23 XP Hyd de Bitmain llega a 8,9 J/TH, se vende a 25,5 dólares por TH y se distribuye desde noviembre; la eficiencia de los modelos insignia bajó de 17,5 J/TH de la S21 en 2023 a 8,9 J/TH, cerca de un 49 % en tres años. Un nuevo entrante que distribuyera en volumen en 2027–2028 necesitaría unos 8 J/TH o menos para sostenerse estimación. Los aspirantes se marchan: Canaan declaró unos ingresos de 31,9 millones de dólares en el segundo trimestre de 2026 con una pérdida bruta de 29,3 millones y previó solo 11–15 millones para el tercer trimestre; las ventas de equipos propios de Bitdeer a compradores externos en el tercer trimestre cayeron a 0,3 millones de dólares, con un margen bruto del grupo del 3,4 %; Auradine se rebautizó como Velaura AI y se pasó a los chips de IA; y Core Scientific canceló un pedido de unos 15 EH/s de máquinas de 3 nm de la unidad Proto de Block y registró una pérdida de 41,9 millones de dólares.

La demanda también se reduce. El 9 de noviembre el hashprice era de unos 31,89 dólares por PH/s y día, la curva a plazo hasta abril de 2027 promediaba unos 32,7 dólares y las comisiones eran solo el 0,72 % de las recompensas de bloque. A 0,045 dólares por kWh, la S23 XP Hyd se amortiza en unos 3,1 años estimación, cerca del próximo halving, previsto hacia 2028. Más importante es el coste de oportunidad de la energía: los arrendamientos de IA y computación de alto rendimiento rinden hoy unos 1,67–1,9 millones de dólares por megavatio contratado y año, lo que con un resultado operativo neto del 80 % al 85 % equivale a entre 1,5 y 2,9 veces el beneficio bruto de la minería por megavatio (unos 0,56–0,9 millones de dólares) [estimación; las dos bases de megavatios difieren]. Core Scientific dejará de minar en el cuarto trimestre; Bitfarms, ahora Keel, ha cerrado todos sus emplazamientos de minería en EE. UU.; IREN ganó 200 millones de dólares con la nube de IA y solo 28 millones con la minería en un trimestre. **La energía y la capacidad de centro de datos son los activos; las máquinas de minería son consumibles.**

La aritmética del NRE es igual de implacable. La eficiencia de frontera exige hoy procesos de clase 3 nm (según fuentes secundarias, MicroBT usa Samsung SF3, y los chips de Auradine y de Proto son de 3 nm). Las máscaras cuestan unos 10–20 millones de dólares y el NRE completo 30–100 millones o más; a 20–25 dólares por TH y un margen bruto del 30 %, cada 1 EH/s vendido deja unos 6–7,5 millones de dólares de beneficio bruto, así que cubrir un gasto anual en I+D de la escala de Bitdeer, unos 110–160 millones de dólares, exige vender unos 20–25 EH/s al año a compradores externos, cerca del 2 % de la red estimación. La conclusión: **un equipo nuevo no debería entrar en el mercado abierto de mineros de bitcoin.** Tres casos siguen teniendo sentido: la integración vertical con energía barata propia; formatos de nicho como los mineros domésticos y de calefacción (Canaan también lanza productos de calefacción doméstica); y participar en las hojas de ruta de los líderes mediante IP y servicios de diseño.

Chips de entrenamiento de IA: no como equipo pequeño

En los chips de entrenamiento, k es lo más difícil de conseguir. A igual generación y precisión, Ironwood de Google supera a la B300 en eficiencia FP8 a nivel de chip solo entre 1,1 y 1,8 veces (la potencia de Ironwood es una estimación de un periodista) estimación; las 30 a 80 veces que se suelen citar del artículo de la TPU v1 se midieron frente a una K80, que no tenía tensor cores. La verdadera palanca es el formato numérico: pasar de FP8 a FP4 suele dar de 2 a 3 veces el rendimiento con el mismo silicio, y NVIDIA es la que más rápido avanza ahí: declara 50 PF de inferencia NVFP4 para Rubin, en volumen en el segundo semestre de 2026. Los umbrales de coste y suministro son aún más altos: los presupuestos de diseño en 3/2 nm van de 60 a 400 millones de dólares o más; la capacidad de CoWoS en 2026 ronda las 106.000–130.000 obleas al mes, y según la prensa TSMC estudia subidas de precio de hasta el 20 %; con un precio contractual de 15,6 dólares por GB para la HBM3E, solo la HBM de un acelerador de 288 GB cuesta unos 4500 dólares estimación. Este mercado pertenece al codesarrollo con hiperescaladores al estilo de Broadcom (capítulo 11). La conclusión: **los equipos pequeños no deberían construir chips de entrenamiento independientes**; mejor IP, chiplets o servicios de diseño.

Chips de inferencia de IA: cambiar generalidad por k, y compensa

La inferencia es donde los ASIC tienen más margen, porque k puede comprarse renunciando a la generalidad. Etched afirma que su servidor de ocho chips ejecuta Llama-70B a más de 500.000 tokens por segundo, unas 20 veces ocho H100; Taalas graba los pesos del modelo en ROM de máscara y afirma 10 veces menos consumo y 20 veces menos coste de fabricación (ambos dato del proveedor). Las arquitecturas centradas en SRAM evitan la HBM y el CoWoS, y sus nodos son más manejables: Taalas usa TSMC 6 nm, y Groq pasó de 14 nm a Samsung 4 nm. La distribución cripto no ayuda a los chips de inferencia, porque las redes de cómputo descentralizado funcionan casi por completo con NVIDIA; la vía de entrada es la economía de compras de las neoclouds y el gran mercado nacional de inferencia de China, donde los centros de datos con financiación estatal deben usar chips de IA nacionales, las importaciones de H200 se están frenando y Pekín quiere que las grandes empresas de internet ejecuten la inferencia en chips nacionales. La conclusión: **merece la pena construirlos**, siempre que haya un cliente claro y el grado de atadura a modelos y arquitectura sea una elección deliberada. Es además el mejor segundo mercado para un chip de minería PoUW.

Aceleradores ZK y FHE: cubrirse en ZK, posicionarse pronto en FHE

Las GPU ya han alcanzado la prueba ZK en tiempo real: SP1 Hypercube de Succinct probó el 99,7 % de los bloques de Ethereum en menos de 12 segundos con 16 RTX 5090. Cysic afirma que su chip ZK C1 es de 10 a 100 veces más eficiente que los probadores con GPU y que se distribuirá en 2026, pero su demanda procede en gran parte de sus propias recompensas en tokens dato del proveedor; no encontré pruebas públicas de que los ASIC para el rompecabezas de Aleo se usen en pruebas ZK reales. La prueba ZK es entera y determinista, de forma parecida a lo que hacen los chips de minería, pero los sistemas de prueba cambian deprisa (SP1, RISC Zero, Halo2, Plonky2 y otros). La conclusión: **esperar y cubrirse**, con motores MSM, NTT y de hash programables, primero en FPGA y después como ASIC.

FHE es la primitiva más ligada al hardware y la de mayor potencial. Heracles de Intel, en Intel 3 con 197 mm², 176 W y 48 GB de HBM3, ejecuta operaciones FHE entre 1074 y 5547 veces más rápido que un Xeon de 24 núcleos, pero Intel no tiene planes comerciales y el responsable del proyecto se marchó en diciembre de 2025. Niobium construye sobre Samsung 8LPU con SemiFive como socio de diseño (un contrato de desarrollo de unos 6,86 millones de dólares) y ha captado más de 23 millones de dólares; Optalysys, con un enfoque fotónico, captó 31 millones de dólares y dice que aún faltan dos o tres años para un chip comercial completo; y el diseño FPGA de código abierto de Zama alcanza unos 13.000 bootstraps programables por segundo a unos 200 W, la referencia pública que cualquier ASIC tiene que superar con claridad. La conclusión: **posicionarse pronto y validar con activos ligeros**, para una ventana en torno a 2027–2029. Es más probable que la demanda se ancle en la IA cifrada

para empresas y administraciones que en las propias criptomonedas, y FHE necesita HBM, así que su estructura de costes se parece a la de un chip de entrenamiento.

Tipo de chip	De dónde sale k	Demanda y bote	NRE y umbrales	Riesgo de reglas o de atadura	Segundo mercado	Juicio
Chip de minería PoUW al estilo de Pearl	Prescindir de HBM, CoWoS y unidades de propósito general: quizá 3–10 veces en J/op análisis	Ingresos de los mineros de unos 0,46–0,49 mil millones de dólares al año, volátiles	Matriz repetitiva en N7/N6, unos 10–30 millones de dólares análisis	Muy alto: el protocolo decide el formato numérico y la lista blanca de dispositivos	Inferencia de IA, si se construye como GEMM INT8/FP8 general	Ningún chip INT7 dedicado; construir para inferencia y que pueda minar, o esperar a una aritmética canónica
Chip de minería SHA-256 de bitcoin	k ≤ 1 frente a Bitmain	Unos 11,5 mil millones de dólares al año; mercado abierto en manos de dos empresas	Máscaras de clase 3 nm de 10–20 millones de dólares; NRE completo de 30–100 millones o más	Bajo (algoritmo estable); competencia alta y riesgo de halving	Casi ninguno (un solo algoritmo)	Los nuevos entrantes, fuera del mercado abierto; solo energía propia, formatos de nicho o IP
Chip de entrenamiento de IA	Matrices sistólicas de la misma generación, solo 1,1–1,8 veces; la palanca es el formato	El mayor (Broadcom, unos 58 mil millones de dólares de ingresos en semiconductores de IA en el ejercicio 2026)	Diseño en 3/2 nm de 60–400 millones de dólares o más; CoWoS y HBM	La cadencia anual de NVIDIA	—	Los equipos pequeños hacen IP, chiplets y servicios de diseño, no chips independientes
Chip de inferencia de IA	Renunciar a la generalidad: 10–20 veces declarado dato del proveedor	Grande y fragmentado; el mercado nacional de China continental es grande por sí mismo	Las arquitecturas SRAM en N6/N7–N4 son manejables	Atadura a modelos y arquitectura	Es el mercado principal	Merece la pena: firmar al cliente y luego elegir el grado de atadura
Acelerador ZK	10–100 veces las GPU, declarado dato del proveedor	Pequeño, impulsado sobre todo por recompensas en tokens	N7/N6 y luego N5; primero FPGA	Los sistemas de prueba cambian deprisa	Limitado fuera de ZK	Esperar y cubrirse: motores MSM y NTT programables
Acelerador FHE	1074–5547 veces un Xeon (Heracles)	Aún sin formar; anclado en empresas y administraciones	Necesita HBM; estructura de costes cercana a la de los chips de entrenamiento	Estándares y ecosistema sin asentar	Computación cifrada, IA privada	Posicionarse pronto, validar con activos ligeros; ventana 2027–2029

FIG. 07 **Cinco tipos de chip, cinco respuestas**

De dónde sale la ventaja de coste k, en qué mercado puede venderse el silicio si cambian las reglas y el veredicto

Chip de minería PoUW al estilo de Pearl

AHORA NO

k: sin HBM ni unidades generales: 3–10× por op. (inferido)

Segundo mercado: inferencia de IA, si es GEMM INT8/FP8 general

Apostar por un chip de inferencia que pueda minar, no por una pieza INT7 dedicada

Chip de minería SHA-256 de bitcoin

MANTENERSE FUERA

k: ≤1 frente a los líderes

Segundo mercado: casi ninguno

Solo con energía propia, un formato de nicho o como IP

Chip de entrenamiento de IA

IP, NO CHIPS

k: 1,1–1,8× a igual precisión

Segundo mercado: —

Deciden el encapsulado y la HBM; hacer IP, chiplets, servicios

Chip de inferencia de IA

CONSTRUIR

k: renunciar a la generalidad: 10–20× declarado

Segundo mercado: es el mercado principal

Firmar al cliente y luego elegir el grado de atadura

Acelerador de pruebas ZK

ESPERAR Y CUBRIRSE

k: 10–100× las GPU, declarado

Segundo mercado: limitado fuera de ZK

Motores MSM y NTT programables; primero FPGA

Acelerador FHE

POSICIONARSE PRONTO

k: 1074–5547× un Xeon (Heracles)

Segundo mercado: computación cifrada, IA privada

Una ventana 2027–2029; validar con activos ligeros

Fuente: Síntesis del autor de los capítulos 7 y 8; cifras de fabricantes señaladas como tales en el informe.

CAPÍTULO 09

NODOS DE PROCESO

09 Nodos de proceso: que decida el mayor coste a lo largo de la vida útil

La escalera de costes: las obleas de vanguardia cuestan diez veces las maduras

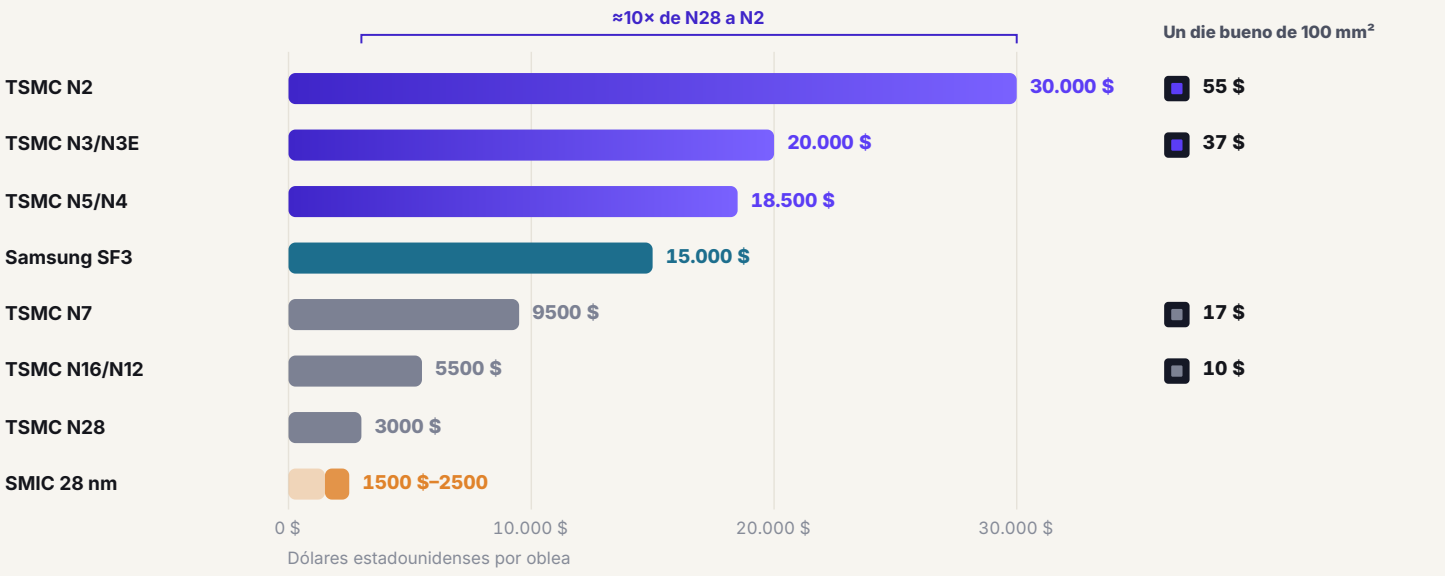
Empecemos por las cuentas. Una oblea de vanguardia cuesta hoy unas diez veces una madura: TSMC N2, unos 30.000 dólares; N3, unos 20.000; N7, unos 9500; y N28, unos 3000; Samsung SF3, unos 15.000, cerca de un 25 % menos que N3; y según la prensa el coste de fabricación de clase 7 nm de SMIC es entre un 40 % y un 50 % superior al de TSMC. Según la prensa, TSMC subirá los precios otro 3 % a 6 % a partir de enero de 2027, sobre todo en 2 nm y 3 nm. Las máscaras y los costes de diseño difieren todavía más, y las fuentes no coinciden, así que la tabla muestra ambos.

Nodo	Precio de la oblea (estimación)	Juego de máscaras (valor del modelo / rango orientativo)	Presupuesto de diseño (AnySilicon)	MPW (por proyecto)
TSMC N2	Unos 30.000 dólares	—	3/2 nm: 60–400 millones de dólares o más	—

Nodo	Precio de la oblea (estimación)	Juego de máscaras (valor del modelo / rango orientativo)	Presupuesto de diseño (AnySilicon)	MPW (por proyecto)
TSMC N3/N3E	Unos 20.000 dólares (18.000–27.000)	15 millones de dólares / 10–20 millones	Igual que arriba	—
Samsung SF3	Unos 15.000 dólares (13.000–17.000)	12 millones de dólares	—	—
TSMC N5/N4	Unos 18.500 dólares (16.000–21.300, baja confianza)	6,5 millones de dólares / 10–20 millones	5/4 nm: 30–200 millones de dólares	—
TSMC N7	Unos 9500 dólares (8100–10.900)	3,5 millones de dólares / 5–10 millones	7/6 nm: 15–100 millones de dólares	Comercial, 100.000–500.000 dólares
TSMC N16/N12	Unos 5500 dólares	1,8 millones de dólares	16/12 nm: 6–40 millones de dólares	Académico, unos 35.000 dólares por mm²
TSMC N28	Unos 3000 dólares	1 millón de dólares / 1–3 millones	28/22 nm: 3–20 millones de dólares	Comercial, 50.000–150.000 dólares; académico, unos 12.000 dólares por mm²
SMIC 28 nm	1500–2500 dólares	—	—	—
SMIC clase 7 nm (N+2/N+3)	Un 40–50 % más que TSMC	—	—	—

FIG. 08 **Una oblea de vanguardia cuesta diez veces una madura**

Precio estimado por oblea de 300 mm según el nodo, y coste de un die bueno de 100 mm² con un rendimiento del 85 %



Fuente: Precios de obleas de Silicon Analysts (estimaciones, N5 de baja confianza), Wccftech (SMIC); los costes por die son cálculos del autor con un rendimiento supuesto.

Después, la física. Las mediciones clásicas de Horowitz muestran que una instrucción programable cuesta unos 70 pJ, mientras que la aritmética en sí cuesta unos pocos picojulios: **la aritmética es barata y la sobrecarga es cara**, y esa es la razón de ser de las matrices sistólicas y de los chips dedicados. Reducir a la mitad el tamaño de las estructuras reduce aproximadamente a la mitad la energía por operación, pero la tensión de alimentación se

ha estancado en torno a 0,9–1,1 V. Una oblea N2 cuesta entre un 50 % y un 67 % más que una N3, así que, salvo que la densidad aumente más de unas 1,5 veces, el coste por transistor sube en vez de bajar estimación: los nodos punteros tienen que amortizarse con eficiencia y rendimiento. Para un die de 100 mm² con un rendimiento del 85 %, el coste de cada die bueno es de unos 55 dólares en N2, 37 en N3, 29–34 en N5, 17 en N7 y 10 en N16 [estimación; el rendimiento es un supuesto].

Elegir el nodo según el tipo de chip y el equipo

Mi regla es: **que la mayor partida del coste del chip a lo largo de su vida útil elija el nodo.** En los chips de cómputo que funcionan a plena carga día y noche (minería, ZK), la electricidad domina el coste de vida útil, así que, una vez estable el algoritmo, deberían pasar a N5/N4 y N3/SF3; mientras el algoritmo no esté asentado o falte dinero, conviene validar primero en MPW de nodo maduro. En los chips de entrenamiento el cuello de botella es el encapsulado y la memoria (la HBM y el CoWoS suponen cerca del 63 % de la lista de materiales de una H100), así que el nodo es secundario. En los chips de inferencia la primera decisión es la arquitectura de memoria: los diseños centrados en SRAM pueden quedarse en N6/N7 a N4, mientras que los diseños con HBM heredan la estructura de costes de los chips de entrenamiento. FHE está limitado por el ancho de banda de memoria y no puede evitar la HBM.

Tipo de chip	Startup o equipo pequeño	Empresa a escala	Vía de China continental	Puntos clave
Chip de minería PoUW por multiplicación de matrices (residente en SRAM, sin HBM)	Prototipo en MPW de 28 nm o 16/12 nm; primer volumen en N7/N6; N5/N4 cuando el algoritmo y los ingresos sean estables; dies de 50–150 mm² en encapsulados flip-chip estándar	N3/N3E o Samsung SF3	SMIC N+1/N+2 es viable (no hacen falta HBM ni CoWoS), con un coste por die bueno un 40–50 % mayor	Domina la energía; las reglas del protocolo son el mayor riesgo
Chip de minería SHA-256	No perseguir la frontera; los formatos de nicho pueden usar nodos maduros	Clase 3 nm; los líderes pasan de TSMC 5 nm a 2 nm y de Samsung SF3 a 2 nm (fuentes secundarias)	Difícil igualar la eficiencia del nodo puntero	El precio de la energía y el ciclo del halving deciden los márgenes
Chip de entrenamiento de IA	Ningún chip independiente: IP, chiplets, servicios de diseño	N3P y luego N2/A16; 2–8 chiplets; asegurar pronto CoWoS y HBM	SMIC N+2/N+3 con HBM nacional (el modelo de Huawei); FP8 por chip de unas 0,2–0,5 veces el de NVIDIA, compensado con la escala del sistema	El encapsulado y la HBM son el cuello de botella
Chip de inferencia de IA	Centrado en SRAM o cableado, N6/N7 o N4/N5/SF4; solo INT8, FP8 y FP4	N3/N2 con HBM (contexto largo, mezcla de expertos), o SRAM en varios chiplets	SMIC N+2 con LPDDR o GDDR; la memoria de clase HBM es el punto de estrangulamiento	El grado de atadura arquitectónica es la decisión central
Acelerador ZK	Empezar en FPGA o MPW de 16/12 nm; después N7/N6 y N5	N5/N3	SMIC N+1/N+2 es viable	Entero y determinista; fácil de verificar
Acelerador FHE	FPGA (frente a la referencia de código abierto) o MPW de 16/12 nm; el volumen necesita HBM	N5/N3 con HBM	Limitado por el acceso a la HBM	La velocidad la fija el ancho de banda de memoria, no la aritmética

Avanzado o maduro, cuatro tipos de fundición y chiplets

Opción	Principales puntos fuertes	Principales puntos débiles	Más adecuado para
Nodos punteros (N3/N2, SF3/SF2, Intel 18A)	Cada nodo completo reduce aproximadamente a la mitad la energía por operación; SRAM densa	Obleas 3–10 veces, máscaras 5–15 veces y presupuestos de diseño 10–20 veces los de los nodos maduros; plazos de más de 50 semanas; en N2 el coste por transistor puede subir	Chips de cómputo con algoritmos estables que funcionan día y noche; chips de entrenamiento
Nodos maduros e intermedios (N7/N6, N16/12, N28)	Máscaras del orden del millón de dólares, MPW baratas, iteración rápida, muchas fundiciones (TSMC, UMC, GlobalFoundries, SMIC)	Eficiencia poco competitiva a plena carga día y noche	Prototipos, mercados de nicho, inferencia en el borde
TSMC	El mejor rendimiento y ecosistema	Los precios más altos; capacidad de N2/N3 ajustada; largas colas para el CoWoS	Equipos que persiguen la frontera de eficiencia con compromisos de volumen
Samsung Foundry	Obleas cerca de un 20–25 % más baratas (SF3 a unos 15.000 dólares frente a unos 20.000 de N3); la usan la segunda generación de Groq y otros	El negocio de fundición pierde dinero y aspira al equilibrio a finales de 2027	Chips de cómputo sensibles al coste
Intel Foundry	Chips complejos como Heracles probados en Intel 3	Sin precios externos públicos ni datos de clientes para 18A; el negocio de fundición pierde dinero	Evaluar ecosistema y soporte caso por caso
Fundiciones de China continental (SMIC y otras)	Cerca de un gran mercado nacional; obleas de 28 nm a 1500–2500 dólares	Capacidad de clase 7 nm de unas 45.000–60.000 obleas al mes, absorbida en gran parte por Huawei; rendimientos de cerca de un tercio de los de TSMC; costes un 40–50 % más altos; solo DUV	Chips de inferencia, minería y ZK para el mercado nacional
Chiplets	Superan los límites de retícula y de rendimiento; estándar en los sistemas con HBM (MI355X, Ironwood, Trainium3)	Mayor coste de encapsulado y dependencia del CoWoS; las pruebas de known-good die añaden coste	Chips de entrenamiento y de inferencia grandes; innecesarios para chips pequeños de minería o ZK

CAPÍTULO 10

UN PRESUPUESTO POR ETAPAS

10 Del modelo al silicio: un presupuesto por etapas para un primer chip

Pruebas antes que máscaras

Un chip full-custom es una serie de apuestas, cada una mayor que la anterior, y la disciplina consiste en comprar pruebas antes de comprar máscaras. Uso cuatro puertas análisis. La puerta 0 es un modelo: un estudio de arquitectura que estima J/op y $/op a partir de la aritmética, la SRAM y la E/S, frente al hardware de referencia del capítulo 7; cuesta semanas y decenas de miles de dólares. La puerta 1 es un prototipo en FPGA que ejecuta la carga real de extremo a extremo, shares de minería o tokens de inferencia, y demuestra la ruta de datos y la integración con el protocolo; cuesta unos meses y unos cientos de miles de dólares. La puerta 2 es una oblea multiproyecto: una plaza comercial cuesta unos 50.000–150.000 dólares en N28 y 100.000–500.000 dólares en N7, y el silicio medido sustituye a las estimaciones de energía del modelo. La puerta 3 es el juego completo de máscaras y el primer lote de producción: máscaras de cerca de 1 millón de dólares en N28, 3,5 millones en N7 y 15 millones en N3, dentro de un presupuesto de diseño de 15–100 millones de dólares en 7/6 nm. Cada puerta necesita una cifra que detenga el proyecto, como una k inferior a 3 frente al hardware marginal, o un hashrate de amortización h* que la ventana no pueda alcanzar.

FIG. 09 **Pruebas antes que máscaras: un presupuesto por etapas para un primer chip**

Gasto acumulado en dólares estadounidenses de un proyecto N7/N6 a través de cuatro puertas, escala logarítmica; cada puerta tiene una cifra que detiene el proyecto



Fuente: Precios de obleas, MPW y máscaras como en el capítulo 9 (Silicon Analysts, AnySilicon, valores de modelo); los plazos y el gasto de cada puerta son supuestos ilustrativos del autor.

La escalera es empinada. En un proyecto N7/N6, las tres primeras puertas juntas cuestan unos 1,5 millones de dólares y unos once meses; la cuarta cuesta más de diez veces más estimación. Por eso el orden importa más que el nodo: si el protocolo cambia su aritmética, o el comprador se retira, un equipo que se detuvo en la puerta 2 pierde el precio de un prototipo, no el de un tape-out.

Cuánto cuesta cada chip, y qué tiene que cumplirse

Proyecto	Puerta 2 (prototipo)	Puerta 3 (nodo y presupuesto de diseño)	Un die bueno de 100 mm²	Qué tiene que cumplirse para amortizarlo
Acelerador pensado para inferencia que pueda minar	MPW en 16/12 nm o N7, 0,1–0,5 millones de dólares	N6/N7, unos 15–30 millones de dólares estimación	Unos 17 dólares en N7, antes del encapsulado y la memoria	Un cliente de inferencia firmado cubre la mayor parte del NRE; los ingresos de la minería son un extra, no el escenario base
Minero INT7 dedicado a Pearl (no recomendado ahora)	MPW en 28 nm, 0,05–0,15 millones de dólares	N7/N6, unos 20 millones de dólares	Unos 17 dólares en N7	Unos 4,9 EH/s durante un año entero con p = 1, o 9,8 EH/s con p = 0,5 (capítulo 7)
Acelerador ZK	FPGA y después MPW en 16/12 nm	N7/N6 y después N5	Unos 10 dólares en N16, 17 en N7	Que los sistemas de prueba se asienten en núcleos MSM y NTT que un diseño fijo pueda servir
Acelerador FHE	FPGA frente a la referencia de código abierto	N5/N3 con HBM, 30 millones de dólares o más	Dominado por la HBM y el encapsulado	Un comprador de IA cifrada a gran volumen, que todavía no existe

De ahí se siguen dos consecuencias comerciales. Primera, la eficiencia más barata es la arquitectónica: mantener los operandos en SRAM y prescindir de la HBM y el CoWoS elimina la mayor parte del coste de una GPU (cerca del 63 % de la lista de materiales de una H100) antes de elegir ningún nodo. Segunda, el tiempo es un coste: un equipo nuevo necesita entre 18 y 24 meses desde el inicio hasta el volumen y los plazos en la vanguardia superan las 50 semanas, así que un protocolo que cambie su aritmética cada pocos meses puede dejar obsoleto un chip antes de que la puerta 3 se amortice.

Tres reglas para un equipo pequeño

En conjunto, estas cifras dan tres reglas análisis. Primera, **quedarse en la puerta 1 o 2 hasta que haya un comprador concreto o una versión de protocolo congelada**; ni el precio del token ni el hashrate sustituyen a ninguno de los dos. Segunda, **comprar k primero con la arquitectura y después con el nodo**: una matriz residente en SRAM en N7 puede superar a una pieza de propósito general en N3 para una carga concreta, con una fracción del coste de las máscaras. Tercera, **anotar el hashrate o el volumen de amortización, con p, antes del tape-out**, y fijar los pagos por hitos contra una versión de protocolo fija, para que sea el cliente, y no la casa de diseño, quien asuma el riesgo de reglas.

CAPÍTULO 11

11 OPORTUNIDADES
Dónde están las oportunidades, y qué haría yo

Chips a medida: la demanda procede abrumadoramente de la IA

En 2026 la demanda de chips a medida procede abrumadoramente de la IA. Los ingresos de Broadcom en semiconductores de IA en el tercer trimestre de su ejercicio 2026 fueron de 16,7 mil millones de dólares, un 221 % más que un año antes, unos 58 mil millones en el año, el 73 % procedente de XPU para seis clientes (según una transcripción automática de la conferencia de resultados). Los ingresos de Global Unichip en el segundo trimestre subieron un 128 % interanual, con los ingresos de producción llave en mano un 188 % más y un margen bruto del 21,5 %. El reparto de ingresos de Alchip por nodo en el primer trimestre fue del 21 % en 3/2 nm, el 63 % en 7/5 nm y el 16 % en 12 nm o más. Los nuevos pedidos de VeriSilicon para 2026 llegaron a 8,24 mil millones de RMB (unos 1,16 mil millones de dólares) a finales de abril, el 91,37 % de ellos para computación de IA. El mercado tiene tres niveles: arriba, codesarrollos plurianuales por miles de millones; en medio, proyectos de 7 nm a 2 nm con un NRE de unos pocos millones a más de cien millones de dólares (5–30 millones de dólares en nodos intermedios, 30–100 millones o más en la vanguardia); y una larga cola de ASIC de nodo maduro con un NRE de 0,5–5 millones de dólares. Los clientes cripto y de minería son un nicho pequeño y cíclico, pero son hoy el «primer cliente» más rápido y que menos permisos requiere de los que pagan por cómputo.

Las capacidades de un equipo de chips de minería que mejor se transfieren —funcionamiento a baja tensión cerca del umbral, matrices de cómputo replicadas masivamente, suministro de potencia y refrigeración de clase kilovatio y tape-outs rápidos en nodos punteros— son justo las que necesitan los chips de inferencia y los aceleradores ZK y FHE. Los mercados de capitales lo valoran: Velaura AI (antes Auradine) captó 110 millones de dólares en agosto con una valoración de más de 1 mil millones. El contraejemplo es la unidad de chips de IA para el borde de Canaan, que tuvo unos ingresos de 0,9 millones de dólares en el ejercicio 2024 mientras consumía el 15 % de los gastos operativos, y se cerró en junio de 2025. La diferencia está en una vía propia al mercado, no en reutilizar los canales de la minería.

Desarrollo de IA: las horas son un commodity; el valor está en las juntas

En la gama baja, los servicios de desarrollo de IA se han convertido en un commodity: las tarifas típicas en Clutch son de 25–49 dólares la hora, y de 50–99 en Estados Unidos, Canadá y mercados similares; un equipo de desarrollo blockchain cuesta de unos 15.000 a más de 90.000 dólares al mes. Vender horas no tiene foso defensivo. El valor está en las juntas entre campos: kernels y compiladores para chips a medida, minería integrada en el servicio de inferencia (el plug-in de vLLM de Pearl es un ejemplo), inferencia verificable que combina TEE con comprobaciones puntuales con ZK, y pagos e identidad para agentes.

Sistemas IA × cripto: firmar al comprador de cómputo antes que el token

Para quien diseñe una cadena o una red de cómputo nativa de IA, la caja de herramientas de verificación de 2026 ya ha tomado forma. Por coste por unidad de trabajo, de menor a mayor: muestreo estadístico (la prueba de muestreo de Hyperbolic), delegación arbitrada (Verde de Gensyn, que solo necesita un verificador honesto pero exige ejecución exacta bit a bit), TEE (velocidad casi nativa, pero confianza en el fabricante del chip), redundancia completa y zkML general. El diseño práctico es por capas: TEE o muestreo para la baja latencia, juegos de

disputa o comprobaciones puntuales con ZK para la finalidad; un artículo de diciembre de 2025 simuló la computación confidencial de la H100 con pruebas de fraude y comprobaciones puntuales aleatorias con ZK al 99 % del rendimiento centralizado y con un coste marginal de 0,07 dólares por consulta. La filosofía de hardware decide la seguridad y la distribución: una PoUW nativa de GPU está protegida al principio por GPU en la nube alquilables, así que el coste de un ataque está ligado a los precios de alquiler, y Qubic muestra que la prueba de trabajo sobre hardware de propósito general puede ser capturada en bloque por incentivos externos.

El coste y el tiempo no son la barrera. Pearl pasó del artículo (abril de 2025) a la mainnet en unos 12 meses, como bifurcación de btcd, con nodo completo, monedero, cliente ligero y minero desde el lanzamiento. La mayoría de las auditorías DeFi cuestan 25.000–100.000 dólares, las auditorías empresariales o multicadena 150.000–250.000 dólares o más, los circuitos ZK entre un 80 % y un 120 % más, y los proyectos con valor real en riesgo presupuestan 150.000–500.000 dólares al año para seguridad. Las barreras reales son la distribución y la liquidez. Algunos fundadores cuentan que Binance les ha pedido entre el 0 % y el 3,5 %, y supuestamente el 8 %, de la oferta de tokens para listarlos (Binance niega lucrarse con los listados); los seis proyectos revisados por Cointelegraph que usaron acuerdos de creación de mercado de «préstamo de tokens más opción de compra» cayeron, y el creador de mercado de Movement vendió unos 38 millones de dólares en tokens el día del listado. Las alternativas son la liquidez on-chain, como en el caso de Hyperliquid, que superó los 10 mil millones de dólares de valor de mercado sin estar listada en Binance y cuyos mercados perpetuos HIP-3 superaron los 3,2 mil millones de dólares de interés abierto en junio, y las ventas públicas a través de grandes plataformas, como cuando Monad captó 269 millones de dólares a través de Coinbase. Un valor totalmente diluido alto con un free float bajo fue el predictor más claro de la caída tras el lanzamiento en 2025–2026.

El diseño del token forma parte del diseño del sistema. La emisión debe seguir a la demanda de pago, no adelantarse a ella: las redes de este informe con los múltiplos de subsidio más sanos vinculan las recompensas a un trabajo que paga un comprador externo, y las peores pagan hashrate o stake sin nada detrás. Una regla práctica es publicar, antes del lanzamiento, el múltiplo de subsidio que la red se propone en su primer, segundo y tercer año, y recortar la emisión cuando no se cumpla análisis.

Dirección	Quién compra, y disposición a pagar	Ventana	Capacidades clave	Riesgo principal
Acelerador pensado para inferencia que pueda minar	Neoclouds, despliegues privados, redes PoUW	2027–2028 (según las bifurcaciones a FP8 y FP4)	GEMM INT8, FP8 y FP4; arquitectura SRAM; integración con el protocolo	Admisión por el protocolo, atadura arquitectónica
Chip de inferencia SRAM o cableado	Neoclouds, centros de datos de China continental, borde	Ahora	Compiladores y adaptación de modelos; diseño en N6–N4	Los modelos cambian demasiado deprisa
Chip de minería (de nicho o con integración vertical)	Mineros con energía propia; mineros domésticos y de calefacción	Ahora, acotado por el halving y el hashprice	Matrices de baja tensión, eficiencia, cadena de suministro de volumen	Guerras de precios de los líderes, precios de la energía
Aceleración de pruebas ZK	Redes de probadores, rollups	Esperar a que converjan los sistemas de prueba	MSM y NTT programables	La demanda depende de las recompensas en tokens
Aceleración FHE	IA cifrada para empresas y administraciones; cadenas de privacidad	2027–2029	NTT, sistemas con HBM, ingeniería criptográfica	Demanda aún sin formar; ciclos de venta largos

Dirección	Quién compra, y disposición a pagar	Ventana	Capacidades clave	Riesgo principal
Refuerzo de TEE e inferencia verificable	Redes de inferencia confidencial, DePIN	Ahora	Cifrado de memoria con integridad, raíces de atestación independientes, comprobaciones puntuales con ZK	Dependencia de los ecosistemas de los fabricantes
Cadena o red de cómputo nativa de IA	Compradores de inferencia, mineros, desarrolladores	Ahora	Diseño de la verificación, diseño de la emisión, distribución	Caída tras el lanzamiento, liquidez del token

Tres caminos que tomaría, cada uno con un primer paso

Si tuviera que elegir, tomaría como mucho tres caminos, cada uno con un primer paso pequeño que lo ponga a prueba antes de comprometer dinero análisis.

- **Camino A: un acelerador pensado para inferencia que también pueda minar.** Primer paso: un estudio de arquitectura de cuatro semanas que modele J/op y $/op de una matriz GEMM INT8/FP8 residente en SRAM frente a la RTX 5090 y la H100, tanto en el trabajo de Pearl como en inferencia de transformers, seguido de una validación en FPGA o en MPW de nodo maduro con un NRE del orden del millón de dólares.
- **Camino B: servicios de silicio a medida con lo cripto como primer cliente y la IA como segundo mercado.** Primer paso: convertir el modelo de NRE y de punto de equilibrio de este informe en un estudio de viabilidad de pago para proyectos de chips de cómputo, con pagos por hitos y aceptación contra una versión de protocolo fija.
- **Camino C: una red nativa de IA que firme a su comprador de cómputo antes que su token.** Primer paso: una nota de diseño sobre verificación por capas (TEE más muestreo más comprobaciones puntuales con ZK) y sobre una aritmética canónica neutral respecto al fabricante, redactada con un comprador de inferencia concreto antes de cualquier diseño del token.

CAPÍTULO 12

12 INDICADORES
Qué vigilar, de noviembre de 2026 a 2027

Cuándo	Qué vigilar	Por qué importa
En cualquier momento	La altura de activación de FP8 en Pearl	Decide si el hardware entero y las tarjetas que no son de NVIDIA siguen admitidos, y si la cadena se divide
Mensual	La proporción de bloques de Pearl con trabajo de IA	La mejor prueba de si lo «útil» se está haciendo realidad
Mensual	Si los proveedores de inferencia trasladan los ingresos de la minería a sus clientes	La demanda de pago se encuentra con la emisión: el modelo que Pearl necesita
Trimestral	Los ingresos verificables de Bittensor frente a sus emisiones	Si la mayor red de trabajo útil está cerrando su brecha de subsidio
Trimestral	La proporción de flujos x402 iniciados de verdad por agentes	Si los pagos de agentes pasan de la señalización al comercio
Enero de 2027	La subida de precios del 3 % al 6 % de TSMC, según la prensa, sobre todo en 2 nm y 3 nm	El coste de los tape-outs en nodos punteros
Primer trimestre de 2027	FP4 en Pearl, según lo contado en el AMA de los fundadores	Un nuevo desplazamiento del trabajo hacia los formatos más nuevos de NVIDIA
Finales de 2027	El objetivo de Samsung Foundry de alcanzar el equilibrio	Si una segunda fuente más barata sigue sana
Abril de 2028	El próximo halving de bitcoin	Reduce a la mitad el bote de los mineros SHA-256

CAPÍTULO 13

13 Preguntas que no hizo pero debería haber hecho

PREGUNTAS QUE HACER

Las preguntas siguientes son las que yo haría antes de comprometer dinero o silicio en cualquier proyecto de IA × cripto.

¿Quién controla la función de trabajo? Si un equipo pequeño puede cambiar la aritmética en una altura fijada en el código, la vida de su hardware es decisión de ese equipo. Pida un compromiso de estabilidad por escrito, o diseñe hardware con un segundo mercado.

¿Cuál es el múltiplo de subsidio? Divida las emisiones (o el valor de mercado) entre los ingresos pagados desde fuera de la red. Un múltiplo superior a 10 significa que la red compra actividad, no que vende un servicio.

¿Cuánto cuesta la siguiente puerta, y qué cifra le hace parar? Una oblea multiproyecto cuesta decenas o cientos de miles de dólares; un juego de máscaras cuesta millones. Decida la cifra de parada antes de gastar.

¿Cuál es el hashrate de amortización, con p? Divida el NRE entre R × (1 − 1/k) × T × p. Si el resultado es una parte grande de la red, el chip es una apuesta sobre las reglas, no sobre el silicio.

¿Cuál es el segundo mercado del chip? Si el protocolo se bifurca o el token cae, ¿en qué otro mercado puede venderse el silicio? Las GPU sobrevivieron al Merge de Ethereum porque lo tenían; los ASIC de un solo algoritmo, no.

¿El valor acaba en el token o en el capital? En 2026, las empresas rentables de IA × cripto son sobre todo empresas de capital con distribución cripto; decida cuál de las dos está construyendo antes de captar dinero.

¿Quién es el primer cliente de pago, con nombre? Todos los casos duraderos de este informe, de Grass a las subredes de cómputo de Bittensor, empezaron con un comprador concreto de fuera de la red.

Son áreas que me gustaría explorar junto con otros, mediante intercambio y colaboración técnicos: chips a medida para minería, entrenamiento e inferencia de IA, y aceleradores; desarrollo de IA; y sistemas IA × cripto de la cadena al nodo y al chip. Puede escribirme a i@tanshuai.com.

Limitaciones de este informe

Es una investigación documental basada en documentos públicos hasta el 13 de noviembre de 2026. Para esta edición no hice entrevistas primarias ni mediciones independientes. Varias cifras clave son declaraciones de empresas o fabricantes, como los ingresos de Aethir, los ingresos de Grass, el rendimiento de Etched y de Taalas y

la eficiencia de Cysic; están señaladas como tales. Las métricas de la cadena de Pearl proceden de un nodo completo independiente y de los pools, y cambian cada hora; he usado las lecturas en torno a las 07:13 UTC del 13 de noviembre y doy los rangos del mismo día. Los costes de obleas, máscaras y diseño son estimaciones de terceros y valores de modelo que no coinciden entre sí; muestro rangos. Mis cálculos de punto de equilibrio de los ASIC y el presupuesto por etapas usan supuestos ilustrativos (k, T, p, NRE y costes de las puertas) que cada lector debería sustituir por los suyos.



Sobre el autor

Tan Shuai es tecnólogo y fundador, con más de veinte años de trabajo en chips, hardware y software de sistemas, y 20 patentes concedidas. Acumula muchos años de I+D y dirección de chips fabless, incluido el diseño full-custom de rutas de datos ASIC para chips de minería de criptomonedas y chips multimedia para cámaras, que es la lente de este informe: energía por operación, coste por oblea, qué permite fijar un protocolo en silicio y quién paga. Su trabajo le ha llevado por Silicon Valley, Nueva York, Chicago, Taiwán, la Gran Área de la Bahía (Shenzhen y Hong Kong), Vietnam y Malasia. Hoy trabaja como investigador y desarrollador de IA.

Debate técnico: i@tanshuai.com · tanshuai.com

Notas y aviso legal

Este informe es una investigación documental independiente basada en documentos públicos disponibles hasta el 13 de noviembre de 2026. Se ofrece solo con fines informativos y de aprendizaje y no constituye asesoramiento de inversión ni de ingeniería. Trata solo de tecnología y negocio. Los precios, las métricas de las cadenas y las reglas de los protocolos cambian; verifique en la fuente primaria todo aquello en lo que se apoye. Las cifras marcadas como `dato del proveedor`, `estimación` o `análisis` deben leerse como tales. Los precios de los criptoactivos son muy volátiles.

© 2026 Tan Shuai. Puede compartir este informe sin modificaciones y citando la autoría.

Fuentes

Todas consultadas entre el 10 y el 13 de noviembre de 2026. Agrupadas por capítulo.

Pearl: protocolo, código y datos de la cadena

- Komargodski y Weinstein, «Proofs of Useful Work from Arbitrary Matrix Multiplication», arXiv 2504.09971: https://arxiv.org/abs/2504.09971
- Pearl Research Labs, whitepaper de INT: https://pearlresearch.ai/research/int-whitepaper
- Código fuente de Pearl: bloque génesis: https://github.com/pearl-research-labs/pearl/blob/main/node/chaincfg/genesis.go ; parámetros de consenso y alturas de las bifurcaciones: https://github.com/pearl-research-labs/pearl/blob/main/node/chaincfg/params.go ; contabilidad del hashrate: https://github.com/pearl-research-labs/pearl/blob/main/zk-pow/src/api/sanity_checks.rs ; guía de actualización para la bifurcación de semilla con sal: https://github.com/pearl-research-labs/pearl/blob/main/docs/salted-seed-fork-upgrade-guide.md ; parámetros de la rama 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, minería en coma flotante: https://github.com/pearl-research-labs/pips/blob/main/PIPS/pip-0003.md
- prlstats (nodo completo independiente): red: https://prlstats.com/api/network ; series: https://prlstats.com/api/series ; portada y modelo de ingresos de los mineros: https://prlstats.com/api/home ; oferta minada: https://prlstats.com/api/mined ; pools: https://prlstats.com/api/pools ; tenedores: https://prlstats.com/api/holders ; supuestos de hardware y de sobrecoste: https://prlstats.com/api/assumptions
- Kryptex Pool: red de Pearl: https://pool.kryptex.com/prl/about-coin ; hashrate: https://pool.kryptex.com/prl/hashrate ; RTX 5090: https://pool.kryptex.com/de/device/gpu/nvidia/nvidia-geforce-rtx-5090 ; RTX 4090: https://pool.kryptex.com/device/gpu/nvidia/nvidia-geforce-rtx-4090 ; «Pearl FP8 fork»: https://pool.kryptex.com/articles/pearl-fp8-fork-en
- CoinGecko, Pearl: https://api.coingecko.com/api/v3/coins/pearl-2 ; https://www.coingecko.com/en/coins/pearl-research ; historial de precios: https://api.coingecko.com/api/v3/coins/pearl-2/market_chart?vs_currency=usd&days=365&interval=daily ; búsqueda del ticker «Perle»: https://api.coingecko.com/api/v3/search?query=perle
- SimeonOnSecurity, «Mining Pearl (PRL) in 2026»: https://simeononsecurity.com/articles/pearl-prl-mining-guide-2026/
- Hashrate Index, «Pearl (PRL): Inside the AI-Compute Cryptocurrency», 2 de junio de 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, alianza con Pearl Research Labs: https://together.ai/blog/together-ai-partners-with-pearl-research-labs ; página del modelo: https://together.ai/models/gemma-4-31b-it-pearl
- Basu, censo de instancias de minería de Pearl, arXiv 2606.04819: https://arxiv.org/html/2606.04819v2
- Blockworks, boletín 0xResearch, 28 de octubre de 2026: https://blockworks.com/newsletter/0xresearch/issue/post_daf0d74a-2bf4-41e4-86d3-2a1275475a24
- Datawallet, «Pearl Explained»: https://www.datawallet.com/crypto/pearl-explained
- Coingabbar sobre la información de TheMarker acerca de la financiación de Pearl Research Labs: https://www.coingabbar.com/en/pearl-research-labs-funding-130m-stealth-round-themarker

Cadenas de trabajo útil e historia

- Grothe et al., análisis de seguridad de Gridcoin, USENIX WOOT 2017: https://www.usenix.org/system/files/conference/woot17/woot17-paper-grothe.pdf
- Falsificación de pruebas de aprendizaje, arXiv 2208.03567: https://arxiv.org/abs/2208.03567
- hoge.gg, verificación y confianza en la inferencia descentralizada: https://hoge.gg/decentralized-inference-2026-verification-trust/ ; ingresos y tesorería de Bittensor: https://hoge.gg/bittensor-quantum-wallet-revenue-tao-treasury-2026/
- Kaspa: código de la matriz de kHeavyHash: https://github.com/kaspanet/rusty-kaspa/blob/master/consensus/pow/src/matrix.rs ; lanzamiento: https://coinranking.com/coin/V8GxkwWow+kaspa-kas/about ; anuncio de IceRiver KS1/KS2: https://www.iceriver.io/2023/03/14/1677644580/ ; historia de GPU a ASIC: https://captainmining.com/blogs/crypto-news/the-evolution-of-kaspa-mining-from-gpus-to-asic-powerhouses ; hashrates de GPU: https://coincalculators.io/algorithms/kheavyhash ; KS7: https://www.asicminervalue.com/miners/bitmain/antminer-ks7-40th ; red: https://minerstat.com/coin/kas ; 1 EH/s en 2024: https://ourcryptotalk.com/news/kaspa-reaches-1-ehs-in-hashrate ; Crescendo: https://ourcryptotalk.com/news/kaspa-crescendo-live
- Aleo: preguntas frecuentes sobre la mainnet: https://aleo.org/post/aleo-mainnet-faq/ ; prueba descentralizada: https://aleo.org/post/decentralized-proving-advantages/ ; ASIC: https://whattomine.com/coins/432-aleo-zksnark/asics ; Blockworks, «State of Aleo Q3 2025»: https://app.blockworks.com/report/state-of-aleo-q3-2025 ; ARC-46: https://chaincatcher.com/en/article/2144326
- Qubic: The Block sobre la reorganización de Monero: https://www.theblock.co/post/366535/monero-faces-chain-reorganization-fears-after-qubic-says-it-controls-51-of-hashrate ; Halborn: https://www.halborn.com/blog/post/explained-the-monero-51-percent-attack-august-2025 ; minería de Dogecoin: https://hackernoon.com/qubic-is-now-mining-dogecoin-on-its-ai-compute-infrastructure ; CoinGecko: https://www.coingecko.com/en/coins/qubic

- Documentación del consenso de 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 y las bifurcaciones de resistencia a los ASIC: https://lbry.com/news/forking-for-asic-resistance ; bifurcación de Sia: https://hotfix.coindesk.com/markets/2018/11/01/sia-completes-hard-fork-as-bricked-miners-form-rival-blockchain ; precios de Obelisk: https://www.cryptoglobe.com/latest/2018/11/obelisk-hardware-skyrockets-in-selling-price-after-sia-hard-fork/ ; los mineros tras el Merge: https://hotfix.coindesk.com/business/2022/11/11/ex-ethereum-miners-token-hop-to-stay-alive-after-the-merge

Sectores de IA × cripto

- DefiLlama: Render: https://defillama.com/protocol/render-network-bme ; Livepeer: https://defillama.com/protocol/livepeer
- Cointelegraph sobre los ingresos DePIN según Messari: https://cointelegraph.com/news/depin-revenues-grow-despite-token-losses
- Aethir, balance de 2025: https://aethir.com/blog-posts/aethirs-2025-wrap-up-decentralized-gpu-cloud-milestones
- CoinGecko, categoría de IA: https://www.coingecko.com/en/categories/artificial-intelligence
- Akash, precios de alquiler de la H100: https://akash.network/the-bid/h100-rental-price-2026-cost-per-hour/ ; informe del T1 de 2026: https://akash.network/blog/akash-network-q1-2026-report/
- io.net, acuerdo de 8 millones de dólares y tokens diarios: 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, contrato de 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 sobre INTELLECT-3: https://www.implicator.ai/prime-intellects-intellect-3-open-source-ambition-meets-centralized-reality/
- SiliconANGLE, financiación de 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, financiación de Nous: https://runtimewire.com/article/nous-research-75-million-round-1-5-billion-valuation
- ownyourmind, Gensyn: https://ownyourmind.ai/projects/gensyn/ ; Phemex, Ritual (solo titular): https://phemex.com/news/article/ritual-launches-ai-layer-1-mainnet-and-ritual-token-price-dives-72-from-high-101259
- Sobrecoste de la computación confidencial, arXiv 2409.03992: https://arxiv.org/abs/2409.03992v2 ; informe de Phala de 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, medición de los pagos de agentes de IA: https://www.trmlabs.com/trm-tech-blog/whos-actually-paying-measuring-ai-agent-payments-onchain ; Forkast sobre la 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 de Tempo: https://www.theblock.co/amp/post/394131/tempo-mainnet-goes-live-with-machine-payments-protocol-for-agents ; BlockEden sobre el aGDP de Virtuals: https://blockeden.xyz/blog/2026/04/21/agdp-agent-gdp-virtuals-protocol-ai-blockchain-valuation-primitive-tvl-displacement/ ; The Block sobre ElizaOS: https://theblock.co/post/410774/eliza-labs-native-token-dead
- Dealroom, ingresos de Grass: https://dealroom.co/news/157794-grass-backed-by-multicoin-projects-75m-in-2026-data-revenue/ ; CryptoSlate, Fetch.ai y Ocean: https://cryptoslate.com/cryptos-flagship-ai-pact-fracture-fetch-sues-ocean-over-263m-fet-community-sales/ ; Fortune sobre World: https://fortune.com/2026/06/01/internet-according-to-sam-altman-seeing-orb/
- Galaxy, capital riesgo cripto: https://www.galaxy.com/insights/research/crypto-blockchain-venture-capital-q3 ; Crypto Briefing (solo titular): https://cryptobriefing.com/crypto-vc-investment-q3-2026-galaxy/ ; Blockhead sobre Tiger Research: https://www.blockhead.co/2026/10/14/bitcoin-rose-43-in-q3-crypto-venture-funding-did-not-follow/ ; ForkLog sobre los lanzamientos de tokens de 2025: https://forklog.com/en/85-of-tokens-launched-in-2025-fell-below-initial-prices/

Economía de los ASIC y tipos de chip

- Hashrate Index, resumen del 9 de noviembre de 2026: https://hashrateindex.com/blog/hashrate-index-roundup-november-9-2026/
- AnySilicon, coste de diseño de ASIC por nodo en 2026: https://anysilicon.com/news/how-much-does-asic-design-cost-in-2026-real-price-ranges-by-process-node/ ; Silicon Analysts, guía de costes de tape-out para fabless: https://siliconanalysts.com/analysis/fabless-startup-tapeout-cost-guide ; datos de mercado: https://siliconanalysts.com/market-data
- MillionMiner, Bitmain S23 XP Hyd: https://millionminer.com/news/bitmain-wdms-2026-s23-xp-hyd-launch
- Khattak y Mikaitis, acumulación FP8 en los tensor cores, arXiv 2512.07004: https://arxiv.org/html/2512.07004v2
- Canaan, resultados del segundo trimestre de 2026: https://finviz.com/news/389337/canaan-inc-reports-unaudited-second-quarter-2026-financial-results ; Bitdeer, resultados del tercer trimestre de 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 y Block: https://theenergymag.com/news/2026-07-29/core-scientific-exit-block-bitcoin-deal-ai ; economía de los arrendamientos de IA: https://theenergymag.com/news/2026-08-06/miner-weekly-ai-pumpability-out-fuel
- Blockworks, Cipher y 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, resultados del T1 del ejercicio 2027: https://globenewswire.com/news-release/2026/11/05/3397020/0/en/iren-reports-q1-fy27-results.html

- Spark, concentración del mercado de ASIC de bitcoin: https://www.spark.money/research/bitcoin-asic-manufacturer-market-concentration
- Google, documentación de 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 ; artículo de la 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, precios de CoWoS: https://trendforce.com/news/2026/10/27/news-tsmc-reportedly-eyes-20-cowos-price-hike-skips-equipment-price-renegotiations-amid-ai-demand-surge
- Menon Lab, comparativa de aceleradores de inferencia de IA: https://themenonlab.blog/blog/ai-inference-accelerators-compared ; Introl, aceleradores de IA más allá de las GPU: https://introl.com/blog/ai-accelerators-beyond-gpus-tpu-trainium-gaudi-cerebras-2025
- StratNews (Reuters), chips de IA nacionales en centros de datos con financiación estatal: https://stratnewsglobal.com/team-sng/china-orders-state-funded-data-centres-to-use-domestic-ai-chips/ ; TNW, importaciones de H200: https://thenextweb.com/news/nvidia-h200-chips-hong-kong-china-import-approval
- Succinct, prueba en tiempo real con 16 GPU: https://blog.succinct.xyz/real-time-proving-16-gpus/ ; productos ZK ASIC de Cysic: https://cysic-docs.pages.dev/hardware-products/zk-asic-products
- Tom's Hardware, Intel Heracles: https://www.tomshardware.com/tech-industry/cyber-security/intels-heracles-chip-computes-fully-encrypted-data-without-decrypting-it-chip-is-1-074-to-5-547-times-faster-than-a-24-core-intel-xeon-in-fhe-math-operations ; IEEE Spectrum: https://spectrum.ieee.org/fhe-intel ; Evertiq, Niobium y SemiFive: https://evertiq.com/design/2026-02-20-semifive-secures-design-win-with-niobium-for-fhe-accelerator ; The Quantum Insider, financiación de 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 en FPGA: https://www.zama.org/post/announcing-hpu-on-fpga-the-first-open-source-hardware-accelerator-for-fhe

Nodos de proceso

- Silicon Analysts, precios de obleas: https://siliconanalysts.com/data/wafer-pricing
- Tom's Hardware, rumores sobre precios de 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, precios de SMIC: https://wccftech.com/smic-5nm-chip-prices-up-to-50-percent-more-expensive/amp/
- EUROPRACTICE, calendario de MPW de TSMC para 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, hoja de ruta de Huawei Ascend: https://www.tomshardware.com/tech-industry/artificial-intelligence/huawei-ascend-npu-roadmap-examined-company-targets-4-zettaflops-fp4-performance-by-2028-amid-manufacturing-constraints ; The Substrate, la cadena de suministro de chips de IA de China: 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

Oportunidades y economía de los lanzamientos

- Webull (Benzinga), conferencia de resultados de Broadcom: https://www.webull.com/news/15511095937917952 ; Quartr, Global Unichip Q2 2026: https://quartr.com/events/global-unichip-corp-3443-q2-2026_F9×6NJQY ; Taiwan News, Alchip: https://www.taiwannews.com.tw/news/6370946 ; Design & Reuse, VeriSilicon: https://www.design-reuse.com/news/202530457-surge-in-custom-chip-business-propels-verisilicon-s-q1-revenue-to-double-with-new-orders-skyrocketing-to-8-2-billion-yuan-increased-r-d-investment-leads-to-wider-net-loss-earnings-insight/ ; The Block, la unidad de IA de Canaan: https://www.theblock.co/post/359250/crypto-miner-canaan-to-shutter-non-core-ai-semiconductor-business-unit-amid-strategic-realignment
- Clutch, precios del desarrollo de IA: https://clutch.co/developers/artificial-intelligence/pricing ; PixelPlex, coste del desarrollo blockchain: https://pixelplex.io/blog/blockchain-development-cost/
- Hyperbolic, prueba de muestreo: 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 ; verificación por capas, arXiv 2512.20176: https://arxiv.org/abs/2512.20176v1
- Sherlock, referencia de precios de auditoría para 2026: https://sherlock.xyz/post/smart-contract-audit-pricing-a-market-reference-for-2026
- Blockworks, criterios de listado de Binance: https://blockworks.com/news/binance-listing-standards-scrutiny ; Cointelegraph, acuerdos con creadores de mercado: 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

PRUEBA DE TRABAJO ÚTIL · CRIPTO × SILICIO

El cómputo como consenso

IA × cripto × chips en 2026, a partir de la prueba de trabajo útil de Pearl

Cada cifra clave de este informe lleva su fuente y su fecha. Cuando el autor deriva una cifra, se marca como estimación; cuando solo la publica el vendedor, se marca como dato del proveedor. Trata solo de tecnología y negocio.

Tan Shuai

Tecnólogo y fundador · Investigador y desarrollador de IA

tanshuai.com · i@tanshuai.com

EDICIÓN PÚBLICA · Datos a 13 de noviembre de 2026

