

מסמך מחקר · מהדורה פומבית

11 · 2026

הוכחת עבודה מועילה · קריפטו × סיליקון

חישוב כקונצנזוס

בינה מלאכותית × קריפטו × שבבים ב־2026, החל מהוכחת העבודה המועילה של Pearl

Pearl · עבודה מועילה · DePIN וסוכנים · כלכלת ASIC · צמתי תהליך · תקציבים מדורגים

המסקנה

Pearl מוכיחה שכפל מטריצות יכול לאבטח בלוקצ׳יין, אבל עדיין לא שהעבודה מועילה. בכל תחום הבינה המלאכותית × קריפטו, הביקוש בתשלום נמצא בהסקה, בנתונים ובפרטיות, ו־ASIC משתלם רק היכן שהקופה גדולה, הכללים יציבים ולשבב יש שוק שני.

תאן שואי

טכנולוג ומייסד · חוקר ומפתח בינה מלאכותית

tanshuai.com · i@tanshuai.com

הנתונים נכונים ל־13 בנובמבר 2026

הסכומים בדולר ארה״ב

תוכן העניינים

תוכן העניינים



◆ **תקציר מנהלים** — המשפט מחזיק, התמריץ לא — 03

01 **היקף ושיטה** — היקף, שיטה וארבעה מבחנים — 06

02 **Pearl · המנגנון** — Pearl: כפל מטריצות כקונצנזוס — 08

03 **Pearl · הכלכלה** — Pearl: היקף, הנפקה ופער התועלת — 10

04 **Pearl · הכללים** — Pearl: סיכון כללים, נומריקה וממשל — 12

05 **רשתות עבודה מועילה** — רשתות עבודה מועילה: לא מטריצות ולא אפס ידיעה עוצרות ASIC — 13

06 **התמונה הרחבה** — שאר הבינה המלאכותית × קריפטו: ביקוש אמיתי אך קטן, והערך עובר להון — 15

07 **כלכלת ASIC** — שאלת ה־ASIC: נוסחה אחת, שלושה קני מידה — 18

08 **שבב אחר שבב** — חמישה סוגי שבבים, חמש תשובות — 20

09 **צמתי תהליך** — צמתי תהליך: לתת לעלות הגדולה ביותר לאורך החיים לבחור — 24

10 **תקציב מדורג** — ממודל לסיליקון: תקציב מדורג לשבב ראשון — 27

11 **הזדמנויות** — היכן ההזדמנויות, ומה הייתי עושה — 28

12 **ציוני דרך** — על מה לעקוב, נובמבר 2026 עד 2027 — 31

13 **שאלות שכדאי לשאול** — שאלות שלא שאלתם אבל הייתם צריכים לשאול — 32

◆ **על המחבר** — תאן שואי — 32

◆ **מקורות** — 33

איורים

איורים



איור 01	בינה מלאכותית × קריפטו בשישה תחומים: היכן הביקוש בתשלום אמיתי	05
איור 02	בעבודת המספרים השלמים של Pearl, ל־GPU תקרת אנרגיה משותפת	08
איור 03	ששת החודשים הראשונים של Pearl: הגיבוב עוקב אחרי המחיר, הכללים משתנים	10
איור 04	מכפיל הסבסוד: פי כמה מההכנסה החיצונית רשת משלמת	14
איור 05	מדוע יעילות מפסיקה להשתלם: ASIC שומר לכל היותר 1 - 1/k מההכנסה	18
איור 06	רוב הקופות קטנות מתקציב תכנון של שבב אחד	19
איור 07	חמישה סוגי שבבים, חמש תשובות	23
איור 08	פרוסה בחזית הטכנולוגיה עולה פי עשרה מפרוסה בצומת בוגר	24
איור 09	לקנות ראיות לפני מסכות: תקציב מדורג לשבב ראשון	27

תגיות בטקסט

- **אומדן** — חישוב או הסקה של המחבר מתוך נתונים מצוטטים
- **טענת החברה** — פורסם על ידי החברה שמוכרת את המוצר, ולא אומת באופן בלתי תלוי
- **ניתוח** — הפרשנות של המחבר לראיות

זהו תרגום של המהדורה האנגלית; בכל מקרה של אי־התאמה, המהדורה האנגלית היא הקובעת.

הנתונים נכונים ל־13 בנובמבר 2026 · הסכומים בדולר ארה״ב



תקציר מנהלים

המשפט מחזיק, התמריץ לא

המסקנה

Pearl מראה ש"חישוב כקונצנזוס" עובד כהנדסה, אבל עדיין לא שהוא מועיל.

≈90 EH/s
רשת Pearl ב־13 בנובמבר 2026, שבה H אחד הוא פעולת הכפלה־צבירה אחת ב־INT7

≈1%
מהבלוקים של Pearl נושאים עבודת AI אמיתית (Blockworks, 28 באוקטובר 2026)

≈70 מיליון דולר
בשנה של הכנסה חיצונית ניתנת לאימות ב־Bittensor, מול הנפקה של יותר מ־300 מיליון דולר

פי 180–175
תגמולי המוכיחים של Aleo ביחס לעמלות שלה ברבעון השלישי של 2025

R × (1 - 1/k)
תקרת הרווח של ASIC שזול פי k מהכורה השולי
אומדן

≈1.5 מיליון דולר
כדי להגיע לסיליקון שנמדד ב־N7 דרך מודל, FPGA ו־MPW, לפני סט מסכות של ≈3.5 מיליון דולר
אומדן

Pearl

ב־13 בנובמבר 2026 הרשת של Pearl פעלה בכ־90 EH/s. בקוד הקונצנזוס שלה "גיבוב" אחד הוא פעולת הכפלה־צבירה אחת במספרים שלמים, כך שהרשת שקולה לתפוקת המספרים השלמים של כ־122,000 מעבדי H100. ובכל זאת רק כ־1% מהבלוקים נושאים עבודת AI אמיתית, ושותפות ההסקה הפומבית היחידה כבר אינה מוצעת ב־API ה־serverless של הספק שלה. הסיבה כלכלית, לא קריפטוגרפית: GPU שכורה בזמן שהוא משרת הסקה מרוויח רק כ־1% עד 27% מקצב הגיבוב של כורה ייעודי, ולכן כורים רציונליים מכפילים מטריצות אקראיות.

ביקוש בתשלום

בכל תחום הבינה המלאכותית × קריפטו, ביקוש בתשלום שניתן לאמת מופיע בשלושה מקומות: מכירת הסקה, מכירת נתונים ומכירת פרטיות. ההכנסה החיצונית הניתנת לאימות של Bittensor היא כ־70 מיליון דולר בשנה, כארבע חמישיות ממנה מתת־רשתות של מחשוב ותשתית. Grass מדווחת על 17 מיליון דולר של הכנסות נתונים במחצית הראשונה של 2026. הסקה חסויה בסביבות ריצה מהימנות (TEE) נמצאת בייצור. לעומת זאת, תשלומים שסוכני AI יזמו בפועל מסתכמים בכ־5,000–11,000 דולר בחודש, ול־zkML ולאימון מבוזר עדיין אין ביקוש משלם שפורסם.

סבסוד ושווי

מכפיל הסבסוד הוא בדיקת הבריאות הכנה ביותר. תגמולי המוכיחים של Aleo היו בערך פי 175–180 מהעמלות שלה; ההנפקה של Bittensor היא בערך פי 4–11 מההכנסה הניתנת לאימות שלה; שווי השוק של Render הוא בערך פי 360 מעמלות ה־on-chain שלה; וההכנסה של כורי Pearl, כ־460 מיליון דולר בשנה, היא ברובה סבסוד בלוקים, כמעט ללא קשר לביקוש AI בתשלום. סבסוד הוא דלק להתנעה, לא מודל עסקי. בינתיים הערך עובר מטוקנים להון: ארבע חברות הון שמשתמשות בתשתית מרוכזת ובהפצה מבית הקריפטו מוערכות בכ־8.5 מיליארד דולר, בערך כמו כל טוקני ה־AI בשכבת הפרוטוקול ובשכבת היישומים יחד.

ASIC

בשבבים, תקרת הרווח של ASIC היא בקירוב ההכנסה ליחידת חישוב כפול (1 - 1/k), כאשר k מבטא פי כמה הוא זול מהחומרה השולית. משום שהביטוי הזה קעור, יעילות מעבר ל־5–10 = k כמעט אינה מוסיפה רווח; מה שמכריע את התוצאה הוא גודל הקופה, כמה זמן הכללים מחזיקים והאם לשבב יש שוק שני. התשובות שלי לפי

סוג שבב: לא לבצע עכשיו טייפ־אאוט לשבב INT7 ייעודי ל־Pearl, ובמקום זאת להמר על מאיץ שמכוון קודם כול להסקה ויכול גם לכרות; להתרחק כשחקן חדש משוק כורי הביטקוין למכירה פתוחה; לא לבנות כצוות קטן שבב אימון עצמאי; לבנות שבבי הסקה כשלקוח חתום; לגדר במאיצי ZK ולהתמקם מוקדם ב־FHE.

צמתים ותקציבים

את צומת התהליך יש לבחור לפי הסעיף הגדול ביותר בעלות השבב לאורך חייו: אנרגיה לשבבים שפועלים מסביב לשעון, מארז וזיכרון לשבבי אימון, ההחלטה בין SRAM ל־HBM לשבבי הסקה. ואת השבב הראשון יש לקנות בשלבים, מודל, FPGA, פרוסה מרובת פרויקטים ורק אז סט מסכות, כך ששינוי בפרוטוקול יעלה את מחירו של אב טיפוס ולא של טייפ־אאוט. ההזדמנויות נמצאות היכן שהתחומים נחצים: הקריפטו כלקוח הראשון, ה־AI כשוק השני.

שאלה	הכרעה	מספרים מרכזיים (המקורות בפרקים)
① **האם Pearl עובדת, והאם העבודה שלה מועילה?**	היא עובדת: בלוקציין שמאובטח בכפל מטריצות בהיקף גדול. עדיין לא מועילה: כמעט כל הכרייה נעשית על מטריצות אקראיות	כ־90 EH/s ב־13 בנובמבר 2026; MAC 1 = H 1 ב־INT7; כ־1% מהבלוקים עם עבודת AI
② **היכן הביקוש האמיתי בתשלום בבינה מלאכותית × קריפטו?**	הסקה, נתונים ופרטיות. תשלומי סוכנים, zkML ואימון מבוזר עדיין לפני ביקוש	ל־Bittensor כ־70 מיליון דולר בשנה של הכנסה ניתנת לאימות; ל־Grass 17 מיליון דולר במחצית הראשונה של 2026; תשלומי סוכנים כ־5,000–11,000 דולר בחודש
③ **האם כדאי לבנות ASIC?**	רק היכן שהקופה גדולה, הכללים יציבים ולשבב יש שוק שני. כיום אף רשת אינה מציעה את שלושתם	תקרת רווח ≈ R × (1 - 1/k); ביטקוין כ־11.5 מיליארד דולר בשנה של הכנסות כורים, Pearl כ־0.46–0.49 מיליארד דולר, Kaspa כ־18 מיליון דולר
④ **אילו שבבים צוות קטן צריך לבנות?**	מאיצים שמכוונים קודם כול להסקה ויכולים גם לכרות; שבבי הסקה ללקוח חתום; עבודה מוקדמת על FHE. לא כורי ביטקוין למכירה פתוחה ולא שבבי אימון	ה־ASIC הטוב ביותר של Kaspa מרוויח 0.86 דולר ביום; תקציבי תכנון ב־3/2 nm של 60–400 מיליון דולר ויותר
⑤ **איזה צומת תהליך?**	העלות הגדולה ביותר לאורך החיים מכריעה: אנרגיה, מארז וזיכרון, או ארכיטקטורת הזיכרון	TSMC N2 כ־30,000 דולר לפרוסה מול כ־3,000 דולר ל־N28; HBM ו־CoWoS כ־63% מעלות ה־H100
⑥ **כמה עולה שבב ראשון, ואיך כדאי לקנות אותו?**	בארבעה שלבים: מודל, FPGA, פרוסה מרובת פרויקטים, ואז מסכות. לעצור לפני המסכות אם הכללים או הקונה לא הוסדרו	משבצת MPW מסחרית 50,000–500,000 דולר; מסכות N7 כ־3.5 מיליון דולר; תקציב תכנון ב־7/6 nm 15–100 מיליון דולר

איור 01

בינה מלאכותית × קריפטו בשישה תחומים: היכן הביקוש בתשלום אמיתי

ביקוש בתשלום ניתן לאימות מול אותות הסבסוד והשווי, נובמבר 2026

01 רשתות עבודה מועילה

אמיתי אך קטן

≈70 מיליון בשנה

הכנסה חיצונית ניתנת לאימות של Bittensor

אות: הנפקה פי 4–11 מההכנסה; ≈1% בלוקי Pearl AI

המנגנונים עובדים; לתועלת דרושים משלמים

02 מחשוב מבוזר

אמיתי אך קטן

2.35 מיליון בשנה

עמלות on-chain של Render, הגבוהות ביותר שניתנות לאימות

אות: שווי שוק ≈פי 360 מהעמלות

בעיקר מכירה חוזרת של הסקה סחורתית

03 אימון והסקה

הרווח בהון

+100 מיליון בשנה

Prime Intellect, חברת הון על GPU מרוכזים

אות: רשת האימון של Gensyn אינה פעילה

ההתקדמות אמיתית; הרווח הולך להון

04 AI ניתן לאימות

מוקדם

398

משתמשים משלמים בענן ה־TEE של Phala, סוף 2025

אות: אין zkML בתשלום שפורסם

TEE בייצור; zkML מוקדם

05 תשלומי סוכנים

לפני ביקוש

5–11 אלף בחודש

תשלומי x402 שסוכנים יזמו בפועל

אות: ≈200 מיליון עסקאות, רובן איתות

המסילות הגיעו קודם; הביקוש לא

06 רשתות נתונים

קונים אמיתיים

17 מיליון

הכנסות נתונים של Grass, המחצית הראשונה של 2026 (דיווח עצמי)

אות: שווי שוק ≈פי 10 מההכנסה השנתית

אחד התחומים המעטים עם קונים אמיתיים

מקור: DefiLlama, SubConnect (דרך hoge.gg), Blockworks, CoinGecko, SiliconANGLE, Dealroom, Phala, TRM Labs; סינתזה של המחבר.



היקף ושיטה

01 היקף, שיטה וארבעה מבחנים

למה הדוח הזה, ולמה עכשיו

באפריל 2026 בלוקצ׳יין חדש, Pearl, הכניס לייצור רעיון ותיק: לתת לכורים להתחרות על בלוקים באמצעות ביצוע כפלי המטריצות שעליהם רצה הבינה המלאכותית, במקום גיבוב. בתוך שישה חודשים הרשת שלה הגיעה לכ־90 EH/s והטוקן שלה לשווי שוק של כ־500 מיליון דולר. באותו זמן, שווקי מחשוב ל־AI, רשתות אימון מבוזר, מסילות תשלום לסוכנים ורשתות נתונים גייסו מיליארדים, וכורי ביטקוין החלו להסב את אתרי החשמל שלהם למרכזי נתונים ל־AI. רציתי לדעת לאילו מהם יש לקוחות אמיתיים, האם סיליקון מותאם הגיוני במקום כלשהו ביניהם, ובאיזה צומת תהליך ובאיזה פאונדרי צוות שבבים יכול להשתמש בפועל.

ארבעה מבחנים לכל פרויקט של בינה מלאכותית × קריפטו

מחקר אבטחה מ־2017 על Gridcoin טען שחישוב מועיל יכול לאבטח בלוקצ׳יין רק אם קשה לפתור אותו, קל לאמת אותו וקל לתגמל עליו. אני מוסיף מבחן רביעי. יחד הם מכריעים את רוב מה שבא בהמשך.

01 **קשה לפתור.** חישוב ישר חייב להיות הדרך הזולה ביותר לייצר הוכחה תקפה; אסור שיהיה קיצור דרך.

02 **קל לאמת.** בדיקת העבודה חייבת לעלות הרבה פחות מביצועה, אחרת הרשת משלמת פעמיים.

03 **קל לתגמל.** הפרוטוקול חייב להיות מסוגל לחשב את התגמול מתוך ההוכחה בלבד, בלי לסמוך על צד שלישי.

04 **מישהו משלם על התוצר.** אם לקוח מחוץ לרשת אינו משלם על התוצאה, העבודה ה״מועילה״ היא סבסוד עם נרטיב.

חידות גיבוב עוברות את שלושת המבחנים הראשונים ונכשלות ברביעי. משימות AI אמיתיות עוברות את הרביעי אבל נכשלות בשני: הוכחת חישוב AI באפס ידיעה עולה פי 100 עד 10,000 מהחישוב עצמו, והוכח שאפשר לזייף בזול שיטות של הוכחת למידה (proof of learning). כפל המטריצות הרועש של Pearl הוא הגשר האמין הראשון שעובר את מבחן האימות ובה בעת שומר על אותה צורה כמו חישוב AI.

איך עבדתי

זהו מחקר שולחני, שנעשה בדרך שבה הייתי בוחן שבב או תוכנית עסקית. קראתי את הקוד של Pearl, את הצעות הפרוטוקול שלה ואת הספר הלבן שלה; נתוני שרשרת מצומת מלא בלתי תלוי; דיווחים של חברות והודעות על תוצאות כספיות; מאמרים בקריפטוגרפיה, בארכיטקטורת מחשבים ובלמידת מכונה; נתוני מחירים של פאונדרים ושל ספקים; ואת העיתונות המקצועית והטכנולוגית. לכל מספר מצוין התאריך שלו. היכן שאני גוזר מספר בעצמי, אני אומר זאת ומסמן **אומדן**. מספרים שרק המוכר מדווח עליהם מסומנים **טענת החברה**; הפרשנות שלי לראיות מסומנת **ניתוח**. הסכומים בדולר ארה״ב. המצב נכון ל־13 בנובמבר 2026.



פרק 02

02

Pearl · המנגנון

Pearl: כפל מטריצות כקונצנזוס

הרעיון, ממשפט לרשת ראשית

באפריל 2025 פרסמו Ilan Komargodski ו־Omri Weinstein את "Proofs of Useful Work from Arbitrary Matrix Multiplication". הם בונים הוכחת עבודה מועילה לכל מכפלת מטריצות בתקורה של פי 1 + o(1) בלבד מכפל נאיבי; הביטחון שלה נשען על הקושי לפתור אצוות של משוואות ליניאריות אקראיות בדרגה נמוכה, וזוהי השערה, לא הוכחה. המאמר מצביע על הקושי האמיתי בתחום: הוכחת עבודה מועילה עם תקורה של גורם קבוע היא טריוויאלית אבל חסרת תועלת; החלק הקשה הוא לשמור על כך שחישוב ישר יישאר הדרך הזולה ביותר לנצח כאשר הכורים בוחרים את הקלטים שלהם בעצמם.

הרשת הראשית של Pearl הושקה ב־27 באפריל 2026, בשעה 09:00 UTC. הכורים מחשבים מכפלה רועשת (A + E)(B + F), שבה הרעש בדרגה נמוכה נגזר מהתחייבות BLAKE3, ואז מסירים את הרעש כדי לשחזר את A·B. מצב הצובר של כל אריח פלט מגובב לכרטיס הגרלה; אריח זוכה נפתח ומוכח בהוכחת אפס ידיעה רקורסיבית של Plonky2 בגודל של פחות מ־60 KB, בתוך תעודת בלוק שמוגבלת ל־65 KB. האימות הוא הוכחה קטנה בגודל קבוע, שאינה תלויה בגודל המטריצה: זול לבדיקה כמו חידת גיבוב, בעוד שהחישוב עצמו יכול להיות מועיל.

הכפלה־צבירה אחת היא "גיבוב" אחד

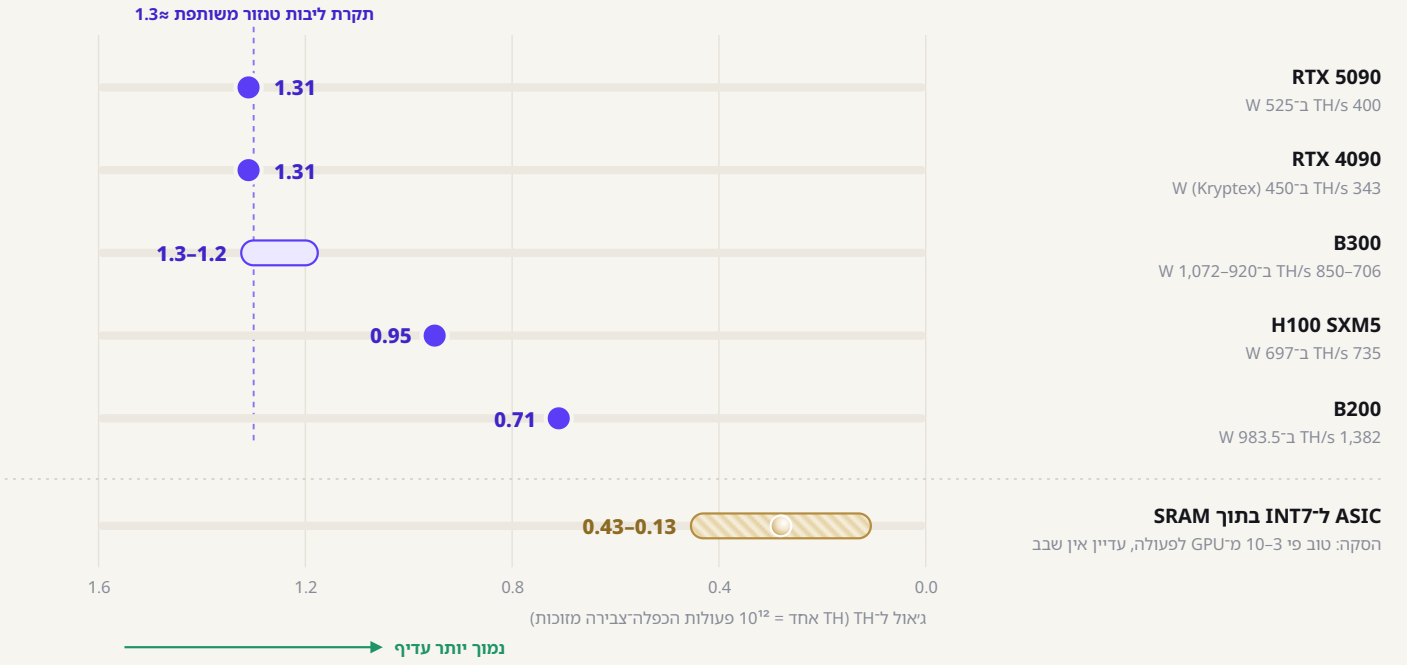
למתכנן שבבים העובדה החשובה ביותר היא היחידה. קוד הקונצנזוס של Pearl מכייל את היעד של כל ניסיון לפי גודל האריח כפול אורך המכפלה הסקלרית, ולכן **1 H הוא בדיוק פעולת הכפלה־צבירה (MAC) מזוכה אחת מסוג INT32 → INT7 × INT7**. מאז בלוק 96,251 קנס דרגה מחייב דרגת רעש של 128 לפחות; בדרגה המינימלית, 1 H עדיין שווה ל־1 MAC. לכן "EH/s" של Pearl פירושו, פשוטו כמשמעו, 18^10 פעולות הכפלה־צבירה במספרים שלמים בשנייה: **לראשונה, תקציב האבטחה של בלוקציין מתומחר בתפוקת הטנזורים של מאיצי AI.** קצבי הגיבוב שנמדדו ב־GPU נופלים ב־60% עד 95% משיא ההכפלה־צבירה הצפופה ב־INT8 של כל כרטיס, מה שמתיישב עם הקריאה הזאת.

חומרה	קצב גיבוב ב־Pearl	הספק	ג'אול לטריליון MAC (J/TH)	מקור (תאריך)
RTX 5090	400 TH/s	525 W	1.31 אומדן	Kryptex (12 בנובמבר)
RTX 4090	כ־305 TH/s (טווח בקהילה 280–330); Kryptex מציינת 343 TH/s	450 W (Kryptex)	1.31 (נתוני Kryptex) אומדן	prlstats (3 באוקטובר); Kryptex
H100 SXM5	735 TH/s	697 W	0.95 אומדן	PeakMiner, דרך prlstats (19 בספטמבר)
B200	1.382 PH/s	983.5 W	0.71 אומדן	PeakMiner, דרך prlstats
B300	850–706 TH/s	1,072–920 W	1.3–1.2 אומדן	prlstats; SimeonOnSecurity

איור 02

בעבודת המספרים השלמים של Pearl, ל־GPU תקרת אנרגיה משותפת

ג׳אול לטריליון פעולות הכפלה־צבירה ב־(J/TH) INT7 בקצב גיבוב שנמדד ב־Pearl; נמוך יותר עדיף



מקור: Kryptex (12 בנובמבר 2026); מבחני PeakMiner ו־SRBMiner דרך SimeonOnSecurity; prlstats; טווח ה־ASIC הוא הסקה של המחבר, לא מדידה.

שני פרטים בטבלה הזאת חשובים למתכנני שבבים. ראשית, כרטיסי RTX 4090 ו־5090 הצרכניים וה־B300 של מרכזי הנתונים מתכנסים לכ־1.3 J/TH, ואפילו הכרטיס הטוב ביותר, ה־B200, מגיע רק לכ־0.71 J/TH. GPU למרכזי נתונים שעולה עשרות אלפי דולרים מוציא כמעט אותה אנרגיה לכל הכפלה־צבירה במספרים שלמים כמו כרטיס גיימינג, מה שמצביע על תקרה משותפת של אנרגיית ליבות הטנזור לכל MAC ב־INT8, והתקרה הזאת היא בדיוק מה ששבב ייעודי יכול לשבור. שנית, ה־B300 איטי מה־B200: NVIDIA קיצצה בחדות את תפוקת ה־INT8 ב־Blackwell Ultra, ולכן עבודת המספרים השלמים צריכה לעבור בנתיב איטי יותר בשבב הזה. סיליקון ה־AI המרכזי מתרחק מ־INT8 לעבר FP8 ו־FP4. כאשר פורמט המספרים שקונצנזוס בוחר מתפצל מהכיוון שאליו החומרה הולכת, הפרוטוקול נאלץ לשנות את כלליו, וזה השורש של פיצול ה־FP8 ושל ההכרעה שלי לגבי ASIC בהמשך הדוח.



03

Pearl · הכלכלה

Pearl: היקף, הנפקה ופער התועלת

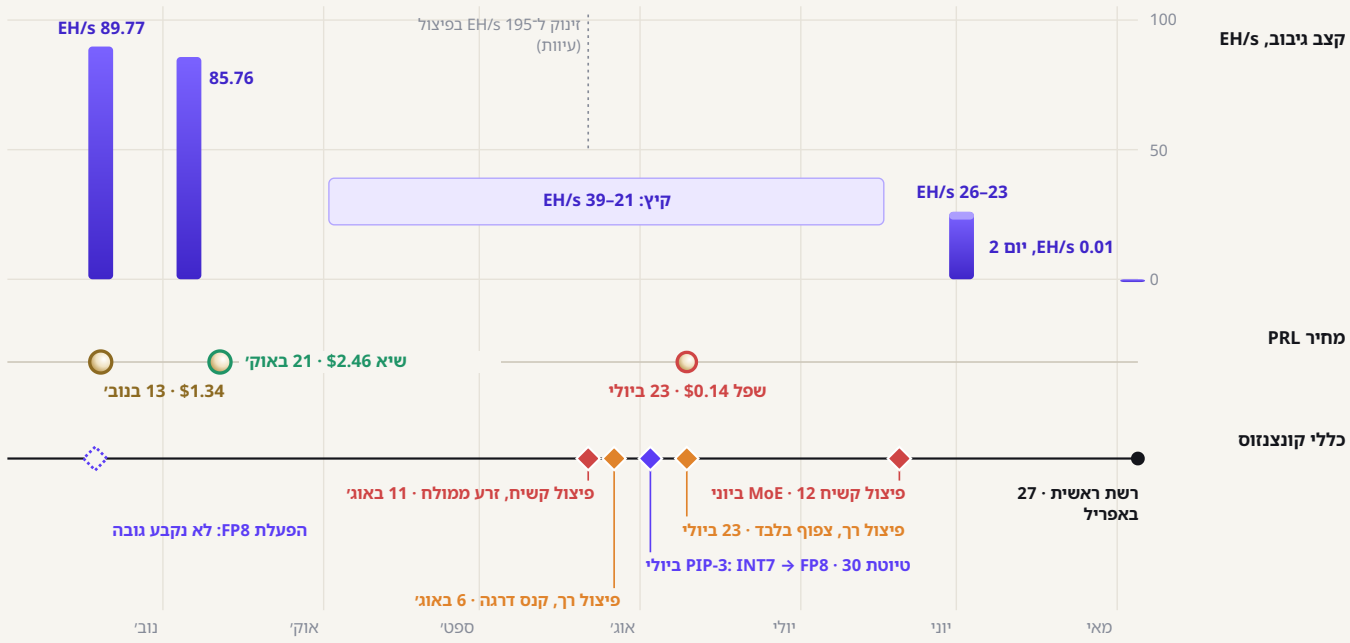
מ־0.01 ל־90 EH/s בשישה חודשים

מדד	ערך (13 בנובמבר 2026)	בסיס ומקור
קצב הגיבוב של הרשת	כ־90 EH/s (בסיס 89.77; טווח באותו יום 88.20–93.22; שיא תוך־יומי ב־12 בנובמבר של 110–123)	צומת מלא בלתי תלוי; prlstats; Kryptex
קושי וגובה בלוק	58,841,982 ו־142,530	prlstats
חומרה שקולה	כ־122,000 מעבדי H100 או 225,000 כרטיסי RTX 5090	לפי 735 ו־400 TH/s לכרטיס אומדן
מחיר, שווי שוק ושווי בדילול מלא	1.32–1.34 דולר; כ־500 מיליון דולר; כ־2.8 מיליארד דולר (דירוג 101 ב־CoinGecko)	CoinGecko
שיא ושפל של כל הזמנים	2.46 דולר (21 באוקטובר) ו־0.1399 דולר (23 ביולי)	CoinGecko
מטבעות שנכרו	377.6 מיליון PRL, 17.98% מהתקרה של 2.1 מיליארד	prlstats
סבסוד בלוק והנפקה יומית	כ־2,173 PRL וכ־0.97–1.03 מיליון PRL	prlstats; Kryptex
היצע חדש ב־12 החודשים הבאים	כ־78% מההיצע הנוכחי	prlstats
הכנסת כורים	כ־1.25 מיליון דולר ביום (מודל), כ־460 מיליון דולר בשנה	prlstats
הכנסה ליחידת קצב גיבוב, R	כ־14–15 דולר ל־PH/s ליום; RTX 5090 מרוויח כ־5–6 דולר ביום	אומדן ; Kryptex
בלוקים שנושאים עבודת AI אמיתית	כ־1% (אומדן מ־28 באוקטובר)	Blockworks
ריכוז בבריכות כרייה	Kryptex מצאה 46.4% מהבלוקים ב־24 השעות האחרונות; שתי הבריכות הגדולות כ־62%	prlstats
ריכוז מחזיקים	מקדם ג׳יני של כתובות 0.969; 100 הכתובות הגדולות מחזיקות 53.0%	prlstats

קצב הגיבוב של Pearl קרוב לפונקציה מפגרת של המחיר שלה. ביום שאחרי ההשקה הרשת פעלה בכ־0.01 EH/s; עד סוף מאי היא הגיעה ל־23–26 EH/s. במהלך הקיץ המחיר ירד בכ־86% מהשיא של מאי, וקצב הגיבוב נע הצידה ב־21–39 EH/s. אחרי שהמחיר עלה בערך פי עשרה בספטמבר ובאוקטובר, קצב הגיבוב הגיע ל־85.76 EH/s ב־27 באוקטובר ולכ־90 EH/s בנובמבר. זינוק ל־195 EH/s ב־11 באוגוסט חל בו בזמן עם פיצול, ויש לקרוא אותו כעיוות של האמידה. האנרגיה גדלה יחד איתו: מחקר אחד העריך כ־112 MW לרשת כולה ב־24 EH/s, ומודל של prlstats מעמיד אותה כיום על כ־190–215 MW.

איור 03 ששת החודשים הראשונים של Pearl: הגיבוב עוקב אחרי המחיר, הכללים משתנים

קצב הגיבוב של הרשת, שלוש נקודות מחיר וכל שינוי קונצנזוס, 27 באפריל עד 13 בנובמבר 2026



מקור: צומת מלא ונתוני prlstats; Kryptex; CoinGecko; params.go ו־PIP-3 של Pearl. הזינוק ב־11 באוגוסט הוא עיוות אמידה בפיצול.

ההנפקה מגבירה את הרפלקסיביות

תקרת ההיצע היא 2.1 מיליארד מטבעות ואין חציות (halvings): סבסוד הבלוק יורד באופן חלק לפי E(t) = S·H / ((t + H)(t + H - 1)) עם H = 650,226. אבל ההנפקה נספרת בבלוקים, לא בזמן, והקושי ההתחלתי היה נמוך מדי והתכוונן לאט. **כ־44.06 מיליון מטבעות נכרו ביום הראשון וכ־120 מיליון בששת הימים הראשונים**, כ־32% מההיצע של היום אומדן. הבלוקים האחרונים, בכ־171–183 שניות, עדיין מגיעים מהר יותר מיעד של 194 שניות. התוצאה היא שההיצע יגדל בכ־78% ב־12 החודשים הבאים, דילול שכורים, יצרני חומרה ומחזיקים צריכים כולם להתמודד איתו.

כלכלת הכורים נגזרת מכך. הכנסת הכורים בכל הרשת היא כ־1.25 מיליון דולר ביום (מודל); ההנפקה היומית של Kryptex כפול המחיר הנוכחי נותנת כ־1.35 מיליון דולר ביום, ולכן **ההכנסה ליחידת קצב גיבוב 15–14 ≈ R דולר ל־PH/s ליום** אומדן. RTX 5090 מרוויח כ־5–6 דולר ביום. H100 או H200 מרוויח כ־11 דולר, בעוד שההשכרה הציבורית הזולה ביותר לפי דרישה היא כ־2.09 דולר לכרטיס לשעה, כ־50 דולר ביום, ולכן שכירת כרטיסים רק כדי לכרות מפסידה הרבה. R משתנה מאוד: לפי הסדרה של prlstats הוא היה כ־10 דולר באוגוסט ובספטמבר ולזמן קצר כ־33 דולר באמצע אוקטובר אומדן. בתוך שבועות מההשקה, ההכנסה היומית המשוערת של RTX 5090 ירדה בחצי, מכ־33.80 דולר לכ־17.19 דולר. שכירת GPU בענן לכרייה השתלמה רק בשבועות הראשונים ובזמן זינוק המחיר באוקטובר.

המשפט מחזיק; התמריץ לא

הבעיה האמיתית היא ביקוש, לא קריפטוגרפיה. האינטגרציה המסחרית הפומבית היחידה של Pearl היא נקודת קצה Gemma-4-31B-it-pearl של Together AI, שהושקה ב־15 במאי במחיר נמוך ביותר מ־25% מהמחיר הרגיל, כשההנחה ממומנת מהערך העתידי הצפוי של ההנפקה. ב־13 בנובמבר, דף המודל הראה שהוא **אינו זמין ב־API ה־serverless של Together**. ביוני, מפקד של 8,012 מופעי כרייה מקוונים, כ־21% מהרשת, הגיע למסקנה שהרשת לא ייצרה שום חישוב AI מועיל כלל (המחבר מכנה את הראיות מרמזות, לא חותכות), והראה שבריכות קיבלו שיתופים שנבנו ממטריצות אקראיות לחלוטין. ב־28 באוקטובר העריכה Blockworks ש**רק כ־1% מהבלוקים נושאים עבודת AI אמיתית**; בלעדיה, כך כתבה, Pearl היא ביטקוין קטן יותר, מאובטח פחות ומתנפח מהר יותר.

את הסיבה אפשר לחשב. התקורה בצד ההסקה באמת קטנה: Pearl מדווחת על עלייה של 5.08% בזמן מקצה לקצה עבור Llama 70B על ארבעה H200, ובדיקה בלתי תלויה מדדה ירידה של 9.2% בתפוקה על GPU אחד מסוג Hopper. צד הכרייה מרוויח מעט מדי. התוסף הרשמי ל־vLLM כורה רק באצוות prefill של 1,024 טוקנים לפחות ולעולם לא בשלבי פענוח (decode). כרטיס שמשרת הסקה מקבל כ־27% מקצב הגיבוב של כורה ייעודי בתעבורה של prefill בלבד, כ־5.4% בתמהיל צ׳אט שהוא 80% פענוח, ורק כ־1% בתעבורה שהפענוח שולט בה. בלי ביקוש להסקה בתשלום, כורה רציונלי מכפיל מטריצות אקראיות; הפרוטוקול מאמת שהמטריצות שהוגשה להן התחייבות הוכפלו ביושר, לא שהן היו מועילות. **המשפט מחזיק; התמריץ לא.** התועלת של Pearl היא בעיה מסחרית מחוץ לשרשרת: מישהו צריך לשלם על ההסקה.



פרק 04

04

Pearl · הכללים

Pearl: סיכון כללים, נומריקה וממשל

ארבעה פיצולים בחודשיים

הפנים האחרות של Pearl הן המהירות שבה הכללים שלה משתנים. מיוני עד אוגוסט 2026 דחפה המעבדה ארבעה שינויי קונצנזוס בגבהים מקודדים בקוד: פיצול קשיח של תערובת מומחים (בלוק 71,935, בסביבות 12 ביוני), פיצול רך למודלים צפופים בלבד (91,630, בסביבות 23 ביולי), פיצול רך של קנס דרגה (96,251, בסביבות 6 באוגוסט) ופיצול קשיח של זרע ממולח (99,000, בסביבות 11 באוגוסט). אחרי פיצול הזרע הממולח, כורים ישנים יכלו לייצר רק שיתופים פסולים; קנס הדרגה קיצץ את קצב הגיבוב האפקטיבי של תצורות בדרגה גבוהה ב־50% עד 87.5%.

הצעד הבא הוא PIP-3, שעדיין בטיוטה: הוא מעביר את הכרייה מ־INT7 ל־FP32 ודורש פרופילי התקנים מדויקים ברמת הביט, **ותומך בשתי ארכיטקטורות בלבד, Hopper ו־Blackwell**. המסמך מכנה את התלות בחומרה רע הכרחי של פרוטוקולי נקודה צפה ומזהיר שכל סובלנות באימות מזמינה התקפות גריסה (grinding). ב־12 בנובמבר, פרמטרי הרשת הראשית בענף fp8-scheme לא קבעו גובה הפעלה. לפי הטיוטה הנוכחית, סדרת RTX 40, ה־A100 וכרטיסי AMD אינם מתקבלים, בעוד שמודל של prlstats בביטחון נמוך מעריך שכ־82% מקצב הגיבוב מגיע מכרטיסים צרכניים. כפיית הפיצול מסכנת בפיצול של השרשרת או בתגובת נגד של הכורים.

טרילמה של פורמט מספרים

Pearl ניצבת מול טרילמה: פורמט מספרים יכול להיות מועיל ל־AI (FP8 או FP4), ניתן לאימות מדויק ברמת הביט בין חומרות, וניטרלי לחומרה, אבל לכל היותר שניים מהשלושה. סכמת ה־INT7 ניתנת לאימות וניטרלית, אבל מתפצלת מהכיוון שאליו הולך סיליקון ה־AI. FP8 עם רשימה לבנה מועיל וניתן לאימות, אבל קושר את הקונצנזוס לספק אחד. FP8 עם סובלנות מועיל וניטרלי, אבל פותח את הדלת להתקפות גריסה. אריתמטיקה בנקודה צפה בליבות טנזור בדרך כלל אינה עוקבת אחר תקן נומרי ואינה ניתנת לשחזור בין דורות; ב־Hopper, למשל, צבירת FP8 שומרת רק 13 סיביות שבר ומקטמת. מנקודת המבט של מתכנן שבבים, ״פרופיל Hopper״ הוא אריתמטיקה מוגדרת לחלוטין ש־ASIC יכול לממש ביט־אחר־ביט, אולי ביעילות רבה יותר מה־GPU. אבל לעשות זאת יהיה עסק באזור אפור, בניגוד לכוונה המוצהרת של הפרוטוקול, עם סיכון ממשל גבוה מאוד ניתוח.

ממשל

הממשל מרוכז. המעבדה כתבה כל הצעת פרוטוקול; פיצולים מופעלים בגבהים מקודדים בקוד ללא איתות של כורים; אין מימוש שני של לקוח; ונכון לתחילת אוקטובר לא פורסמה ביקורת של צד שלישי על הצומת או על מעגלי אפס הידיעה. העיתון הישראלי TheMarker דיווח שהמעבדה גייסה כ־120–130 מיליון דולר בסבב חשאי (כ־55 מיליון דולר הון ועוד קרן תשתית GPU של כ־75 מיליון דולר); Pearl לא אישרה זאת בערוצים רשמיים. ״אין הקצאה למשקיעים בטוקן״ ו״חברה עם בעלי מניות מהון סיכון״ יכולים להיות נכונים שניהם, אבל המתח נשאר: הכנסות ה־AI עשויות לזרום להון בעוד שהדילול נופל על מחזיקי הטוקן.

בסך הכול, Pearl היא הניסוי הנקי ביותר עד כה של "חישוב כקונצנזוס": מנגנון חדשני, ניתן לאימות על השרשרת, עם יחידה שממופה אחד לאחד על חומרת AI. אבל התועלת לא הגיעה, ותקציב האבטחה, קצב הגיבוב והתלהבות הכורים תלויים כולם בטוקן שהביקוש העיקרי לו עוד לא נוצר ושההיצע שלו יגדל בכ־78% ב־12 חודשים. כדאי לעקוב אחר שלושה אותות מ־Blockworks: האם השדרוג נוחת בצורה חלקה, האם חלק הבלוקים עם עבודת AI עולה באופן מהותי, והאם ספק הסקה כלשהו מעביר את הכנסת הכרייה ללקוחותיו.



פרק 05

רשתות עבודה מועילה

05 רשתות עבודה מועילה: לא מטריצות ולא אפס ידיעה עוצרות ASIC

Kaspa: מטריצה אינה מועילה ואינה חסינה ל־ASIC

Kaspa היא מקרה מספר הלימוד. ה־kHeavyHash שלה מציב מכפלת מטריצה־וקטור 64 × 64 על ערכים של 4 סיביות בין שני גיבובים, והמכפלה נזרקת. כ־16 חודשים אחרי ההשקה ההוגנת שלה בנובמבר 2021, IceRiver הכריזה על ה־ASIC הראשונים ל־Kaspa ב־14 במרץ 2023, ועד סוף 2023 הכרייה ב־GPU הסתיימה. השוואה בין RTX 4090 ב־3.8 GH/s ו־300 W לבין KS7 של Bitmain ב־77 J/TH נותנת פער יעילות של **בערך פי 1,000** אומדן. אחרי המשתה באה דחיסה: ב־13 בנובמבר הרשת פעלה ב־358.9 PH/s, כ־64% מתחת ל־1 EH/s שלה מאוגוסט 2024, וה־KS7 הטוב ביותר מרוויח רק **0.86 דולר ביום** ב־0.10 דולר ל־kWh.

Aleo: גם אפס ידיעה אינו עוצר חומרה ייעודית

חידת המוכיחים של Aleo נועדה לחקות את העבודה של יצירת הוכחת SNARK; היא אינה מוכיחה עסקאות של משתמשים, והפרויקט מקבל בברכה ובגלוי ASIC ל־ZK. כחמישה חודשים אחרי הרשת הראשית שלה בספטמבר 2024, מכונות ייעודיות כבר נשלחו. ברבעון השלישי של 2025 קיבלו המוכיחים 31.2 מיליון ALEO (כ־7.8 מיליון דולר), בעוד שהרשת גבתה רק כ־43,000–44,000 דולר בעמלות: **מכפיל סבסוד של כ־175–180**. המחיר ירד מכ־9 דולר במסחר המוקדם ל־0.22 דולר בסוף אותו רבעון, ו־ARC-46 דורש כעת סטייק של 100,000 עד 2.5 מיליון ALEO לכל פתרון, מה שהופך את הכרייה לסף סטייק.

Qubic ו־Bittensor

Qubic משווקת "הוכחת עבודה מועילה" שמאמנת את רשת הנוירונים Aigarth שלה, אבל התוצר הבולט ביותר של כוח החישוב שלה ב־2025–2026 היה בלוקים של שרשראות אחרות. תחילה היא כרתה Monero, העלתה את חלקה מפחות מ־2% באמצע מאי 2025 ליותר מ־25% עד סוף יולי וטענה ל־51% באוגוסט, מה שגרם לארגון מחדש של שישה בלוקים (מקור אחר אומר 18). ב־1 באפריל 2026 היא עברה לכריית Dogecoin עם ASIC, והשתמשה בתמורה לרכישות חוזרות ולשריפות. אין ראיה בלתי תלויה שאימון ה־AI שלה מועיל, ושווי השוק שלה ירד מכ־280–300 מיליון דולר לכ־50 מיליון דולר ב־13 בנובמבר. היא גם חשפה השפעה חיצונית: שרשרת שמשלמת על חישוב כללי יוצרת מאגר גדול, שכיר ומתואם של קצב גיבוב, שאפשר להפנות אותו נגד שרשראות ידידותיות ל־CPU.

Bittensor אינה הוכחת עבודה אלא ניקוד סובייקטיבי משוקלל לפי סטייק: מאמתים מנקדים כורים, והמשקלים נחתכים בחציון המשוקלל לפי סטייק. זוהי הרשת היחידה בקבוצה הזאת עם הכנסה חיצונית בהיקף משמעותי. מדד SubConnect, שסופר רק לקוחות חיצוניים ניתנים לאימות, מראה כ־**70 מיליון דולר בשנה** בתחילת נובמבר 2026 (כ־35 מיליון דולר ב־1 בספטמבר), 78% עד 82% מהם ממחשוב ותשתית; אבל במחירי תחילת נובמבר ההנפקה שלה עולה על 300 מיליון דולר בשנה, כך שההכנסה מכסה רק 9% עד 23%. אומדני ההכנסה משתנים מאוד לפי הבסיס, מכ־17 מיליון דולר (Artemis) ועד 172 מיליון דולר בחישוב שנתי

מתוך סכומים בדיווח עצמי. באפריל, 67 מתוך 128 תת־רשתות שרפו יותר מ־99% מהתמריצים שלהן, וכשהצוות המוביל Covenant עזב בגלל מחלוקת ממשל, TAO ירד ב־24% בתוך שעות.

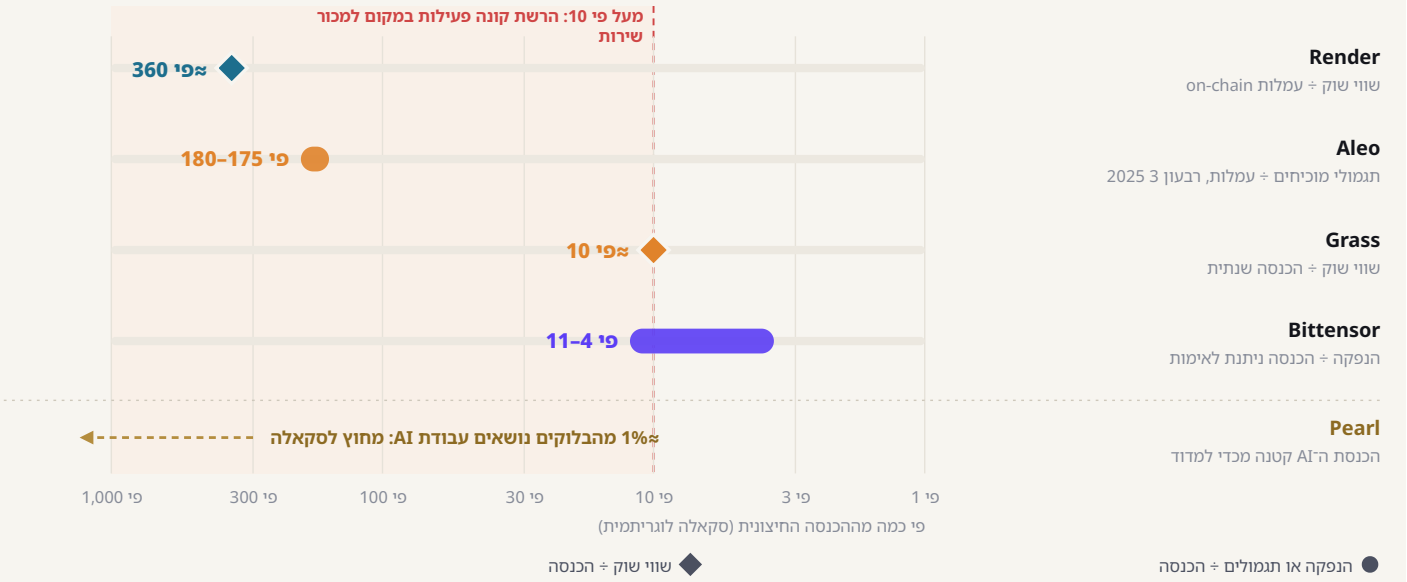
ההיסטוריה נותנת אותה תשובה

שרשראות הראשוניים של Primecoin היו מהירות לאימות, אבל איש לא נזקק להן. Gridcoin הסתמכה על אתרי סטטיסטיקה של צד שלישי כדי לאשר תרומות ל־BOINC, וחוקרים זייפו בלוקים בלי לבצע שום חישוב. הוכחות האחסון של Filecoin עובדות מבחינה טכנית, אבל ברבעון השלישי של 2025 הניצולת הייתה רק 36%, ו־99.5% מעמלות של 793,000 דולר היו קנסות, עם עמלות בסיס של 4,100 דולר בלבד.

פרויקט	עבודה או קונצנזוס	שווי שוק (תאריך)	מסלול חומרה	ביקוש בתשלום מול סבסוד	הכרעה
Pearl	כפל מטריצות רועש (MAC) ב־INT7 ועוד תעודת ZK	כ־500 מיליון דולר (13 בנובמבר)	GPU; טיוטת FP8 תכניס לרשימה לבנה את Hopper ו־Blackwell	כ־1% מהבלוקים עם עבודת AI; הכנסת הכורים ברובה סבסוד	המנגנון האמין ביותר; הביקוש לא הוכח
Kaspa	kHeavyHash (מכפלת מטריצה־וקטור 64 × 64 של 4 סיביות בין גיבובים)	כ־840 מיליון דולר (13 בנובמבר)	ASIC כ־16 חודשים אחרי ההשקה; כיום הכול ASIC	אינה טוענת לתועלת; הכנסת כורים כ־18 מיליון דולר בשנה אומדן	מטריצה אינה מועילה ואינה חסינה ל־ASIC
Aleo	חידת SNARK סינתטית ועוד הוכחת סטייק	כ־118 מיליון דולר (סוף רבעון 3 2025)	מכונות ייעודיות אחרי כחמישה חודשים	תגמולי מוכיחים בערך פי 175–180 מהעמלות	עבודה סינתטית, מונעת בסבסוד
Qubic	“uPoW” (טוענת לאמן את Aigarth)	כ־50 מיליון דולר (13 בנובמבר)	CPU ו־GPU; ASIC ל־Dogecoin מאז אפריל 2026	אין ראיה בלתי תלויה לתועלת; כוח החישוב משמש לכרייה חיצונית ולרכישות חוזרות	צבירת כוח חישוב עטופה בסיפור AI
Bittensor	קונצנזוס Yuma (ניקוד משוקלל לפי סטייק)	כ־2.55 מיליארד דולר (11 בנובמבר)	GPU כלליים; אין מסלול ASIC	כ־70 מיליון דולר בשנה של הכנסה ניתנת לאימות מול יותר מ־300 מיליון דולר של הנפקה	היחידה עם הכנסה חיצונית בהיקף משמעותי; עדיין סבסוד של פי 4–11
Filecoin (להשוואה)	הוכחות שכפול ומרחב־זמן	כ־1.5 מיליארד דולר (סוף רבעון 3 2025)	—	ניצולת 36%; עמלות בסיס רבעוניות 4,100 דולר	מועילה, אבל חסרה ביקוש

איור 04 **מכפיל הסבסוד: פי כמה מההכנסה החיצונית רשת משלמת**

הנפקה או שווי שוק חלקי הכנסה שמשולמת מחוץ לרשת, סקאלה לוגריתמית



מקור: Blockworks (Aleo, רבעון 3 2025), DefiLlama (עמלות Render), Blockworks 0xResearch (Pearl), Dealroom (Grass), CoinGecko, hoge.gg (Bittensor); חישוב של המחבר.

הלקח לחומרה: עבודה קבועה וסדירה שמשולמת בטוקן נזיל מקבלת ASIC בתוך כ־3 עד 16 חודשים, לעתים קרובות מצד מתחרים חדשים לפני השחקנים הוותיקים. טענות "תועלת" החזיקו רק היכן שלקוח חיצוני משלם על התוצר. והגורמים העיקריים לחומרה תקועה הם סבסוד גבוה ביחס להכנסה ושליטה של המפתחים בפונקציית העבודה.

פרק 06

התמונה הרחבה

06 שאר הבינה המלאכותית × קריפטו: ביקוש אמיתי אך קטן, והערך עובר להון

שווקי מחשוב מבוזרים

למחשוב מבוזר יש לקוחות משלמים, אבל הביקוש קטן, מרוכז, וברובו מכירה חוזרת של הסקה על מודלים פתוחים. העמלות שניתן לאמת על השרשרת קטנות: DefiLlama מראה כ־2.35 מיליון דולר עמלות ל־Render בשנה האחרונה וכ־680,000 דולר ל־Livepeer, וכל מגזר ה־DePIN הרוויח ב־2025 רק כ־72 מיליון דולר של הכנסה on-chain. מספרים בדיווח עצמי גדולים יותר אבל מוגדרים ברפיון: Aethir מדווחת על "יותר מ־127.8 מיליון דולר" הכנסות לשנת הכספים 2025 בלי לומר אם זה חיוב ברוטו או החלק של הפרוטוקול. שווי הטוקנים וההכנסות נפרדו בשני הכיוונים: שווי השוק של Render הוא בערך פי 360 מעמלות ה־on-chain שלה, בעוד ששווי השוק של Aethir, כ־87.7 מיליון דולר, נמוך מההכנסה שהיא מדווחת עליה **אומדן**.

הטענה להיות זול ב־70% עד 85% מחזיקה רק מול מחירי הדרישה של ענקיות הענן. ב־3 בנובמבר, המחיר הממוצע המשוקלל של H100 ב־Akash היה כ־**2.73 דולר לשעת GPU** (טווח 3.16–2.04 דולר), נמוך מכ־6.88 דולר ב־AWS ומ־11.06 דולר ב־Google Cloud, אבל גבוה מכ־1.73 דולר ב־Vast.ai, מ־1.99 דולר בענן הקהילתי של RunPod ומכ־2.44 דולר להזמנה לשנה (המחירים בפלטפורמות האחרות כפי שנקראו ב־12 בספטמבר; Akash, המפרסמת, היא צד מעוניין). הביקוש הוא ברובו הסקה סחורתית

שמופצת דרך OpenRouter, כ־1.7 מיליארד טוקנים ביום ב־AkashML ויותר מ־4 מיליארד ב־io.net, בתחרות ישירה עם כל neocloud. הדולרים הגדולים ביותר של "מחשוב מבוזר" חוזרים להתרכז: Axe Compute, שמחזיקה את הטוקן של Aethir כנכס אוצר, חתמה על חוזה לחמש שנים בשווי 1.5 מיליארד דולר להשכרת אשכול משלה של 9,200 מעבדי B300, מול הכנסה של 35,000 דולר ברבעון הראשון של 2026.

אימון והסקה מבוזרים

אימון דרך האינטרנט הודגם ב־8 עד 40 מיליארד פרמטרים. Pluralis-8B התאמן 40 יום על 330 צמתים, ברובם GPU צרכניים, והגיע ל־63% מהיעילות של בסיס השוואה מרוכז של H100, אף שהמאמר שלו מודה שהריצה לא הייתה מבוזרת ונזקקה לצמתים של Pluralis עצמה ולרשות קבלה. Psyche של Nous Research השלימה את ריצת האימון המקדים Consilience של 40 מיליארד פרמטרים. בהיקף החזית העבודה חזרה להתרכז: INTELLECT-3 של Prime Intellect, מודל תערובת מומחים של 106 מיליארד פרמטרים, אומן במשך חודשיים על 512 H200 שמחוברים ב־InfiniBand של 400 Gbps. האילוצים ברורים: צמתים צריכים רוחב פס של 200 Mbit/s לפחות ודחיסת תקשורת של בערך פי 100, והאמון והאימות עדיין תלויים בתפקידים המיוחסים של המפעיל.

הכסף הלך לחברות הון שמוכרות מוצרים מרוכזים. ל־Prime Intellect היו יותר מ־100 מיליון דולר הכנסה שנתית לפני שגייסה 130 מיליון דולר לפי שווי של 1 מיליארד דולר ביולי; Venice, אפליקציית הסקה פרטית, מדווחת על יותר מ־70 מיליון דולר הכנסה שנתית ועל רווחים מאז הרבעון הראשון של 2026 (נתוני החברה), גייסה 65 מיליון דולר לפי שווי של 1 מיליארד דולר ורצה על GPU חכורים. בצד הטוקנים, רשת האימון המרכזית של Gensyn עדיין "תיקבע בהמשך" והטוקן שלה שווה כ־17.2 מיליון דולר; הטוקן של Ritual, שכבה 1 שנבנתה מלכתחילה ל־AI, ירד ב־72% מהשיא שלו אחרי הרשת הראשית (כותרת בלבד).

AI ניתן לאימות

הייצור בחר בסביבות ריצה מהימנות (TEE). המחשוב החסוי של Hopper מוסיף פחות מ־7% תקורה לשאילתות LLM טיפוסיות וכמעט כלום למודלים גדולים; התקורה בייצור של Phala היא כ־0.5% עד 5%, אבל עד סוף 2025 שילמו רק 398 מתוך 10,004 המשתמשים שלה. שורש האמון של TEE הוא יצרן השבב: TEE.fail, שנחשף באוקטובר 2025, השתמש ברכיב יירוט על אפיק DDR5 שעלה פחות מ־1,000 דולר כדי לחלץ מפתחות אישור (attestation) ממכונות Intel TDX ו־AMD SEV-SNP מעודכנות לחלוטין, מה שבתורו מערער את המחשוב החסוי של ה־GPU של NVIDIA, ושני היצרנים אומרים שהתקפות פיזיות נמצאות מחוץ למודל האיום שלהם. זו בדיוק הסביבה של DePIN, שבה למפעילי הצמתים יש גישה פיזית. zkML יכול כיום להוכיח LLM שלם: DeepProve של Lagrange מוכיח את GPT-2 בכ־174 טוקנים לדקה ואת Gemma 3 בכ־86, סדרי גודל רחוק מהסקה בייצור וללא שימוש בתשלום שפורסם.

תשלומי סוכנים ונתונים

x402 היא הדוגמה המזהירה הטובה ביותר. TRM Labs סופרת כ־198.9 מיליון עסקאות ו־52.7 מיליון דולר שנסלקו על ידי מתווכים ידועים מאז מאי 2025, 99.6% ב־USDC; אחרי הסרת תשלומים עצמיים, תזרימי אצווה ומוכרים עם פחות מעשרה קונים, החלק שסוכנים יזמו בפועל הוא רק 0.6% עד 7.5%, **כ־5,000–11,000 דולר בחודש**. Artemis מעריכה שיותר מ־95% מהפעילות הם איתות של פרוטוקול ולא קנייה ומכירה. בינתיים המסילות עוברות תקינה בידי השחקנים הוותיקים: קרן x402 הוקמה תחת Linux Foundation ב־14 ביולי, עם Visa, Mastercard, Stripe, Google, AWS, Cloudflare ו־Coinbase כחברי ליבה, ו־Tempo, שכתבה יחד עם Stripe את Machine Payments Protocol, השיקה את הרשת הראשית שלה במרץ לפי שווי של 5 מיליארד דולר.
פלטפורמות של טוקני סוכנים עדיין פועלות לפי היגיון של משגרי טוקנים (launchpads): Virtuals מדווחת בעצמה על 479 מיליון דולר של "aGDP", ומייסד ElizaOS הכריז על מות הטוקן שלה אחרי הסדר בתביעה ייצוגית. סביר יותר שהערך ייתפס בידי מנפיקי מטבעות יציבים ושחקני התשלומים הוותיקים; התפקיד הייחודי של הקריפטו כאן עשוי להיות זהות ואחריותיות לסוכנים.

נתונים הם אחד מתתי־המגזרים המעטים עם קונים אמיתיים מחוץ לשרשרת. Grass מדווחת על 17 מיליון דולר של הכנסות נתונים ב־2025 ועל 17 מיליון דולר נוספים במחצית הראשונה של 2026, מכוונת ל־75 מיליון דולר לשנה ואומרת שהיא רווחית; GRASS שווה כ־328 מיליון דולר, בערך פי עשרה מההכנסה השנתית של המחצית הראשונה **אומדן**. החפיר שלה הוא היקף האיסוף של רוחב פס מבוזר במגורים, לא הקריפטו עצמו. ל־DAO של נתונים ולשווקי נתונים חסר ביקוש בתשלום גלוי: Vana שווה כ־21.5 מיליון דולר ו־Ocean כ־27.6 מיליון דולר, ו־Ocean מצויה בהתדיינות משפטית עם Fetch.ai. הוכחת אנושיות (proof of

personhood) צוברת אינטגרציות אמיתיות ככל שהסוכנים מתרבים: World אומרת שאימתה כמעט 18 מיליון בני אדם, ו־Zoom, Docusign ו־Tinder תומכות בחלקים של World ID.

ההון כבר הצביע

מגזר	ביקוש בתשלום ניתן לאימות	אותות סבסוד ושווי	הכרעה
רשתות עבודה מועילה	ל־Bittensor כ־70 מיליון דולר בשנה של הכנסה חיצונית ניתנת לאימות; ב־Pearl כ־1% מהבלוקים עם עבודת AI	Aleo פי 175–180; Bittensor פי 4–11; Pearl ברובה סבסוד	למנגנונים יש ערך; התועלת צריכה לקוחות משלמים
שווקי מחשוב מבוזרים	עמלות on-chain כ־0.68–2.35 מיליון דולר בשנה (Livepeer, Render); בדיווח עצמי 4–128 מיליון דולר	שווי השוק של Render בערך פי 360 מהעמלות	אמיתי אך קטן, ברובו מכירה חוזרת של הסקה סחורתית
אימון והסקה מבוזרים	חברות הון: Prime Intellect מעל 100 מיליון דולר בשנה, Venice מעל 70 מיליון דולר (תשתית מרוכזת)	רשת האימון הראשית של Gensyn אינה פעילה; הטוקן של Ritual -72% מהשיא	ההתקדמות הטכנית אמיתית; הרווחים הולכים להון
AI ניתן לאימות	TEE בייצור (398 משתמשים משלמים ב־Phala); אין zkML בתשלום שפורסם	—	TEE מועילים; zkML מוקדם
תשלומי סוכנים	תזרימים שסוכנים יזמו כ־5,000–11,000 דולר בחודש	כ־200 מיליון עסקאות, רובן איתות וספקולציה	קודם המסילות, הביקוש אחר כך
רשתות נתונים	ל־Grass 17 מיליון דולר במחצית הראשונה של 2026 (דיווח עצמי)	GRASS בערך פי 10 מההכנסה השנתית	אחד התחומים המעטים עם קונים אמיתיים

שימו את ששת המגזרים זה לצד זה והמסקנה עקבית: **סבסוד הוא דלק להתנעה, לא מודל עסקי.** ב־13 בנובמבר קטגוריית ה־AI של CoinGecko הייתה שווה בסך הכול כ־18.6 מיליארד דולר, אבל כ־56% מזה הם שכבות 1 כלליות כמו NEAR ו־ICP, ועוד פרויקטים ברמת הקונצנזוס כמו TAO, TAO בסטייקינג (ייתכן שנספר פעמיים יחד עם TAO) ו־Pearl; טוקני AI בשכבת הפרוטוקול ובשכבת היישומים מסתכמים בכ־8 מיליארד דולר או פחות אומדן. זה בערך השווי המשותף של ארבע חברות הון: Tempo (5 מיליארד דולר), Nous (1.5 מיליארד דולר לפי דיווח), Prime Intellect (1 מיליארד דולר) ו־Venice (1 מיליארד דולר). מימון הון הסיכון בקריפטו תקוע על כ־4.5–5 מיליארד דולר לרבעון, כשהעלייה באה מתשתית שקשורה ל־AI, שחלקה בסכומים שפורסמו עלה מ־8.1% ל־18.2% ברבעון השלישי של 2026; ו־84.7% מ־118 הטוקנים שהושקו ב־2025 נסחרים מתחת למחיר ההשקה שלהם. במשפט אחד: **הקריפטו מספק הפצה וסליקה, ההון לוקח את הרווחים**, לפחות ב־2026.



פרק 07

07

כלכלת ASIC

שאלת ה־ASIC: נוסחה אחת, שלושה קני מידה

נוסחה אחת: מדוע יעילות מפסיקה להשתלם אחרי k = 5

יהי R ההכנסה היומית ליחידת חישוב (הכנסת הכורים ברשת חלקי קצב הגיבוב של הרשת), c העלות היומית של יחידת חישוב בחומרה השולית (ב־Pearl, GPU), ו־k פי כמה ASIC זול מהחומרה השולית הזאת (c_ASIC = c/k). כל עוד ה־GPU נשארים הכורים השוליים, כניסה ויציאה חופשיות דוחפות את R לעבר c, ולכן ASIC מרוויח R - c/k ליחידת חישוב:

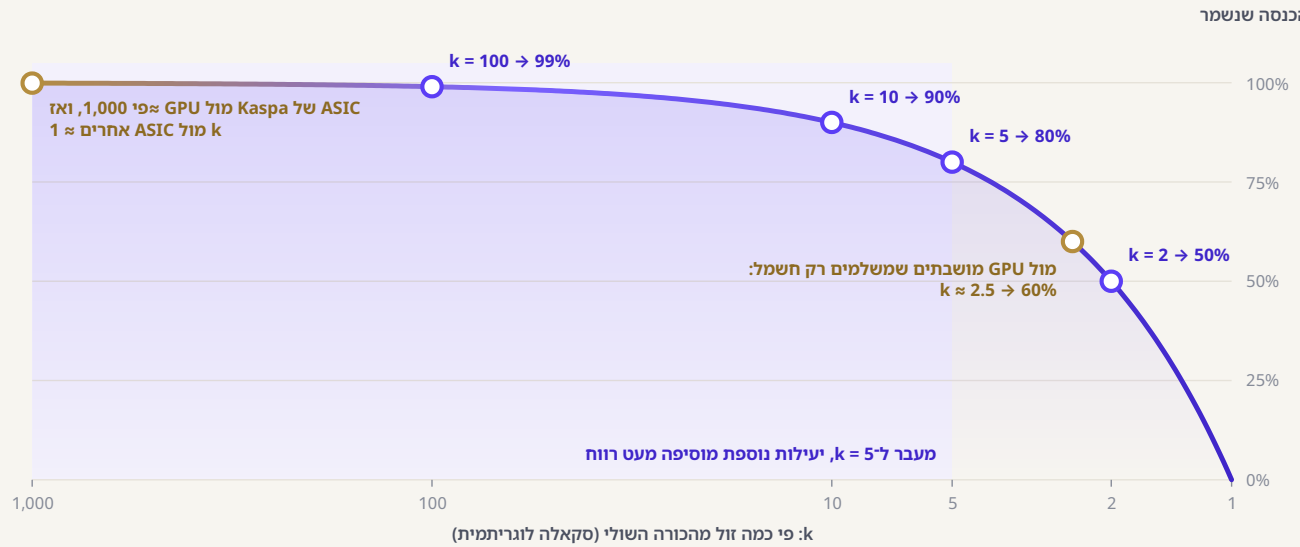
תקרת הרווח של ASIC ≈ R × (1 - 1/k)

ברמת הרשת, תקרת הרווח הגולמי השנתי של צי ASIC היא בקירוב הכנסת הכורים ברשת × (1 - 1/k), ורק עד שה־ASIC מחליפים את ה־GPU ככורים השוליים.

לנוסחה שלוש השלכות. ראשית, (1 - 1/k) קעורה: ב־k = 2 ASIC שומר 50% מההכנסה, ב־k = 5 80%, ב־k = 10 90%, ב־k = 100 99%. **מעבר ל־k = 5–10, יעילות נוספת כמעט אינה מוסיפה רווח**; גודל ה־R וגודל הקופה מכריעים. שנית, נקודת הייחוס מתאפסת: ברגע שה־ASIC הם הכורים השוליים, יש למדוד את k מול ASIC אחרים. ל־ASIC של Kaspa היה יתרון יעילות של בערך פי 1,000 על פני GPU, שקנה רק רווח נפל קצר; כיום ה־KS7 הטוב ביותר מרוויח 0.86 דולר ביום. שלישית, בשרשרת ידידותית ל־GPU כמו Pearl, נקודת הייחוס היא לעתים קרובות לא מחיר הרכישה של GPU אלא **החשמל של GPU מושבת**: RTX 5090 בבעלות, ב־525 W וב־0.10 דולר ל־kWh, עולה כ־1.26 דולר ביום בחשמל ומרוויח כ־5–6 דולר אומדן. מול GPU שהם עלות שקועה, אפילו ל־ASIC שצורך חמישית מהאנרגיה יש k אפקטיבי של כ־2.5 בלבד אם הפחת שלו שווה לחשמל שלו, ותקרת רווח של כ־60% מ־R (הנחה להמחשה).

איור 05 מדוע יעילות מפסיקה להשתלם: ASIC שומר לכל היותר 1 - k/1 מההכנסה

חלק ההכנסה ש־ASIC יכול לשמור מול הכורה השולי, לפי יתרון העלות שלו k



מקור: המודל של המחבר: כשהחומרה השולית קובעת את המחיר, ההכנסה ליחידת חישוב שואפת לעלותה; נתוני Kaspa ו־RTX 5090 מהדוח.

שלושה קני מידה: הקופה, היקף ההחזר והכללים

קנה המידה הראשון הוא **הקופה**. בפריסה לשנה של כ־2,893 BTC של תגמולי בלוקים בשבוע הראשון של נובמבר (כ־222 מיליון דולר), כורי ביטקוין מרוויחים כ־11.5 מיליארד דולר בשנה אומדן. הכורים של Pearl מרוויחים כ־0.46–0.49 מיליארד דולר (במחירים הנוכחיים); תגמולי המוכיחים של Aleo, בפריסה לשנה מהרבעון השלישי של 2025, כ־31 מיליון דולר; והכורים של Kaspa, בתגמול בלוק של 1.945 KAS ועשרה בלוקים בשנייה, כ־18 מיליון דולר אומדן. **שנה שלמה של הכנסות כורים ב־Kaspa היא פחות מהקצה הנמוך של תקציב תכנון של שבב יחיד ב־5/4 nm (כ־30 מיליון דולר).**

קנה המידה השני הוא **קצב הגיבוב הדרוש כדי להחזיר את ה־NRE**:

h* = NRE ÷ [R × (1 - 1/k) × T × p]

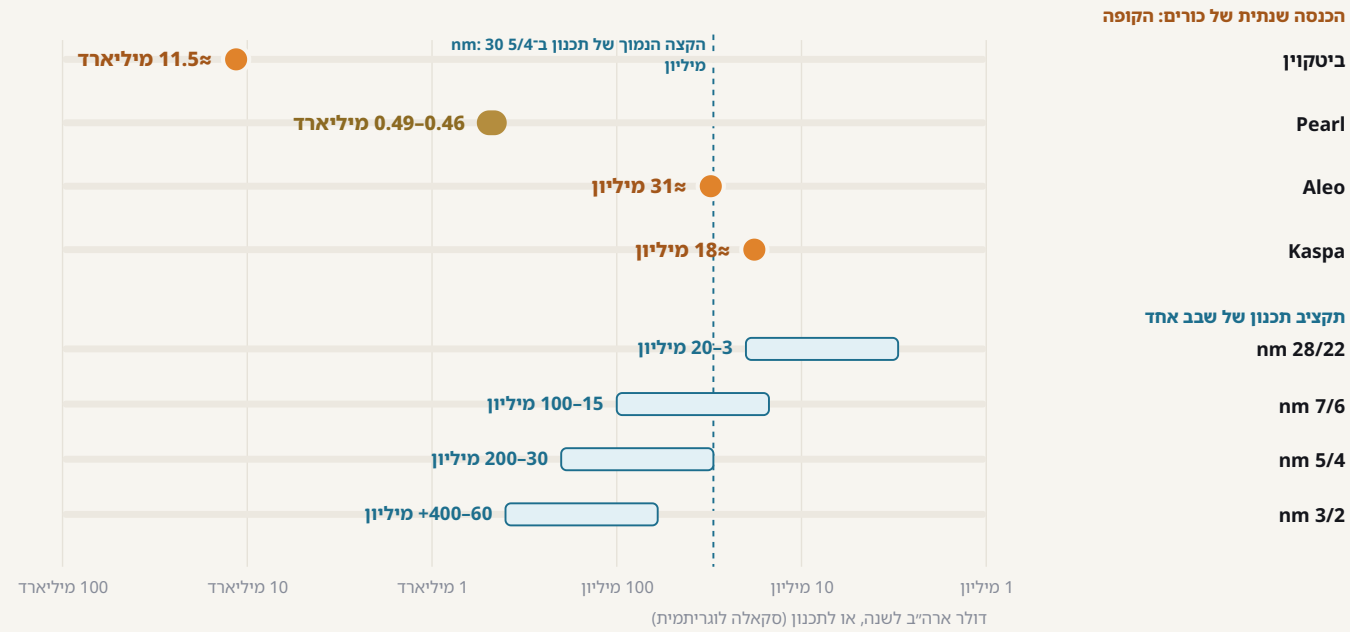
כאשר T הוא החלון המועיל בימים ו־p ההסתברות שהכללים יישארו ללא שינוי בחלון הזה. עבור Pearl, ניקח R = 14 דולר ל־PH/s ליום, k = 5, T = 365 ימים ו־NRE של 20 מיליון דולר (שבב מערך חוזר ב־N7/N6; בקצה הנמוך של הטווחים של 5–30 מיליון דולר לצמתים בטווח הביניים לפי Silicon Analysts ושל 15–100 מיליון דולר ל־ASIC חישוב ב־7/6 nm לפי AnySilicon, משום ששבב כרייה משתמש במעט IP ובמבנה חוזר; ההנחה שלי). עם p = 1, הוא צריך כ־**4.9 EH/s** שפועלים שנה שלמה, כ־5% מהרשת; עם p = 0.5, כ־**9.8 EH/s**, כ־11%; ואם R חוזר לרמת הקיץ של כ־10 דולר, אפילו p = 1 דורש כ־6.9 EH/s אומדן. ההיקף אינו מחוץ להישג יד. הבעיה היא p ו־T.

קנה המידה השלישי הוא **סיכון הכללים**, ולעתים קרובות הוא המכריע. הפיצול נגד ASIC של Monero ב־2018 השאיר את ה־X3 של Bitmain תקוע ולזמן קצר עלה לרשת כמעט מחצית מקצב הגיבוב שלה. הפיצול של Sia באותה שנה הפך לחסרי תועלת את הכורים של Bitmain ושל Innosilicon, בעוד שמחיר היד השנייה של כורה ה־Obelisk, שקשור למפתחים, עלה מ־2,499 דולר ליותר מ־5,000 דולר. קצת יותר מחודש אחרי ה־Merge של Ethereum, פחות מ־20% מכורי Ethereum לשעבר מצאו עבודה חדשה. **מי ששולט בפונקציית העבודה שולט בגורלם של משקיעי החומרה**, ו־Sia מראה ש"עמידות ל־ASIC" יכולה להפוך לחפיר של המפתחים עצמם. חומרה עם שוק שני שורדת: אחרי ה־Merge, ה־GPU עברו ל־AI ולענן, בעוד ש־ASIC של אלגוריתם יחיד הצטופפו ב־Ethereum Classic. גם החלון מצטמצם: ה־ASIC הראשון של Kaspa לקח כ־16 חודשים, של Aleo כחמישה, והראשון שמגיע לחזית היעילות הוא לעתים קרובות מתחרה חדש ולא שחקן ותיק.

רשת	הכנסת כורים שנתית (בקירוב)	חומרה שולית או חומרת ייחוס	יציבות הכללים	k זמין לשחקן חדש	הכרעה
ביטקוין	11.5 מיליארד דולר	Bitmain S23 XP Hyd (8.9 J/TH, 25.5 דולר ל־TH)	גבוהה מאוד	≥1	קופה גדולה, אבל ה־k שייך לשחקנים הוותיקים
Pearl	0.46–0.49 מיליארד דולר (במחירים הנוכחיים)	GPU בבעלות או מושבתים (עלות שולית ≈ חשמל) ו־GPU של מרכזי נתונים	נמוכה מאוד: ארבעה פיצולים בחודשיים; רשימה לבנה של FP8 בהמתנה	אולי פי 3–10 ביעילות ניתוח; פחות מול GPU בעלות שקועה	הקופה גדולה מספיק; סיכון הכללים מטיל וטו
Aleo	כ־31 מיליון דולר (רבעון 3 2025, בפריסה לשנה)	IceRiver AE2 ומכונות ייעודיות אחרות	בינונית: סף סטייק של ARC-46	≈1	קופה קטנה, והסף עולה
Kaspa	כ־18 מיליון דולר	KS7 (77 J/TH)	גבוהה	≈1	הכנסה שנתית מתחת לקצה הנמוך של תקציב תכנון אחד ב־5/4 nm

איור 06 **רוב הקופות קטנות מתקציב תכנון של שבב אחד**

הכנסה שנתית של כורים בארבע רשתות מול עלות התכנון של שבב אחד לפי צומת תהליך, סקאלה לוגריתמית



מקור: Hashrate Index (ביטקוין, נובמבר 2026), prlstats ו־Blockworks (Kryptex (Pearl), Aleo, רבעון 3 2025), minerstat (Kaspa), AnySilicon (תקציבי תכנון, 2026); הפריסה לשנה של המחבר.

הרביע של "קופה גדולה, כללים יציבים ו־$k > 1$" ריק כיום: לביטקוין יש הקופה והכללים היציבים, אבל ה־k שלו שייך לשחקנים הוותיקים; הקופה של Pearl גדולה מספיק אבל הכללים שלה הם הפחות יציבים; ל־Kaspa ול־Aleo יש כללים יציבים מספיק וקופות קטנות מדי.



פרק 08

שבב אחר שבב

08 חמישה סוגי שבבים, חמש תשובות

PoUW בסגנון Pearl: לא שבב INT7 ייעודי; להמר על כורה שיכול לשרת הסקה

הטיעון בעד ASIC אמיתי. רק כ־1% מהבלוקים נושאים כיום עבודת AI, ולכן הטיעון הכלכלי של הספר הלבן, ש־ASIC ייעודי מאבד את היכולת לבצע עבודה מועילה ולכן מרוויח פחות, נמצא כרגע בנקודה החלשה ביותר שלו. החלקים היקרים ביותר של GPU, ה־HBM ומארז ה־CoWoS, הם כ־63% מעלות הייצור של H100, כשה־die הלוגי הוא רק כ־9%. שבב שלא עושה דבר מלבד פעולות הכפלה־צבירה במספרים שלמים של 7 סיביות, מחזיק את האופרנדים שלו ב־SRAM שעל השבב ופועל במתח נמוך, יכול באופן סביר להגיע לאנרגיה לפעולה טובה פי 3 עד 10 מזו של GPU באותו צומת, וליתרון גדול יותר בעלות לפעולה [ניתוח; עדיין אין מדידות]. הטיעון נגד קשה יותר. זמני האספקה של הפאונדרי בצמתים המובילים יכולים לעלות על 50 שבועות, וצוות חדש צריך כ־18 עד 24 חודשים מההתחלה ועד לייצור בנפח **אומדן**; עד שהשבב יגיע, Pearl בהחלט עשויה לעבור ל־FP8 ואחר כך ל־FP4. ומשום ש־PIP-3 מגדיר את האריתמטיקה המתקבלת דרך רשימה לבנה של התקנים, שבב INT7 פשוט יפסיק לעבוד ברגע שהיא תופעל.

לכן המסקנה שלי היא: **לא לבצע עכשיו טייפ־אאוט לשבב INT7 ייעודי ל־Pearl.** יש שני הימורים טובים יותר. האחד הוא מאיץ שמכוון קודם כול להסקה ויכול גם לכרות: בנוי סביב GEMM כללי ב־INT8/FP8 ואופרנדים בתוך SRAM, כך שאם הגישה לכרייה תישלל, הערך נוחת בשוק הסקה אמיתי. השני הוא להצטרף לדיון על הפרוטוקול ולדחוף לאריתמטיקה קנונית ניטרלית לספק ש־ASIC יכול לממש במדויק, יחד עם התחייבות ליציבות האלגוריתם, שהיא התנאי המוקדם לכך שהרביע של "קופה גדולה, כללים יציבים, $k > 1$" יתקיים בכלל. עד אז, אפשר לאמת ארכיטקטורה על FPGA או על פרוסות מרובות פרויקטים (MPW) בצומת בוגר, ולשמור את סיכון ה־NRE בסדר גודל של מיליון דולר. פרק 10 הופך זאת לתקציב מדורג.

כורים בסגנון ביטקוין: ה־k בידי השחקנים הוותיקים, והחשמל הוא הנכס הנדיר

סיכון הכללים של ביטקוין הוא הנמוך ביותר, אבל התחרות דוחפת את ה־k של שחקן חדש אל מתחת ל־1. ה־S23 XP Hyd של Bitmain מגיע ל־8.9 J/TH, נמכר ב־25.5 דולר ל־TH ונשלח מנובמבר; יעילות הדגל ירדה מ־17.5 J/TH ל־S21 ב־2023 ל־8.9 J/TH, כ־49% בשלוש שנים. שחקן חדש שישלח בנפח ב־2027–2028 יצטרך כ־8 J/TH או טוב מזה כדי להחזיק מעמד [אומדן]. המתחרים עוזבים: Canaan דיווחה על הכנסות של 31.9 מיליון דולר ברבעון השני של 2026 עם הפסד גולמי של 29.3 מיליון דולר, ונתנה תחזית של 11–15 מיליון דולר בלבד לרבעון השלישי; המכירות של Bitdeer של מכונות משלה לקונים חיצוניים ברבעון השלישי ירדו ל־0.3 מיליון דולר, עם שולי רווח גולמי קבוצתיים של 3.4%; Auradine שינתה את שמה ל־Velaura AI ופנתה לשבבי AI; ו־Core Scientific ביטלה הזמנה של כ־15 EH/s של מכונות 3 nm מיחידת Proto של Block, ורשמה הפסד של 41.9 מיליון דולר.

גם הביקוש מצטמצם. ב־9 בנובמבר מחיר הגיבוב (hashprice) היה כ־31.89 דולר ל־PH/s ליום, עקומת החוזים העתידיים עד אפריל 2027 עמדה בממוצע על כ־32.7 דולר, והעמלות היו רק 0.72% מתגמולי הבלוקים. ב־0.045 דולר ל־kWh, ה־S23 XP Hyd מחזיר את עצמו בתוך כ־3.1 שנים [אומדן], קרוב לחציה הבאה הצפויה בסביבות 2028. חשובה יותר עלות ההזדמנות של החשמל: חכירות ל־AI ולמחשוב עתיר ביצועים מניבות כיום כ־1.67–1.9 מיליון דולר למגוואט בחוזה לשנה, שבהכנסה תפעולית נטו של 80% עד 85% הן בערך פי 1.5 עד 2.9 מהרווח הגולמי של כרייה למגוואט (כ־0.56–0.9 מיליון דולר) [אומדן; שני בסיסי המגוואט שונים]. Core Scientific תפסיק לכרות ברבעון הרביעי; Bitfarms, כיום Keel, סגרה את כל אתרי הכרייה שלה בארה"ב; IREN הרוויחה 200 מיליון דולר מענן AI ורק 28 מיליון דולר מכרייה ברבעון אחד. **החשמל וקיבולת מרכזי הנתונים הם הנכסים; מכונות הכרייה הן מתכלים.**

חשבון ה־NRE אינו סלחני יותר. יעילות בחזית דורשת כיום תהליכים בדרגת 3 nm (לפי מקורות משניים, MicroBT משתמשת ב־Samsung SF3, והשבבים של Auradine ושל Proto הם 3 nm). מסכות עולות כ־10–20 מיליון דולר ו־NRE מלא 30–100 מיליון דולר ויותר; ב־20–25 דולר ל־TH ובשולי רווח גולמי של 30%, כל 1 EH/s שנמכר מניב כ־6–7.5 מיליון דולר של רווח גולמי, ולכן כיסוי מו"פ שנתי בהיקף של Bitdeer, כ־110–160 מיליון דולר, דורש מכירה של כ־20–25 EH/s בשנה לקונים חיצוניים, כ־2% מהרשת [אומדן]. המסקנה: **צוות חדש לא צריך להיכנס לשוק כורי הביטקוין למכירה פתוחה.** שלושה מקרים עדיין הגיוניים: אינטגרציה אנכית עם חשמל זול משלך; תצורות נישה כמו כורים ביתיים וכורים לחימום (גם Canaan משיקה מוצרי חימום ביתיים); והשתתפות במפות הדרכים של השחקנים הוותיקים דרך IP ושירותי תכנון.

שבבים לאימון AI: לא כצוות קטן

בשבבי אימון, הכי קשה להשיג k. באותו דור ובאותו דיוק, Ironwood של Google מקדים את ה־B300 ביעילות FP8 ברמת השבב רק בערך פי 1.1 עד 1.8 (ההספק של Ironwood הוא אומדן של עיתונאי) [אומדן]; היחס של פי 30 עד 80 שמצוטט לעתים קרובות מתוך מאמר TPU v1 נמדד מול K80, שלא היו לו ליבות טנזור. המנוף האמיתי הוא פורמט המספרים: מעבר מ־FP8 ל־FP4 נותן בדרך כלל פי 2 עד 3 בתפוקה על אותו סיליקון, ו־NVIDIA זזה שם הכי מהר: היא טוענת ל־50 PF של הסקת NVFP4 עבור Rubin, בנפח במחצית השנייה של 2026. ספי העלות וההיצע גבוהים עוד יותר: תקציבי תכנון ב־3/2 nm מגיעים ל־60–400 מיליון דולר ויותר; קיבולת ה־CoWoS ב־2026 היא כ־106,000–130,000 פרוסות בחודש, ולפי הדיווחים TSMC שוקלת העלאות מחירים של עד 20%; במחיר חוזה של 15.6 דולר ל־GB עבור HBM3E, ה־HBM לבדו במאיץ של 288 GB עולה כ־4,500 דולר [אומדן]. השוק הזה שייך לפיתוח משותף בסגנון Broadcom עם ענקיות הענן (פרק 11). המסקנה: **צוותים קטנים לא צריכים לבנות שבבי אימון עצמאיים**; במקום זאת, לבנות IP, צ'יפלטים או שירותי תכנון.

שבבים להסקת AI: להחליף כלליות ב־k, וזה משתלם

ההסקה היא המקום שבו ל־ASIC יש הכי הרבה מרחב, משום שאפשר לקנות k בוויתור על כלליות. Etched טוענת שהשרת שלה עם שמונה שבבים מריץ את Llama-70B ביותר מ־500,000 טוקנים בשנייה, בערך פי 20 משמונה H100; Taalas צורבת את משקלי המודל ב־ROM של מסכה וטוענת להספק נמוך פי 10 ולעלות ייצור נמוכה פי 20 (שתיהן טענת החברה). ארכיטקטורות שמתמקדות ב־SRAM נמנעות מ־HBM ומ־CoWoS, והצמתים שלהן ניתנים לניהול טוב יותר: Taalas משתמשת ב־TSMC 6 nm, ו־Groq עברה מ־14 nm ל־Samsung 4 nm. הפצת הקריפטו אינה עוזרת לשבבי הסקה, משום שרשתות מחשוב מבוזרות רצות כמעט כולן על NVIDIA; הדרך פנימה היא כלכלת הרכש של ה־neoclouds ושוק ההסקה המקומי הגדול בסין, שבו מרכזי נתונים במימון המדינה חייבים להשתמש בשבבי AI מקומיים, היבוא של H200 מוגבל, ובייג'ינג רוצה שחברות אינטרנט גדולות יריצו הסקה על שבבים מקומיים. המסקנה: **שווה לבנות**, כל עוד יש לקוח ברור ומידת הנעילה למודלים ולארכיטקטורה היא בחירה מכוונת. זהו גם השוק השני הטוב ביותר לשבב כרייה PoUW.

מאיצי ZK ו־FHE: לגדר ב־ZK, להתמקם מוקדם ב־FHE

ה־GPU כבר השיגו הוכחת ZK בזמן אמת: SP1 Hypercube של Succinct הוכיח 99.7% מבלוקי Ethereum בפחות מ־12 שניות על 16 כרטיסי Cysic. RTX 5090 טוענת ששבב ה־ZK C1 שלה יעיל פי 10 עד 100 ממוכיחי GPU ושיישלח ב־2026, אבל הביקוש שלו בא ברובו מתגמולי הטוקן שלה עצמה טענת החברה; לא מצאתי ראיה פומבית לכך ש־ASIC לחידה של Aleo משמשים להוכחות ZK אמיתיות. הוכחת ZK היא במספרים שלמים ודטרמיניסטית, קרובה בצורתה למה ששבבי כרייה עושים, אבל מערכות ההוכחה משתנות מהר (SP1, RISC Zero, Halo2, Plonky2 ועוד). המסקנה: **להמתין ולגדר**, עם מנועי MSM, NTT וגיבוב ניתנים לתכנות, תחילה על FPGA ואחר כך כ־ASIC.

FHE הוא הפרימיטיב התלוי ביותר בחומרה, עם פוטנציאל הרווח הגדול ביותר. Heracles של Intel, על Intel 3 עם 176 mm² ,197 W ו־48 GB של HBM3, מריץ פעולות FHE מהר פי 1,074 עד 5,547 מ־Xeon של 24 ליבות, אבל ל־Intel אין תוכניות מסחריות ומוביל הפרויקט עזב בדצמבר 2025. Niobium בונה על Samsung 8LPU עם SemiFive כשותפת תכנון (חוזה פיתוח של כ־6.86 מיליון דולר) וגייסה יותר מ־23 מיליון דולר; Optalysys, בגישה פוטונית, גייסה 31 מיליון דולר ואומרת ששבב מסחרי מלא עדיין רחוק שנתיים עד שלוש; והתכנון בקוד פתוח של Zama ל־FPGA מגיע לכ־13,000 אתחולים מחדש ניתנים לתכנות (programmable bootstraps) בשנייה בכ־200 W, בסיס ההשוואה הפומבי שכל ASIC חייב לעבור בבירור. המסקנה: **להתמקם מוקדם ולאמת בנכסים קלים**, לחלון בסביבות 2027–2029. סביר יותר שהביקוש יעוגן ב־AI מוצפן לארגונים ולממשלות מאשר במטבעות קריפטוגרפיים עצמם, ו־FHE זקוק ל־HBM, כך שמבנה העלויות שלו דומה לזה של שבב אימון.

סוג שבב	מאין בא k	ביקוש וקופה	NRE וספים	סיכון כללים או נעילה	שוק שני	הכרעה
שבב כרייה PoUW בסגנון Pearl	ויתור על HBM, CoWoS ויחידות כלליות: אולי פי 3–10 ב־J/op ניתוח	הכנסת כורים כ־0.46–0.49 מיליארד דולר בשנה, תנודתית	מערך חוזר ב־N7/N6 כ־10–30 מיליון דולר ניתוח	גבוה מאוד: הפרוטוקול קובע את פורמט המספרים ואת הרשימה הלבנה של ההתקנים	הסקת AI, אם נבנה כ־GEMM כללי ב־INT8/FP8	לא שבב INT7 ייעודי; לבנות שבב שמכוון להסקה וניתן לכרייה, או להמתין לאריתמטיקה קנונית
שבב כרייה SHA-256 לביטקוין	1 ≤ k מול Bitmain	כ־11.5 מיליארד דולר בשנה, שוק פתוח שבידי שתי חברות	מסכות בדרגת 3 nm 10–20 מיליון דולר; NRE מלא 30–100 מיליון דולר ויותר	נמוך (אלגוריתם יציב); סיכון גבוה של תחרות וחציה	כמעט אין (אלגוריתם יחיד)	שחקנים חדשים מתרחקים מהשוק הפתוח; רק חשמל עצמי, תצורות נישה או IP
שבב לאימון AI	מערכים סיסטוליים מאותו דור רק פי 1.1–1.8; הפורמט הוא המנוף	הגדול ביותר (ל־Broadcom כ־58 מיליארד דולר הכנסות ממוליכים למחצה ל־AI בשנת הכספים 2026)	תכנון ב־3/2 nm 60–400 מיליון דולר ויותר; CoWoS ו־HBM	הקצב השנתי של NVIDIA	—	צוותים קטנים בונים IP, ציפלטים ושירותי תכנון, לא שבבים עצמאיים
שבב להסקת AI	ויתור על כלליות: פי 10–20 לפי הטענות טענת החברה	גדול ומפוצל; השוק המקומי בסין היבשתית גדול בפני עצמו	ארכיטקטורות SRAM ב־N6/N7–N4 ניתנות לניהול	נעילה למודלים ולארכיטקטורה	זהו השוק העיקרי	שווה לבנות: להחתים את הלקוח, ואז לבחור את הנעילה

סוג שבב	מאין בא k	ביקוש וקופה	NRE וספים	סיכון כללים או נעילה	שוק שני	הכרעה
מאיץ ZK	פי 10–100 מ־GPU לפי הטענות טענת החברה	קטן, מונע בעיקר בתגמולי טוקנים	N7/N6 ואז N5; FPGA קודם	מערכות ההוכחה משתנות מהר	מוגבל מחוץ ל־ZK	להמתין ולגדר: מנועי MSM ו־NTT ניתנים לתכנות
מאיץ FHE	פי 5,547–1,074 מ־Xeon (Heracles)	עדיין לא נוצר; מעוגן בארגונים ובממשלות	זקוק ל־HBM; מבנה עלויות קרוב לשבבי אימון	התקנים והאקוסיסטם לא התגבשו	חישוב מוצפן, AI פרטי	להתמקם מוקדם, לאמת בנכסים קלים; חלון 2027–2029

איור 07 חמישה סוגי שבבים, חמש תשובות

מאין בא יתרון העלות k, לאיזה שוק הסיליקון יכול להימכר אם הכללים ישתנו, והמסקנה

שבב כרייה PoUW בסגנון Pearl

לא עכשיו

k: ללא HBM ויחידות כלליות: פי 3–10 לפעולה (הסקה)

שוק שני: הסקת AI, אם GEMM כללי ב־INT8/FP8

להמר על שבב הסקה שיכול לכרות, לא על רכיב INT7 ייעודי

שבב כרייה SHA-256 לביטקוין

להתרחק

k: ≤1 מול השחקנים הוותיקים

שוק שני: כמעט אין

רק עם חשמל עצמי, תצורת נישה, או כ־IP

שבב לאימון AI

IP, לא שבבים

k: פי 1.1–1.8 באותו דיוק

שוק שני: —

המארז וה־HBM מכריעים; לבנות IP, צ׳יפלטים, שירותים

שבב להסקת AI

לבנות

k: ויתור על כלליות: פי 10–20 לפי הטענות

שוק שני: זהו השוק העיקרי

לחתום עם הלקוח, ואז לבחור את הנעילה

מאיץ הוכחות ZK

להמתין ולגדר

k: פי 10–100 מ־GPU לפי הטענות

שוק שני: מוגבל מחוץ ל־ZK

מנועי MSM ו־NTT ניתנים לתכנות; FPGA קודם

מאיץ FHE

להתמקם מוקדם

k: פי 5,547–1,074 מ־Xeon (Heracles)

שוק שני: חישוב מוצפן, AI פרטי

חלון 2027–2029; לאמת בנכסים קלים

מקור: סינתזה של המחבר לפרקים 7 ו־8; נתוני חברות מסומנים כמו בדוח.



09

צמתי תהליך

צמתי תהליך: לתת לעלות הגדולה ביותר לאורך החיים לבחור

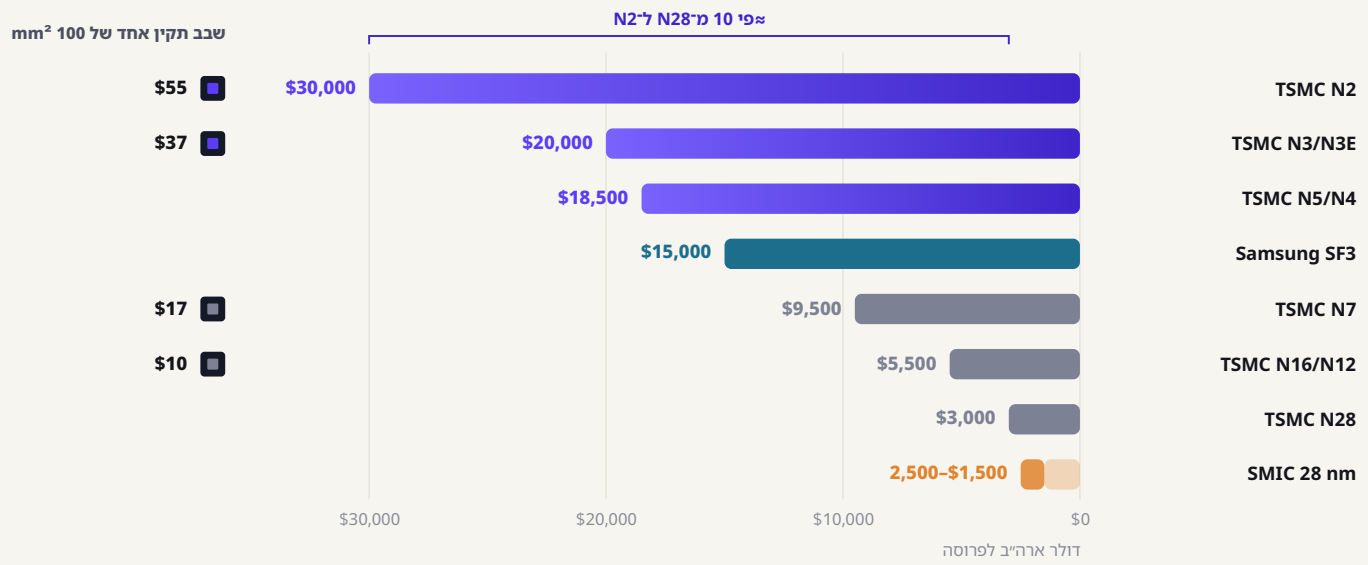
סולם העלויות: פרוסות בחזית הטכנולוגיה עולות פי עשרה מפרוסות בצמתים בוגרים

נתחיל בפנקס החשבונות. פרוסה בחזית הטכנולוגיה עולה כיום בערך פי עשרה מפרוסה בצומת בוגר: TSMC N2 כ־30,000 דולר, N3 כ־20,000 דולר, N7 כ־9,500 דולר ו־N28 כ־3,000 דולר; Samsung SF3 כ־15,000 דולר, כ־25% מתחת ל־N3; ולפי הדיווחים עלות הייצור של SMIC בדרגת 7 nm גבוהה ב־40% עד 50% מזו של TSMC. לפי הדיווחים TSMC תעלה מחירים ב־3% עד 6% נוספים מינואר 2027, בעיקר ל־2 nm ול־3 nm. עלויות המסכות והתכנון שונות עוד יותר, והמקורות חלוקים, ולכן הטבלה מציגה את שניהם.

צומת	מחיר פרוסה (אומדן)	סט מסכות (ערך מודל / טווח מנחה)	תקציב תכנון (AnySilicon)	MPW (לפרויקט)
TSMC N2	כ־30,000 דולר	—	3/2 nm: 400–60 מיליון דולר ויותר	—
TSMC N3/N3E	כ־20,000 דולר (18,000–27,000 דולר)	15 מיליון דולר / 10–20 מיליון דולר	כנ״ל	—
Samsung SF3	כ־15,000 דולר (13,000–17,000 דולר)	12 מיליון דולר	—	—
TSMC N5/N4	כ־18,500 דולר (16,000–21,300 דולר, ביטחון נמוך)	6.5 מיליון דולר / 10–20 מיליון דולר	5/4 nm: 200–30 מיליון דולר	—
TSMC N7	כ־9,500 דולר (8,100–10,900 דולר)	3.5 מיליון דולר / 5–10 מיליון דולר	7/6 nm: 100–15 מיליון דולר	מסחרי 100,000–500,000 דולר
TSMC N16/N12	כ־5,500 דולר	1.8 מיליון דולר	16/12 nm: 40–6 מיליון דולר	אקדמי כ־35,000 דולר ל־mm^2
TSMC N28	כ־3,000 דולר	1 מיליון דולר / 1–3 מיליון דולר	28/22 nm: 20–3 מיליון דולר	מסחרי 50,000–150,000 דולר; אקדמי כ־12,000 דולר ל־mm^2
SMIC 28 nm	1,500–2,500 דולר	—	—	—
SMIC בדרגת 7 nm (N+2/N+3)	40–50% מעל TSMC	—	—	—

איור 08 **פרוסה בחזית הטכנולוגיה עולה פי עשרה מפרוסה בצומת בוגר**

מחיר משוער לפרוסה של 300 mm לפי צומת, ועלות שבב תקין אחד של 100 mm² בתפוקה של 85%



מקור: תמחור פרוסות של Silicon Analysts (אומדנים, N5 בביטחון נמוך), Wccftech (SMIC); עלויות השבבים הן חישוב של המחבר עם תפוקה מונחת.

אחר כך הפיזיקה. המדידות הקלאסיות של Horowitz מראות שהוראה ניתנת לתכנות עולה כ־70 pJ בעוד שהאריתמטיקה עצמה עולה כמה פיקו־ג׳אול: **האריתמטיקה זולה והתקורה יקרה**, וזו הסיבה שמערכים סיסטוליים ושבבים ייעודיים קיימים. הקטנת ממד המאפיין בחצי מקטינה בקירוב בחצי את האנרגיה לפעולה, אבל מתח האספקה נתקע סביב 0.9–1.1 V. פרוסת N2 עולה 50% עד 67% יותר מ־N3, כך שאם הצפיפות אינה עולה ביותר מבערך פי 1.5, העלות לטרנזיסטור עולה במקום לרדת **אומדן**: הצמתים המובילים צריכים להצדיק את עצמם דרך יעילות וביצועים. עבור die של 100 mm² בתפוקה של 85%, העלות של כל die תקין היא כ־55 דולר ב־N2, 37 דולר ב־N3, 29–34 דולר ב־N5, 17 דולר ב־N7 ו־10 דולר ב־N16 [אומדן; התפוקה היא הנחה].

בחירת צומת לפי סוג שבב וצוות

הכלל שלי הוא: **לתת לסעיף הגדול ביותר בעלות השבב לאורך חייו לבחור את הצומת.** בשבבי חישוב שפועלים בעומס מלא מסביב לשעון (כרייה, ZK), החשמל שולט בעלות לאורך החיים, ולכן ברגע שהאלגוריתם יציב הם צריכים לעבור ל־N5/N4 ול־N3/SF3; כל עוד האלגוריתם לא התייצב או שהכסף חסר, יש לאמת קודם על MPW בצומת בוגר. בשבבי אימון צוואר הבקבוק הוא המארז והזיכרון (HBM ו־CoWoS הם כ־63% מרשימת החומרים של H100), ולכן הצומת משני. בשבבי הסקה ההחלטה הראשונה היא ארכיטקטורת הזיכרון: תכנונים שמתמקדים ב־SRAM יכולים להישאר ב־N6/N7 עד N4, בעוד שתכנוני HBM יורשים את מבנה העלויות של שבבי אימון. FHE מוגבל ברוחב הפס של הזיכרון ואינו יכול להימנע מ־HBM.

סוג שבב	סטארט־אפ או צוות קטן	חברה בהיקף	מסלול סין היבשתית	נקודות מפתח
שבב כרייה PoUW של כפל מטריצות (בתוך SRAM, ללא HBM)	אב טיפוס על MPW ב־28 nm או ב־16/12 nm; נפח ראשון ב־N7/N6; N5/N4 אחרי שהאלגוריתם וההכנסה מתייצבים; dies של 50–150 mm² במארזי flip-chip סטנדרטיים	N3/N3E או Samsung SF3	SMIC N+1/N+2 ישים (אין צורך ב־HBM או ב־CoWoS), בעלות גבוהה בכ־40–50% ל־die תקין	האנרגיה שולטת; כללי הפרוטוקול הם הסיכון הגדול ביותר
שבב כרייה SHA-256	לא לרדוף אחרי החזית; תצורות נישה יכולות להשתמש בצמתים בוגרים	דרגת 3 nm; השחקנים הוותיקים עוברים מ־TSMC 5 nm ל־2 nm ומ־Samsung SF3 ל־2 nm (מקורות משניים)	קשה להשתוות ליעילות הצומת המוביל	מחיר החשמל ומחזור החציה קובעים את השוליים

סוג שבב	סטארט־אפ או צוות קטן	חברה בהיקף	מסלול סין היבשתית	נקודות מפתח
שבב לאימון AI	לא שבב עצמאי: IP, צ׳יפלטים, שירותי תכנון	N3P ואז N2/A16; 2–8 צ׳יפלטים; להבטיח CoWoS ו־HBM מוקדם	HBM עם SMIC N+2/N+3 מקומי (המודל של Huawei); FP8 לשבב בערך פי 0.2–0.5 מזה של NVIDIA, שמפוצה בהיקף המערכת	המארז וה־HBM הם צוואר הבקבוק
שבב להסקת AI	ממוקד SRAM או צרוב בחומרה, N6/N7 או N4/N5/SF4; INT8, FP8 ו־FP4 בלבד	N3/N2 עם HBM (הקשר ארוך, תערובת מומחים), או SRAM בריבוי צ׳יפלטים	LPDDR עם SMIC N+2 או GDDR; זיכרון בדרגת HBM הוא נקודת החנק	מידת הנעילה הארכיטקטונית היא ההחלטה המרכזית
מאיץ ZK	להתחיל על FPGA או MPW ב־16/12 nm; אחר כך N7/N6 ו־N5	N5/N3	SMIC N+1/N+2 ישים	במספרים שלמים ודטרמיניסטי; קל לאימות
מאיץ FHE	FPGA (מול בסיס ההשוואה בקוד פתוח) או MPW ב־16/12 nm; נפח מחייב HBM	N5/N3 עם HBM	מוגבל בגישה ל־HBM	רוחב הפס של הזיכרון, לא האריתמטיקה, קובע את המהירות

מתקדם או בוגר, ארבעה סוגי פאונדרי, וצ׳יפלטים

אפשרות	חוזקות עיקריות	חולשות עיקריות	מתאים במיוחד ל־
צמתים מובילים (N3/N2, SF3/SF2, Intel 18A)	כל צומת מלא מקטין בקירוב בחצי את האנרגיה לפעולה; SRAM צפוף	פרוסות פי 3–10, מסכות פי 5–15 ותקציבי תכנון פי 10–20 מאלה של צמתים בוגרים; זמני אספקה מעל 50 שבועות; העלות לטרנזיסטור ב־N2 עשויה לעלות	שבבי חישוב עם אלגוריתמים יציבים שפועלים מסביב לשעון; שבבי אימון
צמתים בוגרים ובטווח הביניים (N7/N6, N16/12, N28)	מסכות בסדר גודל של מיליון דולר, MPW זול, איטרציה מהירה, פאונדרים רבים (TSMC, UMC, GlobalFoundries, SMIC)	יעילות לא תחרותית בעומס מלא מסביב לשעון	אבות טיפוס, שווקי נישה, הסקה בקצה
TSMC	התפוקה והאקוסיסטם הטובים ביותר	המחירים הגבוהים ביותר; קיבולת N2/N3 צפופה; תורים ארוכים ל־CoWoS	צוותים שרודפים אחרי חזית היעילות עם התחייבויות נפח
Samsung Foundry	פרוסות זולות בכ־20–25% (SF3 כ־15,000 דולר מול כ־20,000 דולר ל־N3); בשימוש בדור השני של Groq ואחרים	עסקי הפאונדרי מפסידים כסף, ומכוונים לאיזון עד סוף 2027	שבבי חישוב רגישים לעלות
Intel Foundry	שבבים מורכבים כמו Heracles הוכחו על Intel 3	אין תמחור חיצוני פומבי או נתוני לקוחות ל־18A; עסקי הפאונדרי מפסידים כסף	להעריך את האקוסיסטם והתמיכה לגופו של מקרה
פאונדרים בסין היבשתית (SMIC ואחרים)	קרבה לשוק מקומי גדול; פרוסות 28 nm ב־1,500–2,500 דולר	קיבולת בדרגת 7 nm של כ־45,000–60,000 פרוסות בחודש, שנספגת ברובה בידי Huawei; תפוקות בערך שליש מאלה של TSMC; עלויות גבוהות ב־40–50% בלבד DUV	שבבי הסקה, כרייה ו־ZK לשוק המקומי
צ׳יפלטים	מתגברים על מגבלות הרטיקל והתפוקה; סטנדרט במערכות HBM (MI355X, Ironwood, Trainium3)	עלות מארז גבוהה יותר ותלות ב־CoWoS; בדיקות known-good-die מוסיפות עלות	שבבי אימון ושבבי הסקה גדולים; לא נחוצים לשבבי כרייה או ZK קטנים

פרק 10

10 ממודל לסיליקון: תקציב מדורג לשבב ראשון

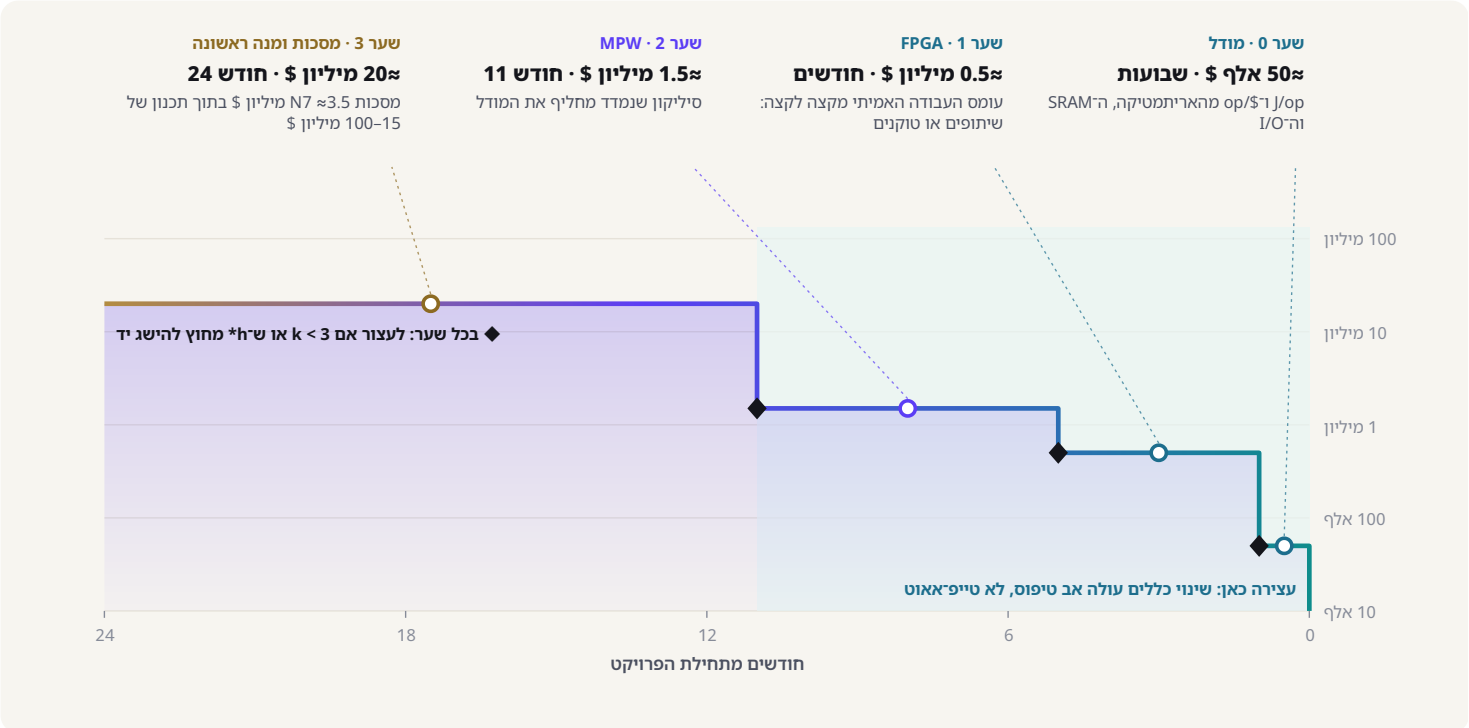
תקציב מדורג

לקנות ראיות לפני מסכות

שבב בתכנון מלא הוא סדרה של הימורים, כל אחד גדול מקודמו, והמשמעת היא לקנות ראיות לפני שקונים מסכות. אני משתמש בארבעה שערים ניתוח. שער 0 הוא מודל: מחקר ארכיטקטורה שמעריך J/op ו־$/op מתוך האריתמטיקה, ה־SRAM וה־I/O, מול חומרת הייחוס של פרק 7; הוא עולה שבועות ועשרות אלפי דולרים. שער 1 הוא אב טיפוס על FPGA שמריץ את עומס העבודה האמיתי מקצה לקצה, שיתופי כרייה או טוקני הסקה, ומוכיח את נתיב הנתונים ואת האינטגרציה בפרוטוקול; הוא עולה כמה חודשים וכמה מאות אלפי דולרים. שער 2 הוא פרוסה מרובת פרויקטים: משבצת מסחרית עולה כ־150,000–50,000 דולר ב־N28 ו־500,000–100,000 דולר ב־N7, וסיליקון שנמדד מחליף את אומדני האנרגיה של המודל. שער 3 הוא סט המסכות המלא ומנת הייצור הראשונה: מסכות של כ־1 מיליון דולר ב־N28, 3.5 מיליון דולר ב־N7 ו־15 מיליון דולר ב־N3, בתוך תקציב תכנון של 100–15 מיליון דולר ב־7/6 nm. כל שער צריך מספר שעוצר את הפרויקט, כמו k מתחת ל־3 מול החומרה השולית, או קצב גיבוב להחזר h* שהחלון אינו יכול להגיע אליו.

איור 09 לקנות ראיות לפני מסכות: תקציב מדורג לשבב ראשון

הוצאה מצטברת בדולר ארה״ב של פרויקט N7/N6 לאורך ארבעה שערים, סקאלה לוגריתמית; לכל שער מספר שעוצר את הפרויקט



מקור: מחירי פרוסות, MPW ומסכות כמו בפרק 9 (Silicon Analysts, AnySilicon, ערכי מודל); התזמון וההוצאה של השערים הם הנחות להמחשה של המחבר.

המדרגות תלולות. בפרויקט N7/N6, שלושת השערים הראשונים יחד עולים כ־1.5 מיליון דולר ונמשכים כאחד־עשר חודשים; הרביעי עולה יותר מפי עשרה אומדן. לכן הסדר חשוב יותר מהצומת: אם הפרוטוקול משנה את האריתמטיקה שלו, או שהקונה נסוג, צוות שעצר בשער 2 מפסיד את מחירו של אב טיפוס, לא של טייפ־אאוט.

כמה עולה כל שבב, ומה חייב להתקיים

פרויקט	שער 2 (אב טיפוס)	שער 3 (צומת ותקציב תכנון)	תקן אחד של die 100 mm²	מה חייב להתקיים כדי להחזיר את ההשקעה
מאיץ שמכוון קודם כול להסקה ויכול לכרות	MPW ב־16/12 nm או ב־N7, 0.1–0.5 מיליון דולר	N6/N7, כ־15–30 מיליון דולר אומדן	כ־17 דולר ב־N7, לפני מארז וזיכרון	לקוח הסקה חתום נושא את רוב ה־NRE; הכנסת הכרייה היא בונוס, לא תרחיש הבסיס
כורה INT7 ייעודי ל־Pearl (לא מומלץ כעת)	MPW ב־28 nm, 0.05–0.15 מיליון דולר	N7/N6, כ־20 מיליון דולר	כ־17 דולר ב־N7	כ־4.9 EH/s לשנה שלמה עם p = 1, או 9.8 EH/s עם p = 0.5 (פרק 7)
מאיץ ZK	FPGA, ואז MPW ב־16/12 nm	N7/N6, ואז N5	כ־10 דולר ב־N16, 17 דולר ב־N7	מערכות ההוכחה מתייצבות על קרנלי MSM ו־NTT שתכנון קבוע יכול לשרת
מאיץ FHE	FPGA מול בסיס ההשוואה בקוד פתוח	N5/N3 עם HBM, 30 מיליון דולר ויותר	נשלט בידי ה־HBM והמארז	קונה ל־AI מוצפן בנפח, שעדיין אינו קיים

מכך נובעות שתי נקודות מסחריות. ראשית, היעילות הזולה ביותר היא ארכיטקטונית: החזקת האופרנדים ב־SRAM והשמטת ה־HBM וה־CoWoS מסירות את רוב העלות של GPU (כ־63% מרשימת החומרים של H100) עוד לפני שנבחר צומת כלשהו. שנית, הזמן הוא עלות: צוות חדש צריך כ־18 עד 24 חודשים מההתחלה ועד לייצור בנפח, וזמני האספקה בחזית עולים על 50 שבועות, כך שפרוטוקול שמשנה את האריתמטיקה שלו כל כמה חודשים יכול להפוך שבב למיושן לפני ששער 3 מחזיר את עצמו.

שלושה כללים לצוות קטן

יחד, המספרים האלה נותנים שלושה כללים ניתוח. ראשית, **להישאר בשער 1 או 2 עד שיש קונה בשמו או גרסת פרוטוקול מוקפאת**; לא מחיר הטוקן ולא קצב הגיבוב מחליפים אף אחד מהם. שנית, **לקנות k קודם בארכיטקטורה ורק אחר כך בצומת**: מערך בתוך SRAM ב־N7 יכול לנצח רכיב כללי ב־N3 בעומס עבודה אחד, בשבריר מעלות המסכות. שלישית, **לרשום את קצב הגיבוב או הנפח להחזר, יחד עם p, לפני הטייפ־אאוט**, ולתמחר תשלומים לפי אבני דרך מול גרסת פרוטוקול קבועה, כך שהלקוח, ולא בית התכנון, נושא את סיכון הכללים.



פרק 11

הזדמנויות

11 היכן ההזדמנויות, ומה הייתי עושה

שבבים מותאמים: הביקוש בא ברובו המכריע מ־AI

ב־2026 הביקוש לשבבים מותאמים בא ברובו המכריע מ־AI. הכנסות Broadcom ממוליכים למחצה ל־AI ברבעון השלישי של שנת הכספים 2026 שלה היו 16.7 מיליארד דולר, עלייה של 221% משנה לשנה, כ־58 מיליארד דולר לשנה, 73% מהן מ־XPU לשישה לקוחות (מתוך תמליל אוטומטי של שיחת התוצאות). הכנסות Global Unichip ברבעון השני עלו ב־128% משנה לשנה, כשהכנסות הייצור מפתח־ליד (turnkey) עלו ב־188% ושולי הרווח הגולמי היו 21.5%. חלוקת ההכנסות של Alchip לפי צומת ברבעון הראשון הייתה 21% ב־3/2 nm, 63% ב־7/5 nm ו־16% ב־12 nm ומעלה. ההזמנות החדשות של VeriSilicon ל־2026 הגיעו עד סוף אפריל ל־8.24 מיליארד יואן (כ־1.16 מיליארד דולר), 91.37% מהן לחישוב AI. לשוק שלוש שכבות: בראש, פיתוח משותף רב־שנתי בשווי מיליארדים; באמצע, פרויקטים מ־7 nm עד 2 nm עם NRE של כמה מיליוני דולרים ועד יותר ממאה מיליון דולר (5–30 מיליון דולר בצמתים בטווח הביניים, 30–100 מיליון דולר ויותר בחזית); וזנב ארוך של ASIC בצמתים בוגרים עם NRE של 0.5–5 מיליון דולר. לקוחות קריפטו וכרייה הם נישה קטנה ומחזורית, אבל הם ה"לקוח הראשון" המהיר ביותר וזה שפחות מכול דורש אישורים, שמשלם כיום על כוח חישוב.

היכולות של צוות שבבי כרייה שעוברות בצורה הטובה ביותר, פעולה במתח נמוך קרוב לסף, מערכי חישוב משוכפלים בהיקף עצום, אספקת הספק וקירור בדרגת קילוואטים, וטייפ־אאוטים מהירים בצמתים מובילים, הן בדיוק מה ששבבי הסקה ומאיצי ZK ו־FHE צריכים. שוקי ההון מתמחרים זאת: Velaura AI (לשעבר Auradine) גייסה 110 מיליון דולר באוגוסט לפי שווי של יותר מ־1 מיליארד דולר. הדוגמה הנגדית היא יחידת שבבי ה־AI לקצה של Canaan, שהכנסותיה היו 0.9 מיליון דולר בשנת הכספים 2024 בעוד שצרכה 15% מההוצאות התפעוליות, ונסגרה ביוני 2025. ההבדל טמון בערוץ עצמאי לשוק, לא בשימוש חוזר בערוצי הכרייה.

פיתוח AI: שעות עבודה הן סחורה; הערך נמצא בחיבורים

בקצה הנמוך, שירותי פיתוח AI הפכו לסחורה: התעריפים הטיפוסיים ב־Clutch הם 25–49 דולר לשעה, ו־50–99 דולר בארצות הברית, בקנדה ובשווקים דומים; צוות פיתוח בלוקצ׳יין עולה כ־15,000 דולר ועד יותר מ־90,000 דולר בחודש. למכירת שעות אין חפיר. הערך נמצא בחיבורים בין התחומים: קרנלים ומהדרים לשבבים מותאמים, כרייה שמשובצת בשירות ההסקה (התוסף של Pearl ל־vLLM הוא דוגמה אחת), הסקה ניתנת לאימות שמשלבת TEE עם בדיקות מדגמיות ב־ZK, ותשלומים וזהות לסוכנים.

מערכות AI × קריפטו: להחתים את קונה המחשוב לפני הטוקן

לכל מי שמתכנן שרשרת או רשת מחשוב שנבנתה מלכתחילה ל־AI, ארגז הכלים של האימות ב־2026 כבר קיבל צורה. לפי סדר העלות ליחידת עבודה, מהנמוכה לגבוהה: דגימה סטטיסטית (הוכחת הדגימה של Hyperbolic), האצלה בפיקוח שופט (Verde של Gensyn, שצריכה מאמת ישר אחד בלבד אבל דורשת ביצוע מדויק ברמת הביט), TEE (מהירות קרובה למקורית, אבל אמון ביצרן השבב), יתירות מלאה ו־zkML כללי. התכנון המעשי הוא בשכבות: TEE או דגימה להשהיה נמוכה, משחקי מחלוקת או בדיקות מדגמיות ב־ZK לסופיות; מאמר מדצמבר 2025 הדמה מחשוב חסוי על H100 עם הוכחות הונאה ובדיקות מדגמיות אקראיות ב־ZK ב־99% מהתפוקה המרוכזת ובעלות שולית של 0.07 דולר לשאילתה. פילוסופיית החומרה מכריעה את האבטחה ואת ההפצה: PoUW מבית GPU מוגן בתחילה על ידי GPU בענן שאפשר לשכור, כך שעלות ההתקפה קשורה למחירי ההשכרה, ו־Qubic מראה שהוכחת עבודה על חומרה כללית יכולה להילכד בסיטונאות בידי תמריצים חיצוניים.

העלות והזמן אינם החסם. Pearl עברה ממאמר (אפריל 2025) לרשת ראשית בתוך כ־12 חודשים, כפיצול של btcd, עם צומת מלא, ארנק, לקוח קל וכורה כבר בהשקה. רוב ביקורות ה־DeFi עולות 25,000–100,000 דולר, ביקורות ארגוניות או רב־שרשרתיות 150,000–250,000 דולר ויותר, מעגלי ZK עוד 80% עד 120%, ופרויקטים עם ערך אמיתי בסיכון מתקצבים 150,000–500,000 דולר בשנה לאבטחה. החסמים האמיתיים הם הפצה ונזילות. מייסדים מדווחים ש־Binance ביקשה 0% עד 3.5%, ולפי הטענות 8%, מהיצע הטוקנים תמורת רישום (Binance מכחישה שהיא מרוויחה מרישומים); כל ששת הפרויקטים ש־Cointelegraph בחנה ושהשתמשו בעסקאות עשיית שוק של "הלוואת טוקנים ועוד אופציית רכש" ירדו, ועושה השוק של Movement מכר טוקנים בכ־38 מיליון דולר ביום הרישום. החלופות הן נזילות על השרשרת, כמו ב־Hyperliquid, שעברה 10 מיליארד דולר של שווי שוק בלי רישום ב־Binance ושבשווקי החוזים הנצחיים HIP-3 שלה הפוזיציות הפתוחות עברו 3.2 מיליארד דולר ביוני, ומכירות פומביות דרך בורסות גדולות, כמו כש־Monad גייסה 269 מיליון דולר דרך Coinbase. שווי גבוה בדילול מלא עם היצע צף נמוך היה המנבא הברור ביותר לירידה אחרי ההשקה ב־2025–2026.

תכנון הטוקן הוא חלק מתכנון המערכת. ההנפקה צריכה לבוא בעקבות ביקוש בתשלום ולא להקדים אותו: הרשתות בדוח הזה עם מכפילי הסבסוד הבריאים ביותר קושרות את התגמולים לעבודה שקונה חיצוני משלם עליה, ואלה עם הגרועים ביותר משלמות על קצב גיבוב או על סטייק שאין מאחוריהם דבר. כלל מעשי הוא לפרסם, לפני ההשקה, את מכפיל הסבסוד שהרשת מכוונת אליו בשנתה הראשונה, השנייה והשלישית, ולקצץ בהנפקה כשהיעד מוחמץ ניתוח.

כיוון	מי קונה, ומוכנות לשלם	חלון	יכולות מפתח	סיכון עיקרי
מאיץ שמכוון קודם כול להסקה ויכול לכרות	neoclouds, פריסות פרטיות, רשתות PoUW	2027–2028 (תלוי בפיצולי FP8 ו־FP4)	GEMM ב־INT8, FP8 ו־FP4; ארכיטקטורת SRAM; אינטגרציה בפרוטוקול	קבלה לפרוטוקול, נעילה ארכיטקטונית
שבב הסקה מבוסס SRAM או צרוב בחומרה	neoclouds, מרכזי נתונים בסין היבשתית, קצה	עכשיו	מהדרים והתאמת מודלים; תכנון ב־N6–N4	המודלים משתנים מהר מדי

כיוון	מי קונה, ומוכנות לשלם	חלון	יכולות מפתח	סיכון עיקרי
שבב כרייה (נישה או אינטגרציה אנכית)	כורים עם חשמל משלהם; כורים ביתיים וכורים לחימום	עכשיו, בגבולות החציה ומחיר הגיבוב	מערכים במתח נמוך, יעילות, שרשרת אספקה בנפח	מלחמות מחירים של השחקנים הוותיקים, מחירי חשמל
האצת הוכחות ZK	רשתות מוכיחים, rollups	להמתין עד שמערכות ההוכחה יתכנסו	MSM ו־NTT ניתנים לתכנות	הביקוש תלוי בתגמולי טוקנים
האצת FHE	AI מוצפן לארגונים ולממשלות; שרשראות פרטיות	2029–2027	NTT, מערכות HBM, הנדסה קריפטוגרפית	הביקוש עוד לא נוצר; מחזורי מכירה ארוכים
הקשחת TEE והסקה ניתנת לאימות	רשתות הסקה חסויה, DePIN	עכשיו	הצפנת זיכרון עם שלמות, שורשי אישור עצמאיים, בדיקות מדגמיות ב־ZK	תלות באקוסיסטמים של היצרנים
שרשרת או רשת מחשוב שנבנתה מלכתחילה ל־AI	קוני הסקה, כורים, מפתחים	עכשיו	תכנון אימות, תכנון הנפקה, הפצה	ירידה אחרי ההשקה, נזילות הטוקן

שלושה מסלולים שהייתי הולך בהם, לכל אחד צעד ראשון

אילו הייתי צריך לבחור, הייתי הולך בשלושה מסלולים לכל היותר, ולכל אחד צעד ראשון קטן שבודק אותו לפני שמתחייבים בכסף ניתוח.

- **מסלול א: מאיץ שמכוון קודם כול להסקה ויכול גם לכרות.** צעד ראשון: מחקר ארכיטקטורה של ארבעה שבועות שממדל J/op ו־$/op של מערך GEMM ב־INT8/FP8 בתוך SRAM מול RTX 5090 ו־H100, גם על העבודה של Pearl וגם על הסקת טרנספורמרים, ואחריו אימות על FPGA או MPW בצומת בוגר עם NRE בסדר גודל של מיליון דולר.
- **מסלול ב: שירותי סיליקון מותאם, עם הקריפטו כלקוח הראשון וה־AI כשוק השני.** צעד ראשון: להפוך את מודל ה־NRE ונקודת האיזון של הדוח הזה למחקר היתכנות בתשלום לפרויקטים של שבבי חישוב, עם תשלומים לפי אבני דרך וקבלה מול גרסת פרוטוקול קבועה.
- **מסלול ג: רשת שנבנתה מלכתחילה ל־AI, שמחתימה את קונה המחשוב שלה לפני הטוקן שלה.** צעד ראשון: מסמך תכנון על אימות בשכבות (TEE ועוד דגימה ועוד בדיקות מדגמיות ב־ZK) ועל אריתמטיקה קנונית ניטרלית לספק, שנכתב עם קונה הסקה אחד בשמו לפני כל תכנון של טוקן.



פרק 12

12

ציוני דרך

על מה לעקוב, נובמבר 2026 עד 2027

מתי	על מה לעקוב	למה זה חשוב
בכל עת	גובה ההפעלה של FP8 ב־Pearl	מכריע אם חומרה של מספרים שלמים וכרטיסים שאינם של NVIDIA יישארו מתקבלים, והאם השרשרת תתפצל
חודשי	חלק הבלוקים של Pearl שנושאים עבודת AI	המבחן הטוב ביותר לשאלה אם ה"מועיל" הופך לאמת
חודשי	האם ספקי הסקה מעבירים את הכנסת הכרייה ללקוחות	ביקוש בתשלום שפוגש הנפקה: המודל ש־Pearl צריכה
רבעוני	ההכנסה הניתנת לאימות של Bittensor מול ההנפקה שלה	האם רשת העבודה המועילה הגדולה ביותר סוגרת את פער הסבסוד שלה
רבעוני	החלק של תזרימי x402 שסוכנים יזמו בפועל	האם תשלומי סוכנים עוברים מאיתות למסחר
ינואר 2027	העלאת המחירים של TSMC ב־3% עד 6% לפי הדיווחים, בעיקר ל־2 nm ול־3 nm	העלות של טייפ־אאוטים בצמתים מובילים
הרבעון הראשון של 2027	FP4 ב־Pearl, כפי שדווח מה־AMA של המייסדים	מעבר נוסף של העבודה לעבר הפורמטים החדשים ביותר של NVIDIA
סוף 2027	היעד של Samsung Foundry להגיע לאיזון	האם מקור שני זול יותר נשאר בריא
אפריל 2028	החציה הבאה של ביטקוין	מקטינה בחצי את הקופה של כורי SHA-256



פרק 13

13 שאלות שלא שאלתם אבל הייתם צריכים לשאול

שאלות שכדאי לשאול

השאלות להלן הן אלה שהייתי שואל לפני שמתחייבים בכסף או בסיליקון לפרויקט כלשהו של בינה מלאכותית × קריפטו.

מי שולט בפונקציית העבודה? אם צוות קטן יכול לשנות את האריתמטיקה בגובה מקודד בקוד, חיי החומרה שלכם הם החלטה שלו. בקשו התחייבות כתובה ליציבות, או תכננו חומרה עם שוק שני.

מהו מכפיל הסבסוד? חלקו את ההנפקה (או את שווי השוק) בהכנסה שמשולמת מחוץ לרשת. מכפיל מעל 10 אומר שהרשת קונה פעילות, לא מוכרת שירות.

כמה עולה השער הבא, ואיזה מספר עוצר אתכם? פרוסה מרובת פרויקטים עולה עשרות או מאות אלפי דולרים; סט מסכות עולה מיליונים. החליטו על מספר העצירה לפני שאתם מוציאים כסף.

מהו קצב הגיבוב להחזר, עם p? חלקו את ה־NRE ב־$p \times T \times (1 - 1/k) \times R$. אם התשובה היא חלק גדול מהרשת, השבב הוא הימור על הכללים, לא על הסיליקון.

מהו השוק השני של השבב? אם הפרוטוקול מתפצל או שהטוקן נופל, לאן עוד הסיליקון יכול להימכר? ה־GPU שרדו את ה־Merge של Ethereum משום שהיה להם שוק כזה; ASIC של אלגוריתם יחיד לא.

האם הערך נוחת בטוקן או בהון? ב־2026 חברות הבינה המלאכותית × קריפטו הרווחיות הן ברובן חברות הון עם הפצה בקריפטו; החליטו מה מהשניים אתם בונים לפני שאתם מגייסים כסף.

מיהו הלקוח המשלם הראשון, בשמו? כל מקרה בר־קיימא בדוח הזה, מ־Grass ועד תתי־רשתות המחשוב של Bittensor, התחיל בקונה בשמו מחוץ לרשת.

אלה תחומים שהייתי רוצה לחקור יחד עם אחרים, דרך חילופי ידע טכני ושיתוף פעולה טכני: שבבים מותאמים לכרייה, לאימון ולהסקה של AI, ומאיצים; פיתוח AI; ומערכות AI × קריפטו מהשרשרת, דרך הצומת ועד השבב. אפשר לפנות אליי בכתובת i@tanshuai.com.

מגבלות הדוח

זהו מחקר שולחני שמבוסס על מסמכים פומביים עד 13 בנובמבר 2026. לא ערכתי ראיונות ראשוניים ולא ביצעתי מדידות בלתי תלויות למהדורה הזאת. כמה מספרים מרכזיים הם טענות של חברות או של יצרנים, כמו ההכנסות של Aethir, ההכנסות של Grass, הביצועים של Etched ושל Taalas והיעילות של Cysic; הם מסומנים ככאלה. מדדי השרשרת של Pearl באים מצומת מלא בלתי תלוי ומבריכות כרייה ומשתנים משעה לשעה; השתמשתי בקריאות סביב 07:13 UTC ב־13 בנובמבר ואני מציין טווחים מאותו יום. עלויות הפרוסות, המסכות והתכנון הן אומדנים של צד שלישי וערכי מודל שאינם מסכימים זה עם זה; אני מציג טווחים. חשבון נקודת האיזון שלי ל־ASIC והתקציב המדורג משתמשים בהנחות להמחשה (k, T, p, NRE ועלויות השערים) שהקורא צריך להחליף בהנחות משלו.



על המחבר

תאן שואי הוא טכנולוג ומייסד עם יותר מעשרים שנות עבודה בשבבים, בחומרה ובתוכנת מערכת, ו־20 פטנטים רשומים. מאחוריו שנים רבות של מו״פ וניהול בתחום שבבי fabless, כולל תכנון נתיבי נתונים בתכנון מלא (full-custom) של ASIC לשבבי כריית מטבעות קריפטוגרפיים ולשבבי מצלמה למולטימדיה, וזו העדשה של הדוח הזה: אנרגיה לפעולה, עלות לפרוסה, מה פרוטוקול מאפשר לקבע בסיליקון, ומי משלם. עבודתו הובילה אותו דרך עמק הסיליקון, ניו יורק, שיקגו, טייוואן, אזור המפרץ הגדול (שנזן והונג קונג), וייטנאם ומלזיה. כיום הוא עובד כחוקר ומפתח בינה מלאכותית.

לדיון טכני: i@tanshuai.com · tanshuai.com

הערות והסתייגות

הדוח הזה הוא מחקר שולחני עצמאי שמבוסס על מסמכים פומביים שהיו זמינים עד 13 בנובמבר 2026. הוא מסופק למטרות מידע ולמידה בלבד ואינו ייעוץ השקעות או ייעוץ הנדסי. הוא עוסק בטכנולוגיה ובעסקים בלבד. מחירים, מדדי שרשרת וכללי פרוטוקול משתנים; אמתו כל דבר שאתם מסתמכים עליו מול המקור הראשוני. נתונים שמסומנים טענת החברה, אומדן או ניתוח יש לקרוא ככאלה. מחירי נכסי קריפטו תנודתיים מאוד.

© 2026 תאן שואי. מותר לשתף את הדוח הזה ללא שינוי ובציון המקור.



מקורות

כל המקורות נצפו בין 10 ל־13 בנובמבר 2026. מקובצים לפי פרק.

Pearl: פרוטוקול, קוד ונתוני שרשרת

- Komargodski ו־Weinstein, "Proofs of Useful Work from Arbitrary Matrix Multiplication", arXiv 2504.09971: https://arxiv.org/abs/2504.09971
- Pearl Research Labs, הספר הלבן של INT: https://pearlresearch.ai/research/int-whitepaper
- קוד המקור של Pearl: בלוק הבראשית: https://github.com/pearl-research-labs/pearl/blob/main/node/chaincfg/genesis.go ; פרמטרי קונצנזוס וגובהי פיצולים: https://github.com/pearl-research-labs/pearl/blob/main/node/chaincfg/params.go ; חישוב קצב הגיבוב: https://github.com/pearl-research-labs/pearl/blob/main/zk-pow/src/api/sanity_checks.rs ; מדריך שדרוג לפיצול הזרע הממולח: https://github.com/pearl-research-labs/pearl/blob/main/docs/salted-seed-fork-upgrade-guide.md ; פרמטרים של ענף fp8-scheme: https://github.com/pearl-research-labs/pearl/blob/fp8-scheme/node/chaincfg/params.go
- הצעות השיפור של Pearl (Pearl Improvement Proposals): https://github.com/pearl-research-labs/pips ; PIP-3, כרייה בנקודה צפה: https://github.com/pearl-research-labs/pips/blob/main/PIPS/pip-0003.md
- prlstats (צומת מלא בלתי תלוי): רשת: https://prlstats.com/api/network ; סדרות: https://prlstats.com/api/series ; דף הבית ומודל הכנסת הכורים: https://prlstats.com/api/home ; היצע שנכרה: https://prlstats.com/api/mined ; בריכות כרייה: https://prlstats.com/api/pools ; מחזיקים: https://prlstats.com/api/holders ; הנחות חומרה ותקורה: https://prlstats.com/api/assumptions
- Kryptex Pool: רשת Pearl: https://pool.kryptex.com/prl/about-coin ; קצב גיבוב: https://pool.kryptex.com/d ; RTX 5090: https://pool.kryptex.com/prl/hashrate ; 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 ; היסטוריית מחירים: https://api.coin gecko.com/api/v3/coins/pearl-2/market_chart?vs_currency=usd&days=365&interval=daily ; חיפוש הסימול "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 ביוני 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, השותפות עם Pearl Research Labs: https://together.ai/blog/together-ai-partners-with-pearl-research-labs ; דף המודל: https://together.ai/model s/gemma-4-31b-it-pearl
- Basu, מפקד של מופעי כרייה של Pearl, arXiv 2606.04819: https://arxiv.org/html/2606.04819v2
- Blockworks, ניוזלטר 0xResearch, 28 באוקטובר 2026: https://blockworks.com/newsletter/0xresearch/issue/post_daf0d74a-2bf4-41e4-86d3-2a1275475a24
- Datawallet, "Pearl Explained": https://www.datawallet.com/crypto/pearl-explained
- Coingabbar על הדיווח של TheMarker על הגיוס של Pearl Research Labs: https://www.coingabbar.com/en/pearl-research-labs-funding-130m-stealth-round-th emarker

רשתות עבודה מועילה והיסטוריה

- Grothe ואחרים, ניתוח אבטחה של Gridcoin, USENIX WOOT 2017: https://www.usenix.org/system/files/conference/woot17/woot17-paper-grothe.pdf
- זיוף הוכחת למידה, arXiv 2208.03567: https://arxiv.org/abs/2208.03567
- hoge.gg, אימות ואמון בהסקה מבוזרת: https://hoge.gg/decentralized-inference-2026-verification-trust/ ; ההכנסות והאוצר של Bittensor: https://hoge.gg/bittens /or-quantum-wallet-revenue-tao-treasury-2026
- Kaspa: קוד המטריצה של kHeavyHash: https://github.com/kaspanet/rusty-kaspa/blob/master/consensus/pow/src/matrix.rs ; השקה: https://coinranking.com/coin/V8GxkwWow+kaspa-kas/about ; ההכרזה של IceRiver על KS1/KS2: https://www.iceriver.io/2023/03/14/1677644580/ ; ההיסטוריה מ־GPU ל־ASIC: https://ca ptainmining.com/blogs/crypto-news/the-evolution-of-kaspa-mining-from-gpus-to-asic-powerhouses ; קצבי גיבוב של GPU: https://coincalculators.io/algorith ms/kheavyhash ; KS7: https://www.asicminervalue.com/miners/bitmain/antminer-ks7-40th ; רשת: https://minerstat.com/coin/kas ; 1 EH/s ב־2024: https://o urcryptotalk.com/news/kaspa-reaches-1-ehs-in-hashrate ; Crescendo: https://ourcryptotalk.com/news/kaspa-crescendo-live
- Aleo: שאלות נפוצות על הרשת הראשית: https://aleo.org/post/aleo-mainnet-faq/ ; הוכחה מבוזרת: 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 על הארגון מחדש ב־51 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 ; כריית Dogecoin: https://hackernoon.com/qubic-is-now-mining-dogecoin-on-its-ai-compute-infrastructure ; CoinGecko: https://www.coingecko.com/en/coins/qubic
- תיעוד הקונצנזוס של Bittensor: https://www.bittensor.com/llms.mdx/docs/internals/consensus/content.md ; FalconX, "State of Bittensor": https://www.falcon x.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 ופיצולי עמידות ל־ASIC: https://lbry.com/news/forking-for-asic-resistance ; הפיצול של Sia: https://hotfix.coindesk.com/markets/2018/11/01/sia-compl etes-hard-fork-as-bricked-miners-form-rival-blockchain ; מחירי Obelisk: https://www.cryptoglobe.com/latest/2018/11/obelisk-hardware-skyrockets-in-sellin /g-price-after-sia-hard-fork ; כורים אחרי ה־Merge: https://hotfix.coindesk.com/business/2022/11/11/ex-ethereum-miners-token-hop-to-stay-alive-after-the-merge

מגזרי הבינה המלאכותית × קריפטו

- DefiLlama: Render: https://defillama.com/protocol/render-network-bme ; Livepeer: https://defillama.com/protocol/livepeer
- Cointelegraph על הכנסות DePIN לפי Messari: https://cointelegraph.com/news/depin-revenues-grow-despite-token-losses
- Aethir, סיכום 2025: https://aethir.com/blog-posts/aethirs-2025-wrap-up-decentralized-gpu-cloud-milestones
- CoinGecko, קטגוריית AI: https://www.coingecko.com/en/categories/artificial-intelligence
- Akash, מחירי השכרה של H100: https://akash.network/the-bid/h100-rental-price-2026-cost-per-hour/ ; דוח הרבעון הראשון של 2026: https://akash.network/blog/akash-network-q1-2026-report
- io.net, עסקת 8 מיליון הדולר והטוקנים היומיים: 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, החוזה של Axe Compute: https://www.datacenterdynamics.com/en/news/axe-compute-secures-15bn-compute-contract-with-unnamed-customer/
- Implicator ; Nous Research, Psyche: https://nousresearch.com/the-next-phase-of-psyche על Pluralis-8B: https://www.alphaxiv.org/abs/2607.13332.md ; INTELLECT-3: https://www.implicator.ai/prime-intellects-intellect-3-open-source-ambition-meets-centralized-reality/
- SiliconANGLE, הגיוס של Prime Intellect: https://siliconangle.com/2026/07/08/prime-intellect-raises-130m-1b-valuation-ai-training-platform/ ; The Block, הגיוס של Venice: https://www.theblock.co/post/406934 ; RuntimeWire Nous: https://runtimewire.com/article/nous-research-75-million-round-1-5-billion-valuation
- Phemex ; Ritual ,ownyourmind, Gensyn: https://ownyourmind.ai/projects/gensyn/ (כותרת בלבד): https://phemex.com/news/article/ritual-launches-ai-layer-1-mainnet-and-ritual-token-price-dives-72-from-high-101259
- תקורת המחשוב החסוי, arXiv 2409.03992: https://arxiv.org/abs/2409.03992v2 ; TEE.fail: דוח 2025 של http s://thehackernews.com/2025/10/new-teefail-side-channel-attack.html ; DeepProve, IACR ePrint 2026/1112: https://eprint.iacr.org/2026/1112 ; Phala: https://phala.com/posts/phala-2025-report
- TRM Labs, מדידת תשלומים של סוכני Forkast: https://www.trmlabs.com/trm-tech-blog/whos-actually-paying-measuring-ai-agent-payments-onchain ; קרן The: https://forkast.news/the-x402-foundation-activated-a-27-year-old-internet-code-200-million-transactions-later-the-real-volume-is-still-tiny/ ; x402 על AI: https://www.trmlabs.com/trm-tech-blog/whos-actually-paying-measuring-ai-agent-payments-onchain ; Block, הרשת הראשית של Tempo: https://www.theblock.co/amp/post/394131/tempo-mainnet-goes-live-with-machine-payments-protocol-for-agents ; BlockEden על ה־aGDP של Virtuals: https://blockeden.xyz/blog/2026/04/21/agdp-agent-gdp-virtuals-protocol-ai-blockchain-valuation-primitive-tvl-displace ; The Block על ElizaOS: https://theblock.co/post/410774/eliza-labs-native-token-dead ment/
- Dealroom, ההכנסות של Fetch.ai ,Grass: https://dealroom.co/news/157794-grass-backed-by-multicoin-projects-75m-in-2026-data-revenue/ ; CryptoSlate, ו־Fortune על Ocean: https://cryptoslate.com/cryptos-flagship-ai-pact-fracture-fetch-sues-ocean-over-263m-fet-community-sales/ ; World: https://fortune.com/2026/06/01/internet-according-to-sam-altman-seeing-orb
- Galaxy, הון סיכון בקריפטו: https://www.galaxy.com/insights/research/crypto-blockchain-venture-capital-q3 ; Crypto Briefing (כותרת בלבד): https://cryptobrief ing.com/crypto-vc-investment-q3-2026-galaxy/ ; Blockhead על Tiger Research: https://www.blockhead.co/2026/10/14/bitcoin-rose-43-in-q3-crypto-venture ; ForkLog: על השקות טוקנים ב־2025: https://forklog.com/en/85-of-tokens-launched-in-2025-fell-below-initial-prices -funding-did-not-follow/

כלכלת ASIC וסוגי שבבים

- Hashrate Index, הסיכום מ־9 בנובמבר 2026: https://hashrateindex.com/blog/hashrate-index-roundup-november-9-2026/
- AnySilicon, עלות תכנון ASIC לפי צומת ב־2026: https://anysilicon.com/news/how-much-does-asic-design-cost-in-2026-real-price-ranges-by-process-node/ ; Silicon Analysts, מדריך עלויות טייפ־אאוט ל־fabless: https://siliconanalysts.com/analysis/fabless-startup-tapeout-cost-guide ; נתוני שוק: https://siliconanalysts.com/market-data
- MillionMiner, Bitmain S23 XP Hyd: https://millionminer.com/news/bitmain-wdms-2026-s23-xp-hyd-launch
- Khattak ו־Mikaitis, צבירת FP8 בליבות טנזור, arXiv 2512.07004: https://arxiv.org/html/2512.07004v2
- Canaan, תוצאות הרבעון השני של 2026: Bitdeer: https://finviz.com/news/389337/canaan-inc-reports-unaudited-second-quarter-2026-financial-results ; תוצאות הרבעון השלישי של 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
- Core Scientific ; TheEnergyMag ו־Block: https://theenergymag.com/news/market-news/velaura-ai-series-a-bitcoin-pivot ; כלכלת החכירות ל־AI: https://theenergymag.com/news/2026-08-06/miner-weekly-ai-pumpa ; https://theenergymag.com/news/2026-07-29/core-scientific-exit-block-bitcoin-deal-ai bility-out-fuel
- Blockworks, Cipher ו־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, תוצאות הרבעון הראשון של שנת הכספים 27: https://globenewswire.com/news-release/2026/11/05/3397020/0/en/iren-reports-q1-fy27-results.html
- Spark, ריכוזיות שוק ה־ASIC לביטקוין: https://www.spark.money/research/bitcoin-asic-manufacturer-market-concentration
- Google, תיעוד TPU7x: https://docs.cloud.google.com/tpu/docs/tpu7x?hl=es ; The Register, Ironwood: https://www.theregister.com/2025/04/10/googles_7th gen_ironwood_tpus_debut/ ; SemiAnalysis, B300: https://inferencex.semianalysis.com/chips/b300 ; מאמר 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, תמחור 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, השוואת מאיצי הסקת AI: https://themenonlab.blog/blog/ai-inference-accelerators-compared ; Introl, מאיצי AI מעבר ל־GPU: https://introl.com/blog/ai-accelerators-beyond-gpus-tpu-trainium-gaudi-cerebras-2025

- StratNews (Reuters), שבבי AI מקומיים במרכזי נתונים במימון המדינה: https://stratnewsglobal.com/team-sng/china-orders-state-funded-data-centres-to-use-do mestic-ai-chips/ ; TNW, יבוא H200: https://thenextweb.com/news/nvidia-h200-chips-hong-kong-china-import-approval
- Succinct, הוכחה בזמן אמת על 16 GPU: https://blog.succinct.xyz/real-time-proving-16-gpus/ ; מוצרי ה־ZK ASIC של Cysic: https://cysic-docs.pages.dev/hardwar e-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 ו־SemiFive: https://evertiq.com/design/2026-02-20-semifive-secures-design-win-with-niobium-for-fhe-accelerator ; The Quantum Insider, הגיוס של Niobium: https://thequantuminsider.com/2025/12/03/niobium-23m-fhe-funding/ ; Optica OPN, Optalysys: https://www.optica-opn.org/home/indu stry/2026/february/optalysys_raises_us_$31_million_to_push_photonics_computing_toward_cloud_infrastructure/ ; Zama, HPU על FPGA: https://www.zam a.org/post/announcing-hpu-on-fpga-the-first-open-source-hardware-accelerator-for-fhe

צמתי תהליך

- Silicon Analysts, תמחור פרוסות: https://siliconanalysts.com/data/wafer-pricing
- Tom's Hardware, שמועות על תמחור TSMC A16: https://www.tomshardware.com/tech-industry/semiconductors/tsmc-could-charge-up-to-usd45-000-for-1-6n m-wafers-rumors-allege-a-50-percent-increase-in-pricing-over-prior-gen-wafers ; Wccftech, התמחור של SMIC: https://wccftech.com/smic-5nm-chip-prices- /up-to-50-percent-more-expensive/amp
- EUROPRACTICE, לוח הזמנים של TSMC MPW ל־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, מפת הדרכים של Huawei Ascend: https://www.tomshardware.com/tech-industry/artificial-intelligence/huawei-ascend-npu-roadmap-exami ned-company-targets-4-zettaflops-fp4-performance-by-2028-amid-manufacturing-constraints ; The Substrate, שרשרת האספקה של שבבי AI בסין: https://ww w.the-substrate.net/p/where-chinas-ai-chip-supply-chain ; Tom's Hardware, Trainium3: https://www.tomshardware.com/tech-industry/artificial-intelligenc e/amazon-launches-trainium3-ai-accelerator-competing-directly-against-blackwell-ultra-in-fp8-performance-new-trn3-gen2-ultraserver-takes-vertical-scali ng-notes-from-nvidias-playbook

הזדמנויות וכלכלת השקות

- Webull (Benzinga), שיחת התוצאות של Broadcom: https://www.webull.com/news/15511095937917952 ; Quartr, Global Unichip Q2 2026: https://quartr.com/ events/global-unichip-corp-3443-q2-2026_F9x6NJQY ; 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 ; The Block, יחידת ה־AI של Canaan: https://www.theblock.co/post/35 -to-8-2-billion-yuan-increased-r-d-investment-leads-to-wider-net-loss-earnings-insight/ 9250/crypto-miner-canaan-to-shutter-non-core-ai-semiconductor-business-unit-amid-strategic-realignment
- Clutch, תמחור פיתוח AI: https://clutch.co/developers/artificial-intelligence/pricing ; PixelPlex, עלות פיתוח בלוקצ׳יין: https://pixelplex.io/blog/blockchain-devel /opment-cost
- Hyperbolic, הוכחת דגימה: 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 ; אימות בשכבות, arXiv 2512.20176: https://arxiv.org/abs/2512.20176v1
- Sherlock, ייחוס לתמחור ביקורות ל־2026: https://sherlock.xyz/post/smart-contract-audit-pricing-a-market-reference-for-2026
- Blockworks, תקני הרישום של Cointelegraph ; Binance, עסקאות עושי שוק: https://cointelegr aph.com/magazine/market-maker-deals-quietly-killing-crypto-projects ; Blockworks, Movement: https://blockworks.co/news/movement-rushi-manche-sus pension-market-maker-web3port ; Nansen, HIP-3: https://nansen.ai/post/what-is-hip-3-hyperliquid ; The Block, Monad: https://www.theblock.co/post/380 094/monad-mainnet-launches

הוכחת עבודה מועילה · קריפטו × סיליקון

חישוב כקונצנזוס

בינה מלאכותית × קריפטו × שבבים ב־2026, החל מהוכחת העבודה המועילה של Pearl

לכל מספר מרכזי בדוח הזה מצוינים המקור והתאריך. מספר שהמחבר גזר בעצמו מסומן כאומדן; מספר שרק המוכר מדווח עליו מסומן כטענת החברה. הדוח עוסק בטכנולוגיה ובעסקים בלבד.



תאן שואי

טכנולוג ומייסד · חוקר ומפתח בינה מלאכותית

tanshuai.com · i@tanshuai.com

מהדורה פומבית · הנתונים נכונים ל־13 בנובמבר 2026