ChatGPT

עובד — ומהן הטכנולוגיות המתקדמות ביותר GPU כיצד בשנת AI 2026 לפריצת חומת הזיכרון במאיצי

FLOPS תמונת המצב: הבעיה כבר איננה רק

לא מספיק GPUאיננה עוד "ה־ AI נכון ל־31 באוגוסט 2026, הדרך המדויקת ביותר לתאר את מגבלת הביצועים של מערכות רוחב פס לזיכרון: **בשרשרת של קירות** מודרניות נתקלות למעשה AI יחיד. מערכות "Memory Wall מהיר", ואפילו לא "יש (**Bandwidth Wall**) מקומי, קיבולת HBM (**Capacity Wall**), תקשורת בין מאיצים (**Communication Wall**), והספק הנצרך בהעברת נתונים (**Power / Data-Movement Wall**). איזה קיר דומיננטי משתנה לפי שלב העבודה: Training, Prefill, Decode, long-context Attention או Mixture-of-Experts. מחקרי מערכות עדכניים על LLM inference מגיעים לאותה מסקנה: bottleneck יכול לעבור בין compute, memory bandwidth, memory capacity ו־synchronization/communication בהתאם ל־batch size, context length, parallelism והחומרה. [1]

החישוב נעשה זול וצפוף יותר מהר מהיכולת להזין אותו בנתונים. המעבר ל־FP8, FP4 ו־INT4 הסיבה הבסיסית היא ש־ עדיין צריכה KV-cache ו־activation, בצורה דרמטית, אך כל מטריצת משקולות Tensor Core מגדיל את קצב פעולות ה־ נעשים מהירים יותר, הבעיה הארכיטקטונית זזה מ"כמה Tensor Cores להגיע פיזית למקום שבו מתבצע החישוב. לכן ככל ש־ אפשר להביא, לאן, באיזו אנרגיה, ובכמה שימוש חוזר?". גם סקירות עדכניות של bytes אפשר לבצע?" אל "כמה MACs ולא על data movement מראות שחלק גדול מהאנרגיה במערכות חישוב מודרניות מוצא על Processing-in-Memory עצמו arithmetic. [2]

כדי להפריד עובדות מסוגים שונים, אשתמש לאורך המחקר בסיווג הבא:

Measured / Demonstrated — שהופעלו ונמדדו. כאשר המדידה היא של היצרן ולא גוף עצמאי link סיליקון, מערכת או אציין זאת.
Research Result — תוצאה ממאמר אקדמי/מערכת מחקרית.
Vendor Claim — נתון מוצר, performance projection או benchmark שהיצרן פרסם ושאין לו אימות ציבורי עצמאי מקביל.
Industry Projection — roadmap, sample או תחזית תעשייתית.
Speculation — מסקנה ארכיטקטונית שלי על בסיס המידע הזמין.

אחת הדרכים הטובות ביותר לראות את המגמה היא ב־Roofline. לדוגמה, NVIDIA מפרסמת ל־Rubin GPU עד 50 PFLOPS של NVFP4 inference ו־22 TB/s של HBM4 bandwidth. זה יוצר נקודת איזון תיאורטית של כ־ **2,273 FLOP/byte**. שאינו כלומר kernel נמצא סביב 8 TB/s ו־ **1,875 FLOP/byte**, עם Blackwell Ultra B300, 15 PFLOPS dense NVFP4 לא יכול להתקרב ל־ HBM שמגיע מ־ byte מצליח לבצע אלפי פעולות על כל peak compute בפורמטים הללו. [3]

מאיץ, נכון ל־2026	Fast memory	Bandwidth מקומי	Peak רלוונטי שפורסם	Roofline משוער knee	סטטוס
NVIDIA Blackwell Ultra B300	288 GB HBM3E	8 TB/s	15 PFLOPS dense NVFP4	~1,875 FLOP/B	**Vendor spec / shipping-generation** [4]

מאיץ, נכון ל־2026	Fast memory	Bandwidth מקומי	רלוונטי Peak שפורסם	Roofline משוער knee	סטטוס
NVIDIA Rubin	288 GB HBM4	22 TB/s	50 PFLOPS NVFP4 inference	~2,273 FLOP/B	**Vendor Claim / preliminary 2026 specs** [5]
NVIDIA Rubin, FP8/FP6 training	288 GB HBM4	22 TB/s	17.5 PFLOPS	~795 FLOP/B	**Vendor Claim / preliminary** [6]
AMD Instinct MI350/MI355X	288 GB HBM3E	8 TB/s עד	מפרסמת עד AMD 10.1 PFLOPS במצבים הכוללים sparsity	לא השוואה ישירה לdense־	**Vendor spec** [7]
Google TPU8t	216 GB HBM + 128 MB SRAM	6.528 TB/s	12.6 PFLOPS FP4	~1,930 FLOP/B	**Vendor spec** [8]
Google TPU8i	288 GB HBM + 384 MB SRAM	8.601 TB/s	10.1 PFLOPS FP4	~1,174 FLOP/B	**Vendor spec** [8]

ההשוואה איננה benchmark בין המאיצים: הפורמטים, הגדרת sparsity, utilization והפעולות הנספרות שונים. היא כן ממחישה את העיקרון הארכיטקטוני: **machine balance־ה של מאיץ 2026 מגיע למאות ואף אלפי FLOP/byte**. לכן Tensor Cores קטן, רחוקים מאוד מהאזור שבו עוד Decode־בbatch נמוך, במיוחד Arithmetic Intensity בעלי workloads מועילים.

מה באמת קורה בתוך GPU מודרני

GPU אינו “CPU עם הרבה cores”. הוא מערכת throughput-oriented שמנסה להחזיק אלפי threads פעילים, להסתיר latency באמצעות multithreading, ולשמור כמה שיותר operands קרוב ליחידות החישוב.

יחידת הבנייה המרכזית של GPU של NVIDIA היא **Streaming Multiprocessor — SM**. בתוך SM נמצאים execution pipelines שונים, register file גדול, scheduler-ים, CUDA cores, Tensor Cores וזיכרון on-chip מהיר המשמש כ־L1/shared-memory complex. בBlackwell Ultra, לדוגמה, NVIDIA 160 מתארת SMs, ארבעה Tensor Cores בכל SM ו־128 CUDA cores לSM; בRubin 224 היא מתארת SMs ו־896 Tensor Cores. [9]

CUDA Cores מטפלים בscalar/vector floating-point וinteger operations רגילות. **Tensor Cores** הם matrix engines ייעודיים: במקום לבצע multiply-add אחד בכל instruction, הם מקבלים tiles שלמים של מטריצות ומבצעים פעולת matrix-multiply-accumulate גדולה יותר. זו אחת הסיבות שAI FLOPS עלו במהירות כה גדולה ביחס לmemory bandwidth. [9]

מודל הביצוע הוא **SIMT — Single Instruction, Multiple Threads**. Threads מאוגדים ל־**warps של 32 threads**; threads מגורם לbranch במקביל. כאשר warp שלthreads עבור ה־instruction מנסה להריץ warp scheduler בגודל 32 והמודל ה־SIMT הם חלק ממודל warp והניצולת יורדת. הגדרת divergence לבחור מסלולים שונים, נוצרת בwarp התכנות הרשמי של CUDA. [10]

ה־**registers** הם on-chip, קרובים מאוד ל־execution units שממוחזרים בתוך kernel: operands המקום המועדף ל־ עצום. אבל register file מוגבל; ככל שכל thread דורש יותר registers, יכולים להיות פחות warps ובעלי bandwidth בו־זמנית, ולכן occupancy עלול לרדת. משאבי registers מחולקים בין ה־ warps הפעילים ב־SM. [11]

נמצאים proximity מתחתיהם מבחינת **Shared Memory ־וL1 Cache**. Shared memory נשלט במפורש על ידי התוכנה/ kernel משמש ל־tiling ולשיתוף operands בין threads באותו block; L1 מנוהל יותר כמו cache. בארכיטקטורות NVIDIA עצמו בנוי מבנקים, ובמודל CUDA מודרניות הם חולקים משאב פיזי מאוחד שחלקו ניתן להקצאה שונה. Shared memory ממופים ל־successive banks; bank conflicts עלולים לסדר גישות שהיו יכולות להתבצע successive 32-bit words הנוכחי במקביל. [12]

ה־**L2 cache** נמצא מחוץ ל־SMs ומשרת את המאיץ כולו. הוא חשוב במיוחד משום שהוא המקום האחרון שבו ניתן ללכוד reuse לפני פנייה ל־HBM. סביבו נמצאים memory partitions/controllers, interfaces ל־HBM והוא־on-chip interconnect שמחבר את ה־SMs, caches, I/O ויתר הבלוקים. הפרטים הפיזיים משתנים מדור לדור, אך היררכיית register → shared/L1 → L2 → off-chip HBM היא הבסיס שעליו מתבצעת אופטימיזציית kernels מודרנית. [13]

מסלול קריאה טיפוסי, באופן סכמטי, הוא:

`HBM DRAM arrays → HBM PHY / memory controller → L2 → on-chip fabric/NoC → SM L1/shared memory → registers → CUDA/Tensor Core`

ובכתיבה:

`Tensor/Core result → registers/shared → L2 → memory controller → HBM`

הנקודה החשובה איננה רק latency. **בכל מעבר כזה נצרכים wires, buffers, PHYs, clocks ואנרגיה**. אם operand משולם פעם אחת. אם HBM ונעשה בו שימוש 100 פעמים, מחיר ההבאה מ־ kernel נשמר ב־register וזו HBM כותב אותו ל־ kernel fusion יכולים לתת ־ikernel משולמים שוב. זו בדיוק הסיבה ש־ data movement הבא קורא אותו שוב, מחיר ה־ kernel משמעותי בלי להוסיף אפילו speedup אחד של transistor arithmetic. [14]

Arithmetic Intensity ־וRoofline Model

Arithmetic Intensity (AI) מוגדר:

$$AI = \frac{\text{FLOPs executed}}{\text{bytes transferred from the limiting memory level}}$$

ה־Roofline בקירוב:

$$Performance \leq \min(P_{peak},\ AI \times BW)$$

כאשר P_{peak} הוא compute throughput ו־BW הוא bandwidth של רמת הזיכרון הנבדקת. NVIDIA משתמשת באותה הגדרה בניתוח שלה Roofline. [15]

אם AI נמוך מנקודת ה־ridge:

$$AI_{ridge} = \frac{P_{peak}}{BW}$$

ה־kernel **memory-bound**. מעליה הוא יכול להפוך ל־compute-bound.

יכולים להתנהג הפוך לחלוטין Tensor Core המשתמשים בדיוק באותו workloads הבדל זה מסביר מדוע שני.

ב־**GEMM גדול** $C = A \times B$, כל tile של A ו־B נטען פעם אחת ל־SRAM/registers ומשמש לעשרות או מאות MACs. גדול הוא לעיתים קרובות GEMM יכול להיות גבוה מאוד ולכן Arithmetic Intensity compute-bound.

ב־**matrix-vector multiply** של Decode, כל weight לעומת זאת, עשוי להיקרא פעם אחת כדי לייצר token אחד. בלבד. אפילו כ־1 FLOP/byte של multiply-add — FLOPs ומייצר בערך שני bytes דורש BF16 2 מסוג weight, בהפשטה INT4 4~ נותן סדר גודל של FLOP/byte לפני metadata/scales. לעומת machine balance 1,000 של מעל FLOP/B, ברור batch קטן־מדוע decode אינו מסוגל "להאכיל" את Tensor Cores.

70 B parameters ב־BF16 כולל כ־140 GB weights. אם כל weight אפשר לראות זאת במספר פשוט. מודל היפותטי של לסרוק אותם בכ־6.36 ms, 157~ כלומר 8 tokens/s. ב־FP4, 35 GB raw weights או ~57 ms, הגבול הוא ~17.5 TB/s לכל היותר. ב־1,22 TB/s, יכול **בגבול פיזיקלי אידיאלי לחלוטין** batch=1 ובו־token נקרא פעם אחת עבור 1.59 לפני KV-cache, synchronization, dequantization, attention, network overhead ו־inefficiency. 22 ב־ms מורידים את גבול הסריקה לכ־ TB/s. אלה אינם benchmarks: הם lower bounds של bandwidth בלבד, 8 נתוני ה־TB/s מגיעים מ־Rubin ו־Blackwell Ultra. [3]

ב־**compute לא פחות משהיא טכנולוגיית memory-wall היא טכנולוגיית quantization** זו גם הסיבה ש־.

פוגעים בכמה קירות בו־זמנית Transformers מדוע

האמירה "Transformers הם memory-bound" נכונה רק בחלק מהמקרים. בפועל, כל שלב שלהם יוצר יחס אחר בין compute, storage ו־communication.

Training כולל GEMMs גדולים מאוד — QKV projections, MLPs, output projections — ולכן במאיץ בודד חלק ניכר מהעבודה יכול להיות compute-bound. אולם ככל שהמודל גדל, weights, gradients, optimizer state ו־activations עוברים את קיבולת ה־HBM, ואז נדרשים tensor/data/pipeline/expert parallelism. בשלב זה הבעיה עוברת בהדרגה מ־HBM ל־interconnect ול־collectives. מחקרי מערכות LLM של issue עדכניים מראים שהמגבלה היא מערכתית ולא רק arithmetic throughput. [1]

Prefill מעבד prompt שלם במקביל. ה־Q/K/V ו־MLP GEMMs גדולים יחסית ולכן מנצלים היטב Tensor Cores. ב־Attention, לעומת זאת, ארוך, attention matrix גדלה ריבועית עם sequence length, וה־I/O של intermediate tensors יכול להפוך לבעיה. FlashAttention נולד בדיוק מכאן: לבצע tiled attention כך ש־Q/K/V וה־partial reductions נשארים ככל האפשר ב־on-chip SRAM/registers במקום לכתוב $S \times S$ intermediates ל־HBM. על FlashAttention-3 דיווח על speedup של כ־1.5–2× לעומת FlashAttention-2, עד כ־740 TFLOPS/s ב־FP16 וכ־1.2 PFLOPS/s ב־FP8 במבחני המחקר שלו. **Research Result.** [16]

Decode מגיע iteration אחד חדש לכל token אחד ומעט התמונה ההפוכה. בכל sequence, ה־weights גדולים, אך עלול לקרוא GPU נמוך. המשמעות היא שה־weights של reuse קטן; לכן batch/token בצד ה־matrix dimension זה אחד הגורמים לכך ש־memory bandwidth, ולא peak FLOPS, קובע latency של interactive inference. Furiosa, ממחישה זאת בחישוב משלה עבור Llama 3.1 8B: 16 כ־GB BF16 weights ו־1.5 TB/s HBM נותנים bandwidth-only ceiling של 94 tokens/s; מדובר בסדר גודל של **בניתוח vendor**, אך העיקרון האריתמטי כללי. [17]

KV Cache יוצר קיר נוסף. עבור Transformer עם L layers, H_{KV} KV heads, head dimension d, precision של b bytes ו־sequence length S:

$$KV\ bytes \approx 2 \times L \times H_{KV} \times d \times b \times S$$

נותנים כ־320 KiB לכל token. 128k tokens ב־BF16 ו־Value. 80 layers, 128 KV heads, head dimension ו־Key ה־2 הוא עבור **יחיד sequence** ל־ GiB זה כבר כ־39.1 TiB, עוד לפני של 32 היה דורש מעל 1.2 Batch. הם “micro-optimizations”; הם לא offload ו־ GQA/MQA, KV quantization, paging, prefix sharing לכן weights. מנגנוני capacity management. מחקרי serving מתארים במפורש את footprint ואת bandwidth cost של KV cache כבעיה מרכזית. [18]

PagedAttention, שהוצג עם vLLM ב־SOSP 2023, מחלק את KV cache ל־ fixed-size blocks שאינם חייבים להיות contiguous בזיכרון, בדומה ל־ virtual-memory paging. המטרה היא להפחית fragmentation ו־ overprovisioning ולאפשר sharing פיזי כלל; הוא פשוט מונע בזבוז bandwidth יעיל יותר. זו דוגמה מצוינת לפתרון שאינו מגדיל capacity ו־ data movement. **Research Result.** [19]

Mixture-of-Experts (MoE) מפחית arithmetic בכך שרק חלק מה־ experts פעילים עבור כל token, אך הוא עלול להחריף שני קירות אחרים. ראשית, כלל ה־ expert weights עדיין צריכים להתגורר איפשהו, כך ש־ capacity עולה. שנית, ב־ Expert Parallelism צריך לשלוח tokens ל־ GPU שמחזיק את expert דפוס — המתאים ולהחזיר אותם **All-to-All**. MoE כאשר expert placement חוצה node/rack מרכזי ב־ bottleneck הוא All-to-All עבודות מ־2025–2026 מראות ש־ EP/TP placement שה־ Megatron עצמה מזהירה בפרסומי NVIDIA צריך לקחת בחשבון את ההבדל הגדול בין NVLink לבין network bandwidth. [20]

לכן, נכון יותר ב־2026 לדבר על “קיר”:

דרך הטיפול העיקרית	דוגמה אופיינית	מה מוגבל	“קיר”
locality, hierarchy, software/hardware co-design	GPU עם Tensor Cores רעבים לנתונים	הפער הכללי בין arithmetic ל־ data supply	**Memory Wall**
HBM4, quantization, SRAM, CIM	batch-1 LLM decode	bytes/s ברמת hierarchy מסוימת	**Bandwidth Wall**
HBM יותר, GQA/KV quantization, DDR/CXL/pooling	KV-cache ארוך, trillion-param MoE	אפשר state כמה להחזיק בזיכרון המהיר	**Capacity Wall**
NVLink/UALink/Ethernet/InfiniBand, topology, optical I/O	All-Reduce, TP, MoE All-to-All	bytes/s + latency בין chips/nodes	**Communication Wall**
locality, 3D stacking, CIM, photonics, lower precision	דרך העברת TB/s DRAM/SerDes	joules/bit והספק כולל	**Power / Data Movement Wall**

כמה באמת עולה להזיז bit

אין מספר pJ/bit אוניברסלי: process node, SRAM size, wire length, access width, PHY, voltage ו־packaging משנים benchmark אחיד, ולא **anchors סדרי גודל והשוואת**, את המספרים משמעותית. לכן טבלת האנרגיה הבאה היא

תנועה	סדר אנרגטי	הערה
Register / very-local SRAM	עד pJ/bitהנמוך ביותר; לרוב תת־low-pJ/bit לפי array/path	הערכות תעשייתיות ;locality היתרון המרכזי הוא בתנאים מסוימים, אך pJ/bit סביב ~0.1 SRAM מציבות זה אינו ערך אוניברסלי. [21]
3D very-near memory path	במודל מחקרי ~0.66–0.88 pJ/bit 2026	**Research Result**, מסחרי HBM לא מוצר. [22]
NVLink-C2C on-package	בדור pJ/bit דיווחה ~1.3 NVIDIA Hopper	**Vendor measured/spec anchor**. [23]
HBM3/HBM3E	הוא סדר low-single-digit pJ/bit גודל מקובל; סקירה מ־2026 HBM3Eל־ pJ/bit מצטטת ~3.4	המספר תלוי מאוד ב־implementation. [24]
Optical I/O של Lightmatter M1000	laser כולל 2.3 pJ/bit	**Vendor measured on EVK/reference platform**, 114.6 Tb/s bidirectional. [25]
Board/rack electrical link	SerDes/גבוה יותר ככל שהמרחק גדלים retiming	זו אחת הסיבות למעבר ל־co-packaged optics; אין מספר יחיד תקף לכל link. [26]

הופכים להספק אדיר pJ/bit אפילו 2–3. **bits המרחק והמספר הכולל של** ההשוואה החשובה יותר מהמספר המדויק היא במודל bit-transferרק עבור אנרגיית ה־ kW הוא כ־1.92 pJ/bit × 2.4 TB/s × 8 bits/byte לדוגמה, 100 .tens of TB/sב־ עצמו נהפך לבעיה תרמית והספקית bandwidth :אינו פתרון חינם "TB/s ל־100 HBM פשוט. זו הסיבה ש"פשוט נגדיל. [27]

הטכנולוגיות שמנסות לפרוץ את הקירות

ניתן לחלק את הפתרונות לשלוש משפחות. אין טכנולוגיה יחידה שפותרת את Memory Wall:

להעביר יותר bytes — HBM4, 2.5D/3D, wider interfaces, optical I/O.
להעביר פחות bytes — quantization, sparsity, compression, FlashAttention, fusion, tiling, KV optimization.
למעבד מלכתחילה bytesלהפסיק להעביר את ה־ — Near-Memory Computing, PIM וCompute-in-Memory.

טכנולוגיה	Bandwidth	Capacity	מפחיתה traffic	משנה מקום חישוב	בגרות ב־2026
HBM3E/HBM4	**כן, מאוד**	כן	לא	לא	Production
HBM4E	כן	כן	לא	לא	Samples / roadmap
CoWoS / 2.5D	מאפשר width גדול	מאפשר יותר stacks/ chiplets	מעט, בגלל links קצרים	לא	Production

טכנולוגיה	Bandwidth	Capacity	מפחיתה traffic	משנה מקום חישוב	בגרות ב־2026
Hybrid bonding / 3D	bandwidth density גבוהה מאוד	אפשרי	כן, דרך proximity	חלקית	Production + rapidly expanding
Quantization	effective BW עולה	effective capacity עולה	**כן**	לא	Production
FlashAttention / fusion	PHY BW לא מגדיל	לפעמים	**כן מאוד**	לא	Production software
KV paging/ compression	effective capacity	**כן**	כן	לא	Production/ research
CXL pooling	מעט יחסית ל־HBM	**כן מאוד**	לא בהכרח	לא	Early production
Near-memory / PIM	internal BW גדול	תלוי	**כן**	**כן**	Emerging
Digital CIM	local effective BW עצום	מוגבל SRAM/3D	**כן מאוד**	**כן, בתוך/ליד array**	Early commercial
Analog CIM	local MAC/W גבוה	מוגבל	**כן מאוד**	**כן**	Early commercial/ edge + research
Optical I/O	**כן מאוד**, בעיקר chip/rack	עקיף	לא	לא	EVK/early deployment
Massive SRAM/ dataflow	**כן מאוד on-chip**	SRAM יקרה בשטח	**כן מאוד**	execution spatially distributed	Commercial niche

HBM3E, HBM4 והלאה

HBM עובד באמצעות stack של DRAM dies המחוברים אנכית ב־TSVs, עם bus רחב מאוד למאיץ דרך package/interposer. HBM3E 1.2–1.1 TB/s של 2024–25 סיפק כבר מעל **stack־ל** במוצרים מובילים. SK hynix החל mass production של HBM3E 1.18 TB/s במרץ 2024 ודיווח על עד 1.2< TB/s למוצר באותה תקופה; Micron מפרסמת TB/s בגרסאות שלה. [28]

HBM4 ב־interface 2,048 I/O pins מוכפל ל־bandwidth חל שינוי גדול נוסף: ה־stack ל־high stack, 12 GB ל־Rubin, 36 עם ברבעון הראשון של 2026 עבור פלטפורמת HBM4 של volume shipment דיווחה על יותר מ־11 Gb/s/pin ויותר מ־2.8 TB/s ל־stack. Samsung הודיעה בפברואר 2026 על commercial HBM4 11.7 עם Gb/s, capability 13 עד Gb/s 3.3~ ועד TB/s ל־stack. **Measured/Product vendor specifications.** [29]

Samsung 2026 כבר סיפקה במאי samples של HBM4E: 14 החברה מתארת Gb/s stable operation, scalability לכיוון 16 Gb/s 3.6 וכ־TB/s לstack־ב sample תצוגות הנוכחי; roadmap אחרות של Samsung 4~ מציגות יעד של TB/s. לכן נכון יותר לסווג כ־ 3.6 TB/s **sample**, 4 ו־TB/s כ־ **roadmap/vendor target**. [30]

HBM5־ל SK hynix. ציבורי סופי ואמין שמצדיק טבלת נתונים כאילו מדובר במוצר specification, עדיין אין, נכון למועד המחקר HBM4E/HBM5־ב custom logic/base dies גבוהים יותר ויותר hybrid bonding, stacks מאותתות על packaging וחברות **Industry Projection**, לא fact. בשלב זה הן HBM5 לכן תחזיות מספריות קשיחות ל־ era. [31]

הבעיה העמוקה היא שגם HBM4 “פתרון סופי אינו”. Rubin 8־ב TB/s קופץ מ־Blackwell Ultra 22־ל TB/s, 2.75~ ×כלומר bandwidth; אבל peak NVFP4 inference 50 עולה מ־15 ל־PFLOPS, כ־3.33×. ה־Roofline knee הוא אפוא אינו יורד — עדיין נמשך. הנתונים הם memory־ל arithmetic זו אינדיקציה לכך שהמרוץ בין FLOP/B. אף עולה מכ־1,875 לכ־2,273 והיחס הוא חישוב מהנתונים הללו vendor specs. [32]

Advanced packaging: מ־CoWoS 3־ל־D

2.5D packaging כגון TSMC CoWoS מאפשר להניח compute dies ו־HBM stacks על silicon interposer או redistribution layer עם אלפי חיבורים רחבים וקצרים. TSMC מתארת את CoWoS כטכנולוגיית integration ל־SoCs ומרחיבה את HBM ו־interposer size מעבר ל־reticle יחיד; CoWoS-S 3.3 מגיע לכ־reticle size ×3 בדורות הנוכחיים, כאשר CoWoS-L/R מיועדים להתרחבות נוספת. **Vendor manufacturing capability.** [33]

3D integration הולך צעד נוסף: dies נמצאים ממש זה על זה. TSMC SoIC משתמש ב־hybrid bonding עם pitches מתחת ל־10 µm, ו־TSMC 3 דיווחה ש־nm stacking נכנס ל־volume production 2025 ב־. Intel Foveros Direct משתמש ב־direct copper-to-copper hybrid bonding, כאשר roadmap 9~ מ־ציבורי כולל מעבר ל־3~ µm pitches לכיוון 5 µm. [34]

Hybrid bonding חשוב מפני ש־bump pitch קטן יותר פירושו **orders-of-magnitude יותר vertical wires ליחידת** שטח. בכך ניתן לשים SRAM/DRAM logic ממש מעל או מתחת ל־compute 3 ולהפוך את זיכרון ה־D למשהו הקרוב יותר ל־on-chip resource מבחינת bandwidth density.

אבל 3D אינו חינם: compute die מתחת ל־memory die יוצר thermal coupling; stacking פוגע ביכולת להוציא חום; bonding yield ו־known-good dies תלוי ב־ yield של stack ו־warpage 3D מסתבכים. לכן סביר שה־ **SRAM/cache, logic-on-memory** ו־**short vertical links**, ולא ב“ערימת Tensor הראשון במאיצי datacenter יתמקד ב־ Cores חמים זה על זה”.

Near-Memory, PIM ו־Compute-in-Memory

המונחים דומים אך אינם זהים.

Near-Memory Computing (NMC) משאיר memory array רגיל, אבל מעביר compute logic אל base die, logic layer או chiplet סמוך. Qualcomm HBC היא דוגמה מעניינת במיוחד: Qualcomm מתארת accelerator logic מ־2026 ב־AI250 “133 TB/s effective memory bandwidth per card”, מה־AI200; מפרסמת ל־LPDDR stack קרוב מאוד ל־ **effective bandwidth**, לא raw DRAM interface bandwidth. זהו **Vendor Claim** חשוב אך החברה מדגישה שמדובר ב־ Rubin. עדיין לא ניתן להשוות אותו ישירות ל־22 TB/s raw HBM של [35]

Processing-in-Memory (PIM) הוא מונח רחב יותר שבו פעולות מסוימות מתבצעות בתוך memory subsystem — למשל כדי לנצל bandwidth פנימי שה־external pins אינם יכולים לחשוף — או ליד banks של stacked DRAM או logic layer.

Compute-in-Memory (CIM/IMC) מכניס את ה־MAC עצמו אל memory array. ב־SRAM CIM, weight bits נשמרים בתאי SRAM וה־wordlines/bitlines משמשים גם לחישוב. **Digital CIM** מבצע arithmetic או bit-serial דיגיטלי קרוב מאוד לתאים; **Analog CIM** מנצל currents/charges/voltages שה־array מבצע vector-matrix multiplication באופן פיזיקלי. ISSCC 2025 הקדיש session שלם ל־CIM, ולא רק רעיון mainstream research direction עדות לכך שהתחום הוא תיאורטי. [36]

היתרון הפוטנציאלי עצום: weight לא צריך לעבור SRAM → MAC בכל operation. החיסרון ב־analog CIM הוא ADC/DAC overhead, quantization/noise, process variation, calibration, limited precision וה־difficulty של general-purpose programming. SRAM CIM גם משלם במחיר area: SRAM הרבה פחות צפוף מ־DRAM, ו־8T/10T CIM cells עדכניות מדגישות בדיוק IEEE הצפיפות יכולה להיות גרועה עוד יותר. סקירות trade-off זה. [37]

EnCharge AI היא אחת הדוגמאות המסחריות המעניינות: היא משתמשת ב־capacitor-based analog in-memory computation עם weights ב־SRAM. ה־EN100 שהוכרז ב־2025 מיועד בעיקר client/edge ומפורסם כ־200+ TOPS. **Vendor Measurement**, לא שנמדדו בסיליקון לעומת מוצרים מתחרים, אך זו MACs על TOPS/W ×החברה טוענת ל־20 datacenter-scale independent benchmark. [38]

d-Matrix הולכת בכיוון אחר — **Digital In-Memory Compute**. Corsair משלב “Performance Memory” בתוך compute fabric, ומפרסם ל־dual-card 4 GB של Performance Memory 300 עם TB/s local bandwidth, לצד עד 512 GB capacity memory; לשרת שמונה cards 1,200 החברה מפרסמת TB/s aggregate Performance Memory bandwidth. חשוב במיוחד: d-Matrix עצמה מסמנת את נתוני ה־LLM performance כ־**preliminary and subject to change**. [39]

Quantization, sparsity וה־compression

הדרך הזולה ביותר לפרוץ memory wall היא לעיתים לא לשנות את ה־memory tensor. כלל, אלא להקטין אותו.

Raw payload:

Format	Bytes/value	BF16 מול traffic הפחתת
BF16 / FP16	2	1×
FP8 / INT8	1	~2×
FP4 / INT4	0.5	~4×

אינם בדיוק פי 4. אבל ההשפעה כפולה: savings בפועל יש packing overhead, scales, metadata, alignment ולכן peak performance שמאיצי 2026 העובדה מתכננים את ה־cache/HBM. נכנסים ב־weights יותר **וגם** weights קטנים יותר מצביעה על כך ש־low precision הפך לחלק מהארכיטקטורה ולא optimization צדדי. Rubin, FP4/NVFP4/MXFP4 סביב בנתוני הביצועים שלהם bit מתחת ל־8 formats כולם כוללים TPU8 ו־MI350. [40]

LLM weights אחרים הראו כבר ב־2023–24 שניתן להוריד low-bit weight-only quantization ומחקרי AWQ מותאמים לחומרה kernels ועם accuracy תוך שמירה טובה יחסית על representations. **Research Result.** [41]

KV-cache quantization חשוב לא פחות: ב־long-context serving ה־cache יכול להיות גדול יותר מה־working-set effective context capacity, אם accuracy יכול כמעט להכפיל/לרבע את ה־FP8/INT8 4-bit ל־BF16 האחר, ולכן מעבר מ־

לא ייצור compression כדי שה־dequantization overhead מפתחים 2024–26 מקובלים. מחקרי fused kernels גדול יותר מהחיסכון overhead. [42]

Sparsity מורידה compute רק כאשר החומרה וה־format מנצלים אותה בפועל. Zero weight שלא מאוחסן לדלג גם על hardware שמאפשרת ל־memory bandwidth; structured sparsity עדיין צורך ב־compressed format storage/loads יכולה לפגוע גם ב־Bandwidth Wall. 2× sparse FLOPS” לכן אינו שקול אוטומטית ל־2× speedup.

Software הוא חלק ממערכת הזיכרון

FlashAttention, tiling, kernel fusion ו־persistent kernels הם עושים דבר ארכיטקטוני: **משנים את מסלול החיים של הנתון**.

במקום:

```
HBM → kernel A → HBM → kernel B → HBM → kernel C
```

מנסים לבצע:

```
HBM → SRAM/registers → A → B → C → HBM
```

FlashAttention-3 מוסיפה ל־IO-awareness גם pipeline אסינכרוני המנצל TMA, Tensor Cores ו־warp specialization ב־Hopper, 2–1.5× improvement ביחס ל־FlashAttention-2 והמאמר מדווח על שלו במבחנים. [43]

NVIDIA Rubin ממשיכה את אותו כיוון בחומרה: Tensor Memory Accelerator ו־mechanisms ל־asynchronous data movement **לחפוף (overlap) load, compute ו־movement**. אלא MMA הם חלק מהארכיטקטורה, משום שהאתגר כבר אינו רק לבצע. [44]

CXL ו־memory pooling

CXL מטפל בעיקר ב־**Capacity Wall**, לא ב־HBM Bandwidth Wall. CXL 3.x מאפשר switching, fabric-attached memory, sharing ו־dynamic capacity; ב־CXL 3.2 Dynamic Capacity Devices עם multi-head ו־2026 כבר הודגמו memory pooling למספר hosts. **Measured / Demonstrated ecosystem interoperability.** [45]

המספרים מדגישים את ההבדל: Rubin 22 מפרסמת TB/s HBM4, בעוד Rubin 256 במערכת PCIe 6 x16 הוא GB/s class. היטב, מדובר בסדר גודל של עשרות פעמים פחות bandwidth ועם latency גבוה יותר. גם אם CXL מנצל את physical link עבור HBM הוא אינו תחליף ל־capacity disaggregation; ל־CXL אידאלי cold KV cache, model spill, shared pools ו־inner-loop GEMM/decode. נתוני Rubin I/O מגיעים מהתיעוד הטכני של NVIDIA. [44]

Network הופך ל־Memory Wall האם ה־rack-scale ל־chip-scale מ־Wall?

יחיד GPU של 2025–26 הוא שהיחידה שאותה ספקי החומרה אופטימיזו כבר איננה AI המעבר המשמעותי ביותר בארכיטקטורת היא **rack או multi-rack compute domain**.

ממחישה את ההיררכיה בצורה כמעט מושלמת Rubin:

- בתוך GPU: **22 TB/s HBM4 bandwidth**.
- GPU-to-GPU scale-up: **3.6 TB/s NVLink 6 per GPU**.
- scale-out networking: כ־**0.4 TB/s per GPU** לפי מפרט NVL72.
- ברמת rack 72 של GPUs: 1,580 TB/s aggregate HBM bandwidth, 260 TB/s NVLink Switch bandwidth ו־28.8 TB/s scale-out. [6]

בקירוב, פר GPU כלומר:

$$HBM : NVLink : ScaleOut \approx 22 : 3.6 : 0.4 \approx 55 : 9 : 1$$

זה נתון קריטי. Byte שנמצא ב־HBM המקומי הוא משאב שונה לחלוטין מ־byte שנמצא אצל GPU אחר, ועוד יותר מ־byte שנמצא מעבר ל־rack.

ב־Blackwell Ultra 8 TB/s HBM היחס דומה: 1.8 TB/s NVLink per GPU. [4]

ה־**NoC hierarchy** של datacenter, נהפכת להמשך של, לפיכך memory hierarchy:

`Registers → SRAM/L1 → L2 → HBM → local GPU fabric → rack fabric → inter-rack network → pooled/remote memory`

מנקודת המבט של compiler/system architect, מקומות שבהם "כולם bit יכול לגור", עם bandwidth, latency ואנרגיה שונים.

All-Reduce הוא traffic pattern מרכזי ב־training. Data-parallel workers צריכים לאחד gradients; tensor-parallel layers דורשים collectives במהלך forward/backward. ב־ring all-reduce, עבור tensor בגודל S ובמספר גדול של ranks, כל rank מעביר בערך $2S$ bytes בסך הכול לאורך reduce-scatter + all-gather. לכן הגדלת GPU FLOPS ללא הגדלת fabric bandwidth יכולה פשוט להזיז את הזמן מ־GEMM ל־communication.

All-to-All של MoE קשה יותר טופולוגית: tokens routed ל־experts מפוזרים, ולכן כל accelerator יכול לשלוח traffic למספר רב של peers. מחקרים מ־2026 מתארים inter-node All-to-All כ־bottleneck חמור ל־expert parallelism, ועבודות אחרות מציעות להחזיק expert communication בתוך node כדי להימנע ממנו. [46]

NVIDIA מפרסמת 130 TB/s aggregate all-to-all bandwidth ל־GB200 NVL72 generation ומקדמת את NVLink fabrics בדיוק כדי להפוך rack ל־scale-up domain. ב־Rubin היא מכפילה את per-GPU NVLink 3.6 ל־3.6 TB/s. אבל הכיוון הארכיטקטוני ברור, **Vendor specs**. [47]

לכן התשובה לשאלה "האם Memory Wall הופך ל־Network Wall?" היא:

Speculation, אבל מבוססת היטב: לא באופן אוניברסלי — אלא **זז החוצה בהיררכיה שה־wall**.

ב־single-GPU batch-1 decode הוא HBM wall.
ב־8-GPU tensor parallelism הוא עשוי להיות NVLink wall.
ב־rack-scale dense training הוא collective wall.
ב־large MoE הוא עשוי להיות All-to-All / topology wall.
ב־multi-rack model with remote KV הוא כבר network + capacity wall.

כלומר, ה־Network Wall אינו מחליף את Memory Wall; הוא **הגרסה הבין־שבבית שלו**.

קיבלה תאוצה עצומה silicon photonics זו בדיוק הסיבה ש־Lightmatter M1000 הוא 4,000 mm² photonic interposer עם **114.6 Tb/s bidirectional optical bandwidth, 256 fibers 2.3 pJ/bit כולל laser**, שכבר reference platform של החברה. זה **Measured/Demonstrated by vendor**, לא rack-scale validation environment שפועל ב־deployment hyperscale בלתי תלוי. [48]

Ayar Labs מציגה TeraPHY optical I/O chiplet עם UCIe ו־**8 Tb/s**. 4 גרסת Tb/s עברה wafer sort, EVT, DVT ו־system integration; 8 גרסת Tb/s עברה engineering validation ו־link testing, כולל multi-day UCIe tests. זה הרבה מעבר לPowerPoint־, אבל עדיין מוקדם מ־mass deployment בתוך מיליוני GPUs. [49]

Celestial AI, שנרכשה על ידי Marvell תמורת 2026 ב־2 בפברואר consideration $3.5 של כ־B, מפתחת Photonic Fabric ברמת package/system/rack; Marvell מציינת שהיישום הראשון מיועד ל־scale-up interconnect all-optical scale-up. ויותר כחלק מארכיטקטורת "network transceiver" פחות כ־optics רואה industry זוהי אינדיקציה חזקה לכך שה־accelerator fabric. [50]

Optical compute, לעומת זאת, נמצא בשלב פחות בוגר משמעותית. Photonic matrix multiplication יכול תאורטית לבצע linear algebra בצורה מקבילית ויעילה, אבל weights/storage, ADC/DAC, non-linear functions, precision, laser efficiency והחיבור לעולם הדיגיטלי נשארים מכשולים. סקירת Nature מראה ש־Communications הם כבר mainstream commercialization בעוד photonic computing עדיין מתפתח. לכן ב־2–5 השנים הקרובות **optical Tensor Core סביר בהרבה מ־optical I/O**. [51]

החברות שמנסות לשבור את מודל GPU + HBM

להלן 13 שחקנים, כולל הסטארטאפים שביקשת ושחקנים חדשים שהפכו רלוונטיים במיוחד ב־2025–26.

חברה	ארכיטקטורה וזיכרון	כיצד הנתונים זזים / מה נשבר	מה הוכח בפועל	מה עדיין Vendor Claim	האם יכולה להתחרות ב־GPU+HBM?
Cerebras	Wafer-Scale Engine; מאות אלפי cores עם **44 GB SRAM on-wafer** ב־WSE-3 family, fabric על כל ה־wafer; external memory systems למודלים גדולים.	מבטלת כמעט chip-to-chip traffic לחלוטין בתוך wafer ומספקת PB/s-class SRAM bandwidth.	WSE-3 הוא silicon deployed; ב־2026 החברה הציגה CS-4 עם WSE-3 שלושה ומערכת Turbo Nexus. [52]	עד 30× inference speed או־10× throughput/W מול GPU הם **Vendor Claims**. [53]	**Speculation:** כן, בנישות גדולות במיוחד inference/training עם dataflow; אך 44 GB SRAM אינה מספיקה למודלי frontier ולכן external model storage/streaming נשאר חיוני.

חברה	ארכיטקטורה וזיכרון	כיצד הנתונים זזים / מה נשבר	מה הוכח בפועל	מה עדיין Vendor Claim	האם יכולה להתחרות ב־GPU+HBM?
Groq	Deterministic LPU/TSP-style architecture, software-scheduled execution ו־large on-chip SRAM.	weights/ activations נשמרים קרוב לחישוב; compiler קובע movement מראש במקום dynamic caches/ schedulers.	LPU deployed ב־GroqCloud; החברה מפרסמת >80 TB/s on-chip SRAM BW. בדצמבר 2025 Groq חתמה עם NVIDIA על non-exclusive technology license ונשארה חברה עצמאית. [54]	performance superiority ו־"10× claims הם vendor claims.	**Speculation:** חזק ל־low-latency inference; SRAM capacity מחייבת scale-out למודלים גדולים ולכן interconnect נהפך לבעיה.
SambaNova	Reconfigurable Dataflow Unit. SN50: **432 MB SRAM + 64 GB HBM2E + עד 512 GB DDR5** לפי chip ל־datasheet/site.	compiler ממפה graph כ־dataflow; hot ל־SRAM operators, weights/KV ל־HBM, large ל־DDR cache/model capacity. [55]	SN40 generation deployed; SN50 הוכרז בפברואר 2026 והחברה ציינה shipping במחצית השנייה של 2026. [56]	performance/ TCO מול B200 ו־support עד 10T parameters הם Vendor Claims.	**Speculation:** מהגישות המאוזנות ביותר, כי אינה מנסה להחליף DRAM ב־SRAM בלבד; programmability/ compiler quality הם קריטיים.
Tenstorrent	Tensix tiled architecture; local SRAM בכל tile, NoC, RISC-V control cores ו־GDDR6 off-chip. Blackhole cards מגיעים עד 32 GB GDDR6.	explicit DMA דרך NoC בין local SRAM, DRAM tiles ו־Ethernet; software חשוף יחסית למיקום הנתונים. [57]	Blackhole cards מוצעים בפועל ומפורסמים כ־in-stock; Wormhole/ Blackhole software stack פתוח. [58]	"infinite scalability" ו־performance/ cost superiority הם vendor positioning.	**Speculation:** יתרון משמעותי ב־openness/cost; HBM-class ללא קשה כרגע BW לנצח high-end GPUs ב־large dense decode.

חברה	ארכיטקטורה וזיכרון	כיצד הנתונים זזים / מה נשבר	מה הוכח בפועל	מה עדיין Vendor Claim	האם יכולה להתחרות ב־GPU+HBM?
d-Matrix	**Digital In-Memory Compute (DIMC)**; compute משולב ב־SRAM-like “Performance Memory”, עם capacity memory חיצוני.	weight MAC נעשה ליד/בתוך storage; dual Corsair card: 4 GB Performance Memory @300 TB/s + עד 512 GB capacity memory. [39]	Corsair silicon/ platform קיים; החברה דיווחה גם על 3D-memory silicon operational במעבדה.	כמעט כל LLM speedup/TCO numbers באתר מוגדרים preliminary; למשל 10× interactive speed הוא **Vendor Claim**. [39]	**Speculation:** אחת האלטרנטיבות הישירות והמעניינות ביותר ל־HBM-centric inference, במיוחד אם 3D DRAM+DIMC יבשיל. Training/ general-purpose עדיין פחות מוכח.
Etched	ב־2026 החברה מתארת rack-scale inference architecture עם **HBM/SRAM hybrid**, Cluster Scale Memory ו־Low Voltage Inference; פחות דגש פומבי כיום על המיתוג המקורי Sohu.	proprietary low-latency fabric יוצר shared memory pool בין chips; SRAM מספק latency, HBM capacity.	החברה אומרת ש־A0 N4P silicon חזר מ־TSMC ושה־customer rack validation מתבצע. [59]	80%+ peak FLOP utilization, SOTA latency/ throughput ו־$1B demand הם **Vendor Claims** ללא public independent benchmark מפורט.	**Speculation:** גבוה משום upside שה־co-design rack+memory מאוד אגרסיבי; גם uncertainty גבוה במיוחד עד הופעת benchmarks ומפרט מלא.
Lightmatter	כיום בעיקר **photonic interconnect**, לא replacement Tensor processor: Passage photonic interposer + optical I/O.	מכל I/O מוציאה שטח ה־interposer במקום רק משפת die; 114.6 Tb/s M1000.	M1000 EVK קיים ופועל ב־validation data center; 2.3 pJ/bit כולל laser. [60]	million-XPU scale and hyperscale economics הם future/vendor claims.	**לא תחליף אלא complement:** יכולה להיות אחת הטכנולוגיות שמאפשרות ל־GPU/ASIC racks להמשיך לגדול.

חברה	ארכיטקטורה וזיכרון	כיצד הנתונים זזים / מה נשבר	מה הוכח בפועל	מה עדיין Vendor Claim	האם יכולה להתחרות ב־GPU+HBM?
Celestial AI / Marvell	Photonic Fabric ל־optical scale-up, package-to-package/ system/rack.	optics מקרב את energy/reach של remote accelerator communication ל־scale-up fabric.	Marvell השלימה את הרכישה בפברואר 2026; technology/ platform קיים ויש hyperscaler engagements לפי Marvell. [61]	commercial large-scale deployment עדיין projection.	**Complement:** אם תצליח, תשנה ולא Network Wall את Tensor Core עצמו.
Ayar Labs	TeraPHY UCIe optical I/O chiplet + external SuperNova laser.	electrical UCIe נכנס ל־chiplet, WDM optics מוציא traffic במטרים עד קילומטרים.	4 Tb/s generation עברה EVT/ DVT/system integration; 8 Tb/s version עברה extensive engineering/ link validation. [49]	economics ו־high-volume GPU integration טרם הוכחו בפומבי.	**Complement**, פוטנציאל גבוה מאוד ל־multi-rack scale-up.
EnCharge AI	Capacitor-based **analog SRAM Compute-in-Memory**.	weight נשאר ב־SRAM array וה־MAC משתמש ב־charge-domain physics, מפחית memory↔MAC traffic.	silicon מחקרי underpinning קיימים; EN100 הוכרז עם 200+ TOPS ל־client/ edge. [38]	20× efficiency ו־10× TCO הם vendor comparisons.	**Speculation:** מבטיח מאוד ב־edge/ inference; datacenter frontier models דורשים capacity, precision ו־software scaling שעוד לא הודגמו.

חברה	ארכיטקטורה וזיכרון	כיצד הנתונים זזים / מה נשבר	מה הוכח בפועל	מה עדיין Vendor Claim	האם יכולה להתחרות בGPU+HBM?
FuriosaAI	Tensor Contraction Processor (TCP); **256 MB on-chip SRAM + 48 GB HBM3, 1.5 TB/s**, 256 TFLOPS BF16/512 TFLOPS FP8, 150W spec.	compiler/dataflow מנסה לשמור intermediates בSRAM ולהשתמש בHBM לweights. [62]	RNGD הוא silicon ממשי; Furiosa הדגימה gpt-oss-120B על שני RNGD ב־5.8 ms/output-token במדידת החברה. [63]	benchmark הוא vendor demo ולא standardized independent test.	**Speculation:** מתחרה אמיתי בefficient inference, אך ecosystem וקנה מידה קטנים בהרבה מCUDA/NVIDIA.
Positron	Transformer-focused Archer accelerators; Atlas כולל 8 chips × **32 GB HBM = 256 GB HBM**.	specialization מפחיתה general-GPU overhead; מערכת משלבת HBM עם host DDR.	Atlas מוגדר באתר כ־“Shipping Today” ב־2026. [64]	>4× perf/W ו־3× lower latency מול Hopper הם vendor benchmarks.	**Speculation:** יכול לנגוס בinference cost, אבל עדיין לא הוצגה architecture בעלת bandwidth/capacity discontinuity ברמת CIM.
Qualcomm HBC	**Near-Memory Compute / High Bandwidth Compute** סביב LPDDR stack; AI250/AI300 roadmap.	computation מתקרב לdata במקום להזרים הכול לcentral tensor core. AI250 מפורסם עם 768 GB/card ו־133 TB/s **effective** bandwidth. [65]	HBC/AI250 הוצגו ב־2026 ויש roadmap רשמי.	18× effective bandwidth מול AI200 ו־6× bandwidth/W מול HBM הם vendor claims; אין raw link BW שקול לHBM לצורך comparison. [66]	**Speculation:** אחת ההתפתחויות החשובות של 2026. אם effective-bandwidth claims יתורגמו לreal LLM benchmarks, היא עשויה להוכיח שLPDDR+near-memory compute יכול לעקוף HBM בdecode.

נקנתה על ידי NVIDIA; Celestial AI חתמה על רישוי טכנולוגיה ל־Groq: מעניין לראות שגם השוק עצמו התחיל להתכנס לrack-scale system architecture. NVLink עצמה הופכת ל־UCIe; NVIDIA מתיישרים עם Marvell; optical chiplets הטכנולוגיות החלופיות מתחילות מ־“alternative world”. מול “GPU world” כלומר אין כרגע **להיספג אל תוך הארכיטקטורה המרכזית**. [67]

Dataflow-centric מול Memory-centric מול Compute-centric

אפשר לזקק כמעט את כל המאבק הארכיטקטוני לשלוש פילוסופיות.

מאפיין	**Compute-centric**	**Memory-centric**	**Dataflow-centric**
דוגמאות	NVIDIA/AMD GPU, TPU בחלקו	d-Matrix, Qualcomm HBC, EnCharge, PIM/CIM	Cerebras, SambaNova, Groq, Tenstorrent
הנחת יסוד	compute array הוא המרכז; memory hierarchy מזינה אותו	data הוא המרכז; צריך להביא אליו compute	graph/data movement הם המרכז; מקמים operations spatially
Compute density	**גבוהה מאוד**	יכולה להיות גבוהה, אך memory cell/ADC תופסים area	בינונית–גבוהה; תלויה ב־spatial utilization
Raw HBM capacity	**מצוינת**	משתנה; SRAM CIM נמוכה, NMC+DRAM גבוהה	חלשה SRAM-only; tiered designs חזקים יותר
Local bandwidth	גבוהה אך מוגבלת ב־HBM	**פוטנציאלית עצומה**	**עצומה on-chip**
Latency	טובה ב־large batch, פחות ב־streaming	נמצא data מצוינת אם ב־local array	ממופה graph מצוינת כאשר היטב
Energy	יעיל arithmetic; יקר movement	**פוטנציאלית הטובה ביותר**	טובה מאוד בזכות locality
Programmability	**הטובה ביותר**, בעיקר CUDA ecosystem	הקשה ביותר, במיוחד analog/PIM	compiler-dependent; פחות flexible מ־GPU
Scalability	מוכחת עד rack/multi-rack	עדיין לא מוכחת באותו קנה מידה	תלויה מאוד ב־fabric/topology
Yield	chiplets/known-good-die משפרים	3D/CIM מוסיפים process complexity	wafer-scale מגדיל defect exposure; redundancy יכולה לפצות
Precision flexibility	**גבוהה**	digital CIM בינונית; analog נמוכה יותר	בינונית–גבוהה לפי design
Cost	HBM + advanced package יקרים	אך דורש HBM עשוי לחסוך integration חדש	אך memory traffic יכול לחסוך גבוה silicon/fabric area
Best fit ב־2026	training + general AI + large-batch inference	memory-bound inference	deterministic / streaming / spatial inference

Compute-centric ניצח עד היום מסיבה שאסור לזלזל בה: flexibility. Transformer של 2026 אינו של 2022. Formats משתנים, attention variants משתנים, MoE routing משתנה, speculative decoding משתנים, kernels משתנים.

Programmable GPU יכול לקבל kernel חדש מחר; hardwired accelerator לכן. CUDA/ עלול להפוך לפחות רלוונטי software ecosystem הוא נכס ארכיטקטוני כמעט כמו HBM.

החיסרון הוא שה־GPU נאלץ "לשלם" על generality. Dynamic scheduling, large register files, caches, instruction machinery, coherent data movement ו־off-chip HBM נותנים flexibility אך יוצרים distance בין weight לבין MAC.

Memory-centric הופך את היחס. d-Matrix שואלת: למה לשלוח terabytes/s של weights ל־MAC אם אפשר להניח HBM-style memory bandwidth לעבור דרך MAC ליד memory? Qualcomm HBC שואלת: למה בכלל להכריח external interface? EnCharge הולכת הכי רחוק ושואלת: למה להזיז weight אפילו מתוך SRAM cell? [68]

החיסרון הוא שה־memory technology עצמה מתחילה לקבוע את architecture. SRAM יקרה בשטח. DRAM process גרוע ללוגיקה מהירה. Analog CIM מתקשה ב־precision וב־conversion. PIM מחייב partitioning חדש של software stack. אם workload משתנה, קשה יותר "לעדכן firmware" מאשר להוסיף CUDA kernel.

Dataflow-centric מנסה לפתור את הבעיה דרך placement וסדר ביצוע של tensor: הוא operator; לא "חוזר לזיכרון" אחרי כל pipeline across processing element אחד לשני. SambaNova מתארת זאת במפורש כ־operator fusion ו־זורם מ־distributed compute/memory, בעוד Tenstorrent חושפת NoC ו־DMA מפורשים יחסית לתוכנה. [69]

Cerebras לוקחת את dataflow לקצה פיזיקלי: במקום לחבר עשרות dies דרך package/rack, wafer שלם היא משאירה רבים; החיסרון הוא package boundaries ולא on-wafer bandwidth. היתרון הוא כ־single fabric. wafer-scale power/cooling וייחודיות manufacturing. [52]

בפועל, הכיוון הסביר ביותר אינו ניצחון "טהור" של אחת השלוש. **Speculation:** accelerator hybrid של סוף העשור יהיה:

```
programmable tensor compute + massive SRAM + HBM/3D memory + explicit dataflow
engines + near-memory operations + optical scale-up
```

כבר מראה התחלה של Rubin כזה convergence: dual large dies עם NV-HBI, HBM4, Tensor Memory Accelerator ו־3.6 TB/s NVLink. TPU8 מוסיפה מאות MB של SRAM ומנועי collectives. SambaNova משלבת SRAM+HBM+DDR. Qualcomm משלבת compute ו־memory. כלומר הארכיטקטורות מתכנסות סביב **heterogeneous memory hierarchy**, ולא סביב סוג זיכרון יחיד. [70]

התחזית לשנים הקרובות והתשובות המרכזיות

HBM4, HBM4E ו־HBM5

Industry Projection: HBM4 יהיה baseline של high-end accelerator generation לא 28–2026, של endpoint. עם HBM4E samples 4–3.6 TB/s שכבר מכוונים ל־>2.8–3.3 TB/s במוצרי Micron/Samsung ל־stack, aggregate package bandwidth של **20–30+ TB/s** נהפך מציאותי במאיצים גדולים. [71]

HBM5 צפוי להישען פחות רק על "faster DRAM cells" ויותר על **hybrid bonding, higher stacks, custom base dies ו־logic-memory co-design**. SK hynix כבר מציגה hybrid bonding ככיוון ל־HBM4E/HBM5 era ומדברת על מעבר בעתיד. מספרי bandwidth מדויקים ל־HBM5 כדי להציג כעובדה אינם מבוססים מספיק ב־2026 ליותר מ־20 layers. [72]

3D stacked SRAM ־וlogic-on-memory

Research Result → likely Industry Direction: ־ככל שhybrid-bond pitch יורד למיקרונים בודדים, SRAM layer מעל compute 3 מחקרי links 0.88–0.66 מ־2026 מדווחים על D systems־או מתחתיו נעשה מעניין במיוחד. קצרי־טווח עם pJ/bit ועל bandwidth-density־גבוה בהרבה מ־HBM-style interfaces. אלה עדיין תוצאות מחקריות ולא מוצר datacenter, אך הן מראות מדוע “L3 cache on top of tensor die” הוא כיוון סביר מאוד. [73]

Speculation: SRAM/embedded cache של GB עד כמה MB שבהם מאות accelerators בשנים 31–2028 נצפה ליותר נמצאים ב־3D, בעוד HBM נשאר כ־capacity tier. SRAM לא תחליף את DRAM בגלל area/leakage/cost, אך היא עשויה לשנות משמעותית כמה פעמים צריך להגיע ל־HBM.

Wafer-scale computing

Measured today: Cerebras ־מוכיחה שwafer-scale הוא בר־ייצור ובהפעלה, לא רק research. [52]

Speculation: wafer-scale ־לא יהפוך בהכרח לmainstream replacement של packaged GPUs, אבל הרעיון הבסיסי — ימשיך להתפשט package/network boundaries ולמנוע scale-up domain־להגדיל את ה. Multi-reticle interposers, giant packages, panel-scale interconnects ־וoptical fabrics הם למעשה וריאציות על אותו רעיון.

Compute-in-Memory

Speculation: Digital CIM ־צפוי להגיע לdatacenter־מהר יותר מ־full analog CIM, משום שהוא שומר על deterministic digital precision ועל toolchain פשוט יותר. d-Matrix 2026־היא הדוגמה הבולטת לכך ב. Analog CIM, כמו EnCharge, צפוי קודם להצליח ב־edge/client ־וfixed/quantized inference, שם power budget־חשוב יותר מ־BF16 training flexibility. [74]

הפריצה שתוכל לשנות באמת את datacenter 3 תהיה אם D DRAM/SRAM־וCIM יתמזגו: memory die צפוף מעל logic die, עם thousands/millions of vertical connections, ־כך שה־external HBM bus מפסיק להיות boundary. מחקרי monolithic-3D PIM על 2026 כבר חוקרים בדיוק מבנים כאלה. **Research Result / early-stage.** [75]

Optical interconnect לעומת optical compute

Industry Projection: Optical I/O הוא מועמד חזק להפוך mainstream ־בscale-up במהלך 31–2027. הסיבה אינה רק bandwidth; copper מאבד efficiency/reach ־ככל שה־SerDes rates ־וrack dimensions גדלים. Lightmatter, Ayar ־וMarvell/Celestial כבר עברו מ־device demos ל־chiplet/interposer/rack-validation stages. [76]

Speculation: optical compute עצמו לא יחליף GPU/Tensor Cores ־כspecialized באותו חלון זמן. הוא עשוי להופיע כ־precision ־וממשיכים להשאיר את conversion, memory, nonlinearity אבל accelerator, או linear operator ־ל־optical communication של maturity תומך בהבחנה בין silicon photonics קריטית. המחקר הרחב על electronics computing. [77]

יהפוך ליחידת התכנון Rack-scale

זו לדעתי התחזית בעלת הביטחון הגבוה ביותר.

NVIDIA כבר מפרסמת Rubin NVL72 לא כאוסף של GPUs אלא כמערכת עם **20.7 TB HBM4 capacity, 1,580 TB/s aggregate HBM bandwidth ו־260 TB/s NVLink Switch bandwidth**. Google מתכננת TPU pods עם זיכרון

־collectives כחלק מהארכיטקטורה; Qualcomm משווקת Dragonfly ברמת rack; Cerebras CS-4 הוא rack-scale; ־SambaNova ו־Etched מתכננות fabric/memory ברמת rack. [78]

לכן "chip architecture" של 2030 תהיה במידה רבה **rack architecture**. Architect יצטרך לחשוב על:

$$\text{FLOPs} + \text{bytes/token} + \text{HBM placement} + \text{KV placement} + \text{expert placement} + \text{collective topology} + \text{J/b}$$

ולא רק על transistor count או MAC array size.

בשנת 2026 AI מהו צוואר הבקבוק האמיתי של מחשוב?

המסקנה: צוואר הבקבוק האמיתי הוא **Data Supply — היכולת להציב את הנתון הנכון ליד יחידת החישוב הנכונה בזמן הנכון, במינימום bytes ובמינימום joules**.

זה אינו רכיב יחיד.

־Training, large GEMM של הבקבוק עדיין יכול להיות Tensor FLOPS.
־Decode, interactive הוא בדרך כלל HBM bandwidth / weight streaming.
־serving, long-context הוא KV-cache bandwidth + capacity.
־MoE, הוא לעיתים expert-memory capacity + All-to-All network.
־training, large-scale הוא יכול להיות collective communication.
־datacenter, וברמת power budget הופך יותר ויותר ל־הוא של הזזת כל אותם bits. [79]

לכן הייתי מדרג את הבעיה ב־2026 כך:

Data movement > local memory bandwidth/capacity > interconnect > raw arithmetic, עבור חלק הולך וגדל של inference workloads.

הדירוג אינו אוניברסלי; training GEMMs גדולים עדיין יכולים להיות compute-bound. אבל העובדה שמאיץ מסוגל Rubin היא אינדיקציה חזקה לכך שעוד HBM bandwidth של byte/s עבור כל NVFP4 FLOPs תאורטית לבצע ~2,273 FLOPS לבדם כבר אינם התשובה. [6]

או שלב ביניים, AI הוא הארכיטקטורה ארוכת הטווח של GPU + HBM האם?

התשובה שלי: GPU + HBM במובנו של 2020–24 הוא כמעט בוודאות שלב "GPU + HBM" לא עומד להיעלם, אך ביניים.

Speculation — confidence high: ־HBM יישארו עמוד בטווח של 2–5 שנים GPU-class programmable compute משום שאין כרגע טכנולוגיה אחרת שמשלבת באותה רמה programmability, mature software, high precision, HBM capacity, manufacturing scale ־multi-rack ecosystem השדרה של frontier training,

בלבד "GPU + HBM" עצמו הולך להשתנות עד שקשה יהיה לקרוא לארכיטקטורה העתידית GPU אבל ה־.

היא תהיה יותר קרובה ל:

$$\text{Tensor compute} + \text{massive/3D SRAM} + \text{HBM4/HBMx} + \text{near-memory engines} + \text{chiplets} + \text{optical/electrical sc}$$

Rubin כבר כולל dual reticle-scale dies, NV-HBI, HBM4, TMA 3.6־ו TB/s per-GPU NVLink. TPU8 משלב מאות MB SRAM עם HBM־וcollective acceleration. SambaNova משלבת SRAM/HBM/DDR. Qualcomm HBC מזיזה compute לתוך memory neighborhood. d-Matrix מזיזה MAC אל memory. Lightmatter/Ayar/Marvell את מזיזות package/rack boundary אל optical domain. כל אחד מהם תוקף **bit מרחק אחר במסלול של אותו**. [80]

לכן השאלה הארכיטקטונית של סוף העשור כנראה לא תהיה "GPU או PIM?" אלא:

> **state כמה עובר אל הזיכרון, כמה programmable tensor compute, כמה מן החישוב נשאר rack ואיזה חלק ממערכת הזיכרון הופך למעשה לרשת אופטית של 3D SRAM/HBM, נשמר ב־שלם?**

אינו "קיר" שבסופו של דבר נפרוץ ונשכח ממנו. הוא כוח שמכריח את המחשב להשתנות. תחילה הוא הביא Memory Wall־ה caches, אחר כך HBM 2.5־ו 3D packaging; עכשיו הוא מביא 3D stacking, CIM, dataflow, quantization־ו optical scale-up. ככל שה־MAC עצמו נעשה זול יותר, **המיקום של הנתונים הופך בהדרגה לדבר היקר ביותר בארכיטקטורה**. [81]

---

1 79 Efficient LLM Inference: Bandwidth, Compute ...
https://arxiv.org/html/2507.14397v1?utm_source=chatgpt.com

2 A Modern Primer on Processing-In-Memory
https://arxiv.org/html/2012.03112v5?utm_source=chatgpt.com

3 5 6 40 78 Rack-Scale Agentic AI Supercomputer | NVIDIA Vera Rubin NVL72
https://www.nvidia.com/en-eu/data-center/vera-rubin-nvl72/

4 9 32 Inside NVIDIA Blackwell Ultra: The Chip Powering the AI Factory Era | NVIDIA Technical Blog
https://developer.nvidia.com/blog/inside-nvidia-blackwell-ultra-the-chip-powering-the-ai-factory-era/

7 AMD Instinct™ MI350 Series GPUs
https://www.amd.com/en/products/accelerators/instinct/mi350.html

8 TPU 8t and TPU 8i technical deep dive | Google Cloud Blog
https://cloud.google.com/blog/products/compute/tpu-8t-and-tpu-8i-technical-deep-dive

10 1.2. Programming Model — CUDA Programming Guide
https://docs.nvidia.com/cuda/cuda-programming-guide/01-introduction/programming-model.html?utm_source=chatgpt.com

11 3.2. Advanced Kernel Programming — CUDA ...
https://docs.nvidia.com/cuda/cuda-programming-guide/03-advanced/advanced-kernel-programming.html?utm_source=chatgpt.com

12 2.3. Writing SIMT Kernels — CUDA Programming Guide
https://docs.nvidia.com/cuda/cuda-programming-guide/02-basics/writing-cuda-kernels.html?utm_source=chatgpt.com

13 CUDA C++ Programming Guide (Legacy)
https://docs.nvidia.com/cuda/cuda-c-programming-guide/?utm_source=chatgpt.com

14 81 Solving computing's memory problem
https://arxiv.org/html/2505.00458v2?utm_source=chatgpt.com

15 Co-Designing AI Model Attention for Fast, Interactive Long- ...
https://developer.nvidia.com/blog/co-designing-ai-model-attention-for-fast-interactive-long-context-inference/?utm_source=chatgpt.com

16 FlashAttention-3: Fast and Accurate Attention with Asynchrony and Low-precision
https://arxiv.org/abs/2407.08608?utm_source=chatgpt.com

17 Bringing HBM3 to RNGD: Challenges and Benefits
https://furiosa.ai/blog/bringing-hbm3-to-rngd-key-challenges-and-benefits?utm_source=chatgpt.com

18 Q-Hitter: A Better Token Oracle for Efficient LLM Inference ...
https://proceedings.mlsys.org/paper_files/paper/2024/hash/bbb7506579431a85861a05fff048d3e1-Abstract-Conference.html?utm_source=chatgpt.com

19 Efficient Memory Management for Large Language Model ...
https://dl.acm.org/doi/10.1145/3600006.3613165?utm_source=chatgpt.com

20 46 DisagMoE: Computation-Communication overlapped MoE ...
https://arxiv.org/html/2605.11005?utm_source=chatgpt.com

21 The Industry is Striving for Custom Memory (Part 2)
https://www.eetimes.com/beyond-bandwidth-the-industry-is-striving-for-custom-memory-part-2/?utm_source=chatgpt.com

22 73 A Full-Stack Performance Evaluation Infrastructure for 3D- ...
https://arxiv.org/html/2604.08044v1?utm_source=chatgpt.com

23 NVIDIA Grace Hopper Superchip Architecture In-Depth
https://developer.nvidia.com/blog/nvidia-grace-hopper-superchip-architecture-in-depth/?utm_source=chatgpt.com

24 High-Bandwidth and Energy-Efficient Memory Interfaces for ...
https://www.researchgate.net/publication/383971948_High-Bandwidth_and_Energy-Efficient_Memory_Interfaces_for_the_Data-Centric_Era_Recent_Advances_Design_Challenges_and_Future_Prospects?utm_source=chatgpt.com

25 48 60 76 Passage M1000 EVK - 3D Photonic Interconnect for AI ...
https://lightmatter.co/products/passage-m1000-evk/?utm_source=chatgpt.com

26 51 77 Roadmapping the next generation of silicon photonics
https://www.nature.com/articles/s41467-024-44750-0?utm_source=chatgpt.com

27 The Memory Wall Has Three Exits - by Anni Sen
https://senanni.substack.com/p/the-memory-wall-has-three-exits?utm_source=chatgpt.com

28 Nvidia supplier SK Hynix begins mass production of next generation memory chip
https://www.reuters.com/technology/nvidia-supplier-sk-hynix-begins-mass-production-next-generation-memory-chip-2024-03-19/?utm_source=chatgpt.com

29 71 Micron in High-Volume Production of HBM4 Designed for ...
https://investors.micron.com/news/press-release/2026/Micron-in-High-Volume-Production-of-HBM4-Designed-for-NVIDIA-Vera-Rubin-PCIe-Gen6-SSD-and-SOCAMM2-03-16-2026/default.aspx?utm_source=chatgpt.com

30 Samsung Electronics Begins Shipment of Industry-First ...
https://news.samsung.com/global/samsung-electronics-begins-shipment-of-industry-first-hbm4e-samples?utm_source=chatgpt.com

31 From HBM to eSSD: How SK hynix is charting the future of ...
https://news.skhynix.com/en/hbm-to-essd/?utm_source=chatgpt.com

33 CoWoS® - Taiwan Semiconductor Manufacturing ...
https://3dfabric.tsmc.com/english/dedicatedFoundry/technology/cowos.htm?utm_source=chatgpt.com

34 TSMC-SoIC® - Taiwan Semiconductor Manufacturing ...
https://3dfabric.tsmc.com/english/dedicatedFoundry/technology/SoIC.htm?utm_source=chatgpt.com

35 Data Center AI Inference Accelerators | Dragonfly
https://www.qualcomm.com/data-center/expertise/ai-accelerators?utm_source=chatgpt.com

36 Session 14 Overview: Compute-in-Memory
https://ieeexplore.ieee.org/iel8/10904417/10904496/10904651.pdf?utm_source=chatgpt.com

37 Special Topic on Energy-Efficient In-/Near-Memory ...
https://ieeexplore.ieee.org/iel8/6570653/10879154/11308150.pdf?utm_source=chatgpt.com

38 News
https://www.enchargeai.com/category/news?utm_source=chatgpt.com

39 68 74 d-Matrix Corsair AI Platform | In-Memory Computing for AI
https://www.d-matrix.ai/product/?utm_source=chatgpt.com

41 AWQ: Activation-aware Weight Quantization for On-Device ...
https://proceedings.mlsys.org/paper_files/paper/2024/file/42a452cbafa9dd64e9ba4aa95cc1ef21-Paper-Conference.pdf?utm_campaign=news&utm_medium=miragenews&utm_source=chatgpt.com

42 Low-Bit Quantization for Efficient and Accurate LLM Serving
https://proceedings.mlsys.org/paper_files/paper/2024/file/5edb57c05c81d04beb716ef1d542fe9e-Paper-Conference.pdf?utm_source=chatgpt.com

43 FlashAttention-3: fast and accurate attention with ...
https://dl.acm.org/doi/10.5555/3737916.3740109?utm_source=chatgpt.com

44 70 80 Inside NVIDIA Rubin GPU Architecture: Powering the Era of Agentic AI | NVIDIA Technical Blog
https://developer.nvidia.com/blog/inside-nvidia-rubin-gpu-architecture-powering-the-era-of-agentic-ai/

45 An Overview of the CXL 3.X Specification
https://computeexpresslink.org/webinars/an-overview-of-the-cxl-3-x-specification-3942/?utm_source=chatgpt.com

47 How NVIDIA GB200 NVL72 and NVIDIA Dynamo Boost ...
https://developer.nvidia.com/blog/how-nvidia-gb200-nvl72-and-nvidia-dynamo-boost-inference-performance-for-moe-models/?utm_source=chatgpt.com

49 Ayar Labs Unveils World's First UCIe Optical Chiplet for AI ...
https://ayarlabs.com/news/ayar-labs-unveils-worlds-first-ucie-optical-chiplet-for-ai-scale-up-architectures/?utm_source=chatgpt.com

50 Marvell to Acquire Celestial AI, Accelerating Scale-up ...
https://investor.marvell.com/news-events/press-releases/detail/1000/marvell-to-acquire-celestial-ai-accelerating-scale-up-connectivity-for-next-generation-data-centers?utm_source=chatgpt.com

52 53 Product - System
https://www.cerebras.ai/cs4?utm_source=chatgpt.com

54 What is a Language Processing Unit?
https://groq.com/blog/the-groq-lpu-explained?utm_source=chatgpt.com

55 RDU | Next-Gen AI Chip for Inference at Scale
https://sambanova.ai/products/rdu-ai-chips?utm_source=chatgpt.com

56 Introducing the SN50 RDU: Purpose-Built for Agentic ...
https://sambanova.ai/blog/introducing-the-sn50-rdu-purpose-built-for-agentic-inference?utm_source=chatgpt.com

57 58 Cards
https://tenstorrent.com/hardware/cards?utm_source=chatgpt.com

59 Etched
https://www.etched.com/?utm_source=chatgpt.com

61 May 28, 2026 - 10-Q: Quarterly report [Sections 13 or 15(d)]
https://investor.marvell.com/sec-filings/all-sec-filings/content/0001835632-26-000019/mrvl-20260502.htm?utm_source=chatgpt.com

62 FuriosaAI RNGD - Furiosa Docs
https://developer.furiosa.ai/v2025.2.0/en/overview/rngd.html?utm_source=chatgpt.com

63 Serving GPT-OSS 120B at 5.8 ms TPOT with RNGD
https://furiosa.ai/blog/serving-gpt-oss-120b-at-5-8-ms-tpot-with-two-rngd-cards-compiler-optimizations-in-practice?utm_source=chatgpt.com

64 Positron | Atlas
https://www.positron.ai/atlas?utm_source=chatgpt.com

65 AI inference at scale with AI250 rack-scale Platform
https://www.qualcomm.com/data-center/products/qualcomm-dragonfly-ai250?utm_source=chatgpt.com

66 Near-Memory Computing and the AI Memory Wall
https://www.qualcomm.com/news/onq/2026/07/qualcomm-hbc-near-memory-computing?utm_source=chatgpt.com

67 Groq and Nvidia Enter Non-Exclusive Inference ...
https://groq.com/newsroom/groq-and-nvidia-enter-non-exclusive-inference-technology-licensing-agreement-to-accelerate-ai-inference-at-global-scale?utm_source=chatgpt.com

69 The Decode Era of AI: Why Dataflow Matters More Than Ever
https://sambanova.ai/blog/why-dataflow-matters-more-than-ever?utm_source=chatgpt.com

72 [Tech Note] Hybrid Bonding: Evolving into a Foundational ...
https://news.skhynix.com/en/tech-note-series-ep2/?utm_source=chatgpt.com

75 PALUTE: Processing-In-Memory Acceleration via Lookup ...
https://arxiv.org/html/2606.08891v1?utm_source=chatgpt.com