

The Hidden Energy Cost of Scaling AI Hardware: A Comparative Analysis of Manufacturing and Operational Energy

Emre Salman

emre.salman@stonybrook.edu

Stony Brook University

Abrar Abdurrob

Stony Brook University <https://orcid.org/0009-0009-6510-4517>

Analysis

Keywords: AI hardware, lifecycle energy, high-bandwidth memory, manufacturing energy, sustainable computing

Posted Date: July 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10120690/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

The Hidden Energy Cost of Scaling AI Hardware: A Comparative Analysis of Manufacturing and Operational Energy

Abrar Abdurrob¹ and Emre Salman^{1*}

¹Department of Electrical and Computer Engineering, Stony Brook
University (SUNY), Stony Brook, New York, 11794, USA.

*Corresponding author(s). E-mail(s): emre.salman@stonybrook.edu;

Abstract

Artificial intelligence infrastructure is scaling rapidly, yet assessments of its environmental footprint have focused almost exclusively on operational electricity consumption. Here we show that this framing omits a large and growing contribution: the energy embodied in manufacturing the hardware itself. Analyzing datacenter AI accelerators, we find that high-bandwidth memory (HBM) fabrication alone accounts for 55–71% of total manufacturing energy because sustaining AI throughput requires stacking dozens of memory dies per device. Projecting vendor roadmaps through 2035, manufacturing energy per accelerator is estimated to grow six- to nine-fold as HBM scales toward 800 GB–1.5 TB. The architectural advances that reduce operational energy per token, like precision quantization, attention compression, and sparse mixture-of-expert activation act on compute and memory-bandwidth utilization rather than on memory capacity, leaving manufacturing energy unaffected by the same innovations that improve operational efficiency. These results demonstrate that hardware sustainability assessments should account explicitly for embodied manufacturing energy alongside operational consumption.

Keywords: AI hardware, lifecycle energy, high-bandwidth memory, manufacturing energy, sustainable computing

1 Introduction

The computational demands of artificial intelligence are increasing at an extraordinary rate. Training compute for frontier AI models has grown by approximately four to five orders of magnitude since 2010, doubling roughly every six months, a rate that outpaces Moore’s Law itself [1, 2]. The infrastructure required to sustain this growth has reached large scales, with training clusters for frontier models now integrating hundreds of thousands of GPUs with aggregate power demands measured in hundreds of megawatts, power equivalent to that of small cities [3]. Looking ahead, the International Energy Agency estimates that global data center electricity consumption could more than double by 2030, driven largely by AI workloads [4].

Given these trends, researchers have focused on quantifying the energy consumed during model training and inference [3, 5], but this captures only part of the lifecycle picture. Modern semiconductor fabrication is extraordinarily energy-intensive. A leading edge fab draws power approaching one gigawatt when fully operational, and the energy required to process a single wafer has increased roughly $3.5\times$ between the 28 nm and 3 nm nodes [6, 7]. This escalation reflects the compounding complexity of advanced transistor architectures, additional metallization layers, and the shift toward multi-patterning and extreme ultraviolet (EUV) lithography. Despite its magnitude, manufacturing energy has received comparatively little attention in the AI sustainability literature. Existing lifecycle analyses include tool-level carbon models such as ECO-CHIP [8] and 3D-Carbon [9], and vendor-reported product carbon footprints such as NVIDIA’s ISO-aligned HGX H100 disclosure [10] do not examine how manufacturing energy scales across hardware generations with the memory-centric architectural trends now defining frontier AI accelerators.

Modern datacenter accelerators integrate large numbers of high-bandwidth memory (HBM) stacks to sustain the memory bandwidth required by large neural network

workloads. The NVIDIA H100, for example, integrates five 8-high HBM3 stacks comprising 45 total dies (40 DRAM and 5 base logic dies), while next-generation platforms such as the NVIDIA Rubin are expected to integrate eight or more 12-high stacks. This stacked memory architecture has a direct manufacturing consequence as each additional memory die contributes fabrication energy, and the aggregate memory die count now substantially exceeds the logic die count on a typical accelerator package. At the same time, sparse architectural innovations, such as mixture-of-experts models, are rapidly becoming the dominant paradigm for frontier AI, improving operational efficiency per token without reducing total memory capacity requirements. However, this gap between manufacturing and operational energy trajectories has not been previously examined.

Here, we present a lifecycle energy analysis of AI accelerators that quantifies both manufacturing and operational contributions across current and projected hardware generations. We show that HBM fabrication accounts for 55–71% of total manufacturing energy in current accelerators, and that this manufacturing footprint is projected to grow six- to nine-fold by 2035 as HBM scales toward 800 GB–1.5 TB per accelerator. We further show that architectural advances improving operational efficiency, such as mixture-of-experts sparsity [11, 12], attention compression [13, 14], and precision reduction [15, 16], act on compute and memory-bandwidth utilization rather than on memory capacity, creating a divergence between the two phases of the life cycle. These findings suggest that hardware sustainability assessments should explicitly account for embodied manufacturing energy and that architectural strategies targeting the memory bottleneck directly may offer lifecycle energy benefits beyond operational efficiency alone. The energy estimation framework, device parameters, and projection assumptions underlying these results are described in Methods.

To quantify the full lifecycle energy of AI accelerators, we analyzed manufacturing and operational contributions across a representative set of current datacenter

accelerators and mobile SoCs, and projected these trends through 2035 using vendor roadmaps and published semiconductor scaling data. The following subsections present these findings, beginning with manufacturing energy and its dominant memory component, followed by operational energy, lifecycle comparisons across platforms, and finally the divergence between manufacturing and operational energy trajectories driven by sparse architectural trends.

2 Results

2.1 HBM fabrication dominates accelerator manufacturing energy

Applying the wafer-level scaling and per-die allocation framework (equations 1–4, Methods), the total manufacturing energy of the NVIDIA H100 SXM is estimated at approximately 146 kWh per device. Logic fabrication on TSMC 4N (5 nm-class) contributes approximately 60 kWh, while the five 8-high HBM3 memory stacks, comprising 45 total stack dies (40 DRAM + 5 base dies), contribute approximately 80 kWh, nearly $1.3\times$ the logic die contribution [10, 17]. Advanced 2.5D packaging via TSMC Chip-on-Wafer-on-Substrate(CoWoS) contributes approximately 5 kWh, modest relative to wafer fabrication [8–10]. Although HBM is fabricated at a less advanced node (10 nm-class) than the logic die, memory dominates manufacturing energy through sheer die count as sustaining the bandwidth demands of modern AI workloads requires stacking large numbers of DRAM dies per accelerator.

Table 1 extends this analysis to a wider set of contemporary datacenter accelerators and mobile SoCs. Among datacenter devices, the total manufacturing energy varies primarily with on-package memory capacity rather than with the logic technology node. Accelerators with larger HBM configurations, such as the AMD Instinct MI300X [18] and the NVIDIA H200 [19], exhibit substantially higher manufacturing energy than the H100 despite similar-generation logic nodes, confirming the dominant contribution

of memory fabrication. The differences in the logic fabrication energy across datacenter GPUs are relatively modest, reflecting both shared 5 nm-class processes and the secondary contribution of logic area relative to the stacked memory.

Mobile SoCs present a markedly different profile, falling in the range of approximately 7–11 kWh per device despite employing equal or more advanced logic nodes. Mobile dies are substantially smaller than datacenter accelerators, improving yield and producing more usable chips per wafer. Mobile devices also integrate minimal external memory, while datacenter accelerators depend on large HBM stacks to sustain throughput, decisively shifting manufacturing energy toward memory-dominated scaling.

2.2 Operational energy across device classes

The operational energy framework (equations 5–17, Methods) is applied to representative devices across both device classes. For the NVIDIA H100 SXM, the 700 W board-level power envelope is decomposed into compute, HBM, and interconnect subsystems. Using vendor-reported HBM3 interface energy efficiencies of 1.7–3.4 pJ/bit [17, 24], peak HBM power is estimated at 45–90 W. NVLink 4 interconnects contribute approximately 5–10 W, and the compute core accounts for the remaining 600–650 W [10]. Under realistic operating conditions, empirical measurements on H100 systems show that active AI training workloads draw 70–80% of rated thermal design power (TDP) on average [25, 26], while inference workloads, which account for the majority of operational compute cycles in deployed AI systems [27, 28], draw lower average power due to the predominance of memory bandwidth-bound token generation over compute-intensive phases [27]. Although these empirical measurements are reported at the node level, they are applicable to per-device estimation as GPUs account for the majority of node power during active AI workloads [27, 29]. Applying the operational energy model (equation 16, Methods) with activity factor $\alpha = 0.6$ –0.8

Table 1 Manufacturing energy for representative datacenter accelerators and mobile systems-on-chip (SoCs). All values are in kWh per device, estimated using the wafer-level scaling and per-die allocation framework (Equations 1–4, Methods §4.3–4.5) with $D_0 = 0.10$ defects/cm² and $\alpha = 2$. HBM3E die area is taken as 121 mm² [20]; HBM3 DRAM die area is 107 mm² and HBM2e die area is 100 mm², consistent with published 16 Gb die densities [17, 21]; HBM base logic die and LPDDR5X die area are each estimated at 50 mm² [21]. Packaging contributes ~ 5 kWh (CoWoS-S, datacenter) and ~ 0.4 kWh (InFO-PoP, mobile) per device, included in Total [8]. Entries marked $\dagger$ include at least one estimated die dimension; see footnotes.

Device	Form factor	Logic node(s)	On-package memory	Logic (kWh)	Memory (kWh)	Total (kWh)	Mem. share
NVIDIA H100 SXM	SXM	5 nm-class (TSMC 4N)	HBM3 (80 GB, 5 $\times$ 8-Hi)	~ 60	~ 80	~ 146	55%
NVIDIA H200 SXM	SXM	5 nm-class (TSMC 4N)	HBM3E (141 GB, 6 $\times$ 8-Hi)	~ 60	~ 110	~ 176	63%
Google TPU v5p †	Module	5 nm-class (TSMC N5, undisclosed)	HBM3 (95 GB, 4 $\times$ 12-Hi)	~ 34	~ 95	~ 134	71%
Intel Gaudi 3 †	OAM	5 nm-class (TSMC N5)	HBM2e (128 GB, 8 $\times$ 8-Hi)	~ 54	~ 119	~ 178	67%
AMD Instinct MI300X	OAM	N5 (XCDs) + N6 (IODs)	HBM3 (192 GB, 8 $\times$ 12-Hi)	~ 83	~ 189	~ 277	68%
Apple A17 Pro †	SoC	3 nm-class (TSMC N3B)	LPDDR5X (8 GB, PoP)	~ 4	~ 3	~ 8	41%
Qualcomm Snapdragon 8 Gen 3	SoC	4 nm-class (TSMC N4P)	LPDDR5X (8–16 GB, PoP)	~ 4	~ 3 –7	~ 7 –11	45%

† **TPU v5p**: Google does not disclose TPU silicon dimensions [22]. Logic die area estimated at ~ 580 mm² (TSMC N5) by scaling from the TPU v4 generation to N5 process density with functional additions for the v5p generation. HBM dominance ($>60\%$) holds across the full plausible die area range (450–700 mm²).

† **Gaudi 3**: Intel does not disclose compute die dimensions. Die area estimated at ~ 495 mm² per die (TSMC N5), derived from the Gaudi 2 predecessor scaled for process node and compute additions (± 80 mm²; ± 15 kWh on total).

† **A17 Pro**: Apple does not disclose die dimensions [23]. Logic die area estimated at ~ 108 mm² (TSMC N3B) from the published transistor count at representative design utilization; consistent with the reported logic energy of ~ 4 kWh.

(fraction of peak power drawn during active computation, reflecting inference-weighted to training-dominated operation), utilization $U = 0.5\text{--}0.85$ (fraction of device lifetime spent actively executing computation, reduced from 100% due to scheduling gaps between jobs, hardware failures and recovery, maintenance, and diurnal demand variation in inference-serving deployments, spanning shared GPU clusters [30] to dedicated AI training deployments [31]), and idle power fraction $\beta \approx 0.1$ (true device idle, transistor leakage and HBM refresh with no active workload, as distinct from within-job reduced-activity states, which are captured by α), the effective operational power ranges from approximately 245–487 W (35–70% of peak). During a three-year service lifetime (26,280 h), the cumulative operational energy at the device-level is approximately 6.4–12.8 MWh, increasing to 9.7–20.5 MWh at the facility level after accounting for datacenter infrastructure overhead through a power usage effectiveness (PUE) of 1.5–1.6, which means 50–60% additional energy beyond the IT equipment load is consumed by cooling, power conversion and other facility systems [5, 32].

For mobile platforms, the Apple A17 Pro SoC is evaluated using a representative duty cycle of 5 hours active and 19 hours connected-standby per day, derived from large-scale telemetry and environmental assessments [23, 33]. Sustained compute power of 0.3–1.0 W during active use [34, 35] and connected-standby power of 0.1–0.2 W [34, 36] yield a daily operational energy of approximately 3–9 Wh and a cumulative three-year total of approximately 4–8 kWh per device, three orders of magnitude below that of a datacenter accelerator (Table 3).

2.3 Lifecycle energy comparison across platforms

Figure 1 compares manufacturing and operational energy across both device classes. For datacenter accelerators, the operational energy dominates at the device level. The operating energy at the facility-level for three-years varies between approximately 10 and 29 MWh per accelerator (Fig. 1(b)), exceeding the manufacturing energy by

Table 2 Operational (compute-phase) energy comparison of representative datacenter accelerators. Facility operational energy is computed over a three-year service lifetime and incorporates datacenter infrastructure overhead through the power usage effectiveness (PUE) metric. Peak power corresponds to the vendor-specified board-level power envelope (TDP or TBP), except for TPU v5p, for which only measured mean chip power has been publicly disclosed. Dynamic memory and interconnect power components are derived from vendor-reported bandwidth specifications and technology-specific energy-per-bit parameters, as defined in Methods. HBM2e entries use $e_{\text{bit}} = 3.4\text{--}4.3$ pJ/bit reflecting the higher I/O energy of the older PHY generation[17]; HBM3/HBM3E entries use $e_{\text{bit}} = 1.7\text{--}3.4$ pJ/bit.

Accelerator	Power Basis	Peak / Mean Power (W)	Memory (type, TB/s)	Interconnect (GB/s)	HBM Dyn. Power (W)	IO Dyn. Power (W)	Facility Energy (MWh, 3 yr)
NVIDIA H100 SXM	Vendor TDP	700	HBM3, 3.35	NVLink 4, 900	46–91	3.6–7.2	9.7–20.5
NVIDIA H200 SXM	Vendor TDP	700	HBM3E, 4.80	NVLink 4, 900	65–131	3.6–7.2	9.7–20.5
AMD Instinct MI350X	Vendor TBP	1000	HBM3E, 8.00	Infinity Fabric, 1075	109–218	4.3–8.6	13.8–29.2
Intel Gaudi 3	Vendor TDP	900	HBM2e, 3.70	24×200 GbE, 600	101–127	2.4–4.8	12.4–26.3
Google TPU v5p	Measured mean	331 [†]	HBM2e, 2.77	ICI, 1200	75–95	4.8–9.6	13.1–13.9

[†]Google does not publicly disclose a TPU v5p TDP; the reported value corresponds to first-party measured mean chip power (excluding host), which directly maps to P_{eff} in our framework[22]. [‡]H100 and H200 share a 700 W TDP; identical facility-energy ranges reflect equal device-level power draw. Efficiency differences between generations are quantified on a per-token basis in Table 5.

Table 3 Operational energy summary of representative flagship mobile SoCs.

Three-year operational energy from the duty-cycle model (Section 4.12). SoC compute power is isolated from display and cellular subsystems using per-rail ODPM measurements[34, 35].

Metric	A17 Pro	Snapdragon 8 Gen 3
Class	iPhone 15 Pro	Flagship Android
Process node	TSMC N3B	TSMC N4P
P_{active} (W)	0.3–1.0	0.4–1.2
P_{standby} (W)	0.1–0.2	0.1–0.3
Daily E_{IT} (Wh)	3.4–8.8	3.9–11.7
E_{IT} (3 yr, kWh)	4–10	4–13

more than one order of magnitude. Deployment-phase utilization and power efficiency are therefore the primary determinants of total lifecycle energy for an individual accelerator.

Manufacturing energy exhibits a different scaling behavior. Logic fabrication contributes within a range of approximately 34–83 kWh across contemporary accelerators despite substantial differences in peak performance (Fig. 1(a)). In contrast, on-package HBM fabrication contributes 80–189 kWh per device, strongly scaling with memory capacity and stack count. Accelerators manufactured at comparable logic nodes but configured with different memory capacities therefore differ in manufacturing energy by up to $2\times$, confirming that memory is the dominant contributor to embodied fabrication energy. Operational energy follows the opposite trend as compute activity dominates, while HBM and interconnect dynamics contribute smaller fractions (Fig. 1(b)).

At deployment scale, manufacturing energy becomes increasingly consequential. Although modest per-device, manufacturing energy scales linearly with fleet size with a deployment of 10,000 accelerators accumulating approximately 1.5 GWh in manufacturing energy. Architectural decisions affecting memory capacity and packaging complexity can therefore materially influence embodied energy at fleet scale.

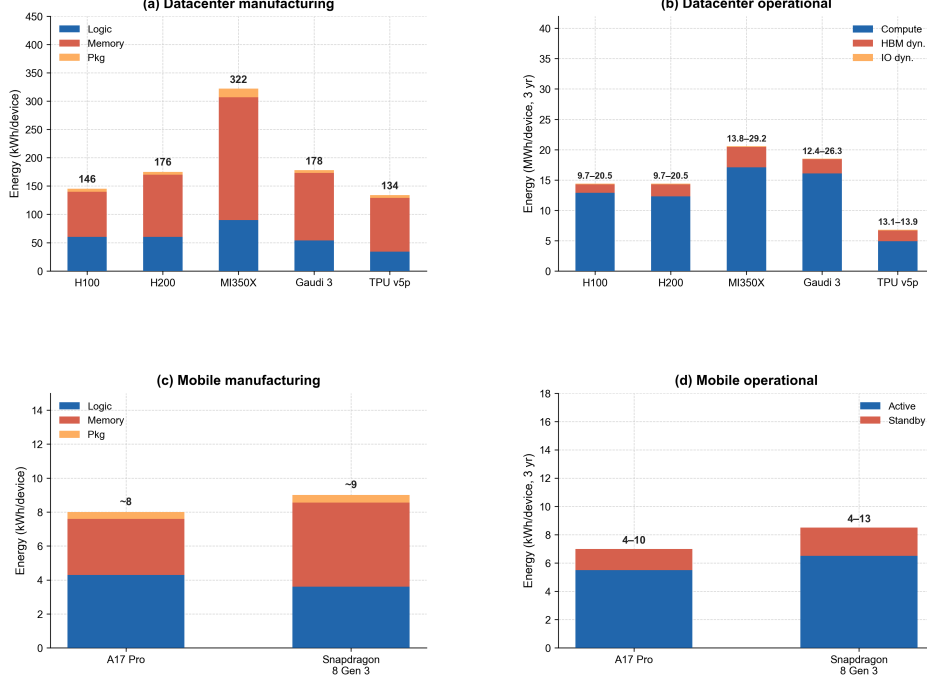


Fig. 1 Manufacturing and operational energy breakdown across datacenter accelerators and mobile systems-on-chip (SoCs). (a) Manufacturing energy for representative datacenter accelerators, decomposed into logic fabrication, on-package memory fabrication, and advanced packaging. (b) Three-year facility operational energy of datacenter accelerators, partitioned into compute (residual), high-bandwidth memory (HBM) dynamic, and interconnect (IO) dynamic contributions using relative power-share allocation; labels indicate the facility-energy ranges reported in Table 2. (c) Manufacturing energy for representative mobile SoCs, shown on a per-device basis. (d) Three-year operational energy of mobile SoCs, decomposed into active and standby components derived from the duty-cycle model. Despite differing absolute scales and units across panels, all values correspond to total energy accumulated over a common three-year service lifetime, enabling direct comparison between manufacturing and operational phases across platforms.

Mobile SoCs exhibit a fundamentally different lifecycle profile. The manufacturing energy of 7–11 kWh per device (Fig. 1(c)) and the three-year operational energy of 4–13 kWh per device (Fig. 1(d)) are of comparable magnitude, a near-parity that stands in sharp contrast to datacenter accelerators, where operational energy exceeds manufacturing energy by two orders of magnitude. For mobile platforms, extending service lifetime and moderating replacement rates therefore offer lifecycle energy leverage comparable to operational efficiency improvements.

2.4 Manufacturing energy outlook through 2035

Three concurrent scaling trends, EUV-driven logic node transitions, aggressive growth of HBM capacity, and increasingly complex advanced packaging, are projected to drive the manufacturing energy per accelerator from approximately 176 kWh today to 1,141–1,660 kWh by 2035 (Fig. 2a), a six- to nine-fold increase. Logic fabrication is projected to grow $\sim 8\text{--}10\times$ through continued node-scaling intensity and multi-chiplet die multiplication, HBM by $\sim 5\text{--}9\times$ from stack-height and die-count scaling, and packaging by $\sim 13\text{--}16\times$ due to expansion of the multi-reticle interposer. These projections are grounded in vendor roadmaps, JEDEC standards, published die specifications, and TSMC packaging disclosures [37–40] (component growth factors shown in Fig. 2b).

Escalating lithography intensity drives logic fabrication energy

EUV lithography has already elevated wafer-level manufacturing energy, with empirical life-cycle analyses showing a two- to three-fold increase in wafer energy when transitioning from 28 nm to 7 nm/5 nm technologies [6, 7]. The subsequent progression from 5 nm to 3 nm to 2 nm further intensifies this trend, as a growing fraction of back-end-of-line (BEOL) layers requires EUV exposure in place of multi-patterning deep ultraviolet (DUV) processing.

IEDM-reported measurements indicate that current 0.33-NA EUV scanners draw approximately 1–1.2 MW and consume 6–8 kWh per exposure at high throughput [41]. The IRDS lithography roadmap projects approximately 6–8 EUV layers at 7 nm, 10–12 at 5 nm, and 18–20 at 3 nm, rising to 22–25 layers for 2 nm logic [40]. Sub-2 nm nodes are projected to require 26–28 EUV layers, unless the selective adoption of High-NA EUV alleviates some double-patterning requirements [40].

High-NA EUV (0.55 NA), exemplified by ASML’s EXE:5200B system, entered production for sub-2 nm logic in 2025 [42]. These tools draw 1.4–1.5 MW and require

a higher dose to maintain stochastic performance, resulting in 7–10 kWh per exposure even while reducing double patterning. Combining the measured exposure energy with the projections of the IRDS layer-count and the Power-Performance-Area-Cost-Environmental (PPACE) scaling [6], the wafer-level manufacturing energy is projected to reach approximately 3–3.5 MWh per wafer for 2 nm nodes and 3.5–4.5 MWh per wafer for sub-2 nm nodes by 2035 [40, 43].

Concurrently, accelerator architectures are transitioning toward multi-chiplet integration, increasing total silicon area even when individual dies remain near reticle limits. NVIDIA’s Rubin GPU (2026) integrates two compute dies along with dedicated I/O chiplets on TSMC N3P (3 nm-class enhanced performance node)[44]. AMD’s MI350X employs eight XCD (Accelerator Complex Die) compute chiplets on TSMC N3P plus two I/O dies on TSMC N6, with confirmed XCD die area of 110 mm² [45, 46]. Although chiplet architectures improve yield economics by avoiding monolithic reticle-scale dies, they do not reduce total manufacturing energy and may instead increase it: the total silicon area in multi-chiplet packages often exceeds that of monolithic predecessors (e.g., H100 at 814 mm²), while each successive node consumes more wafer energy. Combined with continued node advancement through 2 nm (2025–2026), A16-class (2026–2028), and sub-1 nm (~2030), logic fabrication energy per accelerator is projected to grow from ~60 kWh (H200) to ~481–618 kWh by 2035, an approximately 8–10× increase driven by both node scaling and continued die-count growth [6, 40].

HBM die-count scaling dominates embodied energy growth

HBM remains the dominant contributor to per-accelerator manufacturing energy, and its absolute share is projected to grow substantially over the next decade. Unlike logic fabrication, where process improvements can reduce the energy per transistor, historical life-cycle assessments of DRAM demonstrate that the manufacturing energy per die increases with node advancement due to greater process complexity, additional patterning steps, and higher aspect ratio capacitor structures [21, 47]. This

trend is reinforced by the PPACE framework, which quantifies a monotonic increase in wafer-level energy with node scaling [6]. It is important to distinguish this from the well-documented improvement in *operational* power efficiency per bit, for example, Micron’s 1β -to- 1γ transition delivers over 20% lower chip power [48, 49], which reflects reduced switching energy in the fabricated device, not a reduction in the energy consumed during its manufacture.

All major DRAM vendors now ship advanced HBM3E parts: Samsung and Micron offer 24 GB (8-high) and 36 GB (12-high) stacks exceeding 1 TB/s per stack, while SK Hynix provides similar configurations [24, 50, 51]. Current accelerators integrate six to eight stacks, which corresponds to 141–288 GB of total HBM capacity. The NVIDIA H200 employs six 8-high HBM3E stacks totaling 141 GB [19], while the AMD MI350X and NVIDIA Rubin each integrate eight 12-high stacks delivering 288 GB [44, 45].

Roadmaps for HBM4-class products further expand this trajectory. SK Hynix targets HBM4 mass production in 2026, transitioning to a 2048-bit interface with a multi-terabyte per-second bandwidth per stack [52]. Micron has confirmed sampling of 12-high, 36 GB HBM4 parts [53]. Roadmaps extending through HBM5 and HBM5E target 16-high configurations in the 2028–2031 window, with hybrid bonding confirmed for HBM5 generation to enable stack heights beyond what microbump technology permits [54, 55].

The dominant driver of HBM manufacturing energy growth is die-count scaling. The H200 baseline of 54 dies (6 stacks $\times$ 9 dies per stack) at approximately 2.04 kWh per die yields approximately 110 kWh of HBM manufacturing energy. As accelerator configurations scale toward 8–12 stacks of 16–20-high HBM5-class devices by 2035, total HBM die counts reach 128–240 per package. Accounting for a modest 5–20% increase in the per-die energy estimated through the PPACE s-factor scaling applied to projected DRAM node advancement [6, 21, 47], HBM manufacturing energy is projected to reach 595–962 kWh by 2035, a 5–9 $\times$ increase from the H200 baseline,

consistent with the announced JEDEC HBM4 specifications and vendor stack-height roadmaps [20, 38, 53, 54].

HBM remains the largest single contributor to manufacturing energy throughout the projection window, accounting for 49–67% of the total across all scenarios (Fig. 2(c) and Table 4). The temporary reduction in the fractional share of HBM in the 2028–2030 horizon (49–51%) reflects the compounding effect of simultaneous advancement of logic nodes and multiplication of dies at A16-class nodes, before the scaling of the HBM stack-height reasserts memory dominance by 2035 (52–58%).

Advanced packaging amplifies manufacturing energy at system level

Architectural trends in deployed accelerators place stacked DRAM at the center of future manufacturing energy growth. Recent generations of AI systems from Google, NVIDIA, and AMD co-locate multi-die compute clusters with multiple HBM stacks on large 2.5D interposers, optimizing for reduced data-movement cost rather than peak floating-point throughput [10, 22, 58, 59]. Profiling studies of transformer workloads confirm that memory traffic dominates both latency and energy consumption, and that the memory subsystem increasingly governs system-level efficiency [58, 60].

These memory-centric architectures have direct manufacturing implications. TSMC’s CoWoS roadmap now supports interposers exceeding three stitched reticles in size ($>2,500\text{ mm}^2$), with projections approaching $5,000\text{ mm}^2$ using stitched or bridge-based integration schemes [39, 57, 61]. Because advanced packaging involves wafer-level lithography, metallization, and repeated bonding operations, its manufacturing energy scales approximately with interposer area and integration complexity.

Current accelerators such as the H200 employ interposers of approximately 800 mm^2 [61], contributing $\sim 5\text{ kWh}$ to the total manufacturing energy [8–10]. AMD’s MI350X continues the multi-chiplet CoWoS architecture established with the MI300X [62, 63], one of the most integration-intensive commercial AI accelerator packages to date [45]. As multi-reticle interposers become standard, packaging energy is

Table 4 Manufacturing energy projection for AI accelerators showing logic, HBM, and packaging components. Baseline (H200) calibrated to Table 1. Logic projections follow the IRDS node timeline[40] and the multi-chiplet scaling trend already established across current datacenter accelerators[22, 44, 56]: TSMC N3P in production (2025), A16-class targeted late 2026, sub-1 nm $\sim$ 2030, Å-class $\sim$ 2032+[40, 43]; logic energy scales with node via PPACE s-factors[6] and continued die-count growth. HBM projections are grounded in JEDEC HBM4 specifications[38], the SK Hynix AI Summit 2025 roadmap (HBM5 16-Hi targeting 2028–2031; HBM5E/HBM6 beyond)[53, 54], and published analyses of hybrid-bonding adoption for the HBM5 generation[55]. Per-die energy scaled conservatively via PPACE[6, 21, 47]. Packaging projected from CoWoS interposer area scaling[39, 57]. The $\sim$ 2028–2030 and $\sim$ 2035 rows present scenario ranges spanning conservative and aggressive chiplet and HBM scaling trajectories consistent with current multi-vendor trends.

Platform	Year	Logic Node	HBM Configuration			Manufacturing Energy (kWh)				
			Stacks	Height	Capacity	Logic	Memory	Pkg	Total	Mem %
H200	2024	4N (5 nm)	6	8-Hi	141 GB	60	110	5	176	63%
MI350X [†]	2025	N3P + N6	8	12-Hi	288 GB	90	217	15	322	67%
Rubin [†]	2026	N3P	8	12-Hi	288 GB	158	257	25	440	58%
$\sim$ 2028–2030 [‡]	2028–30	A16-class	8	16–20-Hi	$\sim$ 512–640 GB	390	419–455	40–45	848–890	49–51%
$\sim$ 2035 [‡]	2035	sub-1 nm/Å	10–12	16–20-Hi	800 GB–1.5 TB	481–618	595–962	65–80	1,141–1,660	52–58%
Growth (H200 $\rightarrow$ 2035)						$\sim$8–10$\times$	$\sim$5–9$\times$	$\sim$13–16$\times$	$\sim$6–9$\times$	

[†] **MI350X**: XCD die area confirmed at 110 mm² (TSMC N3P) from ISSCC 2026 [46]; IOD area estimated at $\sim$ 600 mm² (TSMC N6) — AMD does not disclose. Sensitivity: $\pm$ 15 kWh over the 550–650 mm² IOD range. **Rubin**: NVIDIA does not disclose die dimensions. Two compute dies estimated at $\sim$ 750 mm² (TSMC N3P) from published transistor counts and process density[44]; two IO dies estimated at $\sim$ 400 mm² (TSMC N6). HBM4 die area estimated at $\sim$ 140 mm² from published 36 GB 12-Hi stack capacity[53] and 10 nm-class die density[38].

[‡] **Projection scenarios**. Logic energy assumes continued multi-chiplet scaling (2–4 compute dies per package) consistent with the trajectory established across current datacenter accelerators from multiple vendors[22, 44, 56], combined with node advancement per the IRDS roadmap[6, 40]. *2028–2030*: Both bounds reflect HBM5, which targets the 2028–2031 window per the SK Hynix roadmap[54]; lower bound uses HBM5 16-Hi with microbump stacking; upper bound uses HBM5 20-Hi with hybrid bonding (confirmed for the HBM5 generation[55]). *2035*: Lower bound — sub-1 nm node ($\sim$ 2030)[40], 10 stacks $\times$ 16-Hi HBM5E[54]; upper bound — Å-class node ($\sim$ 2032+)[40], 12 stacks $\times$ 20-Hi HBM6[54]. Per-die HBM energy increases 5–20% above HBM3E baseline consistent with continued DRAM node advancement[6, 21, 47].

projected to increase from ~ 5 kWh to $\sim 65\text{--}80$ kWh by 2035, a $\sim 13\text{--}16\times$ increase that raises packaging’s share from roughly 3% to $\sim 5\text{--}6\%$ of total manufacturing energy. This projection follows the approximate proportionality between the packaging energy and the interposer area, combined with the CoWoS area scaling trajectory from $\sim 800\text{ mm}^2$ to $\sim 5,000\text{ mm}^2$ by 2035 [8, 39, 57].

Even under optimistic adoption of architectural mitigations such as processing-in-memory and CXL-based memory pooling (discussed in the Discussion), manufacturing energy per accelerator is projected to reach 1,141–1,660 kWh by 2035, a constraint that cannot be addressed through operational efficiency improvements alone.

2.5 Operational energy outlook: sparse architectures and inference scaling

The next phase of AI development is likely to be defined by sparse architectures, inference-time scaling, and increasingly memory-centric system design, trends that alter the trajectory of operational energy growth and create direct links to the manufacturing trends established above.

Mixture-of-experts architectures reduce operational energy without reducing memory capacity

The compute scaling paradigm that drove AI progress from 2017–2024 is approaching structural limits. Kaplan *et al.* established power-law relationships showing that model error decreases predictably as compute budget, parameter count, and training data volume increases [2], and the subsequent Chinchilla analysis showed that for a given compute budget, the optimal balance between model size and training data corresponds to approximately 20 training tokens per parameter [65]. In large language models, text is processed as *tokens*, discrete units typically representing 3–4 characters or common subwords. Under these scaling laws, a 70-billion parameter

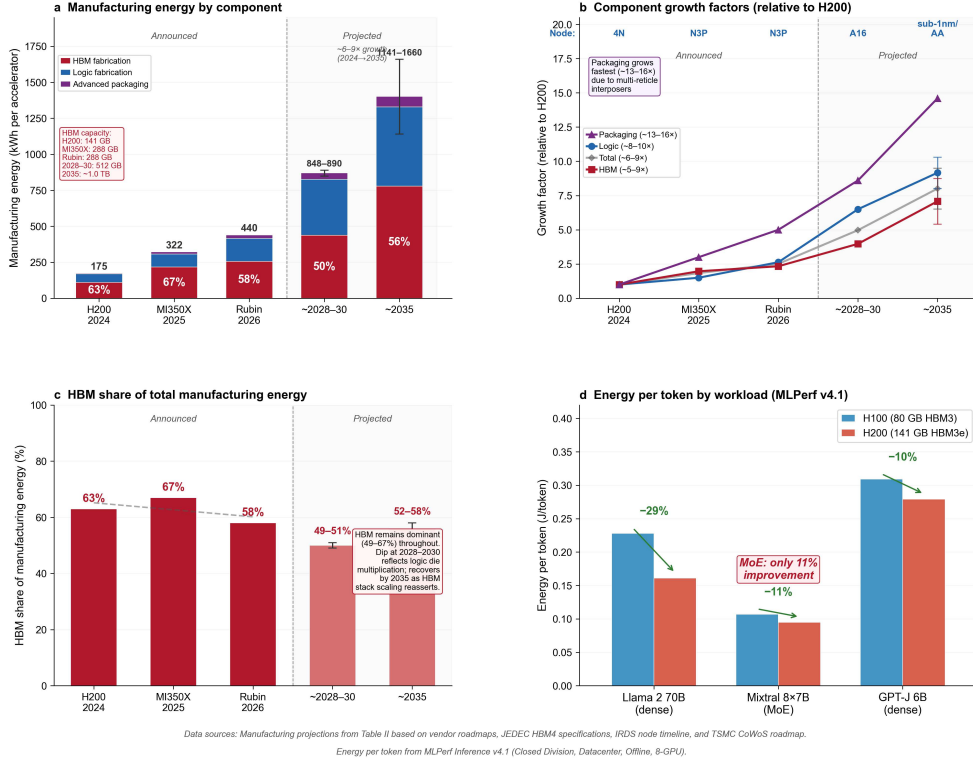


Fig. 2 Manufacturing and operational energy outlook for AI accelerators (2024–2035). (a) Manufacturing energy breakdown by component. H200, MI350X, and Rubin values are confirmed; 2028–2030 and 2035 rows are scenario ranges (Table 4). Percentages indicate HBM share of total. (b) Component growth factors relative to H200. Error bars indicate scenario bounds for the 2035 projection. (c) HBM share of total manufacturing energy across the projection window (49–67%); HBM remains the largest single contributor throughout. (d) Energy per token (MLPerf v4.1 [64]), H100 vs. H200. The H100→H200 transition isolates the effect of the HBM upgrade alone; the logic die is unchanged between generations. Data: Tables 4 and 2; manufacturing projections grounded in vendor roadmaps [38–40, 44, 54].

model requires approximately 1.4 trillion tokens for compute-optimal training, while a 500-billion parameter model requires nearly 10 trillion tokens. However, these relationships implicitly assume unlimited high-quality training data. Villalobos *et al.* estimate that the effective stock of publicly available high-quality text, approximately 3×10^{14} tokens after filtering, may be exhausted between 2026 and 2032 at current consumption rates [66]. Contemporary frontier models are already approaching this constraint: Meta’s Llama 3.1 405B reportedly trained on approximately 15 trillion tokens [66, 67],

neering practical limits of available high-quality data. Beyond this threshold, simply increasing training compute yields diminishing returns, marking an inflection relative to the rapid scaling trajectory observed over the past decade.

The industry response has been twofold. First, a transition toward *inference-time scaling*: Instead of solely increasing parameter counts, emerging reasoning models generate extended intermediate computations before producing final outputs, effectively reallocating computational demand from training to inference [68, 69]. Such models may process 10–100× more tokens per query than conventional architectures, shifting workload balance from training-dominated to inference-dominated regimes, where memory capacity constraints and therefore manufacturing energy become increasingly consequential. Second, the adoption of sparse architectures reduces active compute per token while maintaining large overall parameter counts, as discussed below.

Mixture-of-Experts (MoE) architectures have emerged as the leading approach for frontier-scale models. Unlike conventional dense models, where all parameters participate in every forward pass, MoE architectures partition parameters into parallel subnetworks, or *experts*. For each input, a learned routing mechanism activates only a small subset of experts while the remainder remain idle, enabling very large effective model capacity with substantially reduced active computation per token.

DeepSeek-V3 exemplifies this approach: the model contains 671 billion total parameters distributed across 256 expert subnetworks, yet only 37 billion parameters (approximately 5.5%) are activated for any given input [69]. This sparse activation reduces the effective compute required per token and reportedly lowered total training cost to approximately 2.788 million GPU hours, roughly an order of magnitude less than comparably sized dense models.

However, MoE architectures do not reduce memory capacity requirements. All expert weights must reside in HBM to enable low-latency routing and activation, even when most experts remain idle. A 671-billion parameter MoE model therefore requires

the same HBM capacity as a dense model of equivalent size, despite activating only a fraction of parameters per operation. This creates a distinctive operational profile: compute energy scales with active parameters (approximately 5–10% of the dense model equivalent), while memory capacity scales with total parameters. Consequently, memory bandwidth utilization decreases as fewer bytes are transferred per operation, reducing the benefit of HBM bandwidth improvements. As a result, DeepSeek-V3 achieves energy-per-token comparable to that of a 37-billion parameter dense model despite its 671-billion parameter capacity [69].

MoE architectures improve operational efficiency per token relative to a dense model of the equivalent total parameter count. However, for manufacturing energy, this gain operates on a different lever as it reduces active compute without reducing the total HBM footprint. As frontier models continue scaling in total parameter count even with MoE sparsity keeping active compute manageable, HBM capacity requirements, and therefore manufacturing energy, continue to grow.

Attention compression reduces memory bandwidth demand

Beyond sparsity, architectural innovations are targeting memory bottlenecks within the attention mechanism, the core operation that enables models to relate different elements of an input sequence. During text generation, models produce outputs one token at a time while referencing all previously generated tokens. To avoid redundant computation, implementations cache intermediate values, specifically, the key and value vectors used in attention operations. This *KV cache* grows linearly with sequence length: for each token in context, the model stores two vectors per attention layer. In large models, the KV cache becomes a dominant consumer of HBM capacity. Llama 3.1 405B with conventional grouped-query attention, for example, requires 516 KB of KV cache storage per token in context [70], meaning a 128,000-token context window requires approximately 66 GB of HBM solely for cached attention states.

DeepSeek’s Multi-head Latent Attention (MLA) addresses this bottleneck through low-rank compression of cached values. Instead of storing full key-value pairs, MLA stores a compressed latent representation that can be expanded on-demand during attention computation, reducing per-token cache storage from 516 KB to just 70 KB, a $7.4\times$ reduction [69, 70]. The same 128,000-token context now requires only 9 GB of cache storage. This compression enables longer context windows within fixed HBM capacity budgets, higher batch sizes within existing memory constraints, and reduced memory traffic per generated token, shifting inference workloads from memory-bounded toward more compute-bounded regimes and improving effective hardware utilization.

Empirical energy-per-token measurements across workload types

To quantify how these architectural trends affect operational energy, MLPerf Inference v4.1 benchmark data [64] were analyzed across three representative workloads: Llama 2 70B (memory-bandwidth-limited dense model), Mixtral 8×7 B (MoE architecture with 46.7B total parameters and 12.9B active parameters), and GPT-J 6B (compute-limited small dense model). All measurements correspond to 8-GPU configurations evaluated under the Offline throughput scenario [64]. The energy per generated token is calculated from the measured system throughput and the rated TDP (equation 12, Methods).

Table 5 and Fig. 2(d) presents results comparing NVIDIA H100 and H200 accelerators. The H200 provides 76% more HBM capacity (141 vs. 80 GB) and 43% higher memory bandwidth (4.8 vs. 3.35 TB/s) compared to the H100, yet energy efficiency improvements vary substantially with workload characteristics. For memory bandwidth-limited workloads (Llama 2 70B), a 30% efficiency improvement closely follows the 43% increase in memory bandwidth, confirming that HBM bandwidth dominates energy efficiency for large dense models. For MoE workloads (Mixtral 8×7 B), energy efficiency improves by only 11% despite substantial improvements in HBM,

because sparse activation inherently reduces bandwidth utilization, leaving additional HBM capacity and bandwidth underutilized. For compute-limited workloads (GPT-J 6B), efficiency improves by approximately 10%, as small dense models are limited by arithmetic throughput rather than external memory bandwidth.

Table 5 Energy per token across GPU generations and workload types. Values derived from MLPerf Inference v4.1 throughput results (Closed Division, Datacenter, Offline scenario, 8-GPU systems) and rated TDP. Submission IDs: 4.1-0043 (H100), 4.1-0048 (H200) [64].

Model	Active Params	H100 (J/token)	H200 (J/token)	Efficiency Gain	Limiting Factor
Llama 2 70B	70B (dense)	0.228	0.161	30%	Memory BW
Mixtral 8×7B	12.9B (MoE)	0.107	0.095	11%	Mixed
GPT-J 6B	6B (dense)	0.309	0.279	10%	Compute

Projecting forward to 2028–2030-era accelerators with approximately 512–640 GB HBM capacity and substantially higher bandwidth, memory-bandwidth-limited workloads could approach approximately 0.06 J/token, representing a $\sim 2.5\times$ improvement relative to H200. For MoE-dominated inference workloads, efficiency gains are expected to be substantially smaller due to intrinsic bandwidth underutilization. The architectural levers that drive operational efficiency, such as precision scaling, attention compression, software optimization, and bandwidth-utilization improvements, operate on different variables from those that drive manufacturing energy, which is governed by HBM die count, stack height, logic node complexity, and packaging interposer area (Section 2.4). This divergence between the architectural drivers of the two energy regimes determines the lifecycle trajectory of AI accelerators, as examined in the following subsection.

2.6 The manufacturing–operational energy gap

The architectural trends described above reveal a divergence between manufacturing and operational energy that has not previously been examined in the context of sparse

AI architectures: the same innovations that reduce operational energy per token leave manufacturing energy unaffected, and in some cases accelerate its growth.

Manufacturing energy scales with total HBM capacity. HBM fabrication accounts for 49–67% of accelerator manufacturing energy across the projection window and increases with die count and stack height (Table 4). Although MoE architectures activate only a subset of parameters per token, all expert weights must reside in HBM to enable low-latency routing. Similarly, extended-context inference requires additional memory capacity for attention caches. Together, these factors ensure that manufacturing energy tracks the total HBM capacity, following the projected growth of $6\text{--}9\times$ through 2035 regardless of architectural sparsity.

Operational energy, by contrast, scales with memory bandwidth utilization. MoE sparsity and attention compression reduce the energy per token by decreasing the number of bytes accessed per operation, without reducing HBM capacity requirements. Sparse models still require dense weight storage, and compressed attention mechanisms still consume memory proportional to context length. Manufacturing energy is consumed to fabricate dies that hold the full set of parameters; the operational efficiency benefit accrues only to the subset of dies accessed per forward pass. The same HBM subsystem that dominates the embodied manufacturing energy simultaneously enables operational efficiency gains, but these two effects respond to entirely different architectural levers. Although MoE improves operational efficiency per token relative to a dense model of equivalent total parameter count, this gain does not reduce the HBM footprint, and as frontier models continue scaling in total parameters, both HBM capacity requirements and manufacturing energy continue to grow.

The H100→H200 transition illustrates this divergence: manufacturing energy increases by approximately 21% (146→176 kWh, Table 1), as the HBM capacity increases from 80 to 141 GB, while the operational efficiency of the MoE improves by only 11%, despite a 43% increase in the HBM bandwidth, reflecting how sparse

activation underutilizes the added bandwidth. Manufacturing energy and operational efficiency respond to different properties of the HBM subsystem, capacity and per-bit bandwidth, which scale through different engineering levers. A 2028–2030 era accelerator carries a projected manufacturing energy of 848–890 kWh (Table 4), a $5\times$ increase relative to H200, driven almost entirely by the growth of the HBM die-count and the packaging-area. Operational innovations in precision, attention compression, or sparsity do not reduce this committed embodied cost, because none of these mechanisms acts on capacity. The divergence between manufacturing and operational energy is therefore structural, since the two quantities are driven by independent architectural variables, and no single hardware or software choice addresses both.

3 Discussion

The results presented here establish an asymmetry in AI hardware scaling that has not previously been quantified in the context of sparse AI architectures. Specifically, architectural advances that reduce operational energy per token are largely orthogonal to the forces driving manufacturing energy growth. HBM capacity is the dominant determinant of manufacturing energy and is exactly what MoE architectures leave unchanged. Each generation of more efficient AI hardware is therefore simultaneously more energy-intensive to manufacture, a gap that compounds with every hardware cycle and is not captured by current operational metrics.

3.1 The irreversibility asymmetry

Operational and manufacturing energy differ in a property with important consequences for lifecycle sustainability: Operational energy can be reduced after deployment through software optimization, model compression, and future hardware generations, but manufacturing energy cannot as it is committed at fabrication and

locked in for the life of the device regardless of how efficiently the hardware is subsequently operated, how long it runs, or how capable the models it serves become.

The scale of this embodied debt is growing rapidly. The manufacturing energy per accelerator is projected to reach 1,141–1,660 kWh by 2035, a $6\text{--}9\times$ increase from the H200 baseline. Operational efficiency over the same window is bounded by the saturation of the mechanisms that drive it: precision has progressed through FP16→FP8→FP4 but FP2 is not viable for most production workloads, sparse MoE architectures are limited by the arithmetic intensity of active-expert compute, and software optimization compresses toward the workload-defined performance floor. Operational and manufacturing energies also sit at different points in the lifecycle. Operational energy is re-determined every time the system runs and responds to software, runtime, and workload choices. Manufacturing energy is set once at fabrication and cannot be reduced at deployment. At fleet scale, where hundreds of thousands of accelerators are refreshed on two to three year cycles [3], this divergence accumulates into a material and irreversible lifecycle energy burden.

3.2 Why MoE does not close the gap

MoE lowers operational energy per token by activating only a subset of expert parameters per forward pass, the 11% H100→H200 efficiency gain observed for Mixtral $8\times 7\text{B}$ in Table 5 reflects this directly. The savings do not extend to manufacturing because all expert weights must reside in HBM simultaneously to enable low-latency routing. A 671-billion parameter MoE model requires the same memory footprint as a dense model of equivalent size. As frontier models continue scaling total parameter counts, even as per-token sparsity keeps active compute tractable, HBM requirements grow, and manufacturing energy grows with them.

3.3 The demand rebound

Efficiency gains in computing have historically not reduced total energy consumption. More efficient hardware enables more computation per watt, unlocking new applications, larger models, and higher throughput, pushing utilization back towards the thermal design power envelope [3]. Datacenter accelerators operate near their power ceiling regardless of per-token efficiency, a rebound effect, often termed Jevons’ paradox, that has held across every prior generation of compute scaling [71].

The consequences for manufacturing energy are direct and compounding. If per-device operational consumption is bounded by the power envelope, then growing aggregate demand can only be met by deploying additional hardware. Inference chip sales are projected to increase 4–6 $\times$ over the next five to eight years, driven by reasoning models, MoE architectures, multimodal generation, and retrieval-augmented generation, all of which expand per-query memory requirements [72]. Each additional accelerator carries irreversible embodied energy, and per-device manufacturing energy is itself rising with fabrication complexity (Table 4). Fleet size grows with demand, while per-unit manufacturing energy grows with hardware complexity, producing a multiplicative increase that operational efficiency cannot offset.

Per-token efficiency gains remain essential, as they determine economic viability and the range of applications a given power budget can serve. However, they should not be mistaken for lifecycle energy reductions. The mitigation pathways discussed in the following are selected because they reduce either the total hardware deployed or the manufacturing energy per device.

3.4 HBM is overprovisioned for inference

The manufacturing energy burden of HBM is compounded by a fundamental mismatch between HBM’s design characteristics and the access patterns of LLM inference.

HBM inherits DRAM’s symmetric read/write performance, byte addressability, practically unlimited endurance, and sub-100 ns random access latency. LLM inference uses almost none of these capabilities. During autoregressive decoding, weights are read-only and KV-cache access is append-only with sequential reads, measured read-to-write ratios exceed 1000:1 [73]. Memory accesses are large, sequential, and predictable, byte-level random access is not utilized [72, 73]. The endurance of HBM ($\sim 10^{15}$ write cycles) exceeds the inference requirements by six to nine orders of magnitude for weights and by three to six orders of magnitude for KV-cache writes [73].

Each additional capability, such as the symmetric write path, the random-access circuitry, and the engineering margin for high endurance, represents transistors, process steps, and manufacturing energy that does not serve an important functional purpose for the dominant AI workload. Approximately one-third of an AI accelerator’s operational energy is consumed by HBM [73], and the manufacturing complexity required to deliver these largely unused capabilities contributes to HBM’s lower yield and higher cost relative to commodity DRAM [72–74]. A significant fraction of HBM’s embodied energy is, therefore, committed to capabilities that inference does not require.

3.5 HBM scaling is approaching physical limits

The projections in Table 4 assume continued growth in HBM capacity, and recent industry data confirm that this growth is becoming increasingly difficult to sustain at the individual stack level. The DRAM density scaling has decelerated. The fourfold increase from 8-gigabit dies introduced in 2014 to 32-gigabit dies, which began commercial sampling in 2024, took roughly a decade, while previous generations achieved fourfold gains every three to six years [72]. HBM4, which entered mass production in early 2026, increases the capacity per-layer by only $\sim 30\%$ relative to HBM3e [73]. The number of vertically stacked DRAM dies within a single HBM package, the primary

mechanism to increase the capacity per stack, is not expected to scale beyond 16 due to yield, thermal, and mechanical constraints [73].

This per-stack ceiling is what drives the scaling trajectory in our projections. If neither die density nor die count per stack can grow substantially, future capacity increases must come from adding more HBM stacks per accelerator, more accelerators per system, or more systems per cluster, the pathway reflected in Table 4’s projection of up to 20 stacks per package by 2035. Each of these pathways multiplies the manufacturing energy rather than amortizing it across denser components. Cost data confirm this trajectory, as the normalized cost of HBM3e per gigabyte and per gigabyte-per-second both increased $1.35\times$ from 2023 to 2025 [72, 74], while commodity DDR4 costs fell to $0.54\times$ and $0.45\times$ during the same period [72]. The rise in HBM cost per unit of capacity reflects the increased manufacturing complexity, with more process steps, advanced packaging, and lower yields, all of which translate into higher embodied energy per gigabyte.

Both Cerebras (wafer-scale SRAM) and Groq (full-reticle SRAM) attempted to eliminate HBM entirely, but in both cases, the demands for LLM capacity overwhelmed the SRAM on the chip, forcing the retrofitting of external DRAM [72]. The dependence on HBM is locked in for frontier models, and the manufacturing energy it carries is not an avoidable cost.

3.6 Implications for sustainability reporting

The current environmental reporting frameworks for the AI infrastructure are structured almost exclusively around operational electricity consumption [4, 75]. NVIDIA’s ISO 14067-compliant product carbon footprint disclosures for the H100 and H200 quantify embodied manufacturing emissions at the product level [10], and the Google TPU lifecycle analysis [76] provides a rigorous cradle-to-gate methodology. Yet AI sustainability benchmarks, datacenter energy disclosure requirements under emerging

E.U. and U.S. regulatory frameworks, and industry self-reporting all default to operational energy as the primary metric [4, 77, 78]. Manufacturing energy is systematically absent from all of these frameworks. Recent work by Google has explicitly identified the manufacturing yield and the lifecycle of the device as first-order determinants of embodied carbon, calling for metrics that treat CO₂e and the total cost of ownership as primary optimization targets along with performance [79]. The analysis in this paper supports this direction.

This omission has an important consequence. Fleet decisions that appear sustainable by operational metrics can increase total lifecycle energy. Accelerating hardware refresh cycles to deploy more energy-efficient accelerators reduces operational energy per token but commits additional manufacturing energy with every replacement, an embodied rebound effect that is invisible to operational-only accounting [75, 76]. For mobile platforms, where our results show that manufacturing and operational energy are already comparable in magnitude, both in the range of 4–13 kWh over a three-year device lifetime (Table 3), device longevity is as powerful a sustainability lever as chip efficiency, a finding that gives longer replacement cycles a significant and previously unquantified energy value. Extending lifecycle energy accounting, by following established Life-Cycle-Assessment (LCA) methodology [75, 76, 80], to the fleet and procurement level would make these tradeoffs visible.

3.7 Mitigation pathways

Narrowing the gap between manufacturing and operational energy trajectories requires interventions that reduce the total volume of HBM fabricated per unit of AI capability, not merely improvements in per-token compute efficiency. The demand rebound and HBM overprovisioning described above constrain which approaches are effective: Processing-in-Memory (PIM), Processing-near-Memory (PNM) and MoE improve

operational efficiency but do not reduce total memory deployed, leaving manufacturing energy unchanged. The pathways below target manufacturing energy directly.

High Bandwidth Flash for capacity displacement. High Bandwidth Flash (HBF) stacks flash dies in an HBM-like package to deliver comparable bandwidth at $10\times$ the capacity per stack [72, 81]. A single HBF stack provides ~ 512 GB versus ~ 48 GB for HBM4, while matching the read bandwidth of HBM4 at lower power [72]. That bandwidth is achieved through parallel operation across numerous dies connected to the wide interface, rather than through rapid access to any single die. NAND page reads still take microseconds, while DRAM row access is in nanoseconds. Model weights, frozen during inference and thus immune to flash write-endurance limits, are ideal candidates for HBF storage. Replacing even a portion of the weight-serving HBM with HBF would reduce the number of 3D-stacked DRAM dies fabricated, packaged, and tested per system. Flash density also continues to double every three years, unlike DRAM [72]. Open challenges include software management of page-based read latency, optimal HBM-to-HBF ratios, and whether HBF manufacturing energy per gigabyte is sufficiently lower than HBM to deliver net lifecycle savings. HBF cannot replace all HBM, as KV-cache data, updated every token, still requires DRAM, but it could reshape the capacity scaling trajectory for weight storage.

Managed-Retention Memory for simpler fabrication. Managed-Retention Memory (MRM) trades the long term retention and symmetric write performance that HBM inherits from DRAM for improved read throughput, density, and energy efficiency [73]. By relaxing retention to hours or days, sufficient for weights periodically refreshed from persistent storage, MRM enables memory cell technologies (phase-change memory (PCM) [82], resistive RAM (RRAM) [83], and spin-transfer torque magnetic RAM (STT-MRAM) [84]) that historically failed to displace DRAM because they were forced to compete on non-volatility. These technologies offer read performance on par with DRAM at a potentially higher density and lower cost per

terabyte [73]. For manufacturing energy analysis, the critical property is that MRM does not require 3D DRAM stacking and its yield penalties, and fabrication could use less complex packaging, reducing embodied energy per gigabyte. There are no production MRM devices, as it was only formalized in 2025 as a cross-layer research agenda rather than a finished design [73]. Existing PCM, RRAM, and STT-MRAM devices were designed for non-volatility, rather than write performance, endurance, and density that MRM requires and the proposed memory controllers, dynamic retention control, and software stack remain undeveloped. Even if MRM does not reach production, any memory built without long retention or symmetric write performance would be less manufacturing-intensive per gigabyte than HBM.

CXL-based memory pooling. CXL enables HBM capacity to be shared across multiple accelerators via external memory pools, reducing the total installed HBM required across a fleet. Early deployments demonstrate 50–70% reductions in total installed memory capacity for batch inference workloads [85], a fleet-level manufacturing energy reduction that per-device metrics cannot capture. The remaining challenge is latency, as CXL access times are acceptable for weight storage in offline inference, but remain prohibitive for KV-cache access patterns for latency-sensitive service [72].

Memory footprint-aware model design. If manufacturing energy tracks HBM capacity, then model architectures that minimize total weight and KV-cache footprint, not merely active compute per token, reduce both operational and manufacturing energy simultaneously. DeepSeek’s Multi-head Latent Attention achieves $7.4\times$ KV-cache compression without accuracy loss [69, 70]. Whether structured sparsity, aggressive quantization, or fundamentally different paradigms can achieve large effective weight capacity without proportionally large memory footprints remains an open question.

A note on operational efficiency. Processing in or near memory and related architectural innovations remain valuable for achieving lower energy per operation,

higher throughput within a given power envelope, and economic viability of inference at scale. Samsung’s HBM-PIM demonstrated up to $2.5\times$ performance at lower energy [86, 87]. Processing-in-memory also avoids software sharding constraints by placing the computation on separate dies [58, 72]. These are important advances, but they do not reduce total HBM capacity as a system needs the same memory to hold the same parameters regardless of where computation occurs [72]. Their contribution is to operational sustainability. Treating per-operation efficiency as equivalent to overall resource reduction masks the underlying issues in manufacturing energy use.

The broader principle

The results presented in this paper are not specific to HBM or AI accelerators. In complex semiconductor systems, the energy embedded in manufacturing and the energy consumed during operation are fundamentally decoupled. They respond to different architectural variables, scale at different rates, and require different interventions. This decoupling will expand as EUV layer counts increase, chiplet integration becomes standard, and memory stack heights climb. AI accelerators are where it is currently largest and fastest-growing, but the principle applies wherever hardware complexity scales faster than operational efficiency.

The framework presented here, which quantifies manufacturing energy from wafer-level parameters, allocates it across components, and projects it against operational trends, is transferable to other hardware domains. In general, it is highly important to develop a measurement standard that treats manufacturing energy as a first-class metric alongside operational consumption, a position now supported by independent proposals for evaluation frameworks that include embodied CO₂e as a primary optimization target [72, 79]. Progress measured only in performance-per-watt will continue to undercount the true energy cost of AI scaling. A complete accounting must track both the energy a system consumes and the energy it costs to build it.

4 Methods

This section describes the analytical framework used to estimate manufacturing and operational energy across current and projected AI accelerator platforms. Manufacturing energy is estimated from wafer-level energy data, process node scaling models, and published life-cycle assessments. Operational energy is modeled using subsystem-level power decomposition, vendor-reported specifications, and empirical characterization data from production datacenter deployments. A three-year service lifetime is assumed across all device categories to enable direct comparison between manufacturing and operational energy contributions.

4.1 Manufacturing Energy Framework

Manufacturing energy is defined here as the electrical energy consumed during the fabrication of a functional integrated circuit, encompassing wafer processing, memory fabrication, and advanced packaging. The extraction of raw materials, chemical supply chains, and the energy embodied in equipment are outside the scope of this analysis. The system boundary includes: (i) front-end wafer fabrication, encompassing front-end-of-line (FEOL), middle-of-line (MOL), and back-end-of-line (BEOL) processing [88], (ii) wafer sort and yield losses [89], (iii) advanced packaging technologies (e.g., silicon interposers [61], Embedded Multi-die Interconnect Bridge (EMIB)/Foveros [90], Chip-on-Wafer-on-Substrate (CoWoS) [39, 61]), (iv) on-package memory fabrication, including HBM for GPUs and TPUs [17, 38], and Double Data Rate/Low-Power Double Data Rate (DRAM/LPDDR) for mobile SoCs [91, 92]. This system boundary is consistent with the fab-gate scope adopted in prior semiconductor energy analyses [7, 8, 21].

4.2 Wafer Processing Energy

Semiconductor fabrication dominates manufacturing energy [7]. According to TSMC’s 2024 Environmental, Social, and Governance (ESG) disclosures, the company consumed a total of 27,477 GWh of electricity and shipped 12.9 million 12-inch wafers, corresponding to an average energy intensity of 2,130 kWh per wafer across all process nodes [93]. This figure is consistent with TSMC’s own unit product energy metrics, reporting approximately 37 kWh of electricity per wafer per mask layer, which for 55–60 mask layers in advanced logic yields an estimated 2,000–2,200 kWh per wafer [93].

However, wafer energy scales strongly with process node. The IMEC PPACE analysis quantified that the energy per wafer rises from ~ 1 kWh/cm² at 28 nm to ~ 3.5 kWh/cm² at 3 nm [6]. For a 300 mm wafer, the energy at 28 nm is ~ 700 kWh per wafer and at 3 nm is $\sim 2,470$ kWh per wafer, a $3.5\times$ increase in wafer fabrication energy with scaling. The node-specific wafer energy can be modeled as

$$E_{\text{wafer,node}} = s(n) \times E_{\text{wafer,avg}}, \quad (1)$$

where $E_{\text{wafer,avg}}$ is the foundry average ($\approx 2,130$ kWh/wafer) and $s(n)$ is a node scaling factor (e.g., $s(28 \text{ nm}) \approx 0.33$, $s(5 \text{ nm}) \approx 0.89\text{--}0.93$, $s(3 \text{ nm}) \approx 1.09\text{--}1.16$) [6].

At advanced nodes, EUV lithography is highly power-hungry. An EUV scanner consumes approximately 1 MW of input power, with tool throughputs of 65–170 wafers per hour [41]. Each EUV exposure step therefore consumes approximately 6–14 kWh/wafer [6, 41]. While EUV reduces the number of lithographic passes compared to the multi-patterning deep ultraviolet (DUV) sequences it replaces, node-to-node scaling still increases the total lithographic energy due to additional FEOL patterning steps, MOL layers, and BEOL metal layers [6].

4.3 Per-Die Energy Allocation

The overall wafer energy is allocated across the good dies. The gross dies per wafer (DPW) is estimated using the widely used approximation [88],

$$\text{DPW}_{\text{gross}} \approx \frac{\pi(d/2)^2}{S} - \frac{\pi d}{2\sqrt{S}}, \quad (2)$$

where d is the wafer diameter and S is the die area. To account for defects, the negative-binomial yield model is applied [89],

$$Y_{\text{die}} = \left(1 + \frac{SD_0}{\alpha}\right)^{-\alpha}, \quad (3)$$

where Y_{die} is the die yield, S is the die area (cm^2), D_0 is the defect density (defects/ cm^2), and α is the clustering parameter. The resulting front-end energy consumed per good die is

$$E_{\text{die,FE}} = \frac{E_{\text{wafer,node}}}{\text{DPW}_{\text{gross}} \times Y_{\text{die}}}, \quad (4)$$

where $E_{\text{wafer,node}}$ is the node-specific wafer energy from (1).

Equation (4) captures the two most important scaling effects: wafer energy grows with process node complexity per Eq. (1), and large dies have worse yield per Eq. (3), amplifying per-die energy. To illustrate, we assume representative values of $D_0 = 0.1$ defects/ cm^2 and $\alpha = 2$, consistent with a mature advanced-node process [8, 89].

Effect 1: Node scaling. Consider a fixed die area of $S = 100 \text{ mm}^2$, representative of a mid-size GPU chiplet. At 28 nm, using $s(28 \text{ nm}) \approx 0.33$ from (1), the node-specific wafer energy is $E_{\text{wafer,node}} \approx 0.33 \times 2,130 \approx 703 \text{ kWh}$. Substituting into (2) gives $\text{DPW}_{\text{gross}} \approx 660$, and the yield model in (3) gives $Y_{\text{die}} \approx 0.91$, yielding $E_{\text{die,FE}} \approx 1.2 \text{ kWh/die}$. At 3 nm, using $s(3 \text{ nm}) \approx 1.16$, the wafer energy rises to $E_{\text{wafer,node}} \approx$

$1.16 \times 2,130 \approx 2,471$ kWh, while $Y_{\text{die}} \approx 0.91$, giving $E_{\text{die,FE}} \approx 4.1$ kWh/die—a $\sim 3.5\times$ increase in per-die manufacturing energy from node scaling alone.

Effect 2: Die size and yield amplification. Fixing the node at 3 nm, we compare a small die ($S = 50 \text{ mm}^2$) against a large die ($S = 400 \text{ mm}^2$). For the small die, $\text{DPW}_{\text{gross}} \approx 1,347$ and $Y_{\text{die}} \approx 0.95$, giving $E_{\text{die,FE}} \approx 1.9$ kWh/die. For the large die, $\text{DPW}_{\text{gross}} \approx 153$ and $Y_{\text{die}} \approx 0.69$, giving $E_{\text{die,FE}} \approx 23.2$ kWh/die—an $\sim 12\times$ increase driven primarily by yield loss. These examples underscore that at advanced nodes, die size is the dominant determinant of per-die manufacturing energy, with yield loss amplifying the effect far beyond what node scaling alone would predict.

4.4 Memory Fabrication

Memory fabrication contributes significantly to total manufacturing energy, particularly for DRAM. This is especially relevant for high-performance memory technologies such as HBM, which is critical for datacenter GPUs and TPUs, and LPDDR, which is widely deployed in mobile SoC platforms [17, 38]. Unlike logic wafers, DRAM fabrication has been largely confined to 10 nm-class process nodes, where further scaling is constrained by capacitor architecture and charge retention requirements [94].

The process complexity of DRAM fabrication arises from deep capacitor etching, high- κ dielectric deposition, and multi-patterning lithography, all of which are electrically intensive processes. The same wafer-level energy allocation framework developed for logic is applied here, consistent with findings that electricity consumption per wafer is comparable across device types at equivalent technology nodes [21]. Applying the node scaling model from (1) with $s(10 \text{ nm}) \approx 0.48$, the estimated wafer energy for 10 nm-class DRAM processes is approximately 1,030 kWh per 300 mm wafer. The energy per memory die is then derived using the same die allocation framework as logic, substituting DRAM-specific parameters. For HBM3E the typical die area is

$\sim 121 \text{ mm}^2$ and for HBM3 $\sim 107 \text{ mm}^2$ [50], while LPDDR5-class DRAM dies are significantly smaller, typically in the range of $50\text{--}75 \text{ mm}^2$ depending on density and process node [48, 49].

In datacenter GPUs, memory stacks dominate the total wafer demand. The NVIDIA H100 integrates five 8-layer HBM3 stacks (40 DRAM dies and 5 base dies), translating to roughly 100 kWh of fabrication energy per GPU, as detailed in the Results [10, 17]. In mobile SoCs, LPDDR5/5X memory uses smaller dies and fewer layers, resulting in lower per-die energy; however, wafer-level energy intensity remains comparable due to the shared 10 nm-class process node. Consequently, per-device manufacturing energy in smartphones remains in the tens of Wh, while for datacenter accelerators it scales to hundreds of kWh.

4.5 Advanced Packaging

In modern accelerators, logic and memory dies are co-integrated using advanced 2.5D/3D packaging technologies. Representative examples include TSMC CoWoS used in the NVIDIA H100 [39, 61], Intel EMIB/Foveros used in the Ponte Vecchio GPU [90], and AMD’s multi-chip module (MCM) designs in the Instinct MI250/MI300 series [56, 62]. These packaging flows involve interposer fabrication, die stacking, micro-bump attachment, TSV processing, underfill dispensing, and thermal cure [57, 61].

Direct energy measurements for packaging processes remain limited in the open literature. ECO-CHIP [8] and 3D-Carbon [9] provide validated models confirming that packaging energy is modest relative to wafer fabrication, typically 5 kWh per assembled accelerator package, a figure consistent with NVIDIA’s ISO-aligned HGX H100 Product Carbon Footprint (PCF) report, in which packaging accounted for only 8.6% of the total manufacturing energy footprint [10].

4.6 Operational Energy Framework

The operational energy of any integrated circuit is defined as the cumulative electrical energy consumed over its service lifetime. Datacenter accelerators such as GPUs, TPUs, and custom AI chips operate continuously at power levels ranging from several hundred watts to multiple kilowatts [25, 26]. Their lifetime energy consumption reaches tens of megawatt-hours during active workload execution [26], with additional non-negligible power drawn during idle periods, memory refresh, synchronization, and host-managed infrastructure states. In contrast, mobile SoCs, such as the Apple A-series [95] and Qualcomm Snapdragon platforms [96], operate intermittently at power levels of only a few watts, resulting in a three-year cumulative energy on the order of single-digit kilowatt-hours [33, 36].

To enable a fair comparison across device classes, operational energy is quantified using a unified framework that incorporates subsystem composition, workload activity factors, idle behavior, and infrastructure overhead.

4.7 Energy System Boundaries

To ensure consistent comparison between datacenter accelerators and mobile devices, operational energy is quantified using the following definitions:

1. **Device-level operational energy** (E_{IT}) is the time-integrated electrical power drawn by the processor over its service lifetime:

$$E_{IT} = \int_0^t P_{op}(t) dt \approx P_{eff} \times t, \quad (5)$$

where P_{eff} is the effective long-term operational power and t is the cumulative service time.

2. **Facility-level compute energy** ($E_{facility}$) for datacenter deployments accounts for cooling, power delivery, and infrastructure overheads using the power usage

effectiveness (PUE) metric:

$$E_{\text{facility}} = \text{PUE} \times E_{\text{IT}}. \quad (6)$$

This formulation follows standard datacenter energy accounting practices [5, 32].

For mobile devices, which do not operate in shared facilities, only E_{IT} is reported.

A three-year service life is assumed across all device categories to preserve comparability:

- **Datacenter accelerators** are commonly replaced or redeployed within 3–4 years due to rapid architectural evolution and memory-technology scaling, as reflected in hyperscale operational reports [5].
- **Mobile SoCs** follow the same three-year lifetime assumption, consistent with Apple’s environmental reporting [23] and average consumer device replacement intervals [33].

4.8 Subsystem-Level Power Decomposition

The instantaneous device power is modeled as the sum of compute, memory, and interconnect power:

$$P_{\text{tot}} = P_{\text{core}} + P_{\text{mem}} + P_{\text{IO}}, \quad (7)$$

where P_{core} denotes arithmetic and tensor-operation power, P_{mem} includes DRAM activation, refresh, and I/O-driver switching (HBM for datacenter accelerators, LPDDR for mobile SoCs), and P_{IO} represents network-on-chip fabrics, SerDes links, and external interfaces. Measurements from recent accelerator platforms confirm that data movement and memory activity constitute a dominant fraction of total power under modern AI inference and training workloads [22, 25, 26].

4.9 Peak-Power Subsystem Estimation

Peak subsystem powers are derived from architectural specifications and published characterization data [25, 26]. For each HBM stack indexed by s , the power consumption is modeled as the sum of static and dynamic components:

$$P_{\text{HBM},s} = P_{\text{static},s} + e_{\text{bit},s} \times B_s, \quad (8)$$

where $P_{\text{static},s}$ is the static refresh and bias power, $e_{\text{bit},s}$ is the energy per transferred bit, and B_s is the peak bandwidth. The static component $P_{\text{static},s}$ accounts for DRAM cell refresh, background biasing of the stacked dies, and TSV and logic-layer leakage overhead. As demonstrated in [97], total HBM power does not scale linearly with bandwidth utilization, indicating a non-negligible bandwidth-independent component in addition to dynamic I/O switching energy.

The value of $e_{\text{bit},s}$ is based on vendor-published measurements and technical disclosures [17, 97]. SK Hynix reports an HBM2E I/O energy of 4.3 pJ/bit, with HBM3E achieving a 20% reduction due to improved physical layer (PHY) design and lower-voltage signaling [17]. Micron independently reports that HBM3E delivers $2.5\times$ higher performance-per-watt than HBM2E under comparable conditions [24]. Since energy per bit is inversely proportional to performance-per-watt, this implies $e_{\text{bit,HBM3E}} \approx 4.3/2.5 \approx 1.7$ pJ/bit. Together, these disclosures define a reproducible and technology-consistent range of $e_{\text{bit},s} = 1.7\text{--}3.4$ pJ/bit for HBM3-class stacks adopted in this work.

The total memory peak power is obtained by summing across all HBM stacks:

$$P_{\text{mem,peak}} = \sum_s P_{\text{HBM},s}. \quad (9)$$

Device-specific stack counts and bandwidths are given in the Results section, where Eqs. (8)–(9) are evaluated for each accelerator.

Interconnect power is modeled as a static term plus a bandwidth-scaled dynamic component:

$$P_{\text{IO,peak}} = P_{\text{IO,static}} + \sum_{\ell} e_{\text{bit,IO},\ell} \times B_{\ell}, \quad (10)$$

where $e_{\text{bit,IO},\ell}$ denotes the SerDes energy per transferred bit and B_{ℓ} is the peak bandwidth of link ℓ . The static term $P_{\text{IO,static}}$ reflects clock distribution, bias, and always-on circuitry in multi-lane SerDes PHYs, while the dynamic term captures switching energy proportional to link throughput.

As vendors do not disclose the per-lane energy for NVLink or PCIe PHYs, $e_{\text{bit,IO},\ell}$ is parameterized using measured energy efficiencies from state-of-the-art four-level pulse amplitude modulation (PAM-4) wireline transceivers, which implement the same class of high-speed I/O circuits used in modern accelerator interconnects. A capacitively driven PAM-4 transceiver fabricated in 28 nm CMOS demonstrates an energy efficiency of 0.458 pJ/bit [98], while a 112 Gb/s PAM-4 linear transimpedance amplifier (TIA) reports an efficiency of 0.96 pJ/bit [99]. Together, these measured results bound the representative range of $e_{\text{bit,IO},\ell}$ to 0.5–1.0 pJ/bit. A recent 112 Gb/s DSP-based SerDes receiver in 7 nm FinFET technology further confirms that these PAM-4 PHYs are representative of the signaling techniques employed in contemporary accelerator interconnects [100].

Device-specific SerDes configurations (e.g., NVLink lane counts or PCIe widths) are instantiated in the Results section. The compute subsystem peak power is then obtained as the residual:

$$P_{\text{core,peak}} = P_{\text{tot,peak}} - P_{\text{mem,peak}} - P_{\text{IO,peak}}, \quad (11)$$

ensuring consistency with the device-level thermal design power (TDP) envelope.

4.10 Energy per Token

Energy per generated token is calculated from measured system throughput and rated TDP:

$$E_{\text{token}} = \frac{P_{\text{TDP}} \times N_{\text{GPU}}}{\text{Throughput}_{\text{tokens/s}}}, \quad (12)$$

where P_{TDP} is the rated board-level thermal design power, N_{GPU} is the number of GPUs in the system, and $\text{Throughput}_{\text{tokens/s}}$ is the measured inference throughput under the MLPerf Offline scenario[64].

4.11 Active–Idle Power Model for Datacenter Accelerators

Peak power does not reflect practical operational behavior [25, 26]. In real deployments, datacenter accelerators transition between an active computational state and idle or partially idle states due to synchronization, scheduling gaps, and system-level orchestration overhead. Even during idle periods, accelerators continue to draw significant power.

The effective power of subsystem i is modeled as a time-weighted average of active and idle states:

$$P_{i,\text{eff}} = U P_{i,\text{active}} + (1 - U) P_{i,\text{idle}}, \quad (13)$$

where U is the time-averaged utilization (fraction of time executing computation). During active phases, subsystem power is expressed as a fraction of peak power:

$$P_{i,\text{active}} = \alpha_i P_{i,\text{peak}}, \quad (14)$$

where $\alpha_i \in (0, 1]$ is a workload-dependent activity factor. Measurements on H100- and A100-based systems report $\alpha_i \approx 0.6\text{--}0.8$ [25, 26].

Idle power is explicitly retained in the model. It represents true device idle, dominated by transistor leakage and HBM refresh with no active workload, and is distinct

from within-job reduced-activity states, which are captured by α . Because transistor leakage and DRAM refresh together account for only a small fraction of the peak power envelope in modern GPUs, β is assumed to be approximately 0.1. For a compact representation, idle power is normalized as:

$$\beta = \frac{P_{\text{idle}}}{P_{\text{peak}}}, \quad (15)$$

yielding the device-level form:

$$P_{\text{eff}} = [U\alpha + (1 - U)\beta]P_{\text{peak}}. \quad (16)$$

Utilization values of $U \approx 0.5\text{--}0.6$ are consistent with national-scale surveys of production datacenters [5, 32].

4.12 Mobile SoCs: Duty-Cycle Operational Energy

In contrast to datacenter accelerators, mobile SoCs do not operate continuously at elevated power levels. Their lifetime energy profile is characterized by intermittent bursts of computation with extended low-power standby periods [33, 36]. Consequently, operational energy is determined by user activity patterns and duty cycles rather than sustained high utilization.

The device-level energy E_{IT} of a mobile SoC is modeled as the sum of energy consumed during active and standby phases:

$$E_{\text{IT}} = P_{\text{active}} t_{\text{active}} + P_{\text{standby}} t_{\text{standby}}, \quad (17)$$

where P_{active} and P_{standby} denote the average SoC power during active use and standby, respectively, t_{active} and t_{standby} represent the cumulative time spent in each

state over the device lifetime, and $t_{\text{active}} + t_{\text{standby}} = t$ is the total service lifetime consistent with (5).

Idle and Connected-Standby Power (P_{standby})

Even during idle operation, modern mobile devices consume non-negligible power due to memory refresh, network connectivity, and background system processes. Empirical measurements have shown that connected-standby power remains in the sub-watt regime under display-off, radio-idle conditions [36]. Contemporary operating system frameworks further formalize these low-power states: the Android BatteryStats and On-Device Power Measurement (ODPM) infrastructure provide per-rail energy accounting across active and idle modes, confirming connected-standby consumption on the order of hundreds of milliwatts [35].

Active (Compute) Power (P_{active})

During active operation, mobile SoC power consumption is constrained by thermal limits, battery capacity, and Dynamic Voltage and Frequency Scaling (DVFS). Empirical device-level measurements under compute-intensive workloads indicate sustained power levels in the low-watt range, with short-lived transients suppressed by DVFS [36]. Large-scale environmental and product life-cycle assessments similarly model smartphone use-phase energy based on sustained average power rather than peak excursions [23].

Accordingly, mobile SoC active power is modeled using representative ranges bounded by sustained interactive workload measurements, rather than peak power envelopes. Device-specific active power parameters are provided in the Results section.

Data availability

The data supporting the findings of this study are available within the article. All input parameters used in the manufacturing and operational energy estimates were obtained from publicly available sources, including vendor product disclosures, JEDEC standards, IRDS roadmap documents, MLPerf Inference v4.1 benchmark results, and peer-reviewed life-cycle assessment publications, as cited throughout the manuscript and Methods. Derived values reported in Tables 1–5 are reproducible from these inputs using the equations provided in Methods.

Code availability

The energy estimation framework described in Methods (Equations 1–17) was implemented in Python. The code used to generate the manufacturing energy projections (Table 4, Fig. 2) and the operational energy estimates (Tables 2, 3, 5) is available from the corresponding author upon reasonable request.

Author contributions

A.A. developed the analytical framework, performed the manufacturing and operational energy analysis, generated the figures and tables, and wrote the manuscript. E.S. supervised the project, contributed to the conceptual framing and interpretation of results, and revised the manuscript. Both authors reviewed and approved the final version.

Competing interests

The authors declare no competing interests.

Funding

The authors received no specific funding for this work.

Additional information

Correspondence and requests for materials should be addressed to E.S.

References

- [1] Sevilla, J. *et al.* *Compute trends across three eras of machine learning*, 1–8 (IEEE, 2022). URL <http://dx.doi.org/10.1109/IJCNN55064.2022.9891914>.
- [2] Kaplan, J. *et al.* *Scaling laws for neural language models* (2020). URL <https://arxiv.org/abs/2001.08361>. arXiv:2001.08361.
- [3] Patterson, D. *et al.* *Carbon emissions and large neural network training* (2021). URL <https://arxiv.org/abs/2104.10350>. arXiv:2104.10350.
- [4] International Energy Agency. *Energy and ai*. Tech. Rep., IEA, Paris (2025). URL <https://www.iea.org/reports/energy-and-ai>.
- [5] Shehabi, A. *et al.* *2024 united states data center energy usage report*. Tech. Rep. LBNL-2001637, Lawrence Berkeley National Laboratory (2024). URL <http://dx.doi.org/10.71468/P1WC7Q>.
- [6] Garcia Bardon, M. *et al.* *Dtco including sustainability: Power-performance-area-cost-environmental score (ppace) analysis for logic technologies*, 41.4.1–41.4.4 (2020).
- [7] Boakes, L. *et al.* *Cradle-to-gate life cycle assessment of cmos logic technologies*, 1–4 (2023).

- [8] Sudarshan, C. C., Matkar, N., Vrudhula, S., Sapatnekar, S. S. & Chhabria, V. A. *Eco-chip: Estimation of carbon footprint of chiplet-based architectures for sustainable vlsi*, 671–685 (2024).
- [9] Zhao, Y., Zhao, Y. K., Wan, C. & Lin, Y. C. *3d-carbon: An analytical carbon modeling tool for 3d and 2.5d integrated circuits*, DAC '24 (Association for Computing Machinery, New York, NY, USA, 2024). URL <https://doi.org/10.1145/3649329.3658482>.
- [10] NVIDIA. Hgx h100 product carbon footprint summary, iso 14067 lca. Tech. Rep. (2023).
- [11] Fedus, W., Zoph, B. & Shazeer, N. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.* **23** (2022).
- [12] Shazeer, N. *et al.* Outrageously large neural networks: The sparsely-gated mixture-of-experts layer (2017). URL <https://arxiv.org/abs/1701.06538>. arXiv:1701.06538.
- [13] Dao, T., Fu, D. Y., Ermon, S., Rudra, A. & Ré, C. Flashattention: Fast and memory-efficient exact attention with io-awareness (2022). URL <https://arxiv.org/abs/2205.14135>. arXiv:2205.14135.
- [14] Ainslie, J. *et al.* Gqa: Training generalized multi-query transformer models from multi-head checkpoints (2023). URL <https://arxiv.org/abs/2305.13245>. arXiv:2305.13245.
- [15] Micikevicius, P. *et al.* Fp8 formats for deep learning (2022). URL <https://arxiv.org/abs/2209.05433>. arXiv:2209.05433.

- [16] Micikevicius, P. *et al.* Mixed precision training (2018). URL <https://arxiv.org/abs/1710.03740>. arXiv:1710.03740.
- [17] Moon, K.-I., Son, H.-Y. & Lee, K. *Advanced packaging technologies in memory applications for future generative ai era*, 1–4 (2023).
- [18] Advanced Micro Devices, Inc. Amd instinct mi300 series accelerators (2023).
- [19] NVIDIA Corporation. Nvidia h200 tensor core gpu architecture overview (2024).
- [20] Lee, J. *et al.* *13.4 a 48GB 16-high 1280GB/s HBM3E DRAM with all-around power TSV and a 6-phase RDQS scheme for TSV area optimization*, Vol. 67, 238–240 (2024).
- [21] Nagapurkar, S. & Das, S. Economic and embodied energy analysis of integrated circuit manufacturing processes. *Sustainable Computing: Informatics and Systems* **35**, 100727 (2022).
- [22] Jouppi, N., Young, C. *et al.* *Ten years of tpu development* (2023). Includes details on TPU v5e/v5p.
- [23] Apple Inc. Apple environmental progress report 2024. Tech. Rep., Apple Inc., Cupertino, CA, USA (2024).
- [24] Micron Technology. Hbm3e. URL <https://www.micron.com/products/memory/hbm/hbm3e>.
- [25] Newkirk, A. C. *et al.* Empirically-calibrated h100 node power models for reducing uncertainty in ai training energy estimation (2025). URL <https://arxiv.org/abs/2506.14551>. arXiv:2506.14551.

- [26] Latif, I. *et al.* Single-node power demand during ai training: Measurements on an 8-gpu nvidia h100 system. *IEEE Access* **13**, 61740–61747 (2025). URL <http://dx.doi.org/10.1109/ACCESS.2025.3554728>.
- [27] Patel, P. *et al.* *Characterizing power management opportunities for llms in the cloud*, ASPLOS '24, 207–222 (Association for Computing Machinery, New York, NY, USA, 2024). URL <https://doi.org/10.1145/3620666.3651329>.
- [28] Patterson, D. *et al.* The carbon footprint of machine learning training will plateau, then shrink (2022). URL <https://arxiv.org/abs/2204.05149>. [arXiv:2204.05149](https://arxiv.org/abs/2204.05149).
- [29] Hodak, M., Gorkovenko, M. & Dholakia, A. *Towards power efficiency in deep learning on data center hardware*, 1814–1820 (2019).
- [30] Weng, Q. *et al.* *MLaaS in the wild: Workload analysis and scheduling in Large-Scale heterogeneous GPU clusters*, 945–960 (USENIX Association, Renton, WA, 2022). URL <https://www.usenix.org/conference/nsdi22/presentation/weng>.
- [31] Kokolis, A. *et al.* *Revisiting Reliability in Large-Scale Machine Learning Research Clusters*, 1259–1274 (IEEE Computer Society, Los Alamitos, CA, USA, 2025). URL <https://doi.ieeecomputersociety.org/10.1109/HPCA61900.2025.00096>.
- [32] Uptime Institute. Global data center survey 2024. Tech. Rep., Uptime Institute Intelligence, Seattle, WA, USA (2024).
- [33] IEA 4E EDNA Programme. Mobile devices – final report. Tech. Rep., International Energy Agency (IEA) 4E Electronics Devices and Networks Annex (EDNA), Paris, France (2022). URL <https://www.iea-4e.org/publications/>.

- [34] Gupta, A., Heidari, A., Kalipattapu, A., Jain, I. K. & Bharadia, D. *3 w's of smartphone power consumption: Who, where and how much is draining my battery?*, ACM MobiCom '24, 2248–2250 (Association for Computing Machinery, New York, NY, USA, 2024). URL <https://doi.org/10.1145/3636534.3695905>.
- [35] Android Developers. Power and energy monitoring (ODPM) (2024).
- [36] Carroll, A. & Heiser, G. *An analysis of power consumption in a smartphone* (USENIX Association, Boston, MA, USA, 2010).
- [37] JEDEC Solid State Technology Association. JESD270-4A: High Bandwidth Memory (HBM4) DRAM. JEDEC Standard (2025). URL <https://www.jedec.org/standards-documents/docs/jesd270-4a>. Accessed: 2025.
- [38] JEDEC Solid State Technology Association. JESD238A: High Bandwidth Memory (HBM4) DRAM. Standard JESD238A, JEDEC (2024). Specifies 16-high stack configurations, 2048-bit interface.
- [39] TSMC. Cowos (chip-on-wafer-on-substrate) technology. URL <https://3dfabric.tsmc.com/english/dedicatedFoundry/technology/cowos.htm>. Accessed: 4 December 2025.
- [40] IEEE IRDS Lithography Team. International Roadmap for Devices and Systems: Lithography and Patterning. IRDS 2024 Edition, IEEE (2024). URL https://irds.ieee.org/images/files/pdf/2024/2024IRDS_LITHO.pdf. Accessed January 2026.
- [41] Thijssen, T., Van der Net, T., Janssen, T. & Luijten, C. *Euv energy efficiency*, 1–4 (2023).

- [42] ASML. Twinscan exe:5200b — euv lithography system (n.d.). URL <https://www.asml.com/en/products/euv-lithography-systems/twinscan-exe-5200b>.
- [43] for Devices, I. I. R. & Committee, S. I. S. Ird^s[™] 2024: Environment, safety, health sustainability (eshs) – environmental sustainability of the semiconductor facilities (essf). Tech. Rep., IEEE (2024). URL <https://irds.ieee.org/images/files/pdf/2024/2024IRDS.ESHS-ESSF.pdf>.
- [44] NVIDIA Corporation. Inside the NVIDIA Rubin Platform: Six New Chips, One AI Supercomputer (2026). URL <https://developer.nvidia.com/blog/inside-the-nvidia-rubin-platform-six-new-chips-one-ai-supercomputer>.
- [45] AMD. AMD Unveils Vision for an Open AI Ecosystem, Details AMD Instinct MI350 Series. Press Release (2025). URL <https://www.amd.com/en/newsroom/press-releases/2025-6-12-amd-unveils-vision-for-an-open-ai-ecosystem-detail.html>.
- [46] Adaikkalavan, R. AMD Instinct MI355X: CDNA 4 architecture disclosed at ISSCC 2026. Presented at IEEE ISSCC 2026; reported by *Tom’s Hardware* (2026). 110 mm² XCD die area disclosed. <https://www.tomshardware.com/tech-industry/semiconductors/inside-the-instinct-mi355x>.
- [47] Boyd, S. *Life-Cycle Assessment of Semiconductors* (Springer, 2012).
- [48] Micron Technology. Micron Ships World’s Most Advanced DRAM Technology With 1-Beta Node. Press Release (2022). URL <https://investors.micron.com/news-releases/news-release-details/micron-ships-worlds-most-advanced-dram-technology-1-beta-node>. 1 node delivers 15% higher power efficiency vs. 1.

- [49] Micron Technology. 1 (1-gamma) DDR5 DRAM Technology Brief. Tech. Rep. (2025). URL <https://micron.com/1gamma>. 1 node delivers 20% lower power vs. 1, 30% higher bit density.
- [50] Samsung Semiconductor. Hbm3e shinebolt. URL <https://semiconductor.samsung.com/dram/hbm/hbm3e-shinebolt/>.
- [51] SK hynix. Hbm3e. URL <https://product.skhynix.com/products/dram/hbm/hbm3e.go>.
- [52] SK hynix. Sk hynix completes world’s first hbm4 development and readies mass production. URL <https://news.skhynix.com/sk-hynix-completes-worlds-first-hbm4-development-and-readies-mass-production/>.
- [53] Micron Technology. Micron ships hbm4 to key customers to power next-gen ai platforms. URL <https://investors.micron.com/news-releases/news-release-details/micron-ships-hbm4-key-customers-power-next-gen-ai-platforms>.
- [54] SK hynix. Sk hynix redefines its vision at sk ai summit 2025: From ai memory provider to “creator”. URL <https://news.skhynix.com/sk-hynix-redefines-its-vision-at-sk-ai-summit-2025-from-ai-memory-provider-to-creator/>.
- [55] Bailey, B. HBM4 sticks with microbumps, postponing hybrid bonding. Semiconductor Engineering (2026). URL <https://semiengineering.com/hbm4-sticks-with-microbumps-postponing-hybrid-bonding/>. Accessed March 2026.
- [56] Smith, A. *et al.* 11.1 amd instincttm mi300 series modular chiplet package – hpc and ai accelerator for exa-class systems, Vol. 67, 490–492 (2024).

- [57] Hou, S. Y., Lee, C. H., Wang, T.-D., Hou, H. C. & Hu, H.-P. *Supercarrier redistribution layers to realize ultra large 2.5d wafer scale packaging by cowos*, 510–514 (2023).
- [58] Mutlu, O. *et al.* A modern primer on processing in memory (2025). URL <https://arxiv.org/abs/2012.03112>. arXiv:2012.03112.
- [59] Gómez-Luna, J. *et al.* Benchmarking a new paradigm: Experimental analysis and characterization of a real processing-in-memory system. *IEEE Access* **10**, 52565–52608 (2022).
- [60] Gómez-Luna, J. *et al.* An experimental evaluation of machine learning training on a real processing-in-memory system (2023). URL <https://arxiv.org/abs/2207.07886>. arXiv:2207.07886.
- [61] Huang, P. K. *et al.* *Wafer level system integration of the fifth generation cowos®-s with high performance si interposer at 2500 mm²*, 101–104 (2021).
- [62] Mandalapu, C. S. *et al.* *3.5d advanced packaging enabling heterogenous integration of hpc and ai accelerators*, 798–802 (2024).
- [63] Smith, A. *et al.* *Realizing the amd exascale heterogeneous processor vision : Industry product*, 876–889 (2024).
- [64] MLCommons. MLPerf Inference v4.1 Datacenter Results. <https://mlcommons.org/benchmarks/inference-datacenter/> (2024). Submission IDs: 4.1-0043 (H100), 4.1-0048 (H200).
- [65] Hoffmann, J. *et al.* Training compute-optimal large language models (2022). URL <https://arxiv.org/abs/2203.15556>. arXiv:2203.15556.

- [66] Villalobos, P. *et al.* Will we run out of data? limits of llm scaling based on human-generated data (2024). URL <https://arxiv.org/abs/2211.04325>. arXiv:2211.04325.
- [67] Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A. & Kadian, A. The llama 3 herd of models (2024). URL <https://arxiv.org/abs/2407.21783>. arXiv:2407.21783.
- [68] OpenAI. Learning to reason with llms. <https://openai.com/index/learning-to-reason-with-llms/> (2024).
- [69] DeepSeek-AI. Deepseek-v3 technical report (2025). URL <https://arxiv.org/abs/2412.19437>. arXiv:2412.19437.
- [70] Zhao, C. *et al.* *Insights into deepseek-v3: Scaling challenges and reflections on hardware for ai architectures*, SIGARCH '25, 1731–1745 (ACM, 2025). URL <http://dx.doi.org/10.1145/3695053.3731412>.
- [71] Sorrell, S. Jevons’ paradox revisited: The evidence for backfire from improved energy efficiency. *Energy Policy* **37**, 1456–1469 (2009). URL <https://www.sciencedirect.com/science/article/pii/S0301421508007428>.
- [72] Ma, X. & Patterson, D. Challenges and research directions for large language model inference hardware. *arXiv preprint arXiv:2601.05047* (2026). URL <https://arxiv.org/abs/2601.05047>.
- [73] Legtchenko, S. *et al.* Managed-retention memory: A new class of memory for the ai era (2025). URL <https://arxiv.org/abs/2501.09605>. arXiv:2501.09605.
- [74] Lee, P. & Yang, J. Global semiconductors. Research Report, Citi Research (2024). Citi Research report on the global semiconductor industry.

- [75] Gupta, U. *et al.* Chasing carbon: The elusive environmental footprint of computing (2020). URL <https://arxiv.org/abs/2011.02839>. arXiv:2011.02839.
- [76] Schneider, I. *et al.* Life-cycle emissions of ai hardware: A cradle-to-grave approach and generational trends (2025). URL <https://arxiv.org/abs/2502.01671>. arXiv:2502.01671.
- [77] European Parliament and Council of the European Union. Directive (EU) 2023/1791 on energy efficiency (2023). URL <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32023L1791>. Official Journal of the European Union, L 231. Mandates annual reporting of datacenter energy performance (PUE, energy consumption, water footprint) for facilities with ≥ 500 kW installed IT power demand.
- [78] U.S. Securities and Exchange Commission. The enhancement and standardization of climate-related disclosures for investors (2024). URL <https://www.sec.gov/rules/final/2024/33-11275.pdf>.
- [79] Vahdat, A., Ma, X. & Patterson, D. New computer evaluation metrics for a changing world. *Commun. ACM* **67**, 31–33 (2024). URL <https://doi.org/10.1145/3637867>.
- [80] ISO. Iso 14040:2006 environmental management — life cycle assessment — principles and framework. Tech. Rep. (2006).
- [81] SanDisk. High-bandwidth flash (hbf) fact sheet (2025). URL https://documents.sandisk.com/content/dam/asset-library/en_us/assets/public/sandisk/collateral/company/Sandisk-HBF-Fact-Sheet.pdf. Accessed: 2026-04-10.

- [82] Wong, H.-S. P. *et al.* Phase change memory. *Proceedings of the IEEE* **98**, 2201–2227 (2010).
- [83] Wong, H.-S. P. *et al.* Metal–oxide rram. *Proceedings of the IEEE* **100**, 1951–1970 (2012).
- [84] Apalkov, D. *et al.* Spin-transfer torque magnetic random access memory (stt-mram). *J. Emerg. Technol. Comput. Syst.* **9** (2013). URL <https://doi.org/10.1145/2463585.2463589>.
- [85] Compute Express Link Consortium. Advantages of cxl memory sharing for emerging applications (q2-2025 webinar). PDF report. URL https://computeexpresslink.org/wp-content/uploads/2025/06/CXL_Q2-2025-Webinar_FINAL.pdf.
- [86] Samsung Electronics. Samsung develops industry’s first high bandwidth memory with ai processing power. URL <https://semiconductor.samsung.com/news-events/news/samsung-develops-industrys-first-high-bandwidth-memory-with-ai-processing-power/>.
- [87] Samsung Electronics. Hbm-pim: Cutting-edge memory technology to accelerate next-generation ai. URL <https://semiconductor.samsung.com/news-events/tech-blog/hbm-pim-cutting-edge-memory-technology-to-accelerate-next-generation-ai/>.
- [88] Wolf, S. *Silicon Processing for the VLSI Era, Volume 1: Process Technology* (Lattice Press, 2004).
- [89] Stapper, C. H. Modeling of defects in integrated circuit photolithographic patterns. *IBM Journal of Research and Development* **28**, 461–475 (1984).

- [90] Mahajan, R. *et al.* *Embedded multi-die interconnect bridge (emib) – a high density, high bandwidth packaging interconnect*, 557–565 (2016).
- [91] Liu, C., Lin, S. J. & Lewis, C. Life cycle assessment of dram in taiwan’s semiconductor industry. *Journal of Cleaner Production* **18**, 419–425 (2010). URL <https://www.sciencedirect.com/science/article/pii/S0959652609003266>.
- [92] Hu, A., Chang, Y. & Wang, C. Energy and environmental impact of 300 mm dram fabrication in taiwan. *Energy* **35**, 3788–3794 (2010).
- [93] TSMC. 2024 esg report and 2024 annual report. Tech. Rep., Taiwan Semiconductor Manufacturing Company (2025).
- [94] Belmonte, A. *et al.* Disrupting the DRAM roadmap with capacitor-less IGZO-DRAM technology. *Nature Reviews Electrical Engineering* (2025). URL <https://www.nature.com/articles/s44287-025-00162-w>.
- [95] Apple Inc. Apple a18 pro chip technical specifications (2024).
- [96] Qualcomm Technologies, Inc. Qualcomm snapdragon 8 gen4 and snapdragon elite socs (2024).
- [97] Larimi, S. S. N. *et al.* Understanding power consumption and reliability of high-bandwidth memory with voltage undervolting (2020). URL <https://arxiv.org/abs/2101.00969>. arXiv:2101.00969.
- [98] Kim, S. *et al.* A 0.458-pj/bit 24-gb/s/pin capacitively driven pam-4 transceiver with pam-based crosstalk cancellation for high-density die-to-die interfaces. *IEEE Journal of Solid-State Circuits* **59**, 3730–3740 (2024).
- [99] Li, H., Balamurugan, G., Jaussi, J. & Casper, B. *A 112 gb/s pam4 linear tia with 0.96 pj/bit energy efficiency in 28 nm cmos*, 238–241 (2018).

- [100] Guo, H., Yao, Y., Ni, J. & Gao, X. A 112 gbps dsp-based pam4 serdes receiver with a wide band equalization tuning afe in 7 nm finfet. *Journal of Semiconductors* **46**, 062204 (2025). URL <https://doi.org/10.1088/1674-4926/25030001>.