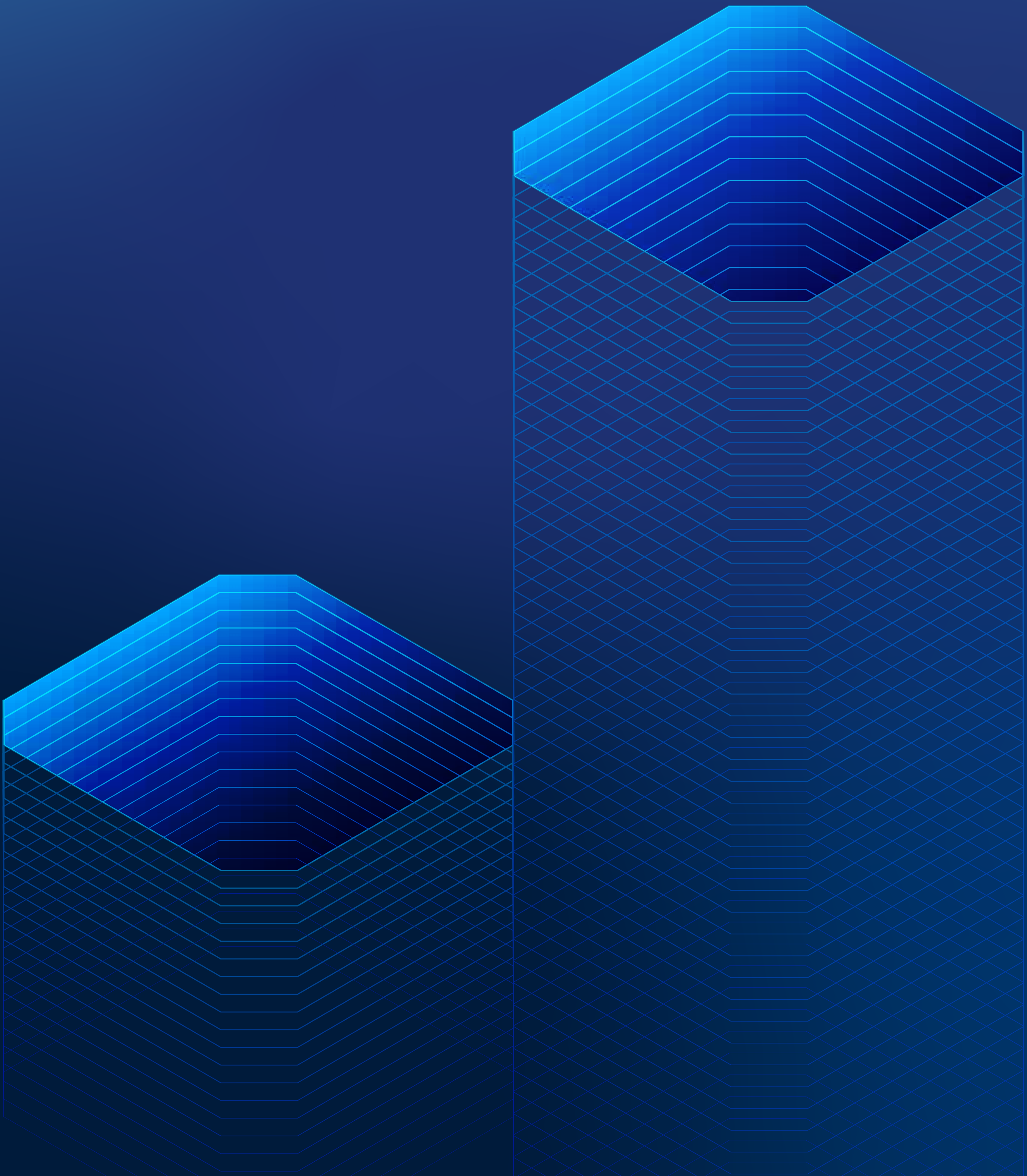
Appendix A.

Methodology: A Bottom-up Analysis of Energy Consumption and Carbon Footprint in AI Chip Production



Appendix A.

Methodology: A Bottom-up Analysis of Energy Consumption and Carbon Footprint in AI Chip Production

Author: Alex de Vries

To assess how artificial intelligence (AI) affects the electricity demand for semiconductor manufacturing, we apply a bottom-up approach in which we first determine how demand for AI hardware affects the production output of key components such as logic chips, memory chips, and data storage chips. By combining the estimated changes in production output with the respective electricity requirements of the production process, it is possible to obtain an aggregated estimate for AI-related electricity consumption in semiconductor manufacturing. Additionally, by assessing where and by whom AI-related hardware is being produced, it is possible to assess the electricity mix fueling the production process. This chapter provides a detailed outline of this approach, along with the data sources utilized to obtain the various estimates.

1. Bottom-up approach

The first step in the bottom-up approach is to assess how the demand for AI hardware is affecting the production output of the semiconductor manufacturing industry. For the years 2023 and 2024, this assessment should commence with a focus on key bottlenecks in the manufacturing process of AI chips, because these limitations can help indicate the overall impact of AI hardware demand on the production output of the semiconductor industry. During these two years, the demand for AI chips skyrocketed as an increasing number of companies were training and deploying generative AI models. As such, the demand for AI chips that could facilitate the computational requirements of these AI models grew significantly. Nvidia, which was estimated to account for a 98% market share in data-center GPU shipments in 2023,[65] was reported to have reached its sales capacity limit for its best graphics processing unit (GPU) for AI workloads (the H100) for 2023 by August of the same year.[66] Market analysts have mentioned that a key bottleneck for Nvidia was the limited availability of advanced packaging technology known as chip-on-wafer-on-substrate (CoWoS), of Nvidia's supplier TSMC,[67] which co-packages processors such as graphics processing units (GPUs) with high bandwidth memory (HBM). CoWoS packaging has been used for almost all HBM-based devices during these years, while all AI accelerators made use of HBM (see Appendix Box 1). TSMC chair, Mark Liu, even acknowledged CoWoS packaging as the most limiting factor, stating the following in September 2023: "It is not the shortage of AI chips, it is the shortage of our CoWoS capacity." He added that the shortage was expected to persist for another one-and-a-half years.[68]

[65] HPCWire, "Nvidia Shipped 3.76 Million Data-Center GPUs in 2023, According to Study," June 10, 2024, accessed February 10, 2025, https://www.hpcwire.com/2024/06/10/Nvidia-shipped-3-76-million-data-center-gpus-in-2023-according-to-study/.

[66] Barron's, "Nvidia's Best AI Chips Sold Out Until 2024, Says Leading Cloud GPU Provider," August 9, 2023, accessed February 10, 2025, https://www.barrons.com/articles/Nvidia-ai-chips-coreweave-cloud-6db44825.

[67] SemiAnalysis, "Quarterly Ramp for Nvidia, Broadcom, Google, AMD, AMD Embedded (Xilinx), Amazon, Marvell, Microsoft, Alchip, Alibaba T-Head, ZTE Sanechips, Samsung, Micron, and SK Hynix," July 5, 2023, accessed February 10, 2025, https://semianalysis.com/2023/07/05/ai-capacity-constraints-cowos-and/.

[68] Tom's Hardware, "TSMC: Shortage of Nvidia's AI GPUs to Persist for 1.5 Years," September 8, 2023, accessed February 10, 2025, https://www.tomshardware.com/news/tsmc-shortage-of-Nvidias-ai-gpus-to-persist-for-15-years.

Appendix Box 1
AI memory challenges and a necessity for HBM/CoWoS

As generative AI models have rapidly grown larger and more complex over time, due to model size and performance being linked to each other,[69] their need for large amounts of dynamic random access memory (DRAM) has also increased. While both the training and inference phases of AI models require memory, the training phase typically has the highest demand for memory, as memory is primarily utilized for storing model weights (directly proportional to the model size) and key-value caching (i.e., storing key and value matrices from previous steps and reusing them to generate subsequent tokens).[70] In particular, the rate at which processors can read and store information in memory (i.e., memory bandwidth) has become a bottleneck for AI model size. While the number of floating point operations per second that processors can deliver has dramatically increased over the past eight years, increasing by more than 45,000% over different generations, the improvements in memory bandwidth have, however, not kept up with the same pace. Over the same time period, the memory speeds over different generations of AI hardware have increased by only around 1,000%.[71] This growing imbalance is known as the memory wall. HBM overcomes memory bandwidth limitations, was designed to address the memory wall challenge, as, contrary to traditional DRAM, it is made to be co-located with a processor within the same package.[72] Doing so enables a wider memory bus and, in turn, a higher bandwidth than could be achieved with traditional DRAM. HBM and the packaging technology to integrate HBM and processors within the same package (chip-on-wafer-on-substrate; CoWoS) have, therefore, become essential for making and operating large-scale AI models.

Various analysts have reported further details regarding TSMC's exact CoWoS capacity: In July 2023, TSMC's CoWoS capacity was limited to only 8,000 300-mm wafers per month, with plans to increase this figure to 11,000 wafers per month by the end of the same year and up to 16,600 wafers per month by the end of 2024.[73] In April 2024, it was reported that TSMC's CoWoS capacity had reached 13,000 wafers per month by the end of 2023 and was on track to reach 35,000 wafers per month by the end of 2024.[74] By October 2024, it was reported that TSMC was expected to reach a CoWoS capacity of 35,000 to 40,000 wafers per month, with further projections that this capacity could potentially double by the end of 2025.[75] For the remainder of this study we based our annual estimated CoWoS capacity on the highest of the aforementioned capacities in their respective years to better capture the limits of the environmental impact of AI chip manufacturing. This means that for 2023 we assumed a total CoWoS capacity of (13,000 * 12 =) 156,000 wafers. For 2024 we assumed this capacity had increased to (40,000 * 12 =) 480,000 wafers. To illustrate how CoWoS capacity causes a major production bottleneck, it should be considered that a single one of these wafers can only deliver 9 packaged Nvidia H100 chips.[76] Assuming a packaging yield of 99%,[77] we cahn find that an annual capacity of 156,000 CoWoS wafers in 2023 reaches a maximum of 1,389,960 packaged H100 chips using:

Equation 1

$$PU = W * DPW * PY$$

Wherein:

$$PU = \text{packaged units}$$
$$W = 300 \text{ mm wafer}$$
$$PY = \text{packaging yield}$$
$$DPW = \text{dice per wafer}$$

[69] Ananthaswamy A., "In AI, Is Bigger Always Better?" *Nature* 615,202–205 (March 8, 2023), https://doi.org/10.1038/d41586-023-00641-w.

[70] UnfoldAI, "GPU Memory Requirements for Serving Large Language Models," 2024, accessed February 10, 2025, https://unfoldai.com/gpu-memory-requirements-for-llms/.

[71] SemiAnalysis, "The Memory Wall: Past, Present, and Future of DRAM," September 3, 2024, accessed February 10, 2025, https://semianalysis.com/2024/09/03/the-memory-wall/.

[72] Chen, V. et al, "Overcoming Design Challenges for High Bandwidth Memory Interface with CoWoS," in *2022 IEEE International Symposium on Electromagnetic Compatibility & Signal/Power Integrity (EMCSI)*, Spokane, WA, USA, August 1-5, 2022, https://doi.org/10.1109/EMCSI39492.2022.10050234.

[73] Tom's Hardware, "TSMC Accelerates Expansion of Advanced Packaging Facilities: Report," July 15, 2023, accessed February 10, 2025, https://www.tomshardware.com/news/tsmc-accelerates-expansion-of-advanced-packaging-facilities-report.

[74] TweakTown, "TSMC's Top 3 Customers in 2023: Apple with 25%, Nvidia with 11%, and AMD with 7%," April 18, 2024, accessed February 10, 2025, https://www.tweaktown.com/news/97457/tsmcs-top-3-customers-in-2023-apple-with-25-Nvidia-11-and-amd-7/index.html.

[75] TrendForce, "TSMC's CoWoS Capacity Doubles for Two Years, Still Insufficient—Positive Outlook for Suppliers," October 21, 2024, accessed February 10, 2025, https://www.trendforce.com/news/2024/10/21/news-cowos-capacity-doubles-for-two-years-still-insufficient-positive-outlook-for-suppliers/.

[76] TheElec, "Samsung at a Crossroad as HBM Deal with Nvidia Yet to Be Sealed," June 20, 2024, accessed February 10, 2025, https://www.thelec.net/news/articleView.html?idxno=4882.

[77] Moomoo Technologies Inc, "Nvidia's New Chip Delayed? Don't Panic, the Impact Isn't That Big," 2024, accessed February 10, 2025, https://www.moomoo.com/news/post/41772359/Nvidia-s-new-chip-delayed-don-t-panic-the-impact.

Note that 300-mm (12 inch) wafers are assumed as the default wafer size for all our calculations, because all of the relevant production facilities identified in Section 1.4 use silicon wafers of this size,[78,79,80,81] as well as TSMC's CoWoS packaging process.[82] A total of 480,000 of these wafers in 2024 would translate to a maximum of 4,276,800 packaged H100 chips in a full year. It was previously reported that Nvidia was set to ship 550,000 H100 units in 2023 while planning to deliver 1.5 to 2 million H100 units in 2024.[83] Omdia was expecting Nvidia to fulfill 650,000 H100 orders in 2023.[84] It can, therefore, be assumed that Nvidia's H100 alone occupied a significant share of TSMC's CoWoS capacity in both 2023 and 2024.

However, Nvidia's H100 does not reflect the entire market for AI chips. Hence, it is still necessary to examine the volumes of alternative AI chips and the related supply chain impact. The starting point of this analysis is a roadmap of flagship AI chips, as presented by Nvidia and AMD,[85,86] depicted in Appendix Table 1:

Appendix Table 1 Flagship AI chip roadmap for Nvidia and AMD, in 2023 and 2024

Brand	Name	Released	2023 Q1	2023 Q2	2023 Q3	2023 Q4	2024 Q1	2024 Q2	2024 Q3	2024 Q4
Nvidia	A100	Q2, 2021	■	■	■	■	■	■	■	■
Nvidia	H100	Q1, 2023	■	■	■	■	■	■	■	■
Nvidia	H200	Q2, 2024						■	■	■
Nvidia	B100/B200	Q3, 2024							■	■
AMD	MI250X	Q4, 2021	■	■	■	■	■	■	■	■
AMD	MI300X	Q3, 2023				■	■	■	■	■
AMD	MI325X	Q4, 2024								■

Appendix Table 1 shows that Nvidia is estimated to have continued to dominate the data-center GPU market in 2024. AMD's market share is estimated at only 5–7%,[87] while Intel's market share is expected to remain at only 1%.[88] We can make several assumptions about the allocation of CoWoS capacity to various AI chips that were relevant in 2023 and 2024. A little over 43% of TSMC's maximum CoWoS capacity in 2023 (156,000 wafers) may have been used for Nvidia's H100, as 600,000 units (an average of the estimated 550,000 and 650,000 H100 shipments in 2023) would require a total of 67,340 packaging wafers after correcting for the packaging yield of 99%. To find this value we solve Equation 2 for W:

Equation 2

$$W = \frac{PU}{DPW*PY}$$

[78] TSMC, "TSMC Fabs, " accessed February 10, 2025, https://www.tsmc.com/english/aboutTSMC/TSMC_Fabs.

[79] EETimes, "Micron Closes Elpida Acquisition," July 31, 2013, accessed February 10, 2025, https://www.eetimes.com/micron-closes-elpida-acquisition/.

[80] The Elec, "SK Hynix wins large order for HBM from Broadcom," December 20, 2024, accessed February 10, 2025, https://www.thelec.net/news/articleView.html?idxno=5084.

[81] Samsung, "Manufacturing that Pushes Your Innovations Forward," accessed February 10, 2025, https://semiconductor.samsung.com/foundry/manufacturing/manufacturing-sites/.

[82] Tom's Hardware, "TSMC Explores Using 510×515 mm Rectangular Silicon Wafers — Tripling the Usable Area of Current 300mm Diameter Tech," June 20, 2024, accessed February 10, 2025, https://www.tomshardware.com/tech-industry/tsmc-explores-using-510×515-mm-rectangular-silicon-wafers-tripling-the-usable-area-of-current-300mm-diameter-tech.

[83] WCCFTech, "Nvidia Aims at Shipping Millions of AI GPUs by 2024, Working to Diversify Supply Chain," August 24, 2023, accessed February 10, 2025, https://wccftech.com/Nvidia-aims-at-shipping-millions-of-ai-gpus-by-2024-working-to-diversify-supply-chain/.

[84] Tom's Hardware, "Nvidia Sold Half a Million H100 AI GPUs in Q3 Thanks to Meta, Facebook—Lead Times Stretch Up to 52 Weeks: Report," November 28, 2023, accessed February 10, 2025, https://www.tomshardware.com/tech-industry/Nvidia-ai-and-hpc-gpu-sales-reportedly-approached-half-a-million-units-in-q3-thanks-to-meta-facebook.

[85] Nvidia, "Investor Presentation," 2023, accessed February 10, 2025, https://s201.q4cdn.com/141608511/files/doc_presentations/2023/Oct/01/ndr_presentation_oct_2023_final.pdf.

[86] WCCFTech, "AMD Instinct AI Accelerator Lineup Gets MI325X Refresh in Q4, 3nm MI350 'CDNA 4' in 2025, CDNA MI400 'CDNA Next' in 2026," June 2, 2024, accessed February 10, 2025, https://wccftech.com/amd-instinct-ai-accelerator-lineup-mi325x-refresh-q4-3nm-mi350-cdna-4-2025-cdna-mi400-cdna-next-2026/.

[87] Investopedia, "What You Need to Know Ahead of AMD's Advancing AI Event," October 9, 2024, accessed February 10, 2025, https://www.investopedia.com/what-you-need-to-know-ahead-of-amd-advancing-ai-event-8723294.

[88] CNBC, "Nvidia Dominates the AI Chip Market, but There's More Competition Than Ever," June 2, 2024, accessed February 10, 2025, https://www.cnbc.com/2024/06/02/Nvidia-dominates-the-ai-chip-market-but-theres-rising-competition-.html.

Broadcom, Google's partner for producing its tensor processing units (TPUs), reportedly took 10% of TSMC's CoWoS capacity,[89] which, therefore, comes down to a maximum of 15,600 packaging wafers. Given Nvidia's dominant 98% estimated market share in data-center GPU shipments in 2023, we assume that TSMC's remaining CoWoS capacity (a maximum of 156,000 – 15,600 – 67,340 = 73,060 wafers) for 2023 was allocated to Nvidia's A100. As the A100 package dimensions are almost the same as that of the H100, we assumed that (using Equation 1) 650,964 A100 units (rounded down) were produced in 2023.

The situation in 2024 was slightly more complex due to increasing competition and the fact that Nvidia discontinued the A100 while introducing newer devices such as the H200 and the next-generation Blackwell GPUs (e.g., the B100 and B200). The latter generation is also the first to make use of TSMC's CoWoS with local silicon interconnect (CoWoS-L) capacity,[90] adding additional complexity to the capacity analysis, because Nvidia's other devices rely on CoWoS with silicon interposer (CoWoS-S). Large monolithic silicon interposers are used in CoWoS-S, which simplifies the analysis. In CoWoS-L, these monolithic interposers are replaced with reconstituted interposers that consist of local silicon interconnect chiplets and global redistribution layers. However, while CoWoS-L is likely to become the primary CoWoS process in the near future,[91] the transition from CoWoS-S to CoWoS-L has limited relevance for analysis that covers 2024, as analysts did not expect many of these Blackwell GPUs to be produced in 2024 due to TSMC's limited CoWoS-L capacity. In September 2024 Analysts from Morgan Stanley estimated that Nvidia would produce around 450,000 of these Blackwell devices in the last quarter of 2024,[92] using about 32,000 CoWoS-L wafers to do so.[93] If we remove these wafers from TSMC's assumed total CoWoS capacity of 480,000 wafers, TSMC's maximum CoWoS-S capacity in 2024 would come down to 448,000 wafers. In terms of distribution, we assume Broadcom is taking a constant 10% of TSMC's total CoWoS capacity (i.e. a maximum of 48,000 wafers) based on Broadcom's estimated share in 2023, while Nvidia and AMD take up the rest.

Industry analysis estimated that AMD would be shipping around half a million of its MI300X in 2024.[94] This device requires an extra step in the production process because it relies on system-on-integrated-chips (SoIC) packaging, for which TSMC has a capacity of 5,000 to 6,000 wafers per month.[95] SoIC is used to vertically integrate accelerator complex dice (XCD) and input/output dice (IOD) before packaging these with HBM in CoWoS. Because Nvidia and AMD share the commonality of pushing chip designs to their maximum size, it can also be assumed that the yield on a single CoWoS wafer is similar for their devices. It would, therefore (using Equation 2), require 56,117 packaging wafers to produce half a million MI300X units. For Nvidia's H100, with 1.75 million estimated shipments in 2024 (an average of the estimated 1.5 million and 2 million H100 shipments in 2024), 196,408 (rounded down) packaging wafers would be required. Again, we assume the remainder of TSMC's maximum CoWoS-S capacity (i.e. 147,475 packaging wafers per month as the result of 480,000 – 32,000 – 48,000 – 56,117 – 196,408) was allocated to Nvidia's H200. Appendix Table 2, below, summarizes our best estimation of the number of AI chips produced in 2023 and 2024 per device type, which forms the foundation of our bottom-up approach.

[89] Tom's Hardware, "AMD and Nvidia GPUs Consume Lion's Share of TSMC's CoWoS Capacity,' June 13, 2023, accessed February 10, 2025, https://www.tomshardware.com/news/amd-and-Nvidia-gpus-consume-lions-share-of-tsmc-cowos-capacity.

[90] 3DInCites, "IFTLE 607: Why Nvidia's Blackwell Is Having Issues with TSMC CoWoS-L Technology," October 2, 2024, accessed February 10, 2025, https://www.3dincites.com/2024/10/iftle-607-why-Nvidias-blackwell-is-having-issues-with-tsmc-cowos-l-technology/

[91] DigiTimes, "Global Capacity of CoWoS Packaging," August 29, 2024, accessed February 10, 2025, https://www.digitimes.com/reports/item.asp?id=20240829RS400.

[92] Tom's Hardware, "Nvidia Expected to Produce 450,000 Blackwell AI GPUs in Q4—Potential $10B in Revenue for the Chipmaker," September 25, 2024, accessed February 10, 2025, https://www.tomshardware.com/pc-components/gpus/Nvidia-expected-to-produce-450000-blackwell-ai-gpus-in-q4-potential-dollar10b-in-revenue-for-the-chipmaker.

[93] DigiTimes, "Global CoWoS and CoWoS-like Packaging Capacity Demand to Surge 113% on Year in 2025, Says DIGITIMES Research," October 25, 2024, accessed February 10, 2025, https://www.digitimes.com/news/a20241025VL210/cowos-demand-packaging-2025-capacity.html.

[94] WCCFTech, "AMD To Ship 'Half a Million' Instinct MI300X AI Accelerators in 2024," October 10, 2024, accessed February 10, 2025, https://wccftech.com/amd-ship-half-a-million-instinct-mi300x-ai-accelerators-2024/.

[95] TrendForce, "TSMC's SoIC Demand Heats Up, Reports Suggest Significant Capacity Expansion," January 18, 2024, accessed February 10, 2025, https://www.trendforce.com/news/2024/01/18/news-tsmcs-soic-demand-heats-up-reports-suggest-significant-capacity-expansion/.

Appendix Table 2 **Estimated device production in 2023 and 2024**

GPU	Year	Estimated unit production
Nvidia A100	2023	650,964
Nvidia H100	2023	600,000
Nvidia H100	2024	1,750,000
Nvidia H200	2024	1,314,000
Nvidia B100/B200	2024	450,000
AMD MI300X	2024	500,000

1.1 In-house solutions and other components

Section 1 covered how Broadcom is one of the parties in addition to Nvidia and AMD to use TSMC's CoWoS supply. While this information was used to correct the unit production estimates for Nvidia and AMD, no further analysis has been performed on the supply chain impact of the TPUs that Broadcom provides to Google. The reason for this lack is that these devices are inherently more difficult to analyze. Multiple big tech companies, such as Google, are developing their own AI accelerators to reduce their reliance on Nvidia,[96] but these accelerators are typically not for sale. For example, Google's TPUs can only be rented on Google's cloud services. It makes the task of simply obtaining relevant, complete, and accurate product specifications relatively difficult. Moreover, these specifications may become available relatively late (if they are disclosed at all) compared to those belonging to the devices sold to the general public. In-house solutions are, therefore, not within the scope of this analysis. The same idea applies to other components such as central processing units (CPUs), double data rate random access memory (DRR RAM), and not-AND (NAND) storage, which are typically used with AI accelerators in complete server systems. These components may be more generic than the AI chips and HBM covered previously, making the impact of AI demand on the manufacturers of these components harder to isolate and analyze.

[96] Verdict, "Explainer: Amazon to Unveil DIY AI Chips as Big Tech Hedges Against AI Chip Supply Squeeze," November 27, 2024, accessed February 10, 2025, https://www.verdict.co.uk/explainer-amazon-to-unveil-diy-ai-chips-as-big-tech-hedges-against-ai-chip-supply-squeeze/.

1.2 Production process

The estimated AI chip sales can be combined with the respective product specifications to determine the supply chain impact of the various components being packaged with CoWoS. The product specifications are provided in Appendix Table 3 (there may be several variants of a certain chip,e.g. Nvidia's H800 and H20 are similar to the H100, but we don't make a distinction between these variants as they also have similar production requirements which makes them interchangeable for the purpose of our calculations over 2023 and 2024):

Appendix Table 3 **Selected data-center GPU specifications obtained through SemiAnalysis,[97,98] WCCFTech[99] and Tom's Hardware[100]**

GPU	Nvidia A100	Nvidia H100	Nvidia H200	Nvidia B100/B200	AMD MI300X
GPU die size	826 mm²	814 mm²	814 mm²	~814 mm²*	115 mm²
GPU dice	1	1	1	2	8 (4 IOD)
Manufacturing process	7nm (TSMC)	4nm (TSMC)	4nm (TSMC)	4nm (TSMC)	5/6nm (TSMC)
Memory interface	HBM2	HBM3	HBM3e	HBM3e	HBM3
Memory size	80 GB	80 GB	141 GB	192 GB	192 GB
Memory stacks	5	5	6	8	8
Memory dice per stack	8+1	8+1	8+1	8+1	8+1

*The exact size of these models was not disclosed by the time of release, while it is estimated that their die size should be between the previous values and the EUV reticle limit of 858mm^2.

Notably, both Nvidia and AMD rely on TSMC for their GPUs, accelerator chiplets (XCDs), and IO dice (IODs). Given TSMC's dominance in the production of advanced logic for AI accelerators, its significant involvement is to be expected.[101] The landscape of HBM manufacturing appears more diverse with more relevant parties (SK hynix, Micron, and Samsung). Whereas Nvidia previously relied solely on SK hynix for its HBM until the first quarter of 2024, it was announced that Micron would provide the HBM for Nvidia's H200,[102] while it would also be used with Nvidia's B200. AMD reportedly obtains its HBM from Samsung,[103] but Samsung has otherwise fallen behind in the HBM market due to disappointing HBM yields.[104]

[97] SemiAnalysis, "AI Capacity Constraints—CoWoS and HBM Supply Chain," July 5, 2023, accessed February 10, 2025, https://semianalysis.com/2023/07/05/ai-capacity-constraints-cowos-and/.

[98] SemiAnalysis, "AMD MI300—Taming the Hype—AI Performance, Volume Ramp, Customers, Cost, IO, Networking, Software," June 12, 2023, accessed February 10, 2025, https://semianalysis.com/2023/06/12/amd-mi300-taming-the-hype-ai-performance/.

[99] WCCFTech, "Nvidia's H100 GPUs & the AI Frenzy; a Rundown of Current Situation," 2023, accessed February 10, 2025, https://wccftech.com/Nvidia-h100-gpus-a-rundown-of-current-situation/.

[100] Tom's Hardware, "Nvidia's Next-Gen AI GPU Is 4X Faster Than Hopper: Blackwell B200 GPU Delivers up to 20 Petaflops of Compute and Other Massive Improvements," 2024, accessed February 10, 2025, https://www.tomshardware.com/pc-components/gpus/Nvidias-next-gen-ai-gpu-revealed-blackwell-b200-gpu-delivers-up-to-20-petaflops-of-compute-and-massive-improvements-over-hopper-h100.

[101] SemiWiki, "No! TSMC Does Not Make 90% of Advanced Silicon," 2024, accessed February 10, 2025, https://semiwiki.com/semiconductor-manufacturers/tsmc/342934-no-tsmc-does-not-make-90-of-advanced-silicon/.

[102] TrendForce, "Micron Begins Mass Production of HBM3e for Nvidia's H200," February 27, 2024, accessed February 10, 2025, https://www.trendforce.com/news/2024/02/27/news-micron-begins-mass-production-of-hbm3e-for-Nvidias-h200/.

[103] WCCFTech, "Samsung Receives Huge Order of HBM3 Memory to Power AMD MI300X GPUs," August 23, 2024, accessed February 10, 2025, https://wccftech.com/samsung-receives-huge-order-of-hbm3-memory-to-power-amd-mi300x-gpus/.

[104] WCCFTech, "Samsung Lowers HBM Production Capacity Amid 'Sluggish' Business & Failure to Grab Market Opportunities," October 15, 2024, accessed February 10, 2025, https://wccftech.com/samsung-lowers-hbm-production-capacity/.

Using the above specifications, we can estimate the wafer demand related to device production. It can be illustrated by further examining Nvidia's H100, which contains five stacks of HBM3 (along with one dummy die) surrounding a single GPU die measuring 814 mm^2 in size. Assuming a perfect yield of 57 dice per 300-mm wafer (see Appendix Table 4), 600,000 packaged GPUs would require 10,633 logic wafers. This value is derived using equation 3:

Equation 3

$$W = \frac{\frac{PU*DPU}{PY}}{DPW*DY}$$

Wherein: DY=die yield
DPU=dice per packaged unit

Since the H100 die size of 814 mm^2 is quite large, we expect the number of good dice per wafer to be significantly less than 57. Considering the fact that Nvidia uses a customized version of TSMC's 5-nm process for its H100, for which a defect density of 0.1 per cm2 has been reported,[105] we can assume a die yield of 46.4%, whereas PY is assumed to be 99%[106] (see Appendix Table 4). As a single package requires only a single GPU die, the DPU is set to one. It raises the demand for logic wafers to the following:

$$W = \frac{\frac{600{,}000*1}{99\%}}{57*46.4\%} = 22{,}915$$

The memory requirements are significantly higher due to the fact that the five HBM3 stacks on Nvidia's H100 are eight-high die stacks. A single 300-mm wafer would deliver a maximum of 476 of these dice (see Appendix Table 4), while 40 are necessary for a single H100. For this reason, 600,000 packaged H100 units require at least 50,929 DRAM wafers following the previous equation with DPU equal to 40, DPW equal to 476 and a DY of 100%.

However, the yields for HBM are affected not only by die yields but also by packaging yields in the process of stacking DRAM dice before packaging the stacked DRAM in CoWoS. We must, therefore, amend Equation 4 as follows:

Equation 4

$$W = \frac{\frac{PU*DPU}{PY}}{DPW*DY*MPY}$$

Wherein: MPY = memory packaging yield

[105] AnandTech, "'Better Yield on 5nm Than 7nm': TSMC Update on Defect Rates for N5," August 25, 2020, accessed February 10, 2025, https://www.anandtech.com/show/16028/better-yield-on-5nm-than-7nm-tsmc-update-on-defect-rates-for-n5.

[106] Moomoo Technologies Inc, "Nvidia's new chip delayed? Don't panic, the impact isn't that big," August 5, 2024, accessed February 17, https://www.moomoo.com/news/post/41772359/Nvidia-s-new-chip-delayed-don-t-panic-the-impact.

The other input variables remain the same (PU, DPU, PY, DPW and DY). As SK hynix, the sole supplier of HBM to Nvidia until the first quarter of 2024,[107] disclosed that it was close to achieving its target yield of 80% for HBM3e chips,[108] we now assume DY multiplied with MPY is equal to 80%. We then find the following:

$$W = \frac{\frac{600{,}000*40}{99\%}}{476*100\%*80\%} = 63{,}662$$

Since each stack of HBM sits on a base logic die, another 7,576 logic wafers are required in the production process (assuming similar dimensions and yields per wafer). The only value that changes in the previous equation is DPU, which is only one per HBM stack. It leads to five per packaged H100, giving the following:

$$W = \frac{\frac{600{,}000*5}{99\%}}{476*80\%} = 7{,}958$$

Altogether, it comes down to an estimated production requirement of 22,915 wafers for GPU and 71,620 wafers for memory to ultimately produce 600,000 of Nvidia's packaged H100 units. Given the dominance of Nvidia's H100 within the AI chip landscape, this example reflects most of the impact the demand for AI had on the total production output of the semiconductor manufacturing industry in 2023. Additionally, multiplying these figures by 2.92, reflecting estimated H100 shipments in 2024, would also capture a significant part of the impact that the demand for AI had on the total production output of the semiconductor manufacturing industry in this year. To capture the wafer demand impact of other devices being considered, we will use Equation 4 while applying the following assumptions:

Appendix Table 4 **Assumed yields on a 300-mm wafer for Nvidia GPUs, AMD XCDs, and IODs and multiple types of HBM**

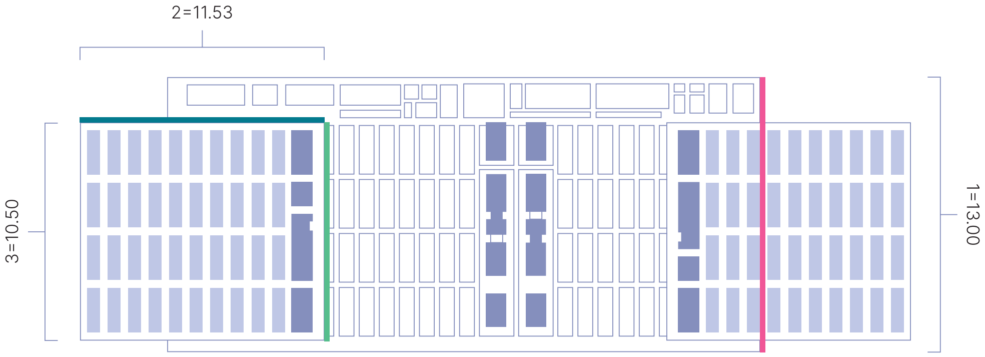
Wafer purpose	Die dimension	DPW	DY	Good dice per wafer
Nvidia GPUs	25.59×32.16mm	57	46.4%	26.4
AMD XCD	10.5×11.5mm	477	88.7%	423.1
AMD IOD	13×29mm	140	69.4%	97.2
HBM3/3e	11×11mm	476	80.0%	380.8
HBM2	7.75×11.87mm	634	80.0%	507.2

[107] Reuters, "Nvidia Supplier SK Hynix Says HBM Chips Almost Sold Out for 2025," May 2, 2024, accessed February 10, 2025, https://www.reuters.com/technology/Nvidia-supplier-sk-hynix-says-hbm-chips-almost-sold-out-2025-2024-05-02/.

[108] TrendForce, "SK Hynix Revealed Progress for HBM3e, Achieving Nearly 80% Yield," May 24, 2024, accessed February 10, 2025, https://www.trendforce.com/news/2024/05/24/news-sk-hynix-revealed-progress-for-hbm3e-achieving-nearly-80-yield/.

Appendix Table 4 suggests that there are several calculators available that can help to determine DPW and DY values given a die dimension, wafer diameter and defect density. For this study we used iSine's die yield calculator[109] with the die dimensions as listed in Appendix Table 4, a wafer diameter of 300 mm, a defect density of 0.1 per square centimeter[110] and default assumptions on scribe and edge loss. AMD disclosed the MI300X IOD dimensions,[111] while we can estimate the XCD dimensions to be 10.5 mm x 11.5 mm based on the same deck (see Appendix Figure 1).

Appendix Figure 1 **Estimated XCD dimensions of AMD's MI300X in millimetre**



In a similar way we can estimate the dimensions of a Nvidia H100 GPU die at 25.59 mm x 32.16 mm.[112] The dimensions are close to the maximum size (26 mm x 33 mm) that lithography equipment can print[113] and all Nvidia GPUs under consideration have a similar surface area close to the reticle limit (see Appendix Table 3), we also apply the same yield estimate for all Nvidia GPUs. HBM2 and HBM3e die dimensions are obtained from JEDEC[114] and Micron.[115] Since SK hynix disclosed it was reaching an 80% yield on HBM3e,[116] we apply this as the fixed value for HBM yield, rather than the calculated yield.

Using the equations and assumptions detailed above, we estimate the wafer demand related to each device type, providing us with the following results, displayed in Appendix Table 5:

Appendix Table 5 **Estimated wafer demand for selected device types**

GPU	Year	Logic wafers	Memory wafers	Memory base
Nvidia A100	2023	24,862	51,856	6,482
Nvidia H100	2023	22,915	63,662	7,958
Nvidia H100	2024	66,836	185,680	23,210
Nvidia H200	2024	50,184	167,303	20,913
Nvidia B100/200	2024	37,810	84,034	10,504
AMD MI300X	2024	30,342	84,882	10,610

[109] iSine, "Die Yield Calculator," accessed February 10, 2025, https://isine.com/resources/die-yield-calculator/.
[110] AnandTech., "'Better Yield on 5nm than 7nm': TSMC Update on Defect Rates for N5," August 25, 2020, accessed February 10, 2025, https://www.anandtech.com/show/16028/better-yield-on-5nm-than-7nm-tsmc-update-on-defect-rates-for-n5.
[111] AMD, "AMD Instinct MI300X Accelerator: Packaging and Architecture Co-Optimization," 2024, accessed February 10, 2025, https://vlsi24.mapyourshow.com/mys_shared/vlsi24/handouts/JFS2-4_Smith.pdf.
[112] HotHardware, "Nvidia H100 Hopper GPU with 80GB HBM3 Gets Pictured and It's a Beast," May 6, 2022, accessed February 10, 2025, https://hothardware.com/news/h100-hopper-gpu-gets-pictured.
[113] ISDI, "The Art of CMOS Stitching: Challenges and Solutions," October 30, 2024, accessed February 10, 2025, https://www.isdicmos.com/news-events/2024/10/30/the-art-of-cmos-stitching-challenges-and-solutions.
[114] AnandTech, "JEDEC Publishes HBM2 Specification as Samsung Begins Mass Production of Chips," January 20, 2016, accessed February 10, 2025, https://www.anandtech.com/show/9969/jedec-publishes-hbm2-specification.
[115] Micron, "Introducing Memory Built for AI Innovation," 2023, accessed February 10, 2025, https://www.micron.com/content/dam/micron/global/public/documents/products/product-flyer/hbm3e-product-brief.pdf.
[116] TrendForce, "SK Hynix Revealed Progress for HBM3e, Achieving Nearly 80% Yield," May 24, 2024, accessed February 10, 2025 https://www.trendforce.com/news/2024/05/24/news-sk-hynix-revealed-progress-for-hbm3e-achieving-nearly-80-yield/.

1.3 Electricity demand

After obtaining insights relating to how demand for AI affects the production output of the semiconductor manufacturing industry, the next step is to estimate the electricity consumption related to the identified changes. The easiest way to do so is by multiplying the estimated changes in total wafer output with the electricity consumption for producing the wafers, accounting for the differences in the processes for producing logic, DRAM and NAND wafers. For the production of logic wafers, we refer to the research conducted by Bardon et al. (2022),[117] in which the total electricity consumption for manufacturing (per square centimeter of wafer production) for full process flows (i.e., the full front end of line (FEOL), middle end of line (MEOL) and back end of line (BEOL) processes for logic technologies was evaluated (Appendix Table 6). It should be noted that these electricity consumption values not only cover the electricity consumption of the tools used in the manufacturing processes but also that of the facility equipment (e.g., cooling and vacuum pumps) supporting these processes.

Appendix Table 6 **Total electricity consumption for manufacturing logic technologies**[118]

Technology node (nm)	Total manufacturing electricity per cm^2 for full process flows (kWh/cm^2)
28	0.943
20	1.146
14	1.134
10	1.465
8 (EUV)	1.639
7 (EUV)	2.098
6	2.767
5	2.871
3	3.273

[117] Garcia Bardon M. et al., "DTCO Including Sustainability: Power-Performance-Area-Cost-Environmental Score (PPACE) Analysis for Logic Technologies," in *2020 IEEE International Electron Devices Meeting (IEDM),* San Francisco, CA, USA, December 12–18, 2020, https://doi.org/10.1109/IEDM13553.2020.9372004.

[118] Garcia Bardon M. et al., "DTCO Including Sustainability: Power-Performance-Area-Cost-Environmental Score (PPACE) Analysis for Logic Technologies," in *2020 IEEE International Electron Devices Meeting (IEDM),* San Francisco, CA, USA, December 12–18, 2020, https://doi.org/10.1109/IEDM13553.2020.9372004.

The values from Appendix Table 6, above, can be directly applied to the matching TSMC technology process indicated in the previous section (while considering that a 300-mm wafer has a surface area of 706.95 cm²). Note that Nvidia uses customized versions of TSMC's 5-nm process for some of its devices. Because TSMC includes these processes under its 5-nm family[119] we assume that the respective 5-nm electricity intensity applies in this case. In general, HBM and DRAM processes do not shrink to the same levels, having been limited to 10-nm processes in recent history.[120] SK hynix made use of a 1-βnm process for its HBM3e.[121] However, it is still a 10-nm-class process. NAND faces similar scaling challenges.[122] Additionally, while the manufacturing processes for the different types of chips vary, research suggests an identical contribution of electricity in the cumulative energy demand for manufacturing different chip types with a similar technology node.[123] We can, therefore, use the same electricity intensities listed above for different chip types. Moreover, for the HBM base dice we also apply the same electricity intensity as the rest of the HBM stacks.

With the estimates for the electricity consumption of wafer production, we can use the estimated changes in wafer production output due to the demand for AI, as outlined in the previous section, to find the increase in electricity consumption in semiconductor manufacturing due to AI:

Equation 5

$$\Delta EW_p = \Delta W_p E_p$$

Wherein: ΔW_p=*change in wafer output W for specific technology process p*

Ep=electricity intensity $\left[\frac{kWh}{wafer} \right]$ *of wafer production for the full process flow of technology process p*

ΔEW_p=*change in total electricity consumption*

relating to the change in wafer output W for technology process p

Therefore, we can sum up the changes in electricity consumption of wafer production for each technology process ***p(p=1,…,n)*** to find the combined total change in electricity consumption of semiconductor manufacturing due the demand for AI hardware ΔET using the following formula:

Equation 6

$$\Delta ET = \sum_{p=1}^{n} \Delta EW_p$$

[119] TSMC, "Advanced Technologies for HP," 2024, accessed February 10, 2025, https://www.tsmc.com/english/dedicatedFoundry/technology/platform_HPC_tech_advancedTech.

[120] Blocks&Files, "Why DRAM Is Stuck in a 10nm Trap," April 13, 2020, accessed February 10, 2025, https://blocksandfiles.com/2020/04/13/dram-is-stuck-in-a-10nm-process-trap/.

[121] WCCFTech, "SK hynix's 1bnm Process to Power Next-Gen DDR5 RDIMM & HBM3E DRAM Solutions," May 30, 2023, accessed February 10, 2025, https://wccftech.com/sk-hynixs-1bnm-process-to-power-next-gen-ddr5-rdimm-hbm3e-dram-solutions/.

[122] Parat, K. and Goda, A., "Scaling Trends in NAND Flash," in *2018 IEEE International Electron Devices Meeting (IEDM),* San Francisco, CA, USA, December 1–5, 2018, https://doi.org/10.1109/IEDM.2018.8614694.

[123] Nagapurkar, P., and Das S., "Economic and Embodied Energy Analysis of Integrated Circuit Manufacturing Processes," *Sustainable Computing: Informatics and Systems* 35 (June 11, 2022): 100771, https://doi.org/10.1016/j.suscom.2022.100771.

In our case we only consider logic, DRAM and NAND wafer production (with the latter only for future electricity consumption scenarios). The results of applying this approach are shown in Appendix Table 7:

Appendix Table 7 **Estimated electricity consumption for manufacturing AI-related wafers**

GPU	Year	Logic wafers (GWh)	Memory wafers (GWh)	Memory base (GWh)
Nvidia A100	2023	36.9	53.7	6.7
Nvidia H100	2023	46.5	65.9	8.2
Total	**2023**	**83.4**	**119.6**	**15.0**
Nvidia H100	2024	135.7	192.3	24.0
Nvidia H200	2024	101.9	173.3	21.7
Nvidia B100/B200	2024	76.7	87.0	10.9
AMD MI300X	2024	61.6	87.9	11.0
Total	**2024**	**375.8**	**540.5**	**67.6**

1.4 Electricity mix and carbon intensity

After examining the AI-related electricity consumption of semiconductor manufacturing, we consider the electricity mix powering this part of semiconductor manufacturing, allowing us to observe the role of fossil fuels in the manufacturing process and establish the related carbon emissions. For this analysis, we are interested in both the location-based and market-based emissions of AI-related semiconductor manufacturing. The former captures the carbon intensity of the power grids where manufacturing takes place, whereas the latter reflects the carbon emissions of specific purchased electricity (e.g. including renewable energy contracts). Thus, we must attribute the estimated electricity consumption to both geographic locations as well as specific manufacturers. A challenge here is that, while extensive information can be found on worldwide semiconductor fabrication plants and their production capacity, it is typically not clear what part of the production capacity is being used for AI-related purposes. However, we know that popular GPUs such as Nvidia's A100 and H100 rely on TSMC's advanced manufacturing process (technology node <= 7nm, see Appendix Table 3).

The manufacturing process significantly narrows the focus on the relevant fabs for logic chips. TSMC indicates Fab 18, located in Southern Taiwan Science Park, as its main 5-nm production facility, which includes the customized 4 N production process that Nvidia uses for its H100 GPU.[124] Additionally, Fab 15, located in Central Taiwan Science Park, is key to the 7-nm process leveraged in Nvidia's A100 GPU.[125] Moreover, SK hynix, as mentioned earlier, was the sole supplier of HBM to Nvidia until the first quarter of 2024,[126] which also narrows the focus in terms of the relevant DRAM production. SK hynix's main facilities for HBM production have been Fabs M10 and M16 in Incheon, South Korea,[127] with M10 only recently converting some of its capacity to HBM products.[128] It was announced that Micron will be supplying the HBM3e for Nvidia's H200. Micron's main HBM production facility is located in Hiroshima, Japan.[129] The same HBM3e will also be used for Nvidia's B200.[130] AMD partnered with Samsung for the HBM of the MI300X.[131] The HBM manufacturing capacity of the latter company is located in Pyeongtaek and Hwaseong, South Korea.[132] Appendix Table 8 summarizes how we attribute the estimated wafer demand for each product line under consideration, as outlined in the previous section, to the aforementioned manufacturers and manufacturing locations:

Appendix Table 8 **Device manufacturers and manufacturing locations**

GPU	Nvidia A100	Nvidia H100	Nvidia H200	Nvidia B100/ B200	AMD MI300X
GPU/XCD/IOD Manufacturer	TSMC	TSMC	TSMC	TSMC	TSMC
Manufacturing location	Taiwan	Taiwan	Taiwan	Taiwan	Taiwan
Memory manufacturer	SK hynix	SK hynix	Micron	Micron	Samsung
Memory manufacturing location	South Korea	South Korea	Japan	Japan	South Korea

[124] TSMC, "5nm Technology," 2024, accessed February 10, 2025, https://www.tsmc.com/english/dedicatedFoundry/technology/logic/l_5nm.

[125] TSMC, "N7+ Technology," 2024, accessed February 10, 2025, https://www.tsmc.com/english/campaign/N7plus.

[126] Reuters, "Nvidia Supplier SK Hynix Says HBM Chips Almost Sold Out for 2025," May 2, 2024, accessed February 10, 2025, https://www.reuters.com/technology/Nvidia-supplier-sk-hynix-says-hbm-chips-almost-sold-out-2025-2024-05-02/

[127] TweakTown, "RAM SK hynix Preparing to Expand DRAM Capacity (HBM, DRAM) by 80K Wafers per Month at M16, M10 Fabs," August 14, 2024, accessed February 10, 2025, https://www.tweaktown.com/news/99896/sk-hynix-preparing-to-expand-dram-capacity-hbm-by-80k-wafers-per-month-at-m16-m10-fabs/index.html.

[128] DigiTimes, "SK Hynix Converts Part of M10 Production Lines to Produce HBM Products," July 5, 2024, accessed February 10, 2025, https://www.digitimes.com/news/a20240705PD203/sk-hynix-production-hbm-plant-2024.html.

[129] TrendForce, "Samsung, SK Hynix and Micron Ramp Up HBM Production, Reportedly Doubling Output Next Year," July 10, 2024, accessed February 10, 2025, https://www.trendforce.com/news/2024/07/10/news-samsung-sk-hynix-and-micron-ramp-up-hbm-production-reportedly-doubling-output-next-year/.

[130] TheNextPlatform, "Micron Is Fashionably Late to the HBM Party, but Not Too Late," December 19, 2024, accessed February 10, 2025, https://www.nextplatform.com/2024/12/19/micron-is-fashionably-late-to-the-hbm-party-but-not-too-late/.

[131] WCCFTech, "Samsung Receives Huge Order of HBM3 Memory To Power AMD MI300X GPUs," https://wccftech.com/samsung-receives-huge-order-of-hbm3-memory-to-power-amd-mi300x-gpus/. August 23, 2023, accessed February 10, 2025

[132] TrendForce, "Samsung, SK Hynix and Micron Ramp Up HBM Production, Reportedly Doubling Output Next Year," July 10, 2024, accessed February 10, 2025, https://www.trendforce.com/news/2024/07/10/news-samsung-sk-hynix-and-micron-ramp-up-hbm-production-reportedly-doubling-output-next-year/.

We have a good sense of direction in terms of where AI hardware is being manufactured and can attribute the electricity consumption (from Appendix Table 7) to the relevant regions involved in the manufacturing process of each component (Appendix Table 8).The results are shown in Appendix Table 9, below:

Appendix Table 9 **Estimated electricity consumption for AI-related manufacturing per region**

Region	Year	Electricity consumption (GWh)
Taiwan	2023	83.4
South Korea	2023	134.6
Total	**2023**	**218.0**
Taiwan	2024	375.8
South Korea	2024	315.2
Japan	2024	292.8
Total	**2024**	**983.9**

For the location-based approach we apply the electricity mix for the respective regions, displayed in Appendix Table 10, below:

Appendix Table 10 **Electricity generation mix of South Korea,[133] Taiwan,[134] and Japan[135]**

South Korea

Electricity generation sources	Value	Year	Units	Percentage
Coal	184,927,212	2023	MWh	31.4%
Oil	1,486,504	2023	MWh	0.3%
Natural gas	157,749,022	2023	MWh	26.8%
Biofuels	12,576,607	2023	MWh	2.1%
Nuclear	180,494,096	2023	MWh	30.7%
Hydro	3,716,437	2023	MWh	0.6%
Solar PV	29,288,018	2023	MWh	5.0%
Wind	3,382,450	2023	MWh	0.6%
Tide	437,567	2023	MWh	0.1%
Other sources	13,988,591	2023	MWh	2.4%

Taiwan

Electricity generation sources	Value	Year	Units	Percentage
Coal	119,157,090	2023	MWh	42.2%
Oil	3,775,616	2023	MWh	1.3%
Natural gas	111,629,599	2023	MWh	39.5%
Biofuels	238,612	2023	MWh	0.1%
Waste	3,499,451	2023	MWh	1.2%
Nuclear	17,801,952	2023	MWh	6.3%
Hydro	7,014,148	2023	MWh	2.5%
Geothermal	23,164	2023	MWh	0.0%
Solar PV	12,908,690	2023	MWh	4.6%
Wind	6,238,287	2023	MWh	2.2%

[133] EPSIS, "Generation Output by Energy Source," 2024, accessed February 10, 2025, https://epsis.kpx.or.kr/epsisnew/selectEkgeGepGesGrid.do.

[134] MOEAEA, "Electricity Statistics," 2024, accessed February 10, 2025, https://www.esist.org.tw/database/search/electric-generation.

[135] METI, "General Energy Statistics," 2024, accessed February 10, 2025, https://www.enecho.meti.go.jp/statistics/total_energy/results.html.

Japan

Electricity generation sources	Value	Year	Units	Percentage
Coal	280,400,000	2023	MWh	28.5%
Oil	71,600,000	2023	MWh	7.3%
Natural gas	324,100,000	2023	MWh	32.9%
Biofuels	40,100,000	2023	MWh	4.1%
Nuclear	84,100,000	2023	MWh	8.5%
Hydro	74,800,000	2023	MWh	7.6%
Geothermal	3,400,000	2023	MWh	0.3%
Solar PV	96,500,000	2023	MWh	9.8%
Wind	10,500,000	2023	MWh	1.1%

Because we intend to determine the location-based carbon emissions related to electricity consumption, the carbon intensity of generating electricity in gCO_2/kWh on each of the aforementioned power grids is listed in Appendix Table 11, below. With respect to Appendix Table 11, note that 2023 was the most recent available year at the time of writing this research report, hence these values were applied to both 2023 and 2024.

Appendix Table 11 **Carbon intensity of electricity generation in South Korea,[136] Taiwan,[137] and Japan[138]**

Region	Year	gCO_2/kWh
Taiwan	2023	494
South Korea	2023	431
Japan	2023	451

[136] Ember, "Energy Institute—Statistical Review of World Energy (2024). Carbon Intensity of Electricity Generation," 2024, accessed February 10, 2025, https://ourworldindata.org/grapher/carbon-intensity-electricity.

[137] MOEAEA, "Electricity Carbon Emission Factor," 2024, accessed February 10, 2025, https://www.moeaea.gov.tw/ecw/populace/content/SubMenu.aspx?menu_id=114.

[138] METI, "General Energy Statistics," 2024, accessed February 10, 2025, https://www.enecho.meti.go.jp/statistics/total_energy/results.html.

Using the electricity mix, we can estimate the location-based carbon emissions related to the estimated electricity consumption, which is shown in Appendix Table 12, below:

Appendix Table 12 **Estimated carbon emissions of AI-related manufacturing in kilotons (kt) of CO_2 equivalent per region**

Region	Year	Electricity consumption (GWh)	CO_2eq emissions (kt)
Taiwan	2023	83.4	41.2
South Korea	2023	134.6	58.0
Total	**2023**	**218.0**	**99.2**
Taiwan	2024	375.8	185.7
South Korea	2024	315.2	135.9
Japan	2024	292.8	132.1
Total	**2024**	**983.9**	**453.6**

We were also interested in the market-based carbon emissions of AI-related semiconductor manufacturing, we examined manufacturer environmental reports to obtain the relevant electricity mix and emission factors. Unfortunately, manufacturers do not all provide the same type of information regarding carbon emissions related to electricity consumption. Both TSMC and Samsung disclose market-based and location-based Scope 2 emissions, but SK Hynix only reports location-based Scope 2 emissions and the combined market-based Scope 1 & 2 emissions. Micron only reports market-based Scope 2 emissions. This means it is not possible to do a complete market-based analysis. We also cannot fully compare our location-based outcomes. When data are available, it is often only provided at a higher level than AI-related manufacturing. Chapter 3 results suggest that AI-related manufacturing represents a relatively small fraction of the total semiconductor manufacturing electricity consumption in 2023 and 2024, meaning that company-level information is probably not representative for this specific sub-activity. Therefore, market-based emissions were removed from the scope of our analysis.

2. Scenario analysis

For this research, we are not only interested in the potential impact of AI in semiconductor manufacturing during 2023 and 2024. Rather, we project forward to the year 2030. However, the impact of AI on semiconductor manufacturing by 2030 cannot be determined by simply extrapolating the current demand for AI hardware, as hardware architectures are likely to evolve. For this reason, we examine several scenarios for the future development of AI-related wafer supply and demand. Management consulting firms such as McKinsey have publicly assessed several scenarios,[139] a summary of which has been provided in Appendix Table 13, below.

Appendix Table 13 **McKinsey's estimated AI-related logic, DRAM, and NAND wafer demand in the year 2030**

2030 WSPY (million)	Logic	DRAM	NAND
Conservative	1.2	7.1	1.7
Base	2.4	13.6	4
Ambitious	3.6	21.0	7.9

The three McKinsey scenarios can be translated into electricity requirements by combining the aforementioned figures with the estimated electricity intensity of wafer production, as discussed in Section 1.3. Notably, we keep the relevant electricity intensity of wafer production constant over time. For example, we assume the current estimated electricity intensity of the 5-nm process as the electricity intensity of producing advanced logic wafers in 2030. While the electricity intensity of the manufacturing process is likely to improve over time, we also note that AI accelerators may soon adopt more advanced manufacturing processes. Companies such as Nvidia and AMD are already planning to use TSMC's 3-nm process,[140] which is more electricity intensive than the 5-nm process (see Appendix Table 6). Moreover, TSMC has already commenced the trial production for its new 2-nm process.[141] Keeping electricity intensities constant, therefore, allows us to balance between these potential future developments.

We apply a similar approach when it comes to the electricity mix and carbon intensity of AI-related semiconductor manufacturing. It means that the manufacturers and locations that are currently emerging as key drivers for manufacturing AI hardware will continue to assume that role in the upcoming years.

Appendix Table 14 **Estimated electricity consumption for AI-related chip manufacturing in 2030 per scenario**

2030 TWh	Logic	DRAM	NAND	Total
Conservative	2.44	7.35	1.76	11.55
Base	4.87	14.09	4.14	23.10
Ambitious	7.31	21.75	8.18	37.24

[139] McKinsey & Company, "Generative AI: The Next S-Curve for the Semiconductor Industry?" March 29, 2024, accessed February 10, 2025, https://www.mckinsey.com/industries/semiconductors/our-insights/generative-ai-the-next-s-curve-for-the-semiconductor-industry.

[140] WCCFTech, "TSMC 3nm Set to Witness Massive Adoption from AI Tech Giants; Nvidia Rubin, AMD Instinct MI355X & Intel Falcon Shores," October 14, 2024, accessed February 10, 2025, https://wccftech.com/tsmc-3nm-witness-massive-adoption-ai-tech-giants-Nvidia-rubin-amd-mi355x-intel-falcon-shores/.

[141] The Chosun Daily, "Samsung and TSMC Locked in Intense 2nm Chip Competition," December 16, 2024, accessed February 10, 2025, https://www.chosun.com/english/industry-en/2024/12/16/SMAOH7NISJAZXBL7LFX3BX2YVU/.

Keeping constant the electricity mix and carbon intensity of logic and dynamic random-access memory (DRAM) manufacturing as per the estimated 2024 values, we find the expected carbon emissions in metric megatonnes of CO_2 ($MMtCO_2$), related to the expected electricity consumption, as shown in Appendix Table 15:

Appendix Table 15 **Estimated CO_2 emissions for AI-related chip manufacturing in 2030 per scenario**

2030 $MtCO_2$	Logic	DRAM	NAND	Total
Conservative	1.20	3.24	0.78	5.22
Base	2.41	6.21	1.83	10.44
Ambitious	3.61	9.58	3.61	16.80

Appendix Table 15 shows the carbon intensity of not-AND (NAND, i.e., flash memory) manufacturing is set to the same value as DRAM manufacturing, considering that the dominant companies involved in HBM manufacturing (SK hynix, Micron, and Samsung) also control the lion's share of the NAND market.[142] The accuracy of this forecast may be improved by examining the expected power grid carbon intensity forecasts for the year 2030 as each of the regions involved has its own climate pledges for the coming years and may succeed in decarbonizing the respective power grids to a certain extent. Official sources, however, provide targets for different years while also using different reference periods, meaning that the data necessary to make this assessment are not available and/or are too inconsistent to be usable at this time. At the same time, the scenario analysis shows that the potential development in power demand as a result of AI-related semiconductor manufacturing may be significant enough to negate some of the decarbonizing efforts of certain regions.

[142] TrendForce, "NAND Flash Market Landscape to Change?" March 14, 2024, accessed February 10, 2025, https://www.trendforce.com/news/2024/03/14/news-nand-flash-market-landscape-to-change/.

Published April 2025 by Greenpeace East Asia