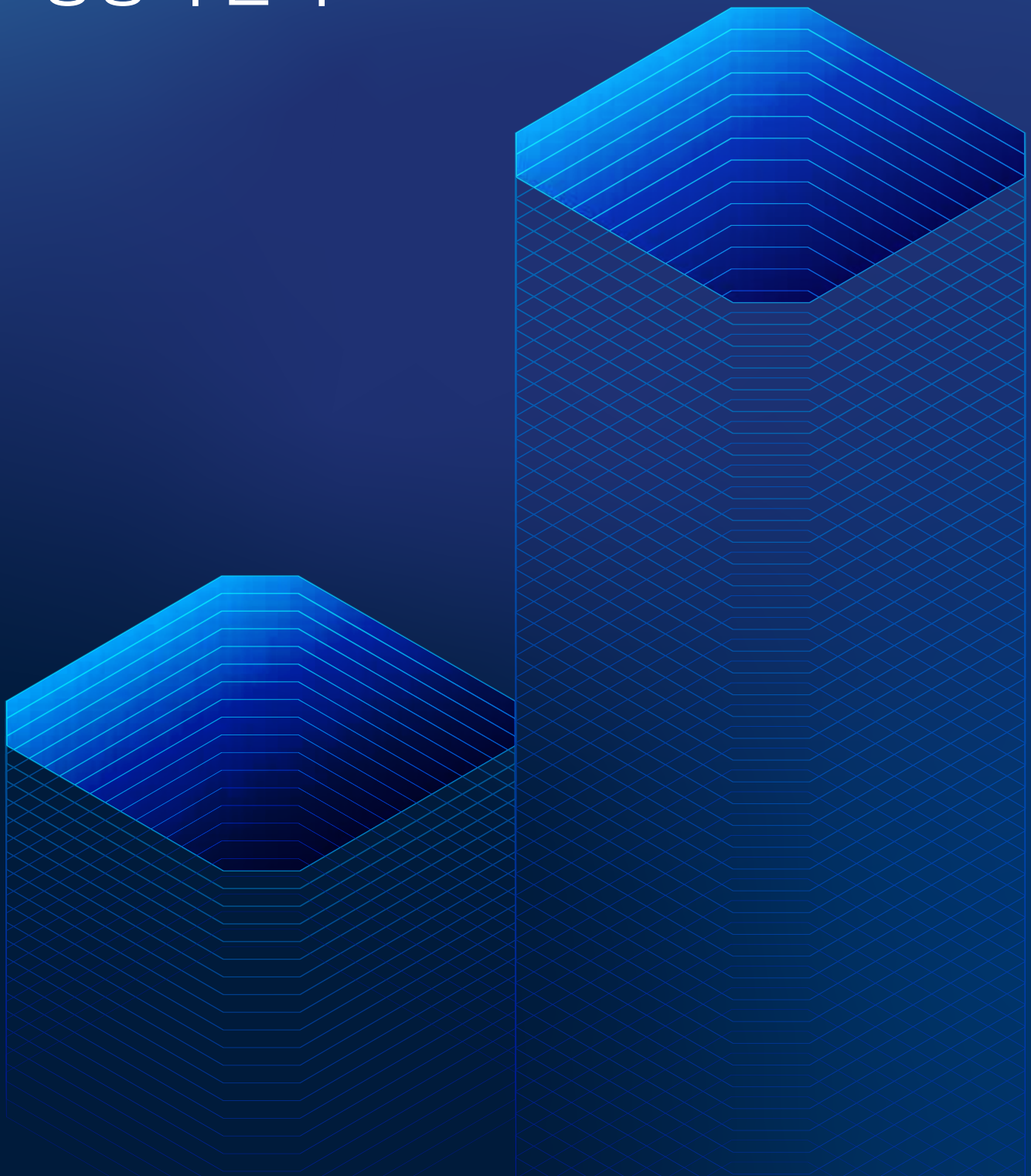
부록 A.

방법론 : AI 칩 생산의 에너지 소비와 탄소 발자국에 대한 상향식 분석



부록 A.

방법론 :
AI 칩 생산의 에너지 소비와 탄소 발자국에 대한 상향식 분석

저자 : 알렉스 드 브리스 Alex de Vries

인공지능(AI)이 반도체 제조 산업의 전력 수요에 미치는 영향을 평가하기 위해서, 우리는 먼저 AI 하드웨어 수요가 로직 칩, 메모리 칩, 데이터 저장 칩과 같은 핵심 구성요소의 생산량에 어떻게 영향을 미치고 있는지를 파악하는 상향식 접근법을 채택했다. 예상되는 생산량 변화와 각 생산 공정에 소요되는 전력 요구량을 결합하면, 반도체 제조 산업에서 AI와 관련한 전체적인 전력 소비량 추정치를 얻을 수 있다. 또 AI와 관련 하드웨어를 누가, 어디에서 생산하는지를 평가함으로써 생산 공정에 공급되는 전력 공급원 구성을 평가할 수 있다. 이 장에서는 이러한 접근법을 상세히 설명하고, 다양한 추정치를 얻는데 활용한 데이터의 출처를 밝힌다.

1. 상향식 접근법

상향식 접근법의 첫 번째 단계는 AI 하드웨어 수요가 반도체 제조 산업의 생산량에 미친 영향을 평가하는 것이다. 그런데 2023년과 2024년은 평가의 초점을 AI 칩 제조 과정의 주요 병목 현상에 맞춰야 한다. 이러한 한계가 AI 하드웨어 수요가 반도체 산업의 생산량에 미치는 전반적인 영향을 나타낼 수 있기 때문이다. 두 해 동안, 점점 더 많은 기업이 생성형 AI 모델을 훈련하고 배포하면서 AI 칩에 대한 수요가 급증했다. 따라서 이들 AI 모델의 연산 요구사항을 충족하는 AI 칩에 대한 수요도 크게 늘었다. 엔비디아는 2023년 기준 데이터센터 그래픽 처리장치(GPU) 출하량에서 시장의 98%를 점유한 것으로 알려졌다.[1] 하지만 같은 해 8월, AI 작업에 사용되는 고사양 GPU인 H100를 더 이상 판매할 수 없는 용량 한계에 이르렀다.[2] 시장 분석가들은 그 원인을 GPU와 같은 프로세서를 고대역폭 메모리(HBM)와 함께 패키징해 엔비디아에 납품하는 TSMC의 과부하에서 찾는다.[3] TSMC의 첨단 패키징 기술인 칩-온-웨어퍼-온-서브스트레이트(CoWoS)의 생산 능력이 제한적이기 때문이다. CoWoS 패키징 기술은 그 동안 거의 모든 HBM 기반 장치에 사용됐으며, AI 가속기도 모두 HBM을 사용했다(부록 박스1. 참조). 2023년 9월, TSMC의 마크 리우 회장은 CoWoS 패키징이 가장 제한적인 부분이라고 인정하며 이렇게 말했다. “AI 칩이 부족한 것이 아니라 CoWoS 생산 능력이 부족한 것이 문제다.” 그는 이러한 부족 현상이 향후 1년 반 가량 지속될 것이라고 전망했다.[4]

1. HPC와이어, “연구에 따르면, 엔비디아는 2023년 376만 개의 데이터센터 GPU를 출하할 것이다”(2024년 6월 10일), 2025년 2월 10일 참조. https://www.hpcwire.com/2024/06/10/nvidia-shipped-3-76-million-data-center-gpus-in-2023-according-to-study/.
2. 배런스, “선도적 클라우드 GPU 업체, ‘엔비디아 최고 AI 칩 2024년 매진’”(2023년 8월 9일), 2025년 2월 10일 참조. https://www.barrons.com/articles/nvidia-ai-chips-coreweave-cloud-6db44825.
3. 세미어낼리시스, “엔비디아, 브로드컴, 구글, AMD, AMD 인베디드(자일링스), 아마존, 마벨, 마이크로소프트, 알칩, 알리바바 T헤드, ZTE 사네칩스, 삼성전자, 마이크론, SK하이닉스의 분기별 램프”(2023년 7월 5일), 2025년 2월 10일 참조. https://semianalysis.com/2023/07/05/ai-capacity-constraints-cowos-and/.
4. 톰스 하드웨어, “TSMC: 1년 반 동안 지속될 엔비디아의 AI GPU 부족”(2023년 8월 8일), 2025년 2월 10일 참조. https://www.tomshardware.com/news/tsmc-shortage-of-nvidias-ai-gpus-to-persist-for-15-years.

부록 박스 1
AI 메모리 문제와 HBM / CoWoS의 필요성

생성형 AI 모델의 크기와 성능은 서로 연결돼 있는데 모델은 시간이 갈수록 커지고 복잡해지면서[5] 대량의 동적 랜덤 액세스 메모리(DRAM)에 대한 수요도 대폭 늘어났다. AI 모델의 학습 및 추론 단계 모두 메모리가 필요하지만, 일반적으로 학습 단계에서 메모리에 대한 수요가 가장 많다. 메모리는 주로 모델 가중치(모델 크기에 비례)와 키-값 캐싱(이전 단계의 키 및 값 행렬을 저장하고, 이를 후속 토큰을 생성하는 데 재사용)에 사용되기 때문이다.[6] 특히, 프로세서가 메모리에 정보를 읽고 저장하는 속도(메모리 대역폭)가 AI 모델의 크기에 병목 현상을 일으킨다. 프로세서가 전달할 수 있는 초당 부동 소수점 연산의 수는 지난 8년 동안 450배 이상 급격히 증가했는데, 메모리 대역폭의 개선은 그 속도를 따라잡지 못했다. 같은 기간 여러 세대의 AI 하드웨어에서 메모리 속도는 약 10배 빨라지는 데 그쳤다.[7] 이처럼 커져가는 불균형이 메모리 벽으로 알려져 있다. HBM은 메모리 대역폭의 한계를 극복하기 위해 설계됐으며, 기존의 DRAM과 달리 동일 패키지 내에 프로세서와 함께 배치될 수 있도록 만들어진다.[8] 이를 통해 더 넓은 메모리 버스(대역폭)가 가능하고, 결과적으로 기존 DRAM보다 더 높은 대역폭을 제공할 수 있다. 따라서 HBM, 그리고 HBM을 프로세서와 동일한 패키지 내에 통합하는 기술(칩-온-웨이퍼-온-서브스트레이트, CoWoS)이 대규모 AI 모델을 만들고 운용하는 데 필수적인 요소가 됐다.

여러 분석가들이 TSMC의 정확한 CoWoS 생산 능력에 관한 세부 내용을 보고했다. 2023년 7월, TSMC의 CoWoS 생산 능력은 한 달에 300mm 웨이퍼 8,000장으로 제한돼 있었다. TSMC는 해당 용량을 같은 해 말까지 월 1만1,000장, 2024년 말까지 최대 1만 6,600장으로 증대하는 계획을 갖고 있었다.[9] 2024년 4월에 보고된 내용에 따르면, TSMC의 CoWoS는 2023년 말 월 1만 3,000장에 도달했다. 또 2024년 말이면 CoWoS가 월 3만 5,000장에 이를 것으로 보고됐다.[10] TSMC가 2024년 10월까지 월 3만 5,000~4만 장의 웨이퍼를 생산하고, 2025년 말까지 이 용량이 다시 두 배로 늘 것이라는 전망도 있다.[11] 이번 연구에서는 AI 칩 제조가 환경에 미치는 영향의 한계를 보다 잘 파악하기 위해, 앞서 언급한 각 연도에서 가장 높은 용량을 기준으로 연간 CoWoS 생산 능력을 계산했다. 즉, 2023년의 CoWoS 총 생산 능력은 웨이퍼 15만 6,000장(=1만3,000*12)이라고 가정했다. 2024년에는 이 생산 능력이 48만 장(4만*12)으로 증가한다고 가정했다. CoWoS 생산 능력이 제조 과정의 주요 병목 현상을 일으키는 원인을 설명하기 위해서는, 이러한 웨이퍼 한 장으로 엔비디아의 패키지된 H100 칩을 9개밖에 얻을 수 없다는 점을 고려해야 한다.[12] 패지징의 수율을 99%라고 가정하면,[13] 2023년 연간 CoWoS 용량(웨이퍼 15만 6,000장)으로 최대 1,389,960개의 H100 패키지 칩 제조가 가능하다고 계산할 수 있다.

공식 1

$$PU = W * DPW * PY$$

기호 :

$$PU = \text{패키지 유닛}$$
$$W = 300\text{mm 웨이퍼}$$
$$PY = \text{패키징 수율}$$
$$DPW = \text{웨이퍼당 다이스}$$

5. 아난타스와미 A., “AI에서 더 큰 것이 항상 더 나은가?”<네이처 615> 202~205쪽(2023년 3월 8일) https://doi.org/10.1038/d41586-023-00641-w.
6. 언폴드AI, “대규모 언어 모델의 GPU 메모리 요구”(2024), 2025년 2월 10일 참조. https://unfoldai.com/gpu-memory-requirements-for-llms/.
7. 세미어낼리시스, “메모리 벽: DRAM의 과거, 현재, 미래”(2024년 9월 3일), 2025년 2월 10일 참조. https://semianalysis.com/2024/09/03/the-memory-wall/.
8. 첸 V. 등, “CoWoS를 통한 고대역폭 메모리 인터페이스의 설계 어려움 극복”(2022 IEEE 국제 전자기 호환성 및 신호/전력 무결성 심포지엄(EMCSI), 미국 워싱턴주 스포캔, 2022년 8월 1~5일). https://doi.org/10.1109/EMCSI39492.2022.10050234.
9. 톰 하드웨어, “보고서: TSMC, 첨단 패키징 시설 확장에 박차를 가하다”(2023년 7월 15일), 2025년 2월 10일 참조. https://www.tomshardware.com/news/tsmc-accelerates-expansion-of-advanced-packaging-facilities-report.
10. 트윅타운, “2023년 TSMC의 3대 고객: 애플(25%), 엔비디아(11%), AMD(7%)”(2024년 4월 18일), 2025년 2월 10일 참조. https://www.tweaktown.com/news/97457/tsmcs-top-3-customers-in-2023-apple-with-25-nvidia-11-and-amd-7/index.html.
11. 트렌드포스, “TSMC, CoWoS 용량이 2년 간 두 배 증가-여전히 불충분하지만 공급 업체에 긍정적 전망”(2024년 10월 21일), 2025년 2월 10일 참조. https://www.trendforce.com/news/2024/10/21/news-cowos-capacity-doubles-for-two-years-still-insufficient-positive-outlook-for-suppliers/.
12. 더일렉, “삼성, 엔비디아와 HBM 계약 여전히 안갯속”(2024년 6월 20일), 2025년 2월 10일 참조. https://www.thelec.net/news/articleView.html?idxno=4882.
13. 무무테크놀로지, “엔비디아의 신제품 지연? 영향은 크지 않을 듯”(2024), 2025년 2월 10일 참조. https://www.moomoo.com/news/post/41772359/nvidia-s-new-chip-delayed-don-t-panic-the-impact.

섹션 1.4에서 명시된 모든 관련 생산 시설이 300mm(12인치)의 실리콘 웨이퍼를 사용하고[14,15,16,17], 또 TSMC의 CoWoS 패키징 공정을 이용하므로[18], 우리는 모든 계산에서 이 크기의 웨이퍼를 기본 웨이퍼로 가정했다. 2024년 한 해 동안 총 48만 개의 웨이퍼로 최대 427만 6,800개의 H100 패키지 칩이 생산될 것이다. 엔비디아는 2023년 55만개 유닛의 H100을 출하할 계획이며, 2024년에는 그 숫자를 150만 내지 200만으로 확대할 계획이라고 보고된 바도 있다[19]. 옴디아는 엔비디아가 2023년 65만 개의 H100 주문을 소화할 것으로 예상했다[20]. 따라서, 엔비디아의 H100 칩 한 종이 2023년과 2024년 TSMC의 CoWoS 생산량에서 상당히 큰 부분을 차지했다고 가정할 수 있다.

그러나, 엔비디아의 H100이 AI 칩 시장 전체를 반영하지는 않는다. 따라서 대체 AI 칩의 물량 및 이들이 관련 공급망에 미치는 영향에 대한 검토가 필요하다. 이번 분석 작업의 출발점은 <부록의 표 1>에 나와 있는 엔비디아와 AMD가 제시한 주력 AI 칩 로드맵이다.[21,22]

부록 표 1 **2023과 2024년, 엔비디아와 AMD의 주력 AI칩 로드맵**

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								

<부록 표1>에 따르면, 엔비디아는 2024년에도 데이터센터 GPU 시장을 계속 지배한 것으로 추정된다. AMD의 시장 점유율은 5~7%에 그쳤고,[23] 인텔 또한 1%에 불과한 점유율을 보였다.[24] 기업들이 2023과 2024년 다양한 AI 칩에 CoWoS의 생산 능력을 할당한 것을 두고 몇 가지 가정을 해 볼 수 있다. 2023년 TSMC의 최대 CoWoS 생산 능력(15만 6,000 웨이퍼) 가운데 43%가 조금 넘는 양이 엔비디아의 H100에 사용됐을 수 있는데, 60만개(2023년 H100 출하량 추정치 55만~65만의 평균)을 출하하려면 패키징 수율 99%를 적용할 때 총 6만 7,340장의 패키징 웨이퍼가 필요할 것이기 때문이다. 이 값을 얻기 위해 우리는 W의 공식2를 이용했다.

공식 2

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

14. TSMC, "TSMC 팹", 2025년 2월 10일 참조. https://www.tsmc.com/english/aboutTSMC/TSMC_Fabs.
15. 이이타임스, "마이크론, 엘피다 인수 완료"(2013년 7월 31일), 2025년 2월 10일 참조. https://www.eetimes.com/micron-closes-elpida-acquisition/.
16. T더일렉, "SK하이닉스, 브로드컴으로부터 대규모 HBM 수주"(2024년 12월 20일), 2025년 2월 10일 참조. https://www.thelec.net/news/articleView.html?idxno=5084.
17. 삼성, "당신의 혁신을 앞당기는 제조 공정", 2025년 2월 10일 참조. https://semiconductor.samsung.com/foundry/manufacturing/manufacturing-sites/.
18. 톰 하드웨어, "TSMC, 510X515mm 직사각형 실리콘 웨이퍼 사용으로 기존 300mm 지름의 실리콘 기술 대비 사용 가능 면적 3배 증대"(2024년 6월 20일), 2925년 2월 10일 참조. https://www.tomshardware.com/tech-industry/tsmc-explores-using-510×515-mm-rectangular-silicon-wafers-tripling-the-usable-area-of-current-300mm-diameter-tech.
19. WCCFTech, "엔비디아, 2024년까지 AI GPU 수백만 개 출하 목표, 공급망 다각화 노력"(2023년 8월 24일), 2025년 2월 10일 참조. https://wccftech.com/nvidia-aims-at-shipping-millions-of-ai-gpus-by-2024-working-to-diversify-supply-chain/.
20. 톰 하드웨어, "보고서: 엔비디아, 메타와 페이스북 덕분에 3분기 H100 AI GPU 50만대 판매. 페이스북-리드 타임 최대 25주 연장"(2023년 11월 28일), 2025년 2월 10일 참조. https://www.tomshardware.com/tech-industry/nvidia-ai-and-hpc-gpu-sales-reportedly-approached-half-a-million-units-in-q3-thanks-to-meta-facebook.
21. 엔비디아, "투자자 프리젠테이션"(2023), 2025년 2월 10일 참조. https://s201.q4cdn.com/141608511/files/doc_presentations/2023/Oct/01/ndr_presentation_oct_2023_final.pdf.
22. WCCFTech, "AMD 인스팅크트 AI 가속기 라인업, 4분기에 MI325X 업그레이드, 2025년 3nm 'CDNA4', 2026년 MI400 'CDNA 넥스트' 출시 예정"(2024년 6월 2일), 2025년 2월 10일 참조. https://wccftech.com/amd-instinct-ai-accelerator-lineup-mi325x-refresh-q4-3nm-mi350-cdna-4-2025-cdna-mi400-cdna-next-2026/.
23. 인베스토피디아, "AMD의 진보된 AI 이벤트에 앞서 알아야 할 것"(2024년 9월 9일), 2025년 2월 10일 참조. https://www.investopedia.com/what-you-need-to-know-ahead-of-amd-advancing-ai-event-8723294.
24. CNBC, "AI 칩 시장 장악한 엔비디아, 그러나 그 어느 때보다 격해지는 경쟁"(2024년 6월 2일), 2025년 2월 10일 참조. https://www.cnbc.com/2024/06/02/nvidia-dominates-the-ai-chip-market-but-theres-rising-competition-.html.

구글의 텐서 프로세싱 유닛(TPU) 생산 파트너인 브로드컴은 TSMC CoWoS 생산 능력의 10%를 차지한 것으로 알려졌다.[25] 따라서 최대 1만 5,600개의 패키징 웨이퍼가 사용됐을 것이다. 엔비디아의 시장 점유율 98%를 감안하면, 2023년 TSMC의 나머지 CoWoS의 생산 능력(최대 156,000 – 15,600 – 67,340 = 73,060 웨이퍼)이 엔비디아의 A100에 할당됐다고 가정할 수 있다. A100 패키지의 크기가 H100의 크기와 거의 동일하기 때문에, 공식1을 이용하면, 2023년 65만 964개(반올림)의 A100이 생산된 것으로 추정된다.

2024년에는 상황이 조금 복잡해졌다. 경쟁이 치열해지고 엔비디아는 H200 및 차세대 블랙웰 GPU(예: B100, B200) 같은 최신 디바이스를 출시하면서 A100을 단종했기 때문이다. 또한 차세대 모델들은 TSMC의 CoWoS-L(로컬 실리콘 인터커넥트를 이용한 CoWoS)의 생산 능력을 활용한 첫 번째 제품들이어서[26], 생산 능력을 분석하는 작업을 한층 복잡하게 만들었다. 엔비디아의 기존 제품들은 CoWoS-S(실리콘 인터포저를 이용한 CoWoS)에 의존하기 때문이다. CoWoS-S에서는 대형 단일 실리콘 인터포저가 사용돼 분석이 간소하다. CoWoS-L에서는 이러한 단일 인터포저가 로컬 실리콘 인터커넥트 칩렛과 글로벌 재분배 레이어로 재구성된 인터포저로 대체된다. 가까운 미래에 CoWoS-L이 주요 CoWoS 공정으로 자리잡을 가능성이 크지만,[27] 2024년에 관한 분석에서 CoWoS-S의 CoWoS-L로의 전환은 유의미하지 않다. 이는 전문가들이 TSMC의 CoWoS-L 생산 능력이 제한적이기 때문에, 2024년 해당 공정을 통해 생산한 블랙웰 GPU가 많지 않을 것으로 예상했기 때문이다. 2024년 9월 모건스탠리의 분석가들은 엔비디아가 2024년 마지막 분기에 약 45만 개의 블랙웰 디바이스를 생산할 것으로 예상했으며,[28] 여기에 약 3만 2,000개의 CoWoS-L 웨이퍼가 사용될 것으로 봤다.[29] TSMC의 총 CoWoS 추정 생산 능력인 48만 웨이퍼에서 이 물량을 제외하면, 2024년 TSMC의 최대 CoWoS 용량은 44만 8,000웨이퍼로 줄어든다. 유통의 측면에서는, 2023년 브로드컴의 예상 점유율을 기준으로 볼 때, 브로드컴이 TSMC의 총 CoWoS 생산 능력의 10%(즉, 최대 4만 8,000 웨이퍼)를 점유한다. 나머지는 엔비디아와 AMD가 차지한다.

업계에서는 AMD가 2024년 약 50만 대의 MI300X를 출하할 것으로 예상했다.[30] 이 장치는 시스템 온 인터그레이티드 칩(SoIC) 패키징에 의존하기 때문에 생산 과정에서 추가적인 단계가 필요하다. TSMC는 해당 공정에 월 5,000~6,000개의 웨이퍼를 할당한다.[31] SoIC는 가속기 복합 다이스(XCD)와 입출력 다이스(IOD)를 수직으로 통합해 이를 CoWoS에서 HBM과 패키징하는 데 사용된다. 엔비디아와 AMD는 칩 설계를 최대 크기로 밀어붙이는 공통점이 있기 때문에, 단일 CoWoS 웨이퍼의 수율도 비슷하다고 가정할 수 있다. 따라서 공식 2를 활용하면, 50만 개의 MI300X를 생산하기 위해서는 5만 6,117개의 패키징 웨이퍼가 필요하다. 2024년 예상 출하량이 175만 개(2024년 H100 예상 출하량 150만~200만 개의 평균)인 엔비디아의 H100의 경우, 19만 6,408개(반내림)의 패키징 웨이퍼를 필요로 한다. 다시 말하지만, 우리는 TSMC의 최대 CoWoS-S 생산 능력의 잔여분(즉, 월 14만 7,475개: 480,000-32,000-48,000-56,117-196,408=147,475)이 엔비디아의 H200 생산에 할당됐다고 가정한다. 아래의 <부록 표2>에는 2023과 2024년 생산된 모델별 AI 칩 생산량에 대한 최선의 추정치가 요약돼 있다. 이들 수치가 상향식 접근법의 기반이 된다.

---

25. 톰스 하드웨어, “AMD와 엔비디아 GPU가 TSMC CoWoS의 생산 능력에서 가장 큰 비중을 차지하는 이유”(2023년 6월 13일), 2025년 2월 10일 참조. https://www.tomshardware.com/news/amd-and-nvidia-gpus-consume-lions-share-of-tsmc-cowos-capacity.
26. 3DinCites, “IFTLE 607: 엔비디아 블랙웰이 TSMC CoWoS-L 기술과 관련한 문제”(2024년 10월 2일), 2025년 2월 10일 참조. https://www.3dincites.com/2024/10/iftle-607-why-nvidias-blackwell-is-having-issues-with-tsmc-cowos-l-technology/
27. 디지타임스, “CoWoS 패키징의 글로벌 생산 능력”(2024년 8월 29일), 2025년 2월 10일 참조. https://www.digitimes.com/reports/item.asp?id=20240829RS400.
28. 톰스 하드웨어, “엔비디아, 4분기 블랙웰 AI GPU 45만개 생산 예상-칩 제조업체 잠재 매출 100억 달러 달성”(2024년 9월 25일), 2025년 2월 10일 참조 https://www.tomshardware.com/pc-components/gpus/nvidia-expected-to-produce-450000-blackwell-ai-gpus-in-q4-potential-dollar10b-in-revenue-for-the-chipmaker.
29. 디지타임스, “디지타임스 리서치, ‘2025년 전 세계 CoWos 및 유사 패키징 용량 수요 전년 대비 113% 급증 전망”(2024년 9월 25일), 2025년 2월 10일 참조. https://www.digitimes.com/news/a20241025VL210/cowos-demand-packaging-2025-capacity.html.
30. WCCFTech, “AMD, 2024년 인스팅크트 MI300X AI 가속기 ‘50만 개’ 출하 예정”(2024년 10월 10일), 2025년 2월 10일 참조. https://wccftech.com/amd-ship-half-a-million-instinct-mi300x-ai-accelerators-2024/.
31. 트렌드포스, “TSMC의 SoIC 수요 증가, 상당한 용량 확대 시사”(2024년 1월 18일), 2025년 2월 10일 참조. https://www.trendforce.com/news/2024/01/18/news-tsmcs-soic-demand-heats-up-reports-suggest-significant-capacity-expansion/.

부록 표 2 2023과 2024년의 모델별 생산량 추정치

GPU	연도	예상 생산량
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 사내 솔루션 및 기타 구성 요소

섹션1에서는 브로드컴이 엔비디아, AMD와 더불어 TSMC의 CoWoS 공급 대상 중 하나라는 사실을 다뤘다. 이러한 내용은 엔비디아와 AMD의 단위 생산량 추정치를 수정하는 데 이용됐고, 브로드컴이 구글에 제공하는 TPU이 공급망에 끼치는 영향에 관한 추가적인 분석은 이뤄지지 않았다. 그 이유는 해당 장치를 분석하는 것이 본질적으로 더 어렵기 때문이다. 구글과 같은 대기업들은 엔비디아에 대한 의존도를 줄이기 위해 자체 AI 가속기를 개발하고 있지만,[32] 일반적으로 이러한 가속기는 판매용이 아니다. 예컨대 구글의 TPU는 구글의 클라우드 서비스에만 사용된다. 따라서 관련성이 있고, 완전하며, 정확한 제품 사양을 얻기가 상대적으로 어렵다. 게다가 이러한 사양은 공개된다고 하더라도, 일반 대중에 판매되는 기기 사양에 비해 상대적으로 늦게 공개될 수 있다. 따라서 사내 솔루션은 이번 보고서의 분석 대상에서 제외했다. 같은 개념이 전체 서버 시스템에서 AI 가속기와 함께 사용되는 중앙 처리 장치(CPU), 이중 데이터 전송률 랜덤 액세스 메모리(DDR DRAM), 낸드(NAND) 스토리지와 같은 다른 구성 요소에도 적용된다. 이런 요소들은 앞서 다룬 AI 칩이나 HBM보다 일반적인 것일 수 있으므로, 이들 장치를 제조하는 업체에 대한 AI 수요의 영향을 분리해 내 분석하기가 더 어렵다.

32. 베르딕트, "설명: 아마존, AI 칩 공급 압박에 대한 헤지 수단으로서 DIY AI 칩 공개"(2024년 11월 27일), 2025년 2월 10일 참조. https://www.verdict.co.uk/explainer-amazon-to-unveil-diy-ai-chips-as-big-tech-hedges-against-ai-chip-supply-squeeze/.

1.2 생산 공정

AI 칩의 예상 판매량을 해당 제품의 사양과 결합해 CoWoS와 함께 패키징되는 다양한 구성 요소가 공급망에 미치는 영향을 파악할 수 있다. 제품 사양은 부록의 표3에 나와 있다.(특정 칩에는 여러 가지 변형이 존재할 수 있다. 예컨대 엔비디아의 H800 및 H20은 H100과 유사하지만, 서로 비슷한 생산 요건을 가지고 있으므로 2023과 2024년의 계산에서 이러한 변형을 구분하지 않았다.)

부록 표 3 세미어낼리시스[33,34], WCCFTech[35], 톰스 하드웨어[36]를 통해 선별된 일부 데이터센터 GPU 사양

GPU	Nvidia A100	Nvidia H100	Nvidia H200	Nvidia B100/B200	AMD MI300X
GPU 다이 크기	826 mm²	814 mm²	814 mm²	~814 mm²*	115 mm²
GPU 다이스	1	1	1	2	8 (4 IOD)
제조 공정	7nm (TSMC)	4nm (TSMC)	4nm (TSMC)	4nm (TSMC)	5/6nm (TSMC)
메모리 인터페이스	HBM2	HBM3	HBM3e	HBM3e	HBM3
메모리 크기	80 GB	80 GB	141 GB	192 GB	192 GB
메모리 스택	5	5	6	8	8
스택당 메모리 다이스	8+1	8+1	8+1	8+1	8+1

*이들 모델의 정확한 크기는 출시 시점까지 공개되지 않았지만, 다이의 크기는 이전 값과 EUV 레티클 한계인 858mm2 사이일 것으로 추정된다.

특히 엔비디아와 AMD는 모두 GPU, 가속기 칩렛(XCD), IO 다이스(IOD)를 TSMC에 의존하고 있다. AI 가속기용 고사양 로직 반도체 시장에서 TSMC의 우월적 지위를 고려하면, TSMC의 상당한 참여는 예상할 수 있다.[37] SK하이닉스, 마이크론, 삼성전자 등 관련업체가 늘어나면서 HBM 제조의 지형은 다양해졌다. 엔비디아는 2024년 1분기까지 HBM을 SK하이닉스에 전적으로 의존했다. 하지만, 마이크론이 엔비디아의 H200에 HBM을 공급하고,[38] 같은 제품이 엔비디아의 B200에도 사용될 것이라고 발표했다. AMD는 삼성전자로부터 HBM을 공급받는 것으로 알려졌지만,[39] 기대에 못 미치는 수율 문제로 인해 삼성전자는 HBM 시장에서 뒤처져 있다.[40]

33. 세미어낼리시스, "AI 용량 제한-CoWoS 및 HBM 공급망"(2023년 7월 5일), 2025년 2월 10일. https://semianalysis.com/2023/07/05/ai-capacity-constraints-cowos-and/.
34. 세미어낼리시스, "AMD MI300-하이프 길들이기-AI 성능, 볼륨 램프, 고객, 비용, IO, 네트워킹, 소프트웨어"(2023년 6월 12일), 2025년 2월 10일 참조. https://semianalysis.com/2023/06/12/amd-mi300-taming-the-hype-ai-performance/.
35. WCCFTech, "엔비디아 H100 GPU와 AI 열풍, 현재 상황 요약"(2023), 2025년 2월 10일 참조. https://wccftech.com/nvidia-h100-gpus-a-rundown-of-current-situation/.
36. 톰스 하드웨어, "엔비디아의 차세대 AI GPU는 호퍼보다 4배 빠르다: 블랙웰 B200 GPU는 최대 20페타플롭의 컴퓨팅 및 기타 대규모 개선 사항을 제공한다"(2024), 2025년 2월 10일 참조. 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.
37. 세미위키, "TSMC가 고급 실리콘의 90% 생산하는 것은 아냐"(2024), 2025년 2월 10일 참조. https://semiwiki.com/semiconductor-manufacturers/tsmc/342934-no-tsmc-does-not-make-90-of-advanced-silicon/.
38. 트렌드포스, "마이크론, 엔비디아 H200용 HBM3e 대량 생산 시작"(2024년 2월 27일), 2025년 2월 10일 참조. https://www.trendforce.com/news/2024/02/27/news-micron-begins-mass-production-of-hbm3e-for-nvidias-h200/.
39. WCCFTech, "삼성, AMD MI300X GPU에 공급할 대규모 HBM3 메모리 수주"(2024년 8월 23일), 2025년 2월 10일 참조. https://wccftech.com/samsung-receives-huge-order-of-hbm3-memory-to-power-amd-mi300x-gpus/.
40. WCCFTech, "삼성, 경영 부진과 시장 기회 포착 실패로 HBM 생산 능력 하향 조정"(2024년 10일 15일), 2025년 2월 10일 참조. https://wccftech.com/samsung-lowers-hbm-production-capacity/.

위의 사양을 이용해 장치 생산과 관련한 웨이퍼 수요를 추정할 수 있다. 814mm² 크기의 단일 GPU 다이를 둘러싼 다섯 개의 HBM3 스택(1개의 더미 다이 포함)이 포함된 엔비디아의 H100을 자세히 살펴보면 이를 설명할 수 있다. 300mm 웨이퍼 하나로 57개의 다이를 완벽히 생산한다고 가정하면(부록 표4 참조), GPU 패키지 60만 개에는 1만633장의 로직 웨이퍼가 소요된다. 공식 3은 이 값을 도출하는 식이다.

공식 3

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

기호 :

$$DY = 다이\ 산출$$

$$DPU = 패키지\ 유닛당\ 다이스$$

H100 다이가 814mm² 로 상당히 크기 때문에 웨이퍼당 얻을 수 있는 양호한 다이스는 57개에 훨씬 못 미칠 것으로 예상된다. 엔비디아가 H100 생산에 cm2당 0.1의 결함 밀도를 보이는 TSMC의 5nm 맞춤형 버전을 이용한다는 사실을 고려하면,[41] 다이 수율은 46.4%, PY는 99%로 가정할 수 있다(부록 표4 참조)[42]. 단일 패키지에는 하나의 GPU 다이만 필요하므로, DPU는 1개로 설정된다. 로직 웨이퍼에 대한 수요는 다음과 같은 식에 따라 증가한다.

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

메모리의 요구사항이 훨씬 높은 이유는 엔비디아 H100에 탑재된 HBM3 스택이 8개의 다이 스택이 쌓인 것이기 때문이다. 하나의 300mm 웨이퍼에는 최대 476개의 다이스가 필요한 반면(부록 표4 참조), H100 한 개에는 40개의 다이스가 필요하다. 이러한 이유로, 60만 개의 H100 유닛 패키지는 앞의 공식에 따라 DPU가 40개, DWP가 476, DY가 100%일 경우 최소 5만929개 DRAM 웨이퍼를 필요로 한다.

그러나, HBM의 수율은 다이의 수율뿐 아니라 적층된 DRAM을 CoWoS로 패키징하기 전 DRAM 다이를 적층하는 과정의 패키징 수율로부터도 영향을 받는다. 따라서 다음과 같이 공식4를 수정해야 한다.

공식 4

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

기호 :

$$MPY = 모리\ 패키징\ 수율$$

41. 아난드테크, "7나노보다 5나노에서 더 나은 수율: N5의 결함율에 대한 TSMC 업데이트"(2020년 8월 25일), 2025년 2월 10일 참조. https://www.anandtech.com/show/16028/better-yield-on-5nm-than-7nm-tsmc-update-on-defect-rates-for-n5.
42. 무무테크놀로지, "엔비디아의 신형 칩 지연? 영향 크지 않을 듯"(2024년 8월 5일), 2025년 2월 17일 참조. https://www.moomoo.com/news/post/41772359/nvidia-s-new-chip-delayed-don-t-panic-the-impact.

다른 입력 변수(PU, DPU, PY, DPW, DY)는 동일하게 유지된다. 2024년 1분기까지 엔비디아에 HBM을 단독으로 공급하는 SK하이닉스가 HBM3e 칩의 목표 수율인 80%에 근접했다고 밝혔으므로,[43] 이제부터 DY에 MPY를 곱한 값이 80%라고 가정한다.[44] 그러면 다음과 같은 결과가 나온다.

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

HBM의 각 스택은 기본 로직 다이 위에 놓이므로, 생산 공정에서 7,576개의 로직 웨이퍼가 추가로 필요하다(웨이퍼당 크기와 수율이 비슷한 것으로 가정). 이전 공식에서 변경되는 유일한 값은 HBM 스택 하나에 한 개씩 포함된 DPU뿐이다. 따라서 H100 패키지마다 5개로 산출되며, 그 식은 다음과 같다.

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

최종적으로 60만 개의 엔비디아 H100 패키지를 생산하기 위해서는 GPU용 웨이퍼 2만 2,915장과 7만 1,620장의 메모리용 웨이퍼가 필요한 것으로 추정된다. AI칩 시장에서 엔비디아 H100의 지배적 위치를 고려할 때, 이 같은 숫자는 2023년 반도체 제조 산업의 총 생산량에 미친 대부분의 AI 수요 영향을 반영한다고 할 수 있다. 또 이들 숫자에 2.92를 곱하면 2024년 H100의 예상 출하량을 얻을 수 있는데, 이 수치가 올해 반도체 제조 산업의 총 생산량에 미친 AI 수요 영향의 상당 부분을 의미한다. 다른 장치들이 웨이퍼 수요 영향을 파악하기 위해, 다음 가정 하에 공식4를 적용하겠다.

부록 표 4 **엔비디아 GPU, AMD XCD, IOD 및 여러 유형의 HBM에 사용되는 300mm 웨이퍼 하나의 수율 추정치**

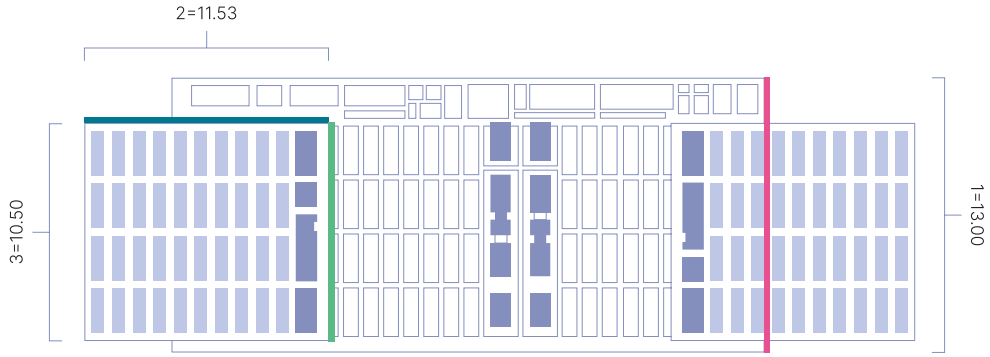
웨이퍼 용도	다이 규격	DPW	DY	웨이퍼당 양품 수
Nvidia GPUs	25.59×32.16mm	57	46.4%	26.4
AMD XCD	10.5×11.5mm	477	88.7%	423.1
AMD IOD	13×29mm	140	69.4%	97.2
HBM3/3e	11×11mm	476	80.0%	380.8
HBM2	7.75×11.87mm	634	80.0%	507.2

43. 로이터, "엔비디아 공급업체 SK하이닉스 2025년 HBM 칩 거의 다 팔려"(2024년 3월 2일), 2025년 2월 10일 참조. https://www.reuters.com/technology/nvidia-supplier-sk-hynix-says-hbm-chips-almost-sold-out-2025-2024-05-02/.

44. 트렌드포스, "SK하이닉스, HBM3e 진전 발표 : 80% 가까운 수율 달성"(2024년 5월 24일), 2025년 2월 10일 참조. https://www.trendforce.com/news/2024/05/24/news-sk-hynix-revealed-progress-for-hbm3e-achieving-nearly-80-yield/.

<부록 표 4>는 다이의 규격, 웨이퍼의 직경 및 결함 밀도 주어졌을 때 DPW 및 DY 값을 결정하는 데 도움이 되는 몇 가지 계산 방식이 있음을 보여준다.[45] 이번 연구에서는 <부록 표 4>에 나열된 다이 규격, 직경 300mm 웨이퍼, cm2당 0.1의 결함 밀도[46], 스크라이브 및 엣지 손실에 대한 기본 가정 등을 바탕으로 iSine의 다이 수율 계산 방식을 사용했다. AMD는 MI300X IOD의 규격을 공개했으므로[47], 동일한 덱을 기준으로 XCD 치수를 10.5mm x 11.5mm로 추정할 수 있다(부록 그림1 참조).

부록 그림 1 **AMD MI300X의 XCD 예상 치수 (단위:mm)**



비슷한 방식으로 엔비디아 H100 GPU 다이의 크기를 25.59mm x 32.16mm로 추정해 볼 수 있다.[48] 이러한 규격은 리소그래피 장비가 출력할 수 있는 최대 크기(26mm x 33mm)에 가깝고,[49] 분석 대상인 엔비디아의 모든 GPU의 표면적은 레티클 한계에 가깝다(부록 표3 참조). 또한 우리는 엔비디아의 모든 GPU에 동일한 수율 추정치를 적용했다. HBM2와 HBM3e 다이의 크기는 JEDEC[50]와 마이크론[51]으로부터 얻었다. SK하이닉스가 HBM3e에서 80%의 수율을 달성했다고 발표했기 때문에,[52] 우리는 계산된 수율 대신 이 수치를 HBM 수율의 고정값으로 적용했다.

위에서 설명한 공식과 가정을 이용해, 우리는 각 모델과 관련한 웨이퍼 수요를 추정했다. 그 결과는 <부록 표5>에 나와 있는 것과 같다.

부록 표 5 **선택된 모델별 웨이퍼 수요량 예상치**

GPU	연도	로직 웨이퍼	메모리 웨이퍼	메모리 베이스
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

45. iSine, “다이 수율 계산 방식”, 2025년 2월 10일 참조. https://isine.com/resources/die-yield-calculator/.
46. 아난드테크, “7나노보다 5나노에서 더 나은 수율: N5의 결함율에 대한 TSMC 업데이트”(2020년 8월 25일), 2025년 2월 10일 참조. https://www.anandtech.com/show/16028/better-yield-on-5nm-than-7nm-tsmc-update-on-defect-rates-for-n5.
47. AMD, “AMD 인스팅크트 MI300X 가속기: 패키징 및 아키텍처 공동 최적화”(2024), 2025년 2월 10일 참조. https://vlsi24.mapyourshow.com/mys_shared/vlsi24/handouts/JFS2-4_Smith.pdf.
48. 핫하드웨어, “80GB HBM3 탑재ehls 엔비디아 H100 호퍼 GPU, 사진으로 본 괴물 같은 성능”(2022년 4월 6일), 2025년 2월 10일 참조. https://hothardware.com/news/h100-hopper-gpu-gets-pictured.
49. ISDI, “CMOS 스티칭 기술: 도전 과제와 솔루션”(2024년 10월 30일), 2025년 2월 10일 참조. https://www.isdicmos.com/news-events/2024/10/30/the-art-of-cmos-stitching-challenges-and-solutions.
50. 아난드테크, “JEDEC, 삼성이 칩 대량 생산을 시작함에 따라 HBM2 사양 발표”(2016년 1월 20일), 2025년 2월 10일 참조. https://www.anandtech.com/show/9969/jedec-publishes-hbm2-specification.
51. 마이크론, “AI 혁신을 위해 설계한 메모리 소개”(2023), 2025년 2월 10일 참조. https://www.micron.com/content/dam/micron/global/public/documents/products/product-flyer/hbm3e-product-brief.pdf.
52. 트렌드포스, “SK하이닉스, ,HBM3e 진전 발표 : 80% 가까운 수율 달성”(2024년 4월 24일), 2025년 2월 10일 참조. https://www.trendforce.com/news/2024/05/24/news-sk-hynix-revealed-progress-for-hbm3e-achieving-nearly-80-yield/.

1.3 전력 수요

AI 수요가 반도체 제조 산업에 어떤 영향을 미치는지 알아봤으니, 다음은 확인된 변화와 관련한 전력 소비량을 추정해 볼 차례다. 가장 쉬운 방법은 총 웨이퍼 생산량의 추정되는 변화에 웨이퍼 생산 시 필요한 전력 소비량을 곱하는 것이다. 이 과정에서 로직, DRAM, NAND 웨이퍼 생산 공정 간의 차이를 고려한다. 로직 웨이퍼 생산의 경우, 로직 기술에 대한 전체 공정의 흐름(즉, 라인의 전체 앞 단(FEOL), 중간 단(MEOL), 마지막 단(BEOL) 공정)에 따른 제조용 총 전력 소비량(웨이퍼 1cm²생산당)을 평가한 바르돈 등의 연구(2022년)[53]를 참고했다(부록 표 6). 이러한 전력 소비량에는 제조 공정에 사용되는 장비의 전력 소비량뿐 아니라, 이러한 공정을 지원하는 시설 장비(예 : 냉각 및 진공 펌프)의 전력 소비량도 포함된다.

부록 표 6 **로직 기술 제조의 총 전력 소비량[54]**

기술 노드 (nm)	전체 공정 흐름의 cm²당 총 전력 사용량(kWh/cm²)
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

53. 가르시아 바르돈 M. 등, “지속가능성을 포함한 DTCO: 로직 기술을 위한 전력-성능-면적-비용-환경 점수(PPACE) 분석”, 2020 IEEE 국제 전자장치 회의(IEDM), 미국 샌프란시스코, 2020년 12월 12~18일. https://doi.org/10.1109/IEDM13553.2020.9372004.

54. 가르시아 바르돈 M. 등, “지속가능성을 포함한 DTCO: 로직 기술을 위한 전력-성능-면적-비용-환경 점수(PPACE) 분석”, 2020 IEEE 국제 전자장치 회의(IEDM), 미국 샌프란시스코, 2020년 12월 12~18일. https://doi.org/10.1109/IEDM13553.2020.9372004.

위 <부록 표6>의 값은 이전 섹션에 나온 TSMC의 기술 공정에 직접 적용할 수 있다(300mm 웨이퍼의 표면적이 706.95 cm2라는 점을 감안할 때). 엔비디아는 일부 제품에 TSMC의 맞춤형 5나노 공정을 이용한다.[55] TSMC가 해당 공정을 5나노 제품군에 포함하기 때문에, 우리는 이 경우 각각 5나노 전력 집약도가 적용된다고 가정했다. 최근 HBM과 DRAM 공정은 10나노 공정으로 제한된 것이 일반적이기 때문에,[56] 해당 공정은 동일한 수준으로 축소되지 않는다. SK하이닉스는 HBM3e에 1-βnm 공정을 사용했다.[57] 그러나 이는 여전히 10나노급 공정이다.[58] NAND도 비슷한 확장성 문제에 봉착했다. 다양한 칩 유형의 제조 공정은 서로 다르지만, 연구에 따르면, 유사한 기술 노드로 다양한 유형의 칩을 제조할 때의 누적 에너지 수요에서 전력이 차지하는 비중은 동일하다.[59] 따라서 다양한 유형의 칩에 대해 위에 나열된 동일한 전력 집약도를 적용할 수 있다. HBM 기반 다이스 또한 나머지 HBM 스택과 동일한 전력 집약도를 적용했다.

웨이퍼 생산 과정의 추정 전력 소비량을 이용하면, 이전 섹션에서 설명한 것과 같이 AI 수요에 따른 웨이퍼 생산량의 변화를 바탕으로 반도체 제조에서 AI로 인한 전력 소비량의 증가를 확인할 수 있다.

공식 5

$$\triangle EW_p = \triangle W_p E_p$$

기호 :

$\triangle W_p$ = *특정 기술 공정 p에 대한 웨이퍼 생산량 W의 변화*

E_p = *기술 공정 p의 전체 공정 흐름을 위한 웨이퍼 생산의 전력 집약도* $\left[\frac{kWh}{wafer}\right]$

$\triangle EW_p$ = *전체 전력 소비량 변화*

기술적 공정 p의 웨이퍼 생산량 W의 변화와 관련

따라서, 각 기술 공정 p(p=1,....,n)에 대한 웨이퍼 생산의 전력 소비량 변화량을 합산하면, AI 하드웨어의 수요에 따른 반도체 제조의 총 전력 소비량의 총 변화 ΔET를 구할 수 있다. 여기 사용되는 식은 다음과 같다.

공식 6

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

55. TSMC, “HP를 위한 진보된 기술”(2024), 2025년 2월 10일 참조. https://www.tsmc.com/english/dedicatedFoundry/technology/platform_HPC_tech_advancedTech.
56. 블록&파일스, “DRAM이 10나노에서 멈춘 이유”2020년 4월 13일, 2025년 2월 10일 참조. https://blocksandfiles.com/2020/04/13/dram-is-stuck-in-a-10nm-process-trap/.
57. WCCFTech, “SK하이닉스의 차세대 DDR5 RDIMM 및 HBM3E DRAM 솔루션을 구동하기 위한 1bnm 프로세서”(2023년 4월 30일), 2025년 2월 10일 참조. https://wccftech.com/sk-hynixs-1bnm-process-to-power-next-gen-ddr5-rdimm-hbm3e-dram-solutions/.
58. 파라트 K., 고다 A., “낸드 플래시의 스케일링 트렌드”, 2018년 IEEE 국제 전자 장비 회의(IDEM), 미국 샌프란시스코, 2018년 12월 1~5일. https://doi.org/10.1109/IEDM.2018.8614694.
59. 나카푸르카르 P., 다스 S., “집적 회로 제조 공정의 경제성 및 구현 에너지 분석”(<지속가능한 컴퓨팅: 정보학 및 시스템 35> 2022년 6월 11일), 100771. https://doi.org/10.1016/j.suscom.2022.100771.

우리는 로직, DRAM, NAND 웨이퍼 생산만 고려한다(후자는 미래의 전력 소비 시나리오에만 해당). 이와 같은 접근법을 적용한 결과는 <부록 표7>에 나와 있다.

부록 표 7 **AI 관련 웨이퍼 제조에 필요한 예상 전력 소비량**

GPU	연도	로직 웨이퍼 (GWh)	메모리 웨이퍼 (GWh)	메모리 베이스 (GWh)
Nvidia A100	2023	36.9	53.7	6.7
Nvidia H100	2023	46.5	65.9	8.2
총합	**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
총합	**2024**	**375.8**	**540.5**	**67.6**

1.4 전력 공급원 구성과 탄소 집약도

AI와 관련한 반도체 제조의 전력 소비량을 조사한 후, 우리는 이 부분에 공급하는 전력 공급원 구성을 고려해 제조 공정에서 화석 연료의 역할을 관찰하고 관련한 탄소 배출량을 확인할 수 있다. 이 분석을 위해 우리는 AI 관련 반도체 제조의 위치 기반 및 시장 기반 배출량 모두에 관심이 있다. 위치 기반 배출량은 제조가 이뤄지는 전력망의 탄소 집약도를 알 수 있게 해 주고, 시장 기반 배출량은 특정한 구매 전력(재생에너지 구매 계약 포함)의 탄소 배출량을 반영한다. 따라서 전력 소비량을 예상할 때에는 특정 업체뿐 아니라 지리적 위치를 고려해야 한다. 그런데 전 세계 반도체 제조 공장 및 생산 시설이 어디에 있는지는 잘 알려져 있지만, 생산 시설 가운데 어느 부분이 AI와 관련한 것인지는 명확하지 않다. 하지만 엔비디아의 A100과 H100 같은 인기 GPU가 TSMC의 첨단 제조 공정(기술 노드 7나노 이하. <부록 표3> 참조)에 의존하고 있다는 것은 잘 알려져 있다.

제조 공정은 로직 칩 관련 제조시설에 대한 초점을 크게 좁혔다. TSMC는 대만 남부 사이언스 파크에 위치한 팹18을 5나노 생산 거점으로 지정했다. 이곳에 엔비디아 H100 GPU 제조에 사용되는 맞춤형 4나노 제조 공정이 있다.[60] 또 대만 중부 사이언스 파크에 위치한 팹15는 엔비디아 A100 GPU에 활용되는 7나노 공정의 핵심 시설이다.[61] 앞서 언급했듯이 SK하이닉스는 2024년 1분기까지 엔비디아에 HBM을 납품하는 유일한 공급업체였기 때문에,[62] 관련 DRAM 생산 시설에 대한 초점도 좁힐 수 있다. SK하이닉스의 주요 HBM 생산시설은 인천에 위치한 팹 M10과 M16이며,[63] M10은 최근에 일부 생산 용량을 HBM 제품으로 전환했다.[64] 마이크론 또한 엔비디아 H200에 HBM3e를 공급할 것이라고 발표했다. 마이크론의 주요 HBM 생산시설은 일본 히로시마에 위치해 있다.[65] 엔비디아 B200에도 동일한 HBM3e가 사용될 예정이다.[66] AMD는 삼성전자와 제휴해 MI300X의 HBM을 개발했다.[67] 해당 HBM의 제조 시설은 한국의 평택과 화성에 있다.68 <부록 표8>은 앞서 설명한 각 제품 라인에 대한 웨이퍼 예상 수요를 제조 업체의 위치와 연계해 보여준다.

부록 표 8 장치 제조사와 제조시설 위치

GPU	Nvidia A100	Nvidia H100	Nvidia H200	Nvidia B100/ B200	AMD MI300X
GPU/XCD/IOD 제조사	TSMC	TSMC	TSMC	TSMC	TSMC
제조시설 위치	대만	대만	대만	대만	대만
메모리 제조사	SK 하이닉스	SK 하이닉스	마이크론	마이크론	삼성
메모리 제조시설 위치	한국	한국	일본	일본	한국

60. TSMC, “5나노 기술”(2024), 2025년 2월 10일 참조. https://www.tsmc.com/english/dedicatedFoundry/technology/logic/l_5nm.
61. TSMC, “N7+ 기술”(2024), 2025년 2월 10일 참조. https://www.tsmc.com/english/campaign/N7plus.
62. 로이터, “엔비디아 공급업체 SK하이닉스, 2025년 HBM 칩 거의 완판”(2024년 5월 2일), 2025년 2월 10일 참조. https://www.reuters.com/technology/nvidia-supplier-sk-hynix-says-hbm-chips-almost-sold-out-2025-2024-05-02/
63. 트위크타운, “RAM SK하이닉스, M16, M10 팹에서 월 8만 장의 웨이퍼로 DRAM 생산능력(HBM, DRAM) 확대 준비”(2024년 8월 14일), 2025년 2월 10일 참조. 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.
64. 디지타임스, “SK하이닉스, M10 생산라인 일부 전환하여 HBM 제품 생산”(2024년 7월 5일), 2025년 2월 10일 참조. https://www.digitimes.com/news/a20240705PD203/sk-hynix-production-hbm-plant-2024.html.
65. 트렌드포스, “삼성, SK하이닉스, 마이크론, 내년 생산량 두 배로 늘릴 것으로 알려진 HBM 생산량 증가”(2024년 7월 10일), 2025년 2월 10일 참조. https://www.trendforce.com/news/2024/07/10/news-samsung-sk-hynix-and-micron-ramp-up-hbm-production-reportedly-doubling-output-next-year/.
66. 더넥스트플랫폼, “마이크론, HBM 파티에 멋지게 늦었지만 너무 늦지는 않았다”(2024년 12월 19일), 2025년 2월 10일 참조. https://www.nextplatform.com/2024/12/19/micron-is-fashionably-late-to-the-hbm-party-but-not-too-late/.
67. WCCFTech, “삼성, AMD MI300X GPU를 구동하기 위한 HBM3 메모리 대량 수주”(2023년 8월 23일), 2025년 2월 10일 참조. https://wccftech.com/samsung-receives-huge-order-of-hbm3-memory-to-power-amd-mi300x-gpus/. August 23, 2023, accessed February 10, 2025
68. 트렌드포스, “삼성, SK하이닉스, 마이크론, 내년 생산량 두 배로 늘릴 것으로 알려진 HBM 생산량 증가”(2024년 7월 10일), 2025년 2월 10일 참조. https://www.trendforce.com/news/2024/07/10/news-samsung-sk-hynix-and-micron-ramp-up-hbm-production-reportedly-doubling-output-next-year/.

AI 하드웨어의 생산지역<부록 표 8>과 전력 소비량<부록 표 7>을 연결 지어 분석했으며, 각 부품의 제조 과정에서 소비된 전력을 지역별로 나누어 정리했다. 그 결과는 아래 <부록 표 9>에 나와 있다.

부록 표 9 **AI 관련 제조 공정의 지역별 전력 소비량 추정치**

지역	연도	전력 소비 (GWh)
대만	2023	83.4
한국	2023	134.6
총합	**2023**	**218.0**
대만	2024	375.8
한국	2024	315.2
일본	2024	292.8
총합	**2024**	**983.9**

우리는 위치 기반 접근법을 위해서 아래의 <부록 표 10>에 표기된 각 지역의 전력 공급원 구성을 적용했다.

부록 표 10 **한국[69], 대만[70], 일본[71]의 발전 공급원 구성**

한국

전력원	값	연도	단위	비중
석탄	184,927,212	2023	MWh	31.4%
석유	1,486,504	2023	MWh	0.3%
천연가스	157,749,022	2023	MWh	26.8%
바이오 연료	12,576,607	2023	MWh	2.1%
원자력	180,494,096	2023	MWh	30.7%
수력	3,716,437	2023	MWh	0.6%
태양광	29,288,018	2023	MWh	5.0%
풍력	3,382,450	2023	MWh	0.6%
조력	437,567	2023	MWh	0.1%
기타	13,988,591	2023	MWh	2.4%

대만

전력원	값	연도	단위	비중
석탄	119,157,090	2023	MWh	42.2%
석유	3,775,616	2023	MWh	1.3%
천연가스	111,629,599	2023	MWh	39.5%
바이오 연료	238,612	2023	MWh	0.1%
폐기물	3,499,451	2023	MWh	1.2%
원자력	17,801,952	2023	MWh	6.3%
수력	7,014,148	2023	MWh	2.5%
지열	23,164	2023	MWh	0.0%
태양광	12,908,690	2023	MWh	4.6%
풍력	6,238,287	2023	MWh	2.2%

69. 한국 전력통계정보시스템, "에너지원별 전력 생산량"(2024), 2025년 2월 10일 참조. https://epsis.kpx.or.kr/epsisnew/selectEkgeGepGesGrid.do.
70. 대만 경제부, "전력 통계"(2024), 2025년 2월 10일 참조. https://www.esist.org.tw/database/search/electric-generation.
71. 일본 경제산업성, "일반 에너지 통계"(2024), 2025년 2월 10일 참조. https://www.enecho.meti.go.jp/statistics/total_energy/results.html.

Japan

전력원	값	연도	단위	비중
석탄	280,400,000	2023	MWh	28.5%
석유	71,600,000	2023	MWh	7.3%
천연가스	324,100,000	2023	MWh	32.9%
바이오 연료	40,100,000	2023	MWh	4.1%
원자력	84,100,000	2023	MWh	8.5%
수력	74,800,000	2023	MWh	7.6%
지열	3,400,000	2023	MWh	0.3%
태양광	96,500,000	2023	MWh	9.8%
풍력	10,500,000	2023	MWh	1.1%

전력 소비와 관련한 위치 기반 탄소 배출량을 파악하기 위해, 우리는 앞서 언급한 각 전력망에서 전기를 생산할 때 나오는 탄소 집약도(gCO2/kWh)를 파악했다. 이는 아래의 <부록 표 11>에 나와 있다. 이 표와 관련해, 본 보고서를 작성할 당시의 가장 최신 자료가 2023년의 자료였으므로 이 값을 2023년과 2024년 모두에 적용했다.

부록 표 11 **한국[72], 대만[73], 일본[74]의 전력 생산 시 탄소 집약도**

지역	연도	탄소집약도 (gCO_2/kWh)
대만	2023	494
한국	2023	431
일본	2023	451

72. 엠버, “에너지 연구소-세계 에너지 통계 검토(2024). 전력 생산의 탄소 집약도”(2024), 2025년 2월 10일 참조. https://ourworldindata.org/grapher/carbon-intensity-electricity.
73. 대만경제부, “전력 탄소 배출 요소”(2024), 2025년 2월 10일 참조. https://www.moeaea.gov.tw/ecw/populace/content/SubMenu.aspx?menu_id=114.
74. 일본 경제산업성, “일반 에너지 통계”(2024), 2025년 2월 10일 참조. https://www.enecho.meti.go.jp/statistics/total_energy/results.html.

우리는 각국의 전력 공급원 구성을 이용해 예상 전력 소비량에 따른 위치 기반 탄소 배출량을 추정할 수 있었다. 그 내용은 아래 <부록 표12>에 나와 있다.

부록 표 12 **AI 관련 제조업의 지역별 탄소 배출량 추정치 (이산화탄소 환산, 단위: kt)**

지역	연도	전력 소비량 (GWh)	이산화탄소 환산 배출량 (단위 : kt)
대만	2023	83.4	41.2
한국	2023	134.6	58.0
총합	**2023**	**218.0**	**99.2**
대만	2024	375.8	185.7
한국	2024	315.2	135.9
일본	2024	292.8	132.1
총합	**2024**	**983.9**	**453.6**

우리는 AI 관련 반도체 제조의 시장 기반 탄소 배출량에 관심이 있어 각 제조업체의 환경 보고서를 조사해 전력 공급원 구성 및 배출 계수를 파악했다. 그러나 안타깝게도 제조업체들이 모두 동일한 유형의 정보를 제공하지는 않았다. TSMC와 삼성전자는 시장 기반 및 위치 기반 스코프 2 배출량을 공개했지만, SK하이닉스는 위치 기반 스코프 2 배출량, 그리고 시장 기반 스코프 1 및 스코프 2 배출량 합계만 보고했다. 마이크론은 시장 기반 스코프 2 배출량만 보고했다. 이러한 편차로 인해 완전한 시장 기반 분석이 불가능했다. 또한 위치 기반 결과와 제대로 비교할 수도 없었다. 데이터가 있다고 하더라도, AI 관련 제조 공정에 관한 내용만 세부적으로 살펴볼 수 없는 경우가 많았다. 3장의 결과는 2023년과 2024년 AI 관련 제조업이 전체 반도체 제조업의 전력 소비량에서 차지하는 비중이 상대적으로 작다는 것을 시사한다. 이는 각 기업이 공개한 정보가 이 분야의 내역을 제대로 보여주지 않고 있을 가능성이 크다는 것을 뜻한다. 따라서 본 연구에서는 시장 기반 배출량 분석은 제외했다.

2. 시나리오 분석

이번 연구에서, 우리는 2023과 2024년 반도체 제조 분야에 AI가 미친 잠재적 영향력뿐 아니라, 2030년까지의 전망에도 관심을 가진다. 그러나 하드웨어 구조가 진화할 가능성이 크기 때문에, 현재의 AI 하드웨어 수요를 추정하는 것만으로 2030년까지 AI가 반도체 제조에 미칠 영향을 파악할 수는 없었다. 따라서 향후 AI 관련 웨이퍼의 공급과 수요의 전망에 대한 몇 가지 시나리오를 살펴본다. 맥킨지와 같은 경영 컨설팅 기업이 여러 시나리오를 공개적으로 평가한 바 있으며,[75] 그 요약 내용은 아래 <부록 표13>에 나와 있다.

부록 표 13 **킨지의 2030년 AI 관련 로직, DRAM, NAND 웨이퍼 수요 추정치**

2030 WSPY (백만)	로직	DRAM	NAND
보수적	1.2	7.1	1.7
기본	2.4	13.6	4
진취적	3.6	21.0	7.9

앞서 언급한 수치를 섹션1.3에서 설명한 대로 웨이퍼 생산의 추정 전력 집약도와 결합해 맥킨지의 세 가지 시나리오에 따른 전력 요구량을 얻을 수 있다. 특히, 우리는 웨이퍼 생산과 관련한 전력 집약도는 시간이 지나도 일정하게 유지된다고 가정한다. 예컨대, 우리는 현재 5나노 공정의 전력 집약도를 2030년 첨단 로직 웨이퍼 생산의 전력 집약도로 가정한다. 제조 공정의 전력 집약도는 시간이 갈수록 개선될 가능성이 높지만, AI 가속기가 곧 보다 진보된 제조 공정을 채택할 수 있다. 엔비디아와 AMD는 이미 5나노 공정보다 전력 집약도가 높은 TSMC의 3나노 공정을 이용할 계획이다(부록 표6 참조).[76] 또한 TSMC는 이미 새로운 2나노 공정의 시험 생산에 착수했다.[77] 따라서 전력 강도를 일정하게 유지하면, 이러한 미래의 잠재적인 개발 사이에서 균형을 맞출 수 있다.

우리는 AI 관련 반도체 제조의 전력 공급원 구성과 탄소 집약도에 대해서도 비슷한 접근 방식을 적용한다. 이는 현재 AI 하드웨어 제조의 핵심 동력으로 부상하고 있는 제조업체와 지역이 향후에도 계속 그 역할을 맡게 될 것임을 의미한다.

부록 표 14 **각 시나리오별 2030년 AI 관련 칩 제조에 따른 예상 전력 소비량**

2030년 (단위: TWh)	로직	DRAM	NAND	총합
보수적	2.44	7.35	1.76	11.55
기본	4.87	14.09	4.14	23.10
진취적	7.31	21.75	8.18	37.24

75. 맥킨지컴퍼니, “생성형 AI: 반도체 산업의 차세대 S-커브?”(2024년 3월 29일), 2025년 2월 10일 참조. https://www.mckinsey.com/industries/semiconductors/our-insights/generative-ai-the-next-s-curve-for-the-semiconductor-industry.

76. WCCFTech, “TSMC 3나노 공정, 엔비디아 루빈 및 AMD 인스팅트 MI355X 그리고 인텔 팔콘 쇼어 등 AI 빅테크로부터 수주할 듯”(2024년 10월 14일), 2025년 2월 10일 참조 https://wccftech.com/tsmc-3nm-witness-massive-adoption-ai-tech-giants-nvidia-rubin-amd-mi355x-intel-falcon-shores/.

77. 조선일보, “삼성과 TSMC, 치열한 2나노 경쟁”(2024년 12월 16일), 2025년 2월 10일 참조. https://www.chosun.com/english/industry-en/2024/12/16/SMAOH7NISJAZXBL7LFX3BX2YVU/.

2024년 추정치에 따라 로직 및 동적 랜덤-어세스 메모리(DRAM) 제조의 전력 공급원 구성과 탄소 집약도를 일정하게 유지하면, <부록 표 15>와 같이 예상 전력 소비량과 관련한 탄소 배출량을 구할 수 있다.

부록 표 15 **2030년 시나리오별 AI 칩 제조에 따른 예상 CO_2 배출량**

2030년 (단위: $MtCO_2$)	로직	DRAM	NAND	Total
보수적	1.20	3.24	0.78	5.22
기본	2.41	6.21	1.83	10.44
진취적	3.61	9.58	3.61	16.80

부록 표 15는 낸드(NAND, 즉 플래시 메모리) 제조의 탄소 집약도를 보여준다. HBM 제조 분야의 지배적 기업(SK하이닉스, 마이크론, 삼성전자)이 NAND 시장의 대부분을 장악하고 있다는 점을 고려해, 탄소집약도를 DRAM 제조와 동일한 값으로 설정했다.[78] 그런데 해당 지역마다 기후 대책을 가지고 있고 각 전력망의 탈탄소화에 어느 정도 성공할 수 있다. 따라서 이러한 전망의 정확도는 2030년 전력망 탄소 집약도 예측을 검토함으로써 개선될 여지가 있다. 그러나 공식적인 자료는 서로 다른 목표시점을 제시하고, 각각의 기준 기간도 다르기 때문에, 이 평가를 위해 필요한 데이터들은 이용하지 못하거나 일관성 결여로 인해 이번 평가에서는 사용할 수가 없었다. 또 시나리오 분석 결과, AI 관련 반도체 제조로 인한 전력 수요의 잠재적 증가는 특정 지역의 탈탄소화 노력을 무력화할 수 있는 것으로 나타났다.

78. 트렌드포스, "NAND 플래시 시장 지형 변하나?"(2024년 3월 14일), 2025년 2월 10일 참조. https://www.trendforce.com/news/2024/03/14/news-nand-flash-market-landscape-to-change/.

그린피스 동아시아 2025년 4월 발간.

그린피스는 평화로운 시위와 창의적인 커뮤니케이션을 통해 글로벌 환경 문제를 폭로하고, 친환경적이고 평화로운 미래에 필수적인 해결책을 모색하는 독립적 캠페인 단체다.