일요일은 AI · 반도체 스터디

AI 인프라와 반도체

통계·전산·AI 배경 학습자를 위한 일곱 회차 자습서.
용어를 외우는 대신 몇 개의 식을 세워 숫자를 직접
넣어 본다.

권보원(대구지방법원 부장판사) 엮음

구성 · 사용법 + 7회차 + 부록 3종
정량 실습 · 32문항, 전체 풀이 수록
인터랙티브 그림 · 12종(화면본. 인쇄본은 같은 자리에 정지 그림)
용어 색인 · 114개
권장 진행 · 8주, 주당 약 2시간

인쇄 · Ctrl+P → PDF로 저장하면 접이식 풀이가 자동으로 모두 펼쳐집니다.
오프라인 · 이 파일 하나로 완결됩니다. 인터랙티브 그림도 인터넷 없이 움직이고, 링크는 인터넷이 있을 때 열립니다.

2.0.1판 · 2026. 10. 22. · 생성형 AI와 함께 엮었습니다. 틀린 곳이 있을 수 있으며 앞으로도 계속 고쳐 나가겠습니다.

차례

이 책의 사용법 7

이 책이 다루는 것 7

회차마다 무엇이 들어 있나 8

8주 진행표 8

매주 같은 그림 한 장을 갱신할 것 9

막혔을 때 9

수치에 관한 일러두기 — 반드시 읽을 것 10

다른 자료와 함께 볼 것 10

AI 반도체 산업 지형 11

학습목표 11

0. 자습 안내 — 혼자 읽는 사람에게 11

1. 직관 잡기 — "AI 반도체"라는 말의 정체 13

2. 밸류체인 — 누가 어디서 돈을 버는가 14

3. 수요의 구조 변화 — 학습에서 추론으로 15

4. 시장 수치를 비판적으로 읽는 법 16

5. 공급망 리스크의 세 축 17

6. 자주 하는 오해 18

7. 토론 질문 19

8. 후속 공부 19

자습 부록 A — 수치 감사 실습과 풀이 19

자습 부록 B — 토론 질문 생각의 실마리 22

자습 부록 C — 핵심 용어 22

자습 부록 D — 자가진단 23

다음 회차로 넘어가기 전에 24

출처 24

AI 칩 — 연산 반도체 26

학습목표 26

0. 자습 안내 — 혼자 읽는 사람에게 26

1. 직관 잡기 — 계산은 싸고, 데이터 배달은 비싸다 28

2. 왜 CPU로는 안 되는가 28

3. GPU의 실체 — matmul 한 줄이 하드웨어에서 하는 일 29

4. TPU와 시스톨릭 어레이 — 재사용을 하드웨어에 새기기 31

5. NPU와 ASIC — 전용화의 경제학 33

6. 루프라인 모델 — 이 책의 공용 도구 34

7. 소프트웨어 락인 — 기술 외적 해자 38

8. 자주 하는 오해 39
9. 정량 실습 40
10. 토론 질문 41
11. 후속 공부 41
자습 부록 A — 정량 실습 풀이 42
자습 부록 B — 토론 질문 생각의 실마리 49
자습 부록 C — 핵심 용어 49
자습 부록 D — 자가진단 50
다음 회차로 넘어가기 전에 51
출처 51

AI 메모리 — DRAM · HBM · HBF 54
학습목표 54
0. 자습 안내 — 혼자 읽는 사람에게 54
1. 직관 잡기 — 메모리 계층은 물리적 타협의 지도 56
2. DRAM의 물리 — 왜 느린가, 왜 안 빨라지는가 57
3. 대역폭을 올리는 두 가지 길, 그리고 막다른 골목 58
4. HBM 구조 정밀 해설 59
5. HBM은 왜 비싼가 — 수율의 곱셈 61
6. HBF (High Bandwidth Flash) — 용량이라는 또 다른 병목 62
7. PIM / PNM — 연산을 메모리로 가져가기 65
8. 자주 하는 오해 66
9. 정량 실습 67
10. 토론 질문 67
11. 후속 공부 67
자습 부록 A — 정량 실습 풀이 68
자습 부록 B — 토론 질문 생각의 실마리 75
자습 부록 C — 핵심 용어 75
자습 부록 D — 자가진단 76
다음 회차로 넘어가기 전에 76
출처 77

팹리스 · 파운드리와 밸류체인 80
학습목표 80
0. 자습 안내 — 혼자 읽는 사람에게 81
1. 직관 잡기 — 왜 설계와 제조가 갈라졌나 82
2. 플레이어별 역할 83
3. 칩 설계 파이프라인 — 소프트웨어 개발과 대응시켜 보기 84
4. "3nm"의 진실 — 노드 이름 읽는 법 86

5. 수율의 경제학 — 이 책의 두 번째 핵심 도구 88
6. ASIC · FPGA · 범용 프로세서 — 정량적 선택 91
7. 자주 하는 오해 92
8. 정량 실습 92
9. 토론 질문 93
10. 후속 공부 93
자습 부록 A — 정량 실습 풀이 94
자습 부록 B — 토론 질문 생각의 실마리 101
자습 부록 C — 핵심 용어 101
자습 부록 D — 자가진단 102
다음 회차로 넘어가기 전에 102
출처 103
전공정 106
학습목표 106
0. 자습 안내 — 혼자 읽는 사람에게 107
1. 직관 잡기 — 반도체 제조는 "조각"이 아니라 "인쇄와 도장" 108
2. 웨이퍼 — 출발점 109
3. 트랜지스터 구조의 진화 — 평면에서 FinFET·GAA로 111
4. 노광 (리소그래피) — 가장 깊이 들어갈 부분 112
5. 식각 (Etching) — 빼는 공정 115
6. 증착 (Deposition) — 더하는 공정 116
7. 이온주입 · 열처리 · CMP 117
8. 검사 · 계측 — CS/AI 전공자의 직접적 접점 118
9. 전체 통합 — 층 하나가 만들어지는 과정 119
10. 자주 하는 오해 120
11. 정량 실습 121
12. 토론 질문 121
13. 후속 공부 121
자습 부록 A — 정량 실습 풀이 122
자습 부록 B — 토론 질문 생각의 실마리 131
자습 부록 C — 핵심 용어 132
자습 부록 D — 자가진단 133
다음 회차로 넘어가기 전에 133
출처 134
후공정과 첨단 패키징 138
학습목표 138
0. 자습 안내 — 혼자 읽는 사람에게 139

1. 직관 잡기 — 후공정의 목적이 바뀌었다 140
2. 왜 무게중심이 옮겨갔나 — 세 가지 압력 141
3. 다이싱 — 자르기도 어려워졌다 143
4. TSV — 실리콘을 관통하는 수직 배선 143
5. 인터포저 — 칩들을 잇는 판 145
6. 본딩 — 붙이는 기술의 세대 교체 147
7. 2.5D와 3D — 정확한 구분 150
8. 이종집적과 칩렛 표준 (UCIe) 155
9. 신호 무결성 — 통신 전공 배경이 있으면 익숙한 영역 156
10. 자주 하는 오해 156
11. 정량 실습 157
12. 토론 질문 158
13. 후속 공부 158
자습 부록 A — 정량 실습 풀이 159
자습 부록 B — 토론 질문 생각의 실마리 168
자습 부록 C — 핵심 용어 169
자습 부록 D — 자가진단 170
다음 회차로 넘어가기 전에 171
출처 171
AI 데이터센터 175
학습목표 175
0. 자습 안내 — 혼자 읽는 사람에게 175
1. 직관 잡기 — 밀도가 바뀌면 모든 것이 바뀐다 177
2. 전력 체인 — 계통에서 트랜지스터까지 178
3. 냉각의 물리 — 왜 공기로는 안 되는가 179
4. 효율 지표를 비판적으로 읽기 181
5. 네트워크 — 스케일업과 스케일아웃 182
6. 거시 수치와 그 불확실성 184
7. 다시 메모리 시장으로 — 마지막 고리 185
8. 자주 하는 오해 186
9. 정량 실습 186
10. 토론 질문 187
11. 후속 공부 188
자습 부록 A — 정량 실습 풀이 188
자습 부록 B — 토론 질문 생각의 실마리 203
자습 부록 C — 핵심 용어 203
자습 부록 D — 자가진단 204

마치며 — 일곱 회차를 한 줄로 꿰기 204
출처 206
부록 I — 공식 요약과 실습 색인 209
1. 연산과 대역폭 (2회차) 209
2. 수율과 원가 (4회차) 210
3. 노광 (5회차) 210
4. 패키징과 열 (6회차) 211
5. 데이터센터 (7회차) 211
6. 정량 실습 색인 (전 32문항) 211
7. 일곱 회차를 한 줄로 꿰기 213
부록 II — 용어 색인 214
부록 III — 인터랙티브 그림 목록 219
만든 이 222
이 책에 대하여 222
어떻게 만들었나 222
판 이력 223
감사의 말 223
이용할 때 224

들어가며

이 책의 사용법

> **먼저 알아 둘 것**
>
> 이 책은 엮은이가 생성형 AI(Claude)와 함께 만들었습니다. 회차마다 1차 출처와 대조하는 검증을 여러 차례 거쳤지만 틀린 서술이 남아 있을 수 있습니다. 수치와 인용은 원문으로 다시 확인해 주십시오. 이 책은 앞으로도 계속 고쳐 나가겠습니다. 지금 보시는 판과 날짜는 표지와 간기에 있습니다(2.0.1판 · 2026. 10. 22.).
>
> 이 책에서 엮은이에게 권리가 있는 부분은 CC BY-NC 4.0(저작자표시-비영리 4.0 국제)으로 공개합니다. 저작자와 라이선스(링크 포함)를 밝히고, 고친 곳이 있으면 고쳤다고 알리면 비영리 목적으로 자유롭게 고쳐 쓰고 나눌 수 있습니다. 받은 사람의 이용을 법적 조건이나 기술적 조치로 제한해서는 안 됩니다. 이 요약은 이해를 돕기 위한 것이고, 조건은 라이선스 원문을 따릅니다. 화면본에 든 인터랙티브 그림의 코드(HTML·CSS·자바스크립트)는 PolyForm Noncommercial 1.0.0으로 따로 공개합니다. 비영리 목적의 이용·수정·재배포는 그 라이선스의 조건에 따라 허용되고, 상업적으로 이용하시려면 미리 저장소 이슈로 연락해 이용 조건을 따로 협의해 주십시오. 원문에서 옮기거나 다시 그린 표·인용·그림은 이 라이선스들의 대상이 아니며 각 원저작자의 권리를 따릅니다. 권리자께서 저장소 이슈로 알려 주시면 고치거나 빼겠습니다.

이 책은 「일요일은 AI」 스터디의 반도체 회차를 위해 엮은 자습서로, AI 인프라와 반도체를 일곱 회차로 다룹니다. 통계·전산·AI 배경은 있지만 반도체를 전공하지 않은 분을 위해 썼고, 용어를 외우게 하는 대신 **몇 개의 식을 세워 숫자를 직접 넣어 보게** 합니다. 회차마다 계산 실습과 전체 풀이가 붙어 있어 혼자서도 답을 검증할 수 있습니다. 화면본(HTML)에서는 식 가까이에 놓인 인터랙티브 그림에서 값을 바꿔 볼 수 있고, 인쇄본(PDF)에는 같은 자리에 처음 상태의 그림이 실려 있습니다. 다른 모임에서 같은 방식으로 읽으시거나 혼자 읽으셔도 좋습니다.

이 책이 다루는 것

일곱 회차는 하나의 질문을 따라간다. **"AI를 돌리는 물리적 설비의 한계는 어디에서 정해지는가."**

회차	다루는 것	그 회차가 세우는 도구
1	AI 반도체 산업의 지형과 수치 읽는 법	정의를 먼저 묻는 습관
2	연산 반도체 (GPU·NPU·TPU·ASIC)	**루프라인 모델**
3	메모리 (DRAM·HBM·HBF)	대역폭과 용량의 분리
4	팹리스·파운드리와 밸류체인	**수율 식**
5	전공정 (노광·식각·증착·계측)	**레일리 판정식**
6	후공정과 첨단 패키징	칩렛 손익, 열저항 사슬
7	AI 데이터센터 (전력·냉각)	종합

2·4회차에서 세운 두 식이 나머지 전부를 지배한다. 시간이 없다면 그 둘만이라도 손으로 풀어 볼 것.

회차마다 무엇이 들어 있나

각 회차는 같은 구조를 따른다.

순서	이름	무엇을 하나
1	핵심 Takeaway	이 회차의 결론을 세 줄로 미리 본다
2	학습목표	끝나고 답할 수 있어야 하는 질문들
3	**0. 자습 안내**	**혼자 읽는 사람을 위한 안내.** 직전 회차에서 가져올 것, 선행 지식 점검, **읽는 순서와 시간 배분**, 막히는 지점의 우회 설명
4	본문	
5	자주 하는 오해	흔히 잘못 알려진 것들
6	정량 실습	손으로 푸는 문제
7	토론 질문	정답이 없는 질문
8	**자습 부록 A**	**정량 실습 전체 풀이.** 접혀 있으니 먼저 풀고 펼칠 것
9	자습 부록 B	토론 질문의 논점 축
10	자습 부록 C	그 회차의 핵심 용어
11	**자습 부록 D**	**자가진단 체크리스트.** ★ 항목은 통과 전에 다음 회차로 넘어가지 말 것
12	다음 회차로 넘어가기 전에	남긴 질문 ↔ 다음 회차의 답

「**자습 부록 A**」**의 풀이는 접혀 있다.** 화면에서는 제목을 눌러 펼치고, 종이로 인쇄하면 자동으로 모두 펼쳐진다.

8주 진행표

일곱 회차지만 **8주로 잡는 편이 낫다.** 5회차(전공정)가 다른 회차의 1.5배 분량이기 때문이다.

주	읽을 것	소요	그 주에 반드시 손으로 풀 것
1	1회차 AI 반도체 산업 지형	100분	수치 감사 3문항
2	2회차 AI 칩	120분	**실습 1·2**
3	3회차 AI 메모리	110분	**실습 1·2**
4	4회차 팹리스·파운드리	110분	**실습 1·2·4**
5	5회차 전공정 전반 (1~4절)	95분	**실습 1·2**
6	5회차 전공정 후반 (5~9절)	65분	**실습 3·4·5**
7	6회차 후공정·첨단 패키징	130분	**실습 1·2·4**
8	7회차 AI 데이터센터 + 종합	190분	**실습 5** (필수)

주당 약 2시간이다. 회차 하나를 한 번에 다 읽기 어려우면 각 회차의 「읽는 순서」 표에 있는 단계 단위로 끊어 읽으면 된다.

매주 같은 그림 한 장을 갱신할 것

자습에서 가장 쉽게 무너지는 것이 **"읽었는데 남지 않는다"** 이다. 대응책은 하나뿐이다. 종이 한 장(또는 화면 한 장)에 아래 골격을 그려 두고, **매주 새로 배운 층을 덧그린다.**

```
[사용자 요청] → [모델·소프트웨어] → [연산 칩] → [메모리] → [패키지]
                                                          ↓
[전력·냉각] ← [데이터센터·랙] ← [서버·보드] ← [제조·공급망] ↵
```

갱신 규칙 세 가지

1. **각 화살표에 숫자 하나를 붙인다.** (예: 연산 칩 → 메모리 화살표에 "리지 포인트 ≈ 295 FLOP/byte")
2. **각 상자에 병목 여부를 표시한다.** 회차가 진행되며 병목 표시가 옮겨 다니는 것이 보여야 한다.
3. **지난주 그림을 지우지 않는다.** 색을 바꿔 덧그린다.

7회차 말미의 「일곱 회차를 한 줄로 꿰기」가 최종 정답지다. **8주차에 그것을 보지 않고 복원할 수 있으면 끝난 것이다.**

막혔을 때

증상	원인	대응
2회차부터 계속 겉돈다	1회차 3절의 "가중치 전체를 읽는다"가 소화되지 않았다	1회차 자습 안내의 **도서관 비유**로 돌아갈 것. 여기가 안 되면 이후가 전부 안 된다
식은 따라가는데 왜 중요한지 모르겠다	숫자를 안 넣어 봤다	부록 A를 **풀이를 덮고** 직접 풀 것. 계산 없이 읽으면 남지 않도록 설계된 책이다
용어가 너무 많다	한꺼번에 외우려 한다	외우지 말고 **부록 II 용어 색인**을 옆에 두고 그때그때 찾을 것
5회차 노광이 안 읽힌다	광학 배경 부족	5회차 자습 안내의 **사진 조리개 비유**부터. 그래도 막히면 3·4절만 읽고 넘어간 뒤 6회차 후 복귀
분량에 압도된다	한 번에 다 읽으려 한다	각 회차 「읽는 순서」 표의 **단계 단위로** 끊어 읽을 것
다 읽었는데 연결이 안 된다	회차별로 끊어 읽고 사슬을 안 만들었다	위의 **누적 그림**을 지금부터라도 시작할 것

수치에 관한 일러두기 — 반드시 읽을 것

이 책의 계산 실습에는 **가정값**이 많이 쓰인다. 결론의 방향을 보이기 위한 예시이지 실측치가 아니다.

기준일은 2026년 10월 8일이다. 양산·출하 현황, 전력 수요 전망, 표준과 제품 사양처럼 시점에 따라 바뀌는 서술은 이 날짜에 1차 출처(제조사 발표·표준 문서·기관 보고서)로 다시 확인했고, 본문에 확인한 시점을 함께 적었다.

가정값이며 그대로 인용하면 안 되는 것

- 웨이퍼 한 장 가격(15,000달러), 전기요금(kWh당 0.08달러), 가속기 노드 가격(25만 달러)
- ASIC·GPU 단가와 NRE, 결함 밀도 D_0, 조립 수율, 인터페이스 오버헤드 비율
- 각 실습에 명시된 그 밖의 설정값

공개 스펙시트 기준이며 세대 교체 시 달라지는 것

- 가속기의 연산 성능과 메모리 대역폭(A100·H100·B200). 모두 **dense(희소성 미적용)** 기준이다. 벤더 자료는 희소성 적용 시 2배 수치를 병기하므로 반드시 확인할 것
- HBM 규격, 레티클 한계, 마이크로범프·하이브리드 본딩 피치

구조적으로 안정적인 것 — 인용해도 되는 것

- 식 자체: 루프라인, 수율 모형, 레일리 판정식, 열저항, PUE 정의
- 그 식에서 나오는 **관계의 방향:** 면적이 커지면 수율이 지수적으로 떨어진다, NA를 키우면 초점심도가 제곱으로 나빠진다, 배치가 작으면 대역폭에 묶인다 등

시장 규모·점유율의 절대 수치는 이 책이 의도적으로 싣지 않았다. 자료마다 정의와 집계 기준이 달라 편차가 크기 때문이다. 1회차 4절이 그 이유를 다룬다.

다른 자료와 함께 볼 것

이 책이 상대적으로 얇은 부분은 아래와 같다.

주제	이 책의 현황
AI 컴파일러·소프트웨어 스택	2회차 7절에서 락인 관점으로만 다룬다. 컴파일 파이프라인 자체는 거의 없다 — **가장 큰 공백**
1차 자료 원문 링크	회차 첫머리 「원자료 바로가기」와 각 회차 출처 절에 링크를 달았다. 유료 표준·논문은 서지와 DOI까지만 있다
도해·이미지	다시 그린 도해와 인터랙티브 그림이 있으나 사진·현미경 이미지는 없다

1회차

AI 반도체 산업 지형

핵심 TAKEAWAY

"AI 반도체"는 단일 제품이 아니라 연산 칩과 메모리 칩을 아우르는 느슨한 범주이며, 이 모호함이 시장 통계를 읽을 때 첫 번째 함정이 된다. 산업은 설계(팹리스) · 제조(파운드리/IDM) · 장비·소재의 3층으로 나뉘고, **이익이 큰 층과 대체 불가능한 층이 일치하지 않는다.** 수요는 학습에서 추론으로 이동 중이며, 이 전환이 메모리 병목을 산업 전체의 중심 문제로 끌어올렸다. 이 회차의 목표는 수치를 외우는 것이 아니라, 어떤 수치가 어떤 정의 위에 서 있는지 확인하는 습관을 들이는 것이다.

원자료 바로가기

KDB미래전략연구소 「AI 반도체 기술 및 산업 동향」(KDI 경제교육·정보센터) · 삼일PwC 「글로벌 반도체 산업 전망 2026」(PDF) · PwC컨설팅 「반도체 슈퍼 사이클의 역설」 · LBNL 2024 미국 데이터센터 전력 보고서 발표 · NVIDIA DGX H100/H200 사용자 안내서 — 전원 사양 · eCFR 15 CFR §734.9 (FDP 규칙)

학습목표

이 회차가 끝나면 다음에 답할 수 있어야 한다.

1. "AI 반도체 시장이 몇 조 달러"라는 기사를 봤을 때, 무엇을 먼저 확인해야 하는가?
2. 밸류체인의 각 층에서 이익률과 대체 불가능성은 어떻게 다른가?
3. 학습(training)에서 추론(inference)으로 수요가 이동하면 하드웨어 요구사항은 어떻게 바뀌는가?
4. 반도체 공급망 리스크는 어떤 유형으로 나뉘고, 각각의 성격이 어떻게 다른가?

0. 자습 안내 — 혼자 읽는 사람에게

이 회차는 읽는 회차다

1~7회차 중 유일하게 계산이 없는 회차다. 대신 이 회차의 진짜 목표는 지식 습득이 아니라 **습관 하나를 몸에 붙이는 것**이다 — 숫자를 볼 때마다 "이건 무엇을 센 숫자인가"를 먼저 묻는 습관. 4절이 그 훈련이고, 나머지 절은 그 습관을 쓸 무대를 깔아 준다.

선행 지식 점검 — 다음 다섯 개를 한 줄로 설명할 수 있는가

막히는 것이 있으면 아래 답을 읽고 넘어가면 된다. 이 회차를 읽는 데는 이 정도면 충분하다.

개념	한 줄 답
파라미터(parameter)	모델이 학습으로 정한 숫자(가중치) 하나하나. "70B 모델"은 그 숫자가 700억 개라는 뜻
FLOP / FLOP/s	FLOP은 부동소수점 연산 **횟수**, FLOP/s는 초당 횟수(**속도**). 앞은 일의 양, 뒤는 처리 능력
메모리 대역폭	메모리에서 칩으로 초당 옮길 수 있는 바이트 수 [byte/s]. 용량(GB)과 완전히 다른 개념
웨이퍼(wafer)	실리콘 원판. 이 위에 칩 수백 개를 한꺼번에 만들고 마지막에 잘라 낸다
팹리스 / 파운드리	설계만 하는 회사 / 남의 설계를 위탁 생산하는 공장

읽는 순서 (총 100분 권장)

혼자 읽을 때는 앞에서부터 순서대로 읽는 것이 최선이 아니다.

순서	절	시간	이유
1	1절 직관 잡기	15분	용어의 범위를 먼저 고정한다
2	**3절 학습→추론 전환**	**25분**	**이 회차의 심장.** 나머지 여섯 회차가 전부 여기서 파생된다
3	2절 밸류체인	15분	3절을 읽고 나면 "왜 이 층에 돈이 몰리나"가 보인다
4	4절 수치 읽기	20분	아래 실습과 함께
5	5·6절 리스크·오해	15분	
6	자습 부록 실습	10분	

절대 건너뛰지 말 것: 3절 「왜 이 전환이 메모리를 왕좌에 앉혔나」. 여기서 나온 "연산강도 $I \approx B$" 한 줄이 2회차의 루프라인, 3회차의 HBM, 6회차의 패키징, 7회차의 전력 문제로 전부 이어진다. **이 문단이 이해되지 않으면 2회차를 읽어도 계속 겉돈다.**

3절이 어려우면 — 우회 경로

3절은 아직 세우지 않은 개념(연산강도, 리지 포인트)을 미리 당겨 쓰기 때문에 첫 독에서 막히는 것이 정상이다. 다음 비유로 감만 잡고 넘어간 뒤, 2회차를 읽고 **다시 돌아와서** 읽으면 풀린다.

> 도서관에서 책 한 권을 통째로 옮겨 와 한 문장만 읽고 돌려보낸다고 하자. 책을 옮기는 데 10분, 읽는 데 1초다. 이때 병목은 "읽는 속도"가 아니라 "옮기는 속도"다. 추론의 디코딩이 정확히 이렇다. 토큰 **하나**를 만들려고 모델 가중치 **전체**를 메모리에서 읽어 온다. 그러니 칩이 아무리 빨라도 소용이 없고, 메모리에서 얼마나 빨리 실어 오느냐가 전부를 결정한다. 해결책은 하나뿐이다 — 책을 한 번 옮겨 온 김에 **여러 사람이 나눠 읽는 것**(배치). 배치가 왜 그렇게 중요한지가 여기서 나온다.

---

1. 직관 잡기 — "AI 반도체"라는 말의 정체

하나의 물건이 아니다

"AI 반도체"라는 말은 최소 네 가지를 가리킨다.

부르는 것	실제 물건	대표 예
좁은 뜻 (연산)	AI 연산에 특화된 가속기 칩	GPU, TPU, NPU, 추론 ASIC
넓은 뜻 (연산+메모리)	위에 AI용 메모리까지 포함	+ HBM
시스템 뜻	가속기 모듈·보드 전체	HGX 보드, 가속기 카드
산업 뜻	AI 수요로 팔리는 반도체 전부	+ 네트워크 칩, 전력반도체, CPU

기사나 보고서가 "AI 반도체 시장 규모 OO억 달러"라고 할 때 넷 중 무엇인지 밝히지 않는 경우가 많다. **정의가 다르면 수치는 몇 배씩 차이난다.** 이 책이 들이려는 첫 번째 습관은 숫자를 볼 때마다 "이건 어느 범위인가?"를 묻는 것이다.

왜 이 산업이 갑자기 이렇게 커졌나 — 세 줄 요약

1. **트랜스포머는 행렬곱 덩어리다.** 대규모 병렬 곱셈-누산으로 환원되는 연산이라 전용 하드웨어의 이득이 극단적으로 크다.
2. **스케일링 법칙이 "더 큰 모델 = 더 좋은 성능"을 경험적으로 뒷받침했다.** 컴퓨트 투입이 곧 성능이라는 믿음이 자본 지출을 정당화했다.
3. **그런데 그 하드웨어를 만들 수 있는 곳이 극소수다.** 수요는 폭발했는데 공급은 물리적·자본적으로 탄력이 없다. 그래서 가격과 이익이 특정 지점에 몰렸다.

이 책은 3번의 "왜 극소수인가"를 일곱 회차에 걸쳐 해부하는 과정이라고 봐도 좋다.

---

2. 밸류체인 — 누가 어디서 돈을 버는가

3층 구조



그림의 1층과 2층 사이에는 얇은 층이 하나 더 있다. **디자인하우스**(설계 서비스 회사)는 팹리스가 넘긴 논리 설계를 특정 파운드리의 공정에 맞춰 물리적으로 구현하고 검증한 뒤 제조로 넘긴다. 파운드리도 이들을 협력망으로 묶어 둔다. TSMC는 VCA(Value Chain Alliance, 이전 이름은 Value Chain Aggregator)의 회원을 독립 설계 서비스 회사로 소개하면서, 이들이 IP 개발·설계 백엔드부터 웨이퍼 제조·조립·테스트까지 단계별 서비스를 맡아 시스템 회사·ASIC 회사·신생 고객의 설계를 양산으로 이끈다고 설명한다(2026년 10월 확인). 플레이어와 파운드리별 제도는 4회차에서 다룬다.

핵심 통찰 — 이익률과 대체 불가능성은 다르다

- **이익률이 가장 높은 곳:** 1층의 선두 팹리스와 2층의 선두 파운드리
- **대체 불가능성이 가장 높은 곳:** 3층. 특히 EUV 노광장비는 **사실상 ASML 단독**이다. 매출 규모는 NVIDIA보다 훨씬 작지만, 이 회사가 멈추면 첨단 로직 전체가 멈춘다.

이 비대칭이 지정학적 통제가 완제품 칩에 그치지 않고 3층(장비)까지 겨냥하는 이유를 설명한다. 완제품 칩 통제는 지금 팔리는 칩의 흐름을 막고, 장비 통제는 그런 칩을 스스로 만들 능력을 묶는다(5절 ③).

한국의 위치

- **강함:** 메모리(DRAM·NAND·HBM). SK하이닉스와 삼성전자가 세계 시장의 큰 몫을 차지한다. 예컨대 SK하이닉스는 2026년 1분기 매출 기준으로 HBM 세계 1위, DRAM(HBM 포함)과 NAND 세계 2위였다(IDC 집계. SK하이닉스가 2026년 7월 미국 상장 때 낸 투자설명서에 실린 수치, 2026-10-08 확인).
- **상대적으로 약함:** 첨단 파운드리 점유율, 팹리스 생태계, 장비·소재 국산화율.
- **구조적 함의:** 메모리는 전통적으로 사이클이 극심한 시장이다. AI 수요가 그 사이클을 구조적으로 바꿨는지, 아니면 진폭만 키운 것인지는 7회차 마지막 토론 주제다.

---

3. 수요의 구조 변화 — 학습에서 추론으로

이후 모든 회차의 문제의식이 이 절에서 나온다.

학습과 추론은 다른 기계를 원한다

항목	학습(TRAINING)	추론(INFERENCE)
연산 패턴	큰 배치의 행렬곱 (GEMM)	디코딩은 행렬-벡터곱에 가까움
연산강도	높음 (수백 FLOP/byte)	**낮음** (배치 크기 수준)
주된 병목	연산 성능, 노드 간 통신	**메모리 대역폭, 용량, 지연시간**
정밀도	BF16/FP8	INT8/FP8/INT4 (양자화 적극적)
메모리 요구	가중치+활성화+옵티마이저 상태	가중치 + **KV 캐시**
지연 민감도	낮음 (배치 처리)	높음 (사용자 대기)
비용 구조	일회성 대규모 투자	**지속적 운영비**

왜 이 전환이 메모리를 왕좌에 앉혔나

학습은 큰 배치를 쓰기 때문에 연산 유닛을 채울 수 있다. 그러나 **추론의 디코딩 단계는 토큰 하나를 만들기 위해 모델 가중치 전체를 메모리에서 읽어야 한다.** 가중치를 한 번 읽어서 하는 연산은 배치 안의 시퀀스 수만큼뿐이다.

거칠게 말해 fp16 가중치 1바이트당 연산량은 배치 크기 B 에 비례한다. 연산강도는 대략 $I \approx B$ 수준이다. 2회차에서 보겠지만 최신 GPU의 리지 포인트는 300 FLOP/byte 근처다. **배치가 300 미만이면 GPU는 대역폭에 묶여 연산 유닛이 논다.**

그런데 지연시간 요구 때문에 배치를 무한정 키울 수 없고, KV 캐시가 메모리를 잡아먹어 배치 크기가 메모리 용량에 의해 제한된다. 결국 추론 서비스의 경제성은 **"얼마나 빠른 칩이냐"가 아니라 "얼마나 넓고 큰 메모리를 붙였느냐"** 로 결정된다.

> **이것이 HBM이 산업의 중심이 된 이유이고(3회차), 첨단 패키징이 병목이 된 이유이며(6회차), 데이터센터 전력 문제가 터진 이유다(7회차).**

---

4. 시장 수치를 비판적으로 읽는 법

실제로 벌어지는 일

같은 해의 "AI 가속기 시장"을 두고 자료마다 수치가 크게 다르다. 이유는 다음과 같다.

1. **범위 차이** — 가속기 칩만인가, 모듈·보드까지인가, AI 서버 전체인가? 서버 전체로 잡으면 CPU·메모리·네트워크·전원이 다 들어가 몇 배가 된다.

   전력 수치에도 같은 함정이 있다. NVIDIA는 H100(SXM)의 최대 열설계전력(TDP)을 「최대 700W, 설정 가능」으로 공시하고, H100 8장을 담은 DGX H100 서버의 최대 입력 전력을 10.2kW로 적는다. 계산하면 칩 8장의 TDP 합은 5.6kW이고, 서버 최대 입력 전력과의 차이는 4.6kW다. CPU, 메모리, 네트워크, 팬, 전원 변환 손실 같은 항목이 이 차이에 든다. 다만 두 값이 모두 상한이라, 실제 작업 중에 칩 밖에서 4.6kW를 쓴다는 뜻은 아니다. 이 회차에서는 내역을 숫자로 나누지 않는다. 다음 세대 DGX B200의 시스템 최대 전력은 약 14.3kW다. 칩에서 서버로, 서버에서 냉각·배전을 포함한 시설로 경계를 넓힐 때마다 숫자가 커진다. 「AI 칩 하나가 쓰는 전력」이라는 말을 보면 어느 경계에서 잰 것인지부터 묻는다. 이 문제는 2회차 실습 4의 전력비 계산에서 다시 나온다. 제조 쪽도 구조가 같다. 웨이퍼 캐파, 양품 다이 수, 패키지 산출량, 출하량은 서로 다른 숫자다.

2. **집계 단위 차이** — 매출액인가, 출하 수량인가, 배포된 연산량(FLOPS)인가? 예를 들어 자체 칩(TPU 등)은 외부 판매가 없어 매출 기준 통계에서 과소평가되고, 배포 FLOPS 기준으로는 훨씬 크게 잡힌다.

3. **발표 주체의 이해관계** — 칩 회사, 시장조사기관, 투자은행은 각각 다른 유인을 가진다.

4. **시점 차이** — 반기마다 전망치가 크게 수정되는 시장이다.

확정할 수 있는 것과 없는 것

이 책은 다음과 같이 구분한다.

구조적으로 확정 가능 (여러 출처 일치)

- NVIDIA가 데이터센터 AI 가속기 매출의 압도적 다수를 점한다. 비율이 얼마이고 어느 쪽으로 움직이는지는 분모를 어떻게 잡느냐에 따라 달라지므로 여기에 넣지 않는다(실습 1).
- 하이퍼스케일러 자체 ASIC(Google TPU, AWS Trainium, Meta MTIA, Microsoft Maia)의 비중이 빠르게 늘고 있고, 그 설계 파트너로 Broadcom·Marvell 등이 부상했다.
- 추론이 AI 컴퓨트 지출에서 차지하는 비중이 학습을 넘어섰거나 넘어서는 중이다.
- HBM은 SK하이닉스·삼성·Micron 3사 과점이며, 매출 기준 선두는 SK하이닉스다(2026년 1분기 IDC 집계로 확인, 2026-10-08). 2·3위의 순서와 격차는 제품 세대가 바뀔 때마다 출렁이므로 여기서 단정하지 않는다.

확정 불가 (자료별 편차 큼 — 범위로만 언급)

- 연도별 시장 규모 절대 수치
- 업체별 점유율 정확한 퍼센트 (예: 같은 회사의 HBM 점유율도 분기 실적인지 연간 전망인지, 매출 기준인지 출하량 기준인지에 따라 수 %p씩 달라진다)
- 데이터센터 전력 소비 총량. 공신력 있는 추정도 폭이 넓다. 미국 로렌스버클리국립연구소(LBNL)는 2028년 미국 데이터센터 전력을 325~580TWh로 추정했다(6절 오해 6)

> **모임에서 읽는다면:** 발제자에게 "이 수치의 출처와 정의를 한 줄로 말해달라"고 요구하는 규칙을 만들면, 이 회차의 학습 효과가 크게 올라간다.

---

5. 공급망 리스크의 세 축

① 지리적 집중

첨단 로직 제조가 대만에 크게 편중되어 있다. 이는 자연재해·지정학 양쪽에서 단일 실패점(single point of failure)이다. 각국의 자국 내 팹 유치 정책(미국 CHIPS법, EU 칩스법, 일본·한국의 지원)은 이 집중을 완화하려는 시도지만, 팹 하나를 짓고 수율을 올리는 데 수년이 걸린다는 점에서 단기 해법이 아니다.

② 단일 공급자 병목

병목 지점	실질 공급자
EUV 노광장비	ASML (사실상 단독)
첨단 패키징(CoWoS 등)	TSMC 중심
EDA 툴	Synopsys, Cadence, Siemens EDA
CPU IP	Arm (라이선스 구조)
고순도 웨이퍼·포토레지스트	일본 업체 편중

여기서 중요한 점은 **"기술적으로 만들 수 있는가"와 "실제로 얼마나 만들 수 있는가"가 다르다는** 것이다. 예컨대 AI 가속기 공급량은 로직 웨이퍼가 아니라 CoWoS 패키징 캐파와 HBM 공급량에 먼저 걸린다. (6회차에서 자세히)

병목은 수요 쪽에서도 사슬을 이룬다. EUV → CoWoS → HBM은 공급 쪽에서 본 병목의 이동이다. 같은 이동을 수요 쪽에서 따라갈 수도 있다. 모델이 커지고 컨텍스트가 길어지면 연산량과 KV 캐시가 함께 늘고, 가중치와 KV 캐시를 담을 HBM의 용량·대역폭 요구가 커진다. HBM을 더 많이 붙이려면 패키지가 커지고 복잡해진다. 그런 가속기를 서버에 빽빽이 꽂으면 열밀도가 오르고, 끝내 랙과 시설의 전력·냉각이 한계에 닿는다. 한 고리의 요구가 늘면 압력이 다음 고리로 넘어간다. 그래서 시스템을 묶는 것은 대개 한 층 안의 성능이 아니라 **층과 층이 만나는 접점이다.** 메모리와 칩, 칩과 패키지, 서버와 시설 사이가 그렇다. 3회차(HBM), 6회차(패키징), 7회차(서버·시설)가 이 고리를 하나씩 맡는다.

③ 정책 리스크

수출통제는 완제품 칩만이 아니라 **장비와 제조 능력까지** 겨냥한다. 역외적용(해외직접생산품규칙 등) 때문에 제3국 기업도 영향을 받는다. 2026년 10월 현재 미국 수출관리규정(EAR)의 해외직접생산품(FDP) 규칙 조문(15 CFR §734.9)에는 첨단 컴퓨팅 FDP 규칙((h)항. 첨단 집적회로와 컴퓨터)과 반도체 제조장비(SME) FDP 규칙((k)항)이 따로 들어 있다. 미국의 특정 기술·소프트웨어로 만든 품목이거나 그런 기술로 지은 공장에서 만든 품목이면, 미국 밖에서 생산했어도 EAR의 적용을 받을 수 있다는 구조다. 조문은 8절 「후속 공부」의 eCFR 링크에서 볼 수 있다.

"슈퍼사이클의 역설"

호황일수록 모두가 병목 공정에 증설 투자를 한다. 팹·장비의 리드타임이 2~3년이라, 투자 결정 시점의 수요 신호와 캐파가 완성되는 시점의 수요가 어긋난다. 이것이 반도체 산업의 고전적 채찍효과(bullwhip effect)이며, 메모리에서 특히 심하다. **AI 수요가 이 사이클을 없앴다는 주장과, 진폭만 키웠을 뿐이라는 주장이 대립한다.** (7회차 마지막 토론)

---

6. 자주 하는 오해

오해 1: "AI 칩 = GPU" → GPU가 현재 지배적이지만 TPU·Trainium 같은 ASIC의 비중이 빠르게 늘고 있다. 더 중요한 것은, AI 시스템의 성능과 원가에서 메모리(HBM)와 패키징이 차지하는 비중이 크다는 점이다.

오해 2: "나노미터 숫자가 작을수록 무조건 좋은 칩" → 노드 이름은 물리 치수가 아니라 마케팅 명칭이다(4회차). 또한 AI 가속기 성능은 노드보다 메모리 대역폭과 패키징에 더 크게 좌우된다.

오해 3: "돈만 있으면 팹을 지으면 된다" → 장비 리드타임, 숙련 인력, 수율 램프업(양산 수율에 도달하기까지의 학습 곡선)이 자본보다 더 큰 제약이다. 팹 건물은 2년이면 짓지만 수율을 경쟁력 있는 수준으로 올리는 데는 그보다 오래 걸린다.

오해 4: "중국이 곧 따라잡는다 / 절대 못 따라잡는다" → 둘 다 단순화다. 성숙 노드(레거시)에서는 이미 상당한 캐파를 확보했고, 첨단 로직에서는 EUV 부재라는 구조적 제약이 있다. 층별로 나눠서 봐야 한다.

오해 5: "AI 반도체 호황은 곧 끝난다 / 영원하다" → 정직한 답은 "수요의 성격이 학습에서 추론으로 바뀌었으므로, 과거 사이클 모형이 그대로 적용될지가 불확실하다"이다. 추론 수요는 학습보다 지속적·운영비적 성격이 강해 사이클을 완화할 수 있지만, 증설 리드타임이라는 구조는 그대로다.

오해 6: "전력 효율이 좋은 칩이 나오면 데이터센터 총전력은 줄어든다" → 칩 한 장의 효율과 시설 전체의 총량은 다른 질문이다. LBNL이 미국 에너지부 지원으로 내놓은 2024년 미국 데이터센터 전력 보고서에 따르면, 미국 데이터센터 전력 사용량은 2014년 58TWh에서 2023년 176TWh로 늘었다. 2023년 미국 전체 전력의 약 4.4%다. 2028년에는 325~580TWh, 미국 전력의 6.7~12%로 추정된다(나머지 경제가 얼마나 성장하느냐에 따라 비율이 달라진다). 2017~2023년 사이에만 두 배 넘게 늘었고, 보고서는 그 주된 원인으로 AI 서버의 증가를 꼽는다. 이 기간에도 칩은 세대마다 전력 효율이 좋아졌다. 그런데도 총량은 세 배로 늘었으니, 효율과 총량은 따로 움직일 수 있다(이 해석은 이 책의 것이고 보고서의 결론으로 옮긴 것이 아니다). 2028년 추정의 폭이 두 배 가까이 벌어져 있다는 점도 눈여겨볼 만하다. 4절이 총량을

「확정 불가」로 분류하고 범위로만 쓰는 이유가 이것이다.

---

7. 토론 질문

1. 밸류체인 3층 중 어디에 투자·정책 자원을 넣는 것이 한 국가 입장에서 가장 효율적인가? 그 판단은 시간 지평(5년 vs 20년)에 따라 어떻게 달라지는가?
2. 하이퍼스케일러가 자체 칩을 만드는 것은 합리적인가? 어떤 조건에서 합리적이고 어떤 조건에서 아닌가? (4회차에서 정량적으로 다시 다룬다)
3. "AI 반도체 시장 규모" 추정치들이 크게 갈리는 것은 측정의 문제인가, 정의의 문제인가? 통계학을 공부한 사람으로서 이 추정치들을 어떻게 다뤄야 하는가?
4. 공급망 병목이 EUV(장비)에서 CoWoS(패키징)로, 다시 HBM(메모리)으로 옮겨가는 현상을 어떻게 이해해야 하는가? 병목은 왜 계속 이동하는가?
5. 메모리 산업의 전통적 사이클(호황→증설→과잉→침체)이 AI 수요로 인해 구조적으로 바뀌었다고 볼 근거가 있는가?

---

8. 후속 공부

- **산업 구조 개괄:** Chris Miller, *Chip War* (2022) — 반도체 산업의 지정학을 다룬 표준 대중서. 이 회차의 배경 독서로 알맞다
- **기술 동향 정기 추적:** SemiAnalysis, TrendForce, Counterpoint Research 등의 산업 분석. 단, 각각의 편향을 감안할 것
- **1차 자료:** TSMC·SK하이닉스·Micron·NVIDIA의 분기 실적 발표 자료(특히 컨퍼런스콜 Q&A). 시장 보고서보다 정확한 경우가 많다
- **한국 관점:** KDI 경제교육·정보센터, 산업연구원(KIET), ETRI 전자통신동향분석(ETRI Trends)의 반도체 산업 분석
- **전력 수요:** LBNL, 2024 미국 데이터센터 전력 보고서 발표 — 6절 오해 6의 수치 출처
- **수출통제 조문:** eCFR 15 CFR §734.9 — 해외직접생산품(FDP) 규칙의 현행 조문

자습 부록 A — 수치 감사 실습과 풀이

이 회차에는 계산 실습이 없다. 대신 4절에서 익힌 "정의를 먼저 묻는 습관"을 실제로 써 보는 세 문제를 둔다. **답을 보기 전에 종이에 먼저 써 볼 것.**

실습 1 — 같은 회사, 다른 숫자

어떤 기사가 "A사의 AI 가속기 시장 점유율이 90%"라고 하고, 다른 기사가 "70%"라고 한다. 두 기사 모두 거짓말을 하지 않았다고 가정할 때, 어떤 정의 차이가 이 간극을 만들 수 있는가? 최소 세 가지를 제시하라.

▾ 풀이

세 가지 이상의 정당한 정의 차이가 있다.

1. **분모에 자체 설계 칩을 넣는가.** Google TPU·AWS Trainium은 외부에 팔지 않으므로 **매출 기준 시장**에는 거의 잡히지 않는다. 자체 칩을 분모에서 빼면 A사 점유율은 올라가고, 배포된 물량·연산량 기준으로 넣으면 내려간다. 이 하나만으로 20%p 차이가 난다.
2. **칩만인가, 시스템까지인가.** 가속기 다이 기준과 서버·랙 기준의 점유율은 다르다. 서버 전체에는 CPU·네트워크·전원이 들어가고 이 항목들의 공급자 구성이 다르다.
3. **학습용만인가, 추론까지인가.** 추론 시장에는 CPU 추론·저가 가속기·엣지 NPU가 섞여 들어와 파편화되어 있다. 학습 시장만 보면 집중도가 훨씬 높다.
4. (추가) **시점.** 분기 출하 기준인가 누적 설치 기준인가. 급성장하는 시장에서는 이 둘이 크게 벌어진다.

얻어 갈 것: 점유율은 분자가 아니라 **분모가 정한다.** 기사에서 분모를 밝히지 않으면 그 숫자는 검증 불가능한 진술이다.

실습 2 — 성장률의 함정

같은 해에 한 기사는 "AI 반도체 시장이 전년 대비 200% 성장했다"고 쓰고, 다른 기사는 "반도체 시장이 전년 대비 25% 성장했다"고 썼다. 두 기사 모두 반도체 산업을 다룬다. 모순인가? 답하기 전에 두 문장이 각각 무엇을 센 숫자인지부터 적어 볼 것.

▾ 풀이

모순이 아니다. **부분과 전체를 섞은 것**이다.

간단한 수치 예로 확인해 보자. 전체 반도체 시장 100 중 AI 관련이 5, 나머지가 95라고 하자. AI가 15로 세 배(+200%)가 되고 나머지가 95 → 110으로 약 16% 성장하면, 전체는 100 → 125로 **25% 성장**이다.

두 문장 모두 참이다. 서로 다른 층위를 말하고 있을 뿐이다.

얻어 갈 것: 성장률은 **기저(base)** 를 밝히지 않으면 의미가 없다. 작은 기저에서 출발한 항목의 성장률은 언제나 극적으로 보인다. 4절에서 "구조적으로 확정 가능"한 것만 본문에 쓰고 절대 수치를 유보한 이유가 이것이다.

실습 3 — 병목의 위치 추론

다음 세 가지 관찰이 동시에 보고되었다고 하자.

- 첨단 로직 웨이퍼 캐파는 여유가 있다.
- 가속기 리드타임은 40주가 넘는다.
- HBM은 이듬해 물량까지 선계약이 끝났다.

병목은 어디인가? 그리고 "웨이퍼 캐파를 늘리자"는 정책 제안은 왜 효과가 없는가?

▾ 풀이

병목은 **웨이퍼 다음 단계**에 있다. 로직 웨이퍼가 남는데 완제품이 안 나온다는 것은 웨이퍼 이후에서 막혀 있다는 뜻이다. 세 관찰이 직접 가리키는 후보는 HBM 공급이다. HBM과 로직 다이를 한 패키지로 묶는 첨단 패키징(CoWoS 등)도 후보지만, 이 세 관찰만으로는 패키징까지 막혔는지 가릴 수 없다. HBM만 모자라도 세 관찰이 똑같이 나오기 때문이다. 패키징 쪽은 CoWoS 캐파의 가동률이나 패키징 대기 물량을 따로 확인해야 한다.

직렬 공정에서 산출량은 **가장 좁은 목**이 결정한다. 여유 있는 단계(웨이퍼)를 더 넓혀도 총산출은 1도 늘지 않고, 재고만 그 앞에 쌓인다. 늘려야 하는 것은 HBM 라인이고, 패키징 대기열이 확인되면 CoWoS 캐파도 함께다.

여기서 한 걸음 더: 그런데 HBM 라인을 늘리면 다음에는 무엇이 막히는가? HBM은 DRAM 다이를 쓰므로 일반 DRAM 캐파와 경쟁하고, TSV 공정과 열압착·MR-MUF 적층 장비가 필요하며(다음 세대에는 하이브리드 본딩 장비까지), 그 장비는 다시 소수 공급자에 묶여 있다. **병목은 제거되는 것이 아니라 이동한다**(5절). 이 구조가 6회차와 7회차의 주제다.

얻어 갈 것: 산업 기사를 읽을 때 "무엇이 남고 무엇이 모자라는가"를 나란히 놓으면 병목의 위치가 드러난다. 한쪽만 보면 언제나 틀린 처방이 나온다.

자습 부록 B — 토론 질문 생각의 실마리

7절의 토론 질문은 정답이 있는 문제가 아니다. 다만 혼자 읽으면 "무엇을 근거로 답해야 하는지"조차 막막할 수 있으므로, **논점의 축**만 제시한다. 결론은 각자 내리는 것이다.

질문	갈리는 지점	어느 쪽으로도 답할 수 있게 하는 근거
1. 3층 중 어디에 투자?	**시간 지평**	5년이면 설계·인력(빠른 회수), 20년이면 장비·소재(대체 불가성이 복리로 쌓임). 한국의 현 위치(메모리 강·장비 약)를 넣으면 답이 또 달라진다
2. 자체 칩은 합리적인가	**물량 × 워크로드 통제력**	4회차 손익분기식 N^* 로 정량화된다. 핵심은 손익분기 물량이 의외로 작다는 것(약 3천~5천 대)이고, 진짜 제약은 개발 2~3년 동안 그 워크로드가 살아남느냐다
3. 측정의 문제인가 정의의 문제인가	**오차의 성질**	측정 오차라면 표본을 늘리면 수렴한다. 정의 차이라면 아무리 정밀하게 재도 수렴하지 않는다. 4절 사례는 후자다 → 통계적으로는 **추정량의 분산이 아니라 추정 대상(estimand)이 다른** 경우
4. 병목은 왜 이동하나	**투자의 시차**	병목이 드러나면 그곳에 자본이 몰리고, 리드타임 2~3년 뒤 그곳이 풀리면 그다음으로 좁은 곳이 드러난다. 이동은 실패가 아니라 정상 작동의 증거다. 수요 쪽에서 보면 모델·컨텍스트 증가가 KV 캐시 → HBM → 패키지 → 서버 열밀도 → 시설 전력으로 압력을 넘긴다(5절 ②). 두 방향을 겹치면, 압력이 도착한 고리 가운데 증설이 가장 느린 곳이 다음 병목이 된다
5. 메모리 사이클은 바뀌었나	**수요의 성질**	학습 수요는 투자성(일회성 자본지출), 추론 수요는 운영성(지속 과금)이다. 후자 비중이 커지면 수요가 평탄해진다는 것이 "바뀌었다" 쪽 논거. 반대쪽 논거는 증설 리드타임 구조가 그대로라는 것 — **수요가 평탄해도 공급이 계단식이면 사이클은 남는다**

자습 부록 C — 핵심 용어

이 회차에서 처음 나온 용어만 모았다. 2회차 이후로 계속 쓰인다.

용어	영문	한 줄 정의
팹리스	fabless	설계만 하고 생산은 위탁하는 반도체 회사
파운드리	foundry	타사 설계를 위탁 생산하는 공장
디자인하우스	design house (design service company)	팹리스의 설계를 특정 파운드리 공정에 맞춰 물리 구현·검증해 제조로 넘기는 설계 서비스 회사
IDM	integrated device manufacturer	설계·제조·판매를 모두 하는 종합 반도체 회사
OSAT	outsourced semiconductor assembly and test	후공정(패키징·테스트)을 전문으로 하는 외주 업체
EDA	electronic design automation	칩 설계용 소프트웨어 도구
HBM	high bandwidth memory	DRAM을 수직 적층해 대역폭을 극대화한 메모리
가속기	accelerator	특정 연산에 특화된 칩(GPU·TPU·NPU 등)의 총칭

용어	영문	한 줄 정의
ASIC	application-specific integrated circuit	용도를 고정해 만든 맞춤형 칩
학습 / 추론	training / inference	모델을 만드는 과정 / 만들어진 모델을 쓰는 과정
연산강도	arithmetic intensity	메모리에서 옮긴 1바이트당 수행하는 연산 수 [FLOP/byte]
KV 캐시	key-value cache	이미 만든 토큰의 중간 결과를 저장해 재계산을 피하는 메모리
양자화	quantization	가중치를 더 적은 비트로 표현해 용량·대역폭을 아끼는 기법
노드	(process) node	"3nm" 같은 공정 세대 명칭. 물리 치수가 아니라 마케팅 명칭
수율	yield	만든 것 중 정상 동작하는 비율
CoWoS	chip-on-wafer-on-substrate	TSMC의 대표적 2.5D 첨단 패키징 방식
하이퍼스케일러	hyperscaler	초대형 클라우드 사업자(Google·AWS·Microsoft·Meta 등)
채찍효과	bullwhip effect	수요 변동이 공급망 상류로 갈수록 증폭되는 현상
단일 실패점	single point of failure	그곳 하나가 멈추면 전체가 멈추는 지점

자습 부록 D — 자가진단

각 항목에 **자기 말로** 답할 수 있으면 다음 회차로 넘어가도 좋다. 두 개 이상 막히면 위 「읽는 순서」의 해당 절로 돌아간다.

☐ "AI 반도체 시장 규모"라는 기사를 봤을 때 확인할 것 세 가지를 즉시 댈 수 있다 → 1·4절

☐ 밸류체인 3층을 그리고, 각 층에서 이익률과 대체 불가능성이 왜 어긋나는지 설명할 수 있다 → 2절

☐ 수출통제가 완제품 칩과 함께 제조장비까지 겨냥하는 이유, 그리고 장비 통제가 칩 통제와 달리 무엇을 묶는지를 한 문장으로 말할 수 있다 → 2·5절

☐ 학습과 추론이 요구하는 하드웨어가 어떻게 다른지 표를 안 보고 네 가지 이상 댈 수 있다 → 3절

☐ **"토큰 하나를 만들려고 가중치 전체를 읽는다"가 왜 메모리를 병목으로 만드는지** 설명할 수 있다 → 3절 ★

☐ 배치 크기를 키우면 왜 효율이 오르고, 왜 무한정 키울 수 없는지 말할 수 있다 → 3절 ★

☐ 공급망 리스크 세 축을 구분하고 각각의 대응 수단이 왜 다른지 설명할 수 있다 → 5절

☐ 병목이 EUV → CoWoS → HBM으로 이동한 현상을 "정상 작동"으로 설명할 수 있다 → 5절

★ 표시 두 개는 **반드시** 통과해야 한다. 2회차 전체가 이 두 문장의 정밀화다.

다음 회차로 넘어가기 전에

이 회차에서 이름만 나오고 정의되지 않은 채 남은 것이 셋 있다. 2회차가 이것들을 정식으로 세운다.

남겨 둔 것	2회차에서
연산강도 I	루프라인 모델의 가로축으로 정식 정의
"리지 포인트 300 FLOP/byte"	$I^* = P_{\text{peak}} / B$ 로 유도하고 실제 칩으로 확인
"배치 300 미만이면 대역폭에 묶인다"	디코딩 상한식으로 정량화

한 줄로 이어지는 질문: 칩이 아무리 빨라져도 메모리가 못 따라오면 소용없다면, **그 한계는 정확히 어디서 갈리는가?**

출처

스터디 커리큘럼이 고른 입문 자료 ("AI 반도체" 항목)

커리큘럼 소주제	링크	이 책에서의 활용
2024 기술 동향	KDB미래전략연구소, 「AI 반도체 기술 및 산업 동향」(2024-08, KDI 경제교육·정보센터 국내연구자료)	국책은행 연구소 시각에서 본 AI 반도체 기술 동향. 2절 밸류체인·한국 위치 부분의 배경
2026 산업 전망	삼일PwC, 「글로벌 반도체 산업 전망 2026 — Semiconductor and beyond」(PDF)	회계법인 관점의 산업 전망. 4절 시장 수치 읽기의 사례 자료로 활용 — 이 보고서의 정의와 집계기준을 직접 확인해 보기를 권함
공급망 리스크	PwC컨설팅, 「반도체 슈퍼 사이클의 역설」	5절 "슈퍼사이클의 역설" 논의의 직접 출처. 호황기 증설이 다음 침체를 만드는 구조

2.0판에서 더한 1차 출처 (2026-10-08 확인)

- TSMC, Value Chain Alliance — 2절 디자인하우스 설명
- SK하이닉스, 미국 상장 투자설명서(Form 424B4, 2026) — SEC EDGAR 원문. 2026년 1분기 매출 기준 HBM 1위(점유율 56.4%), DRAM(HBM 포함) 2위(29.1%), NAND 2위(18.5%). 모두 IDC 집계를 인용한 것이다(2026-10-08 원문 대조)
- NVIDIA H100 제품 사양 · NVIDIA DGX H100/H200 사용자 안내서, Table 3 전원 사양 · NVIDIA DGX B200 사양 — 4절 측정 경계 사례
- LBNL, 「Berkeley Lab Report Evaluates Increase in Electricity Demand from Data Centers」(2025-01-15) — 6절 오해 6
- eCFR, 15 CFR §734.9 Foreign-Direct Product (FDP) Rules — 5절 ③

검증 상태

- 밸류체인 3층 구조와 각 층의 주요 기업 — 업계 표준 분류. 검증 완료
- 디자인하우스와 TSMC VCA의 성격 — TSMC 공식 페이지로 확인(2026-10-08). 옛 이름 「Value Chain Aggregator」는 Internet Archive 2019-07-30·2020-11-16 사본의 페이지 제목으로, 「Value Chain Alliance」는 2022-09-25 사본부터 확인(2026-10-22). 회원 명단은 시점에 따라 바뀌므로 본문에 쓰지 않음
- EUV 노광장비 ASML 단독 공급 — 검증 완료 (복수 출처 일치)
- HBM 3사 과점 구조와 선두(SK하이닉스) — 2026년 1분기 매출 기준 IDC 집계로 재확인(2026-10-08, SK하이닉스 투자설명서). 2·3위 순서는 1차 출처로 확인하지 않았으므로 본문에 쓰지 않는다. **정확한 점유율 수치는 자료별 편차가 커 본문에서 단정하지 않음**
- H100 SXM TDP 「Up to 700W (configurable)」, DGX H100 최대 입력 전력 10.2kW, DGX B200 시스템 최대 전력 약 14.3kW — 제조사 사양 원문 확인(2026-10-08). 5.6kW·4.6kW는 계산값
- 미국 데이터센터 전력 58TWh(2014)·176TWh(2023, 약 4.4%)·325~580TWh(2028 추정, 6.7~12%), 2017~2023년 두 배 이상 증가와 주원인(AI 서버) — LBNL 보도자료 원문 확인(2026-10-08)
- 15 CFR §734.9의 첨단 컴퓨팅 FDP 규칙·SME FDP 규칙 — eCFR 현행 조문 확인(2026-10-05 기준 갱신본, 2026-10-08 열람). (h)항의 품목 범위가 ECCN 3A090·4A090 등 첨단 집적회로·컴퓨터이고 (k)항이 3B001 등 제조장비라는 점은 eCFR API의 2026-10-01 판 조문으로 다시 대조했다(2026-10-08)
- 추론이 AI 컴퓨트 지출의 다수를 차지하는 전환 — 복수 산업분석에서 일치하나 정확한 시점·비율은 자료별 상이. 구조적 추세로만 서술
- AI 가속기 시장 규모 절대 수치 — **미확인·편차 큼. 본문에서 특정 수치 제시하지 않음**
- 업체별 정확한 점유율 퍼센트 — **미확인·편차 큼. 구조적 순위만 서술**

2회차

AI 칩 — 연산 반도체

핵심 TAKEAWAY

가속기 설계의 제1원리는 "연산을 빠르게"가 아니라 "데이터 이동을 줄이기"다. 곱셈 한 번의 에너지보다 DRAM에서 데이터를 한 번 가져오는 에너지가 두 자릿수 배 크기 때문이다. CPU·GPU·NPU·TPU·ASIC의 차이는 ① **명령어당 오버헤드를 얼마나 상각하는가**, ② **가져온 데이터를 몇 번 재사용하는가** 라는 두 축으로 통일적으로 설명된다. 이 회차에서 세우는 **루프라인 모델**은 이후 모든 회차에서 반복 사용하는 공용 도구다. 최신 가속기의 리지 포인트가 300 FLOP/byte 근처인 반면 LLM 디코딩의 연산강도는 그보다 한두 자릿수 낮다는 사실이, 이 책 전체의 문제의식이다.

원자료 바로가기

엘리먼트코리아 「AI칩이란 무엇인가」 · MSAP 「AI 가속기란 무엇인가」 · ScienceON JAKO202412457639314 「국내외 인공지능 반도체에 대한 연구 동향」 · ScienceON JAKO202504350488400 「AI 반도체와 주요 응용 분야에 관한 연구」 · InitNode 「CPU vs GPU vs TPU vs NPU」 · Williams 외, Roofline (CACM 2009) · Jouppi 외, TPU 논문 arXiv:1704.04760 · NVIDIA H100 제품 사양 · NVIDIA Blackwell 아키텍처 기술 개요

학습목표

1. CPU가 AI 연산에 비효율적인 이유를 "클럭이 느려서"가 아닌 에너지 관점에서 설명할 수 있다.
2. GPU 텐서 코어와 TPU 시스톨릭 어레이가 각각 어떤 방식으로 효율을 얻는지 구분할 수 있다.
3. 루프라인 모델을 세우고, 주어진 워크로드가 연산 바운드인지 대역폭 바운드인지 계산으로 판별할 수 있다.
4. 자체 ASIC 개발의 손익분기를 정량적으로 따질 수 있다.

0. 자습 안내 — 혼자 읽는 사람에게

이 회차는 책 전체의 도구를 만드는 회차다

여기서 세우는 루프라인 부등식 하나가 3회차·4회차·6회차·7회차에서 계속 재사용된다. 다른 회차는 한 번 읽고 넘어가도 되지만 **이 회차만은 손으로 계산해 보지 않으면 남지 않는다.** 계산기만 있으면 되고, 실습 1은 5분이면 끝난다.

직전 회차에서 가져오는 것

1회차에서	이 회차에서
"토큰 하나를 만들려고 가중치 전체를 읽는다"	이 문장이 fp16 가중치에서 $I \approx B_{sz}$ 라는 근사식이 된다
"배치 300 미만이면 대역폭에 묶인다"	그 300이 $I^* = P_{peak} / B$ 에서 나오는 것을 확인한다
"이익이 큰 층과 대체 불가능한 층이 일치하지 않는다"	여기서는 배치가 작은 디코딩에서 **성능의 병목**이 연산이 아니라 대역폭에 있음을 본다

선행 지식 점검

개념	한 줄 답	몰라도 되는가
행렬곱 GEMM	행렬 두 개를 곱하는 연산. $n \times n$ 행렬곱은 약 $2n^3$ FLOP	❌ 필수
로그-로그 그래프	양 축에 로그를 취한 그래프. 거듭제곱 관계가 직선으로 보인다	6절 그림을 읽는 데 필요
fp16 / INT8 / INT4	숫자 하나를 각각 2바이트 / 1바이트 / 0.5바이트로 표현	❌ 필수
시스톨릭 어레이	데이터를 격자 사이로 흘려보내며 곱셈-누산을 반복하는 구조	✅ 개념만
CUDA	NVIDIA GPU용 프로그래밍 환경과 그 생태계	✅ 이름만

읽는 순서 (총 120분 권장)

순서	절	시간	비고
1	1~5절 (칩 구조)	35분	빠르게. 지금은 배경이다
2	**6절 루프라인**	**35분**	**정독.** 식을 종이에 직접 옮겨 적을 것
3	자습 부록 실습 1·2	20분	6절을 읽자마자 바로. 미루면 효과가 반감된다
4	7절 소프트웨어 락인	10분	
5	실습 3·4 + 자가진단	20분	

6절이 어려우면 — 세 문장으로 압축한 우회 경로

루프라인은 이름이 거창하지만 내용은 초등학교 산수다.

1. 일을 하려면 **재료를 옮겨 오고**(대역폭 B), **처리해야**(연산 P_{peak}) 한다.
2. 둘 중 **느린 쪽이 전체 속도를 정한다** — 그래서 $\min$ 이다.
3. 어느 쪽이 느린지는 "재료 1kg당 얼마나 오래 처리하는가"(I)로 갈린다. 처리가 오래 걸리는 일이면 연산이, 금방 끝나는 일이면 운반이 병목이다.

$$P_{\text{attainable}}(I) = \min\left(P_{\text{peak}},\ B \times I\right)$$

이 식이 말하는 전부다. 나머지는 실제 숫자를 넣어 보는 일뿐이다.

---

1. 직관 잡기 — 계산은 싸고, 데이터 배달은 비싸다

물류 비유

공장에서 부품 하나를 가공하는 데 1초가 걸린다고 하자. 그런데 그 부품을 창고에서 가져오는 데 3분이 걸린다면, 공장 라인을 아무리 늘려도 생산량은 늘지 않는다. 개선 방법은 두 가지뿐이다.

- **창고를 가까이 옮긴다** (= 온칩 SRAM, HBM을 칩 옆에 붙이기)
- **한 번 가져온 부품으로 여러 작업을 한다** (= 데이터 재사용, 타일링, 배치 키우기)

AI 가속기 설계의 거의 모든 아이디어가 이 둘 중 하나다. 나머지는 전부 세부사항이다.

숫자로 보는 충격

Horowitz의 고전적 정리(ISSCC 2014, 45nm 공정 기준)에 따르면 연산과 메모리 접근의 에너지 차이는 다음과 같다.

연산	대략적 에너지
8비트 정수 덧셈	0.03 pJ
32비트 부동소수 덧셈	0.9 pJ
32비트 부동소수 곱셈	3.7 pJ
32비트 온칩 SRAM 읽기 (8 KB 캐시)	5 pJ
32비트 **DRAM 읽기**	**640 pJ**

위 세 줄(덧셈·곱셈)은 Horowitz 발표문의 그림 1.1.9에서 그대로 옮겼다. 메모리 두 줄은 같은 45nm 표를 인용한 Han 외(2015)의 32비트 접근 값이다. Horowitz의 그림은 메모리를 64비트 접근으로 적는다. 캐시는 8 KB 10 pJ, 32 KB 20 pJ, 1 MB 100 pJ이고 DRAM은 1.3~2.6 nJ이다. SRAM 5 pJ은 8 KB 캐시 값의 절반에 해당하므로, 캐시가 커지면 이 값도 커진다.

즉 **DRAM에서 데이터를 한 번 읽는 비용이 부동소수 곱셈 한 번의 약 170배**다. Horowitz의 DRAM 범위를 32비트로 반 나누어 세면 약 180~350배다. 공정이 미세화되면서 연산 에너지는 계속 줄었지만 오프칩 메모리 접근 에너지는 훨씬 덜 줄었기 때문에, 이 격차는 시간이 갈수록 벌어졌다.

> **여기서 따라 나오는 결론:** 칩 설계자가 최적화해야 할 목적함수는 FLOPS가 아니라 **"유효 연산당 총 에너지"** 이고, 그 지배항은 데이터 이동이다.

---

2. 왜 CPU로는 안 되는가

오해부터 걷어내기

"CPU는 코어가 적어서 느리다"는 설명은 절반만 맞다. 더 근본적인 이유는 **CPU가 지불하는 명령어당 고정 비용**이다.

CPU가 `c = a * b` 한 줄을 실행할 때 실제로 일어나는 일:

1. 명령어 인출 (I-cache 접근)
2. 명령어 디코드
3. 분기 예측·레지스터 리네이밍·비순차 실행 스케줄링
4. 레지스터 파일 읽기
5. **실제 곱셈** ← 이것만이 "일"이다
6. 결과 쓰기, 커밋

1~4와 6이 5보다 훨씬 많은 에너지와 면적을 쓴다. CPU는 **"다음에 무슨 일이 올지 모른다"는 가정** 위에 설계되었고, 그 불확실성에 대비하는 비용이 곧 오버헤드다.

그런데 딥러닝은 다음에 무슨 일이 올지 안다

행렬곱은 데이터 의존적 분기가 없고, 접근 패턴이 정적으로 결정되며, 같은 연산이 수십억 번 반복된다. **불확실성이 없으므로 그 대비 비용을 전부 없앨 수 있다.** 이것이 가속기가 존재하는 이유의 전부다.

오버헤드 항목	CPU	GPU	TPU/ASIC
명령어 인출·디코드	스칼라는 연산마다, SIMD·AMX는 벡터·타일 단위로 상각	워프 단위로 상각	거의 없음 (고정 데이터흐름)
분기 예측	있음 (비쌈)	최소	없음
비순차 실행	있음	없음	없음
캐시 계층	크고 복잡	작고 명시적 제어 가능	스크래치패드(소프트웨어 관리)
제어 로직 면적 비중	매우 큼	중간	매우 작음

위 1~6은 스칼라 명령어 하나를 기준으로 그린 그림이다. 실제 CPU도 명령어 하나에 여러 연산을 싣는다. SIMD 명령어는 벡터 하나를 통째로 처리하고, Intel AMX는 아예 행렬 단위로 간다. Intel AMX 코드 예제 문서에 따르면 AMX는 2차원 레지스터인 타일과 타일끼리 행렬곱을 하는 가속기 TMUL로 이루어지고, 팔레트 1 설정에서는 타일 레지스터 8개가 합쳐서 8 KB이고 타일 하나는 최대 16행 × 64바이트다. 같은 문서는 4세대 Xeon 스케일러블 프로세서와 Xeon Max 시리즈를 사용 대상으로 든다. 그러니 CPU가 연산을 하나씩 순차 처리한다는 설명은 입문용 축약이다. 다만 결론은 바뀌지 않는다. CPU는 범용성을 위한 제어 장치를 그대로 지고 있으므로, 같은 일을 가속기만큼 깊게 상각하지는 못한다.

정리: CPU→GPU→ASIC으로 갈수록 "유연성을 팔아 효율을 산다". 이 거래의 조건이 5절 ASIC 경제학이다.

---

3. GPU의 실체 — `matmul` 한 줄이 하드웨어에서 하는 일

SIMT 실행 모델

GPU는 스레드를 32개 묶음(워프, warp)으로 실행하며 **한 워프의 32스레드가 같은 명령어를 동시에 수행**한다. 명령어 인출·디코드 비용이 32개 연산에 분산되므로 오버헤드가 1/32이 된다.

GPU 코드를 짜 본 독자라면 익숙한 성능 함정이 여기서 나온다.

- **워프 다이버전스:** 같은 워프 안에서 스레드가 서로 다른 분기를 타면 두 경로를 차례로 실행하고, 그동안 그 경로에 해당하지 않는 스레드는 쉰다. 길이가 같은 두 경로로 스레드가 반씩 갈리면 처리량이 절반이 되고, 경로 길이와 갈리는 비율에 따라 손실의 크기가 달라진다. → "GPU에서 `if`를 피하라"는 조언의 실체.
- **메모리 코얼레싱:** 전역 메모리는 32바이트 단위 트랜잭션으로 읽힌다. 워프의 32스레드가 연속된 4바이트 값을 하나씩 읽으면 모두 128바이트라 트랜잭션 네 번으로 끝나지만, 주소가 32바이트 이상씩 흩어지면 스레드마다 한 번씩, 32번이 든다. → 텐서 레이아웃(NCHW vs NHWC)이 성능을 좌우하는 이유.
- **점유율(occupancy):** 메모리 대기 시간을 다른 워프 실행으로 숨기는 구조라, 동시에 상주할 수 있는 워프 수가 중요하다. GPU에도 L1·L2 캐시와 공유 메모리가 있지만, CPU보다 작은 캐시로 지연을 다 줄이지는 못하므로 **지연의 상당 부분을 병렬성으로 감춘다.**

텐서 코어 — 상각의 극단

텐서 코어는 작은 행렬 블록의 곱셈-누산 $D = A \times B + C$ 를 한 명령어로 수행한다. 예컨대 16×16 블록곱이라면 명령어 하나가 수천 번의 MAC을 담당하므로 명령어 오버헤드가 사실상 0에 수렴한다.

계층 구조로 보는 `torch.matmul` 한 줄:

```
torch.matmul(A, B)
  └ cuBLAS/cuBLASLt 커널 선택 (타일 크기·스플릿 전략)
      └ 그리드/블록 분해: 출력 행렬을 타일로 쪼갬
          └ 글로벌 메모리(HBM) → 공유 메모리(SRAM) 타일 로드
              └ 공유 메모리 → 레지스터 프래그먼트
                  └ 텐서 코어 MMA 명령 (블록 단위 곱셈누산)
                      └ 누산 결과를 다시 위로
```

핵심: 이 계층 전체가 "HBM에서 한 번 가져온 타일을 최대한 많이 재사용한다"는 하나의 목적을 위해 존재한다. 타일 크기 T 로 블록곱을 하면 $O(T^2)$ 바이트를 읽어 $O(T^3)$ 연산을 하므로, 연산강도가 $O(T)$ 로 커진다. **타일링은 곧 루프라인의 I 를 키우는 소프트웨어적 방법이다.**

정밀도의 계단

AI 가속기는 정밀도를 낮출수록 처리량이 대략 배수로 늘어난다(FP32 → TF32 → BF16/FP16 → FP8 → FP4). 이유는 두 가지다.

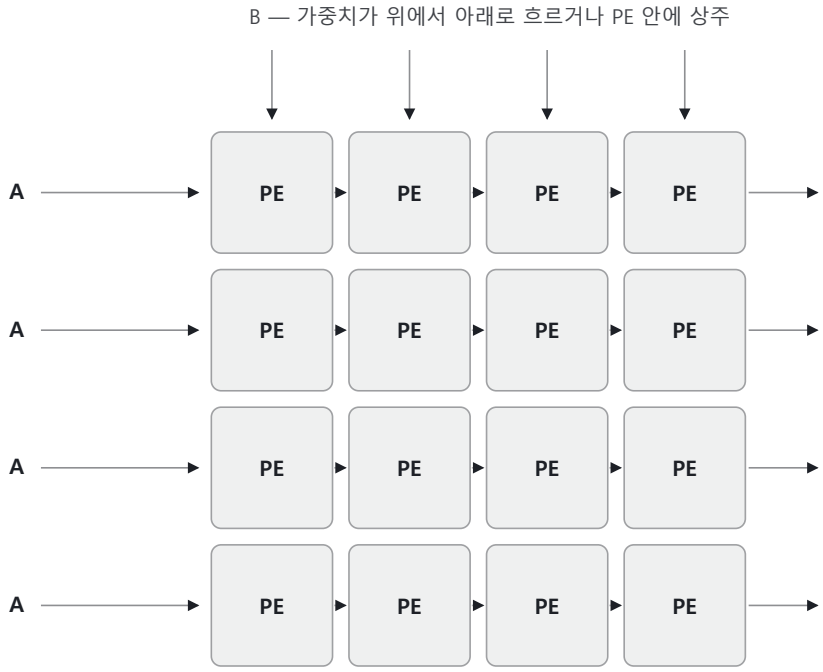
1. **연산기 면적:** 곱셈기 면적은 비트 폭의 대략 제곱에 비례하므로, 비트를 절반으로 줄이면 같은 면적에 약 4배를 넣을 수 있다.
2. **메모리 트래픽:** 데이터가 절반이면 대역폭 요구도 절반이다. **양자화의 진짜 이득은 연산이 아니라 대역폭에서 온다** — 이것이 추론에서 양자화가 그토록 중요한 이유다.

---

4. TPU와 시스톨릭 어레이 — 재사용을 하드웨어에 새기기

구조

시스톨릭 어레이는 처리요소(PE)를 격자로 배열하고, 데이터를 격자 안으로 흘려보내며 **각 PE가 이웃에게 값을 넘겨주는** 구조다. 심장이 피를 밀어내듯(systole) 데이터가 규칙적으로 흐른다고 해서 붙은 이름이다.



활성화가 왼쪽에서 오른쪽으로 흐르고, 각 PE 는 곱셈누산 뒤 이웃에 그대로 넘긴다
— 한 번 읽은 데이터를 어레이 안에서 재사용한다

$N \times N$ 어레이에서 입력 원소 하나는 어레이를 통과하며 N 번 재사용된다. 즉 메모리에서 한 번 읽은 데이터로 $O(N)$ 번 연산한다. **연산강도를 소프트웨어가 아니라 배선 구조가 보장한다.**

Google TPU v1은 $256 \times 256 = 65{,}536$개의 8비트 MAC 유닛을 이 방식으로 배열했다. 어레이 내부 데이터 이동은 인접 PE 간 짧은 배선이므로 에너지가 극히 낮다.

아래 그림은 가중치를 PE에 앉혀 둔 4×4 어레이에 토큰 하나($M = 1$, 디코딩)를 흘릴 때의 넷째 박자다. 화면본에서는 박자를 한 칸씩 넘기거나 토큰 수 M 과 행렬 크기를 바꿔 가며 박자마다 몇 개의 PE가 일하는지 셀 수 있다.



그림 2-1. 4×4 가중치 고정 시스톨릭 어레이에 토큰 하나(M = 1, 배치 1 디코딩)를 흘린 넷째 박자. 활성화가 대각선으로 지나가는 동안 16개 PE 가운데 대각선의 4개만 일한다. — 화면본에서는 이 자리에서 값을 바꿔 볼 수 있다(화면본 2회차 4절).

GPU vs 시스톨릭 어레이 — 무엇을 주고 무엇을 받았나

	GPU	시스톨릭 어레이
재사용 보장	소프트웨어(타일링·커널)가 책임	하드웨어 구조가 보장
유연성	임의 커널 실행 가능	행렬곱 형태에 최적, 그 외엔 비효율
프로그래밍	CUDA 등 성숙한 생태계	컴파일러 의존도 매우 높음
활용률	커널 품질에 크게 좌우	행렬 크기가 어레이에 안 맞으면 급락

시스톨릭 어레이의 약점: 행렬 크기가 어레이 차원의 배수가 아니면 가장자리 PE가 논다. 배치가 작거나 행렬이 가늘고 길면(디코딩!) 활용률이 크게 떨어진다. 이것이 추론 전용 설계가 학습용과 다른 형태를 갖는 이유다.

박자 수로 세면 이렇다. 가중치를 PE에 고정해 두고 $K \times N$ 타일 하나에 토큰 M 개를 흘리면 채우고 비우는 데까지 $M + K + N - 2$ 박자가 걸린다. TPU v1 크기($K = N = 256$)로 계산하면 $M = 1$ 일 때 511박자 동안 일하는 PE는 평균 0.2%이고, $M = 4{,}096$ 이면 약 89%로 오른다. 실제 칩은 다음 타일을 바로 뒤에 이어 넣어 채움·비움 시간을 겹치고, 가중치를 PE에 싣는 시간도 따로 든다. 그래도 M 이 작으면 가중치 하나를 M 번만 쓰고 버리므로 겹쳐서 얻는 이득이 작다. 데이터 흐름도 가중치 고정(weight-stationary) 말고 출력 고정 등 여러 방식이 있다(11절 Sze 외 교과서).

5. NPU와 ASIC — 전용화의 경제학

NPU

모바일·엣지용 신경망 가속기. 목표 함수가 다르다.

- 최적화 대상이 절대 성능이 아니라 **TOPS/W** (전력당 성능)
- 저정밀(INT8/INT4) 중심, 배터리·발열 제약
- SoC 안에 CPU·GPU·ISP와 함께 통합되어 메모리를 공유(LPDDR)
- 사용 패턴이 간헐적이라 지연시간과 웨이크업 오버헤드가 중요

서버 가속기와의 근본 차이: 서버는 처리량을 위해 전력을 쓰지만, 엣지는 전력 예산이 고정되어 있고 그 안에서 최대 성능을 짜낸다. 같은 "AI 칩"이라도 설계 목표가 정반대에 가깝다.

ASIC 손익분기 계산 (실습)

전용 칩을 만들지 사올지의 판단은 정량화할 수 있다.

- C_{NRE} : 비반복 개발비 (설계 인력, EDA 라이선스, IP 라이선스, **마스크셋 제작비**, 검증·시제품). 최신 노드에서는 마스크셋만으로도 수천만 달러 규모이며, 총 NRE는 억 달러 단위에 이를 수 있다.
- c_A : ASIC 단위원가 (웨이퍼비 ÷ 굿다이 수 + 패키징 + 테스트)
- c_G : 대안(상용 GPU)의 단위 취득원가
- N : 예상 물량

여기서는 ASIC 한 대가 GPU 한 대의 일을 같은 처리량·지연·품질로 대신한다고 둔다(1:1 대체). 그래야 같은 물량 N 이 같은 수요를 뜻한다.

총비용 비교: $C_{\text{NRE}} + Nc_A \ < \ Nc_G$

$$N^* = \frac{C_{\text{NRE}}}{c_G - c_A}$$

ASIC의 처리량이 GPU의 r 배라면 같은 수요에 ASIC은 N / r 대가 필요하다. 그때는 c_A 자리에 c_A / r 을 넣는다($r < 1$ 이면 손익분기 물량이 커진다).

그런데 이 식은 절반만 맞다. 실제로는 세 항을 더 넣어야 한다.

1. **시간 축:** 칩 개발에 2~3년이 걸리는데 모델 아키텍처는 그보다 빨리 바뀐다. 완성 시점에 그 워크로드가 여전히 지배적인가?
2. **전력비:** 데이터센터에서 칩의 총소유비용은 취득가보다 수년치 전기요금이 더 클 수 있다. c 에 운영비를 넣으면 전용 칩이 훨씬 유리해진다. 다만 장당으로 세면 손익분기 대수를 가르는 것은 주로 취득가 차이이고 전력비 차이는 몇 %에 그친다(실습 4 풀이). 전력비가 큰 금액이 되는 것은 수만 장을 몇 년 돌리는 규모의 총액에서다.
3. **협상력:** 자체 칩이 있으면 상용 GPU 벤더와의 가격 협상에서 옵션 가치를 얻는다. 실제로 쓰지

않더라도 존재 자체가 가치다.

> **직접 해 보기:** 위 세 항을 넣어 "왜 하이퍼스케일러는 자체 칩을 만들고, 대부분의 스타트업은 안 만드는가"를 설명해 보라. 답의 핵심은 N 과 C_{NRE} 의 규모 차이, 그리고 워크로드가 안정적인지 여부다.

유연성-효율 스펙트럼



FPGA는 흥미로운 중간 지점이다. 하드웨어를 재구성할 수 있어 마스크셋이 필요 없지만, 같은 기능을 ASIC보다 훨씬 큰 면적·전력으로 구현한다. 소량 다품종이나 알고리즘이 유동적인 초기 단계에 적합하다.

---

6. 루프라인 모델 — 이 책의 공용 도구

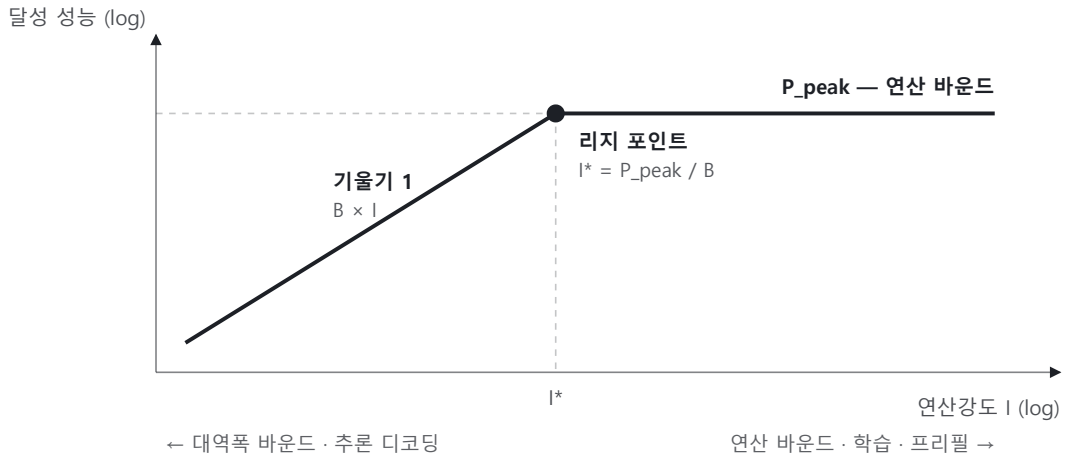
정의

성능 상한은 두 제약의 최솟값이다.

$$P_{\text{attainable}}(I) = \min\left(P_{\text{peak}},\ B \times I\right)$$

- P_{peak} : 칩의 최대 연산속도 [FLOP/s]
- B : 메모리 대역폭 [byte/s]
- I : 연산강도 [FLOP/byte] = (수행 연산 수) / (메모리에서 옮긴 바이트 수)

로그-로그 좌표에 그리면 기울기 1의 직선(대역폭 한계)과 수평선(연산 한계)이 만나는 지붕 모양이 된다. 꺾이는 지점이 **리지 포인트** $I^{*} = P_{\text{peak}} / B$ 다.



실제 칩의 리지 포인트

칩	P_{PEAK} (BF16 DENSE)	B	I^*
NVIDIA H100 SXM	989 TFLOP/s	3.35 TB/s (HBM3, 80GB)	≈ 295 FLOP/byte
NVIDIA B200	2,250 TFLOP/s	7.7 TB/s (HBM3e, 칩 공표 최대 192GB · DGX B200에서는 GPU당 180GB)	≈ 292 FLOP/byte

> **주의:** 벤더 자료는 희소성(structured sparsity) 적용 시 2배 수치를 병기한다. 위는 dense 기준이다. 스펙시트를 볼 때 반드시 확인할 것.
>
> **용량도 칩과 시스템이 다르다.** NVIDIA의 Blackwell 아키텍처 기술 개요(V2.1)는 HGX B200의 GPU 메모리를 「최대 192 GB」로 적는다. 반면 완제품 서버인 DGX B200의 사용자 안내서와 데이터시트는 8장 합계 1,440 GB, 계산하면 GPU당 180 GB로 적는다. 용량을 넣어 계산할 때는 칩 공표값인지 시스템 사양인지를 밝힌다. 대역폭도 문서마다 다르다. 기술 개요 V2.1은 같은 HGX B200 GPU의 대역폭을 7.7 TB/s로, DGX B200 데이터시트는 8장 합계 64 TB/s(계산하면 GPU당 8.0 TB/s)로 적는다. 이 책의 표와 그림은 용량과 같은 출처인 기술 개요의 칩 값(192 GB, 7.7 TB/s)을 쓴다. 8.0 TB/s로 계산하면 $I^* \approx 281$ 이어서, 300 근처라는 아래 결론은 어느 쪽이든 같다. 연산 2,250 TFLOP/s는 같은 표의 희소 4.5 PFLOPS를 반으로 나눈 값이다(표의 dense 칸은 2.2로 적는다).

주목할 점: 세대가 바뀌어도 I^* 가 300 근처에서 거의 변하지 않는다. 연산이 2.3배 늘 때 대역폭도 2.3배 늘려서 **간신히 균형을 유지했다**는 뜻이다. 그리고 그 대역폭을 얻기 위해 HBM3에서 HBM3e로, 다시 HBM4로 가는 막대한 비용이 들어갔다. 3회차의 주제다.

LLM 워크로드를 루프라인에 찍어보기

(가) 학습 / 프리필(prefill) 단계

시퀀스 길이 S, 배치 B_{sz} 의 토큰들을 한꺼번에 처리하므로 큰 GEMM이 된다. 가중치 행렬 $W \in \mathbb{R}^{n \times n}$ 를 M 개 토큰에 적용할 때:

- 연산량 $\approx 2Mn^2$ FLOP

- 메모리 이동 ≈ $2n^2$ 바이트 (fp16 가중치 1회 로드)
- 연산강도 $I \approx M$ (활성화 이동을 셈에서 뺄 때)

활성화를 뺀 것은 근사다. $M \times n$ 입력을 HBM에서 읽고 같은 크기의 출력을 다시 쓰면 이동량이 $2n^2 + 4Mn$ 바이트가 되어 $I = Mn / (n + 2M)$ 이다. 이 값은 M 과 $n / 2$ 가운데 작은 쪽을 넘지 못한다. 그래서 $I \approx M$ 은 n 이 M 보다 훨씬 클 때나, 커널 퓨전으로 활성화를 온칩에 붙잡아 둘 때 맞는다. LLM은 n 이 수천 이상이므로($n = 8{,}192$, $M = 4{,}096$ 이면 $I = 2{,}048$) 어느 쪽으로 세든 $I \gg I^*$ → **연산 바운드.** GPU의 연산 유닛이 잘 채워진다.

(나) 추론 / 디코딩(decode) 단계

토큰을 한 개씩 생성한다. 배치 B_{sz} 개 시퀀스를 동시에 처리해도 각 스텝에서 처리하는 토큰은 시퀀스당 1개다.

- 연산량 ≈ $2B_{sz}n^2$ FLOP
- 메모리 이동 ≈ $2n^2$ 바이트 (가중치를 통째로 다시 읽어야 함)
- 연산강도 $I \approx B_{sz}$

fp16 가중치라면 배치가 300 미만일 때 대역폭 바운드다. 배치 32로 서비스한다면 $I \approx 32$ 이므로 이론 성능의 대략 32 / 295 ≈ 11% 만 쓰는 셈이다. 나머지 89%의 연산 유닛은 메모리를 기다리며 논다.

아래 그림은 H100에서 fp16 가중치로 디코딩을 배치 32로 돌릴 때 워크로드가 지붕 아래 어디에 떨어지는지 보여 준다. 화면본에서는 칩·단계·배치·가중치 정밀도·모델 크기를 바꿔 가며 점이 리지 포인트를 언제 넘는지 볼 수 있다.

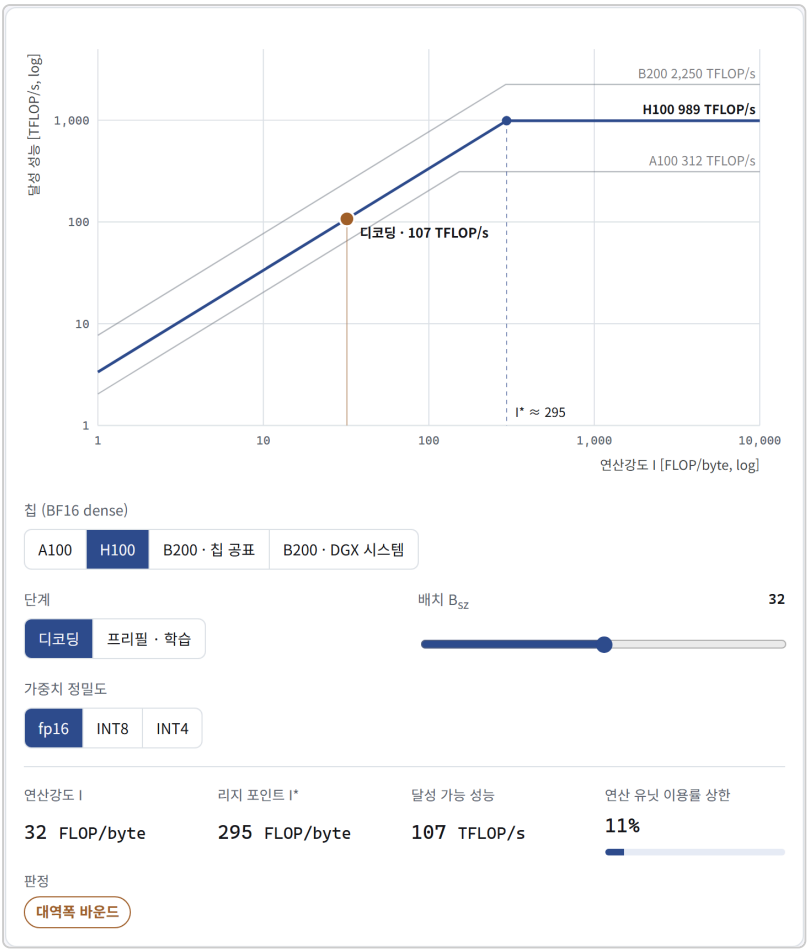


그림 2-2. 루프라인 위의 디코딩 워크로드. H100에서 배치 32, fp16 가중치면 연산강도가 약 32로 리지 포인트(약 295)에 못 미쳐 대역폭 바운드에 놓이고, 연산 유닛 이용률 상한은 약 11%다. — 화면본에서는 이 자리에서 값을 바꿔 볼 수 있다(화면본 2회차 6절).

디코딩 속도의 상한 — 대역폭으로 세는 어림식

가중치 P 개(파라미터), 바이트/파라미터 b 인 모델의 초당 최대 토큰 생성 수는 대역폭만으로 결정된다.

$$\text{tokens/s} \lesssim \frac{B}{P \cdot b} \times B_{\text{sz}}$$

예: 70B 모델을 fp16($b = 2$)으로 H100 한 장에서 배치 1로 돌린다면, 가중치만 140GB라 80GB에 안 들어간다. 2장에 나눠 쓰면 총 대역폭 6.7 TB/s이므로 6.7e12 / (1.4e11) ≈ 48 토큰/초가 이론 상한이다. 이 어림식은 칩 N 장에 가중치를 나눠 얹으면 각 칩이 제 몫을 동시에 읽는다고 보고 대역폭을 N 배로 더한다. INT8로 양자화하면($b = 1$) 가중치가 70GB라 한 장에 겨우 들어간다. 한 장으로 줄이면 대역폭도 3.35 TB/s로 줄어 상한은 그대로 약 48 토큰/초이고, 두 장을 그대로 쓰면 약 96 토큰/초로 두 배가 된다.

이 식은 대역폭 쪽만 센 것이라 연산강도 I 가 리지 포인트 I^* 보다 작을 때만 맞는다. 가중치만 b 바이트로 읽고 연산은 BF16 피크로 한다고 보면 $I = 2B_{\text{sz}} / b$ 이므로, 연산 바운드로 넘어가는 배치는 $B_{\text{sz}} = (b / 2)\, I^*$ 다. H100이면 fp16에서 약 295, INT8에서 약 148, INT4에서 약 74다. 연산까지 같은 저정밀로 해 피크가 함께 늘면 경계는 다시 BF16 기준 I^* 근처로 돌아간다(실습 3 풀이). 연산 바운드에서는 상한이 토큰 하나에 드는 연산 $2P$ FLOP으로 정해진다(배치 전체로 $\text{tokens/s} \lesssim NP_{\text{peak}} / (2P)$, P_{peak} 은 실제로 쓰는 정밀도의 피크). 매 스텝 KV 캐시를 읽는 트래픽도 빠져 있어서, 문맥이 길수록 실제 속도는 이 상한보다 더 낮다. 사용자 체감 지표로 옮기면, 시퀀스 하나

기준으로 이 식은 출력 토큰 사이 간격(TPOT)의 하한 $P \cdot b / B$ 를 준다. 첫 토큰까지의 시간(TTFT)은 이 식이 아니라 프리필이 정한다(8절 끝 「공정한 비교의 조건」).

> **이 어림식 하나로 설명되는 것들:** 양자화가 왜 추론에서 필수인지, 배치를 왜 키우려 하는지, 왜 "가중치가 메모리에 다 들어가느냐"가 결정적인지, 그리고 왜 3회차의 HBF(대용량 메모리)가 논의되는지.

연산강도를 올리는 모든 방법 (통합 정리)

방법	어디서	어떻게 I 를 올리나
타일링·블로킹	소프트웨어	온칩에 올린 타일을 여러 번 재사용
커널 퓨전 (예: FlashAttention)	소프트웨어	중간 결과를 HBM에 쓰지 않고 온칩에 유지
배치 확대	서빙 정책	같은 가중치로 더 많은 시퀀스 처리
양자화	모델	이동 바이트를 줄여 분모 감소
MoE / 희소성	모델 구조	I 를 올리는 방법이 아니다. 토큰 하나가 쓰는 가중치만 읽고 계산하므로 연산과 이동이 함께 줄어 I 는 그대로이고, 토큰당 바이트가 줄어 디코딩 상한이 오른다. MoE는 배치 안의 토큰이 여러 전문가로 흩어지면 전문가 하나를 재사용하는 토큰 수가 줄어 I 가 오히려 낮아진다 (라우팅 오버헤드도 붙는다)
시스톨릭 어레이	하드웨어	구조적으로 재사용 보장
큰 온칩 SRAM	하드웨어	오프칩 접근 횟수 자체를 줄임
HBM 대역폭 증대	하드웨어	I 가 아니라 B 를 올림 (→ 3회차)
PIM	하드웨어	연산을 메모리 쪽으로 이동 (→ 3회차)

짚어 둘 점: 위 목록에서 소프트웨어 항목은 전부 I 를 올리는 것이고, 하드웨어 항목의 절반은 B 를 올리는 것이다. **두 진영이 같은 부등식의 서로 다른 항을 공격하고 있다는 구도를 잡으면 논문·기사가 훨씬 잘 읽힌다.**

---

7. 소프트웨어 락인 — 기술 외적 해자

하드웨어 성능이 비슷해도 실제 채택은 소프트웨어가 결정한다.

- **CUDA 생태계:** cuDNN, cuBLAS, NCCL, TensorRT 등 오랜 기간 쌓인 최적화 커널과 방대한 라이브러리 지원.
- **대안들:** AMD ROCm, Intel oneAPI, Google XLA/JAX, OpenAI Triton, PyTorch 2.x의 컴파일러 스택. Triton처럼 커널을 하드웨어 중립적으로 기술하려는 시도가 락인을 약화시킬 수 있다. 컴파일러 스택은 모델을 연산 그래프로 잡아 중간 표현(IR)에 올린 다음, 하드웨어에 가까운 표현으로 단계적으로 낮춰(lowering) 커널을 만든다. PyTorch 2.14 문서에 따르면 `torch.compile` 에서는 TorchDynamo가 파이썬 코드에서 그래프를 잡고, 기본 컴파일러 TorchInductor가 여러 가속기·백엔드용 코드를 만든다. NVIDIA·AMD·Intel GPU에서는 TorchInductor가 OpenAI Triton을 바탕으로 커널을 짠다.

- **하이퍼스케일러가 자체 칩을 만들 수 있는 이유:** 자기 워크로드를 자기가 통제하므로 컴파일러 스택을 자기 모델에만 맞추면 된다. **범용 소프트웨어 생태계를 만들 필요가 없다는 것이 이들의 결정적 이점이다.**

연산자 커버리지와 폴백

하드웨어는 좋은데 모델이 느린 흔한 이유가 여기 있다. 가속기용 라이브러리가 모든 연산자를 갖추고 있지는 않다. 실행기가 지원하지 않는 연산자를 만나면 그 노드만 CPU로 넘긴다(폴백, fallback). 그러면 입력을 장치 메모리에서 호스트 메모리로 내리고, CPU가 계산한 결과를 다시 장치로 올려야 하며, 옮기는 앞뒤로 장치와 호스트가 서로를 기다린다. 연산 자체는 작아도 이 왕복이 그래프 전체의 지연을 키운다.

널리 쓰는 실행기 둘이 이 동작을 문서에 적어 두었다. ONNX Runtime은 실행 제공자(execution provider) 목록을 우선순위 순으로 받는다. `['CUDAExecutionProvider', 'CPUExecutionProvider']`를 주면 CUDA가 처리할 수 있는 노드는 CUDA에서, 나머지는 CPU에서 실행한다. PyTorch의 Apple GPU 백엔드(MPS)는 환경변수 `PYTORCH_ENABLE_MPS_FALLBACK`을 1로 두면 MPS가 지원하지 않는 연산을 CPU로 돌린다(PyTorch 2.14 문서).

왕복이 얼마나 비싼지는 두 통로의 폭을 나란히 놓으면 보인다. NVIDIA H100 사양표는 H100 SXM의 PCIe Gen5를 128 GB/s, GPU 메모리 대역폭을 3.35 TB/s로 적는다. H100 백서는 128 GB/s가 양방향 합계이고 한 방향은 64 GB/s라고 밝힌다. 공식 표기끼리 나누면 약 26배, 한 방향으로 따지면 약 52배다. 폴백 한 번은 HBM 안에서 끝났을 데이터를 그보다 한 자릿수 이상 좁은 통로로 두 번 옮기는 일이다. 락인의 실체가 커널의 축적이라는 말(부록 B 질문 4)이 여기서 구체적인 모양을 얻는다. 연산자 목록이 짧은 새 칩은 연산기가 빨라도 폴백 때문에 느려질 수 있다.

사례 — DeepSeek-V3, 통신을 연산 뒤에 숨기기

소프트웨어로 하드웨어 제약을 메운 예를 DeepSeek-V3 기술 보고서(arXiv:2412.19437)에서 볼 수 있다. 보고서에 따르면 이 모델은 H800 2,048장 클러스터에서 학습했고, 전체 학습(사전학습 2,664K, 문맥 길이 확장 119K, 사후학습 5K GPU시간)에 2.788M H800 GPU시간이 들었다. 보고서는 GPU시간당 임차료를 2달러로 가정해 557.6만 달러라는 금액을 내는데, 같은 문단에서 이 금액은 공식 학습만 센 것이고 아키텍처·알고리듬·데이터에 관한 사전 연구와 절제(ablation) 실험 비용은 빠졌다고 밝힌다. 그러므로 이 금액을 모델 개발비로 읽으면 안 된다.

기술적으로 볼 것은 통신 처리다. 보고서는 파이프라인 병렬에서 순전파·역전파의 연산 구간과 통신 구간을 겹치는 DualPipe 알고리듬을 설계했고, 노드 사이 전문가(MoE) 통신을 처리하는 데 H800의 SM 132개 가운데 20개를 떼어 두었다. 계산하면 연산 유닛의 약 15%다. 보고서 스스로도 이 방식이 연산 처리량을 제한한다고 적고, 통신을 SM 밖의 전용 하드웨어로 넘겨 달라고 칩 설계자에게 제안한다. 루프라인 밖의 항목인 노드 간 통신(부록 B 질문 1)을 스케줄링으로 연산 뒤에 숨긴 사례이고, 그 대가로 연산 유닛 일부를 내준 사례다.

---

8. 자주 하는 오해

오해 1: "FLOPS가 높으면 빠른 칩이다" → 워크로드의 연산강도가 리지 포인트보다 낮으면 FLOPS는 무의미하다. LLM 추론에서는 대역폭 수치가 더 나은 예측자다.

오해 2: "TOPS/TFLOPS 수치를 그대로 비교하면 된다" → 정밀도(FP32/BF16/FP8/INT8/INT4)와 희소성 적용 여부에 따라 4~8배 차이가 난다. 비교 전에 조건을 맞춰야 한다. 정밀도와 희소성을 맞춰도 아직 부족하다 — 이 절 끝 「공정한 비교의 조건」.

오해 3: "GPU 이용률(utilization) 90%면 잘 쓰고 있다" → `nvidia-smi` 의 이용률은 "커널이 실행 중인 시간 비율"이지 연산 유닛이 채워졌다는 뜻이 아니다. 메모리를 기다리며 도는 커널도 100%로 표시된다. 실제 지표는 **MFU(Model FLOPs Utilization)** 처럼 이론 FLOPS 대비 달성 FLOPS를 봐야 한다. 커널이 아주 짧을 때는 반대쪽 함정도 있다. 커널을 띄우고 기다리는 데 드는 오버헤드가 실행 시간보다 커져, GPU가 커널 사이에서 논다(부록 B 질문 1의 CUDA Graphs 예).

오해 4: "양자화는 연산을 빠르게 한다" → 연산기 처리량도 늘지만, 추론에서 체감 이득의 대부분은 **메모리 트래픽 감소**에서 온다.

오해 5: "ASIC이 GPU보다 무조건 효율적이다" → 워크로드가 고정된 경우에만 그렇다. 어레이 크기와 맞지 않는 행렬, 새로운 연산자(예: 새 어텐션 변형)가 들어오면 활용률이 급락한다.

오해 6: "NPU와 GPU는 같은 것의 크기 차이다" → 최적화 목적함수가 다르다(처리량 vs 전력당 성능). 설계 결정이 정반대로 갈리는 지점이 많다.

공정한 비교의 조건 — 정밀도 말고도 맞출 것

추론 서비스의 속도는 숫자 하나로 잡히지 않는다. 사용자가 느끼는 지연은 셋으로 나뉜다. **TTFT**(time to first token)는 요청을 보낸 뒤 첫 토큰이 나오기까지의 시간으로, 입력 전체를 한꺼번에 처리하는 프리필이 좌우한다. **TPOT**(time per output token)는 출력 토큰 사이 간격의 평균으로, 6절의 디코딩 상한이 좌우한다. **P99는** 요청의 99%가 그 이하로 끝나는 꼬리 지연이다. 평균이 좋아도 꼬리가 길면 서비스 품질 약속을 지키지 못한다.

업계 표준 벤치마크인 MLPerf Inference는 그래서 칩끼리 비교할 때 여러 조건을 함께 묶는다. MLCommons 추론 규칙(2026-10-08에 확인한 master 판, 마지막 수정 2026-08-20)에서 데이터센터용 Llama2-70B 과제를 보면, Closed 부문은 참조 구현과 같은 모델과 전처리를 쓰게 하고(양자화 보정은 허용, 재학습은 금지) 데이터셋은 OpenOrca로 정해 둔다. 정확도는 FP32 기준 ROUGE 점수의 99.9% 이상이어야 하고, 지연 제약은 Conversational 범주에서 TTFT 2,000 ms·TPOT 200 ms, Interactive 범주에서 TTFT 450 ms·TPOT 40 ms다. 요청이 포아송 분포로 무작위로 들어오는 Server·Interactive 시나리오에서는 지연을 99번째 백분위수로 잰다. 낮은 정밀도로 처리량을 올린 칩도 정확도 하한과 지연 제약을 넘어야 유효한 결과가 된다. TOPS 표 한 줄로는 이 가운데 어느 것도 알 수 없다. 규칙은 라운드마다 바뀌므로 인용할 때는 판을 밝힌다.

---

9. 정량 실습

실습 1 — 리지 포인트 계산 스펙시트에서 임의의 가속기를 골라 dense 기준 P_{peak} 와 B 를 찾아 I^* 를 구하라. 세 세대를 비교해 I^* 의 추세를 보라.

실습 2 — 디코딩 상한 7B / 70B / 405B 모델을 fp16, INT8, INT4로 각각 서빙할 때, 특정 가속기에서 배치 1의 이론 최대 토큰/초를 구하라. 실제 서비스 지연 요구(예: 30 토큰/초)를 만족하려면 어떤 조합이

가능한가?

실습 3 — KV 캐시 예산 시퀀스 길이 S, 레이어 L, KV 헤드 차원 총합 d(K·V를 저장하는 헤드 수 × 헤드당 차원. 쿼리 헤드는 세지 않는다), 배치 B_{sz} 일 때 fp16 KV 캐시 크기는 대략 $2 \cdot L \cdot S \cdot d \cdot B_{sz} \cdot 2$ 바이트다. 주어진 메모리 용량에서 가능한 최대 배치를 구하고, 그 배치에서의 연산강도를 루프라인에 찍어라. **용량 제약이 어떻게 대역폭 효율을 제한하는지**가 드러난다.

실습 4 — ASIC 손익분기 가정값(NRE 1억 달러, ASIC 단가 3천 달러, GPU 단가 3만 달러, 3년 전력비 포함)으로 N^* 를 구하고, 모델 수명이 2년일 때 실현 가능한지 논하라.

---

10. 토론 질문

1. 루프라인 모델은 무엇을 설명하지 못하는가? (지연시간, 캐시 계층, 통신, 커널 실행 오버헤드 등 — 모델의 한계를 아는 것이 모델을 아는 것의 절반이다)
2. 리지 포인트가 세대가 지나도 300 근처에 머무는 것은 우연인가, 설계 의도인가? 어느 쪽으로 벗어나면 무슨 일이 생기는가?
3. MoE(전문가 혼합)는 루프라인 관점에서 어떤 효과를 내는가? 총 파라미터는 늘지만 활성 파라미터는 줄어드는 구조를 어떻게 평가해야 하는가?
4. 소프트웨어 락인은 얼마나 지속 가능한가? Triton·XLA 같은 중간 계층이 성숙하면 무엇이 바뀌는가?
5. 추론 전용 칩을 새로 설계한다면, 학습용 칩과 무엇을 다르게 하겠는가? (힌트: 리지 포인트를 어디에 둘 것인가)

---

11. 후속 공부

- **루프라인 원전:** Williams, Waterman & Patterson, "Roofline: An Insightful Visual Performance Model for Multicore Architectures", *CACM* 52(4), 2009 — DOI 10.1145/1498765.1498785
- **에너지 수치 원전:** Horowitz, "Computing's Energy Problem (and what we can do about it)", ISSCC 2014 keynote — DOI 10.1109/ISSCC.2014.6757323. 1절 표의 32비트 메모리 두 값은 Han, Pool, Tran & Dally, "Learning both Weights and Connections for Efficient Neural Networks", NIPS 2015의 그림 1 — arXiv:1506.02626
- **TPU 아키텍처:** Jouppi et al., "In-Datacenter Performance Analysis of a Tensor Processing Unit", ISCA 2017 — TPU v1 설계 근거를 직접 밝힌 논문. **이 회차와 함께 한 편만 읽는다면 이 논문을 권한다.** arXiv:1704.04760
- **가속기 설계 교과서:** Sze, Chen, Yang & Emer, "Efficient Processing of Deep Neural Networks" (Synthesis Lectures, 2020) — 데이터플로우 분류(weight stationary / output stationary 등)의 표준 정리. DOI 10.1007/978-3-031-01766-7
- **커널 수준 최적화:** Dao et al., "FlashAttention" 시리즈 — 커널 퓨전으로 I 를 올리는 대표 사례. 첫 논문은 arXiv:2205.14135

- **GPU 구조:** NVIDIA 아키텍처 백서(H100 백서, Blackwell 아키텍처 기술 개요)와 CUDA Programming Guide의 메모리 계층 장
- **커널 실행 오버헤드:** NVIDIA 기술 블로그 「Getting Started with CUDA Graphs」(2019) — 짧은 커널을 묶어 띄우는 방법과 그 효과를 측정한 예
- **비교의 규칙:** MLCommons MLPerf Inference 규칙 — 과제별 정확도 하한과 TTFT·TPOT 제약

자습 부록 A — 정량 실습 풀이

9절 실습의 풀이다. **반드시 먼저 스스로 풀고 펼칠 것.** 숫자는 공개 스펙시트 기준이며 모두 dense(희소성 미적용) 값이다.

실습 1 풀이 — 리지 포인트 계산

▾ 풀이

$I^* = P_{\text{peak}} / B$ 에 세 세대를 넣는다.

칩	P_{PEAK} (BF16 DENSE)	B	I^*
A100 80GB SXM	312 TFLOP/s	2.04 TB/s	**153** FLOP/byte
H100 SXM	989 TFLOP/s	3.35 TB/s	**295** FLOP/byte
B200	2,250 TFLOP/s	7.7 TB/s	**292** FLOP/byte

계산: 312 / 2.04 = 153, 989 / 3.35 = 295, 2250 / 7.7 = 292.

세대별 증가율로 보면 훨씬 선명하다.

구간	연산	대역폭	I^*
A100 → H100	×3.17	×1.64	×1.93 (악화)
H100 → B200	×2.28	×2.30	×0.99 (거의 그대로)

읽는 법: A100→H100 세대에서는 연산이 대역폭보다 훨씬 빨리 늘어 리지 포인트가 두 배 가까이 올라갔다. 즉 **대역폭 바운드가 되기 더 쉬워졌다.** H100→B200에서는 HBM3e를 붙여 대역폭을 연산과 거의 같은 비율로 늘렸고, 리지 포인트는 제자리에 머물렀다.

함의: I^* 가 300 근처에 머무는 것은 우연이 아니라 **대역폭이 연산을 겨우 따라잡고 있다는 뜻**이고, 그 "겨우"의 대가가 HBM 가격과 첨단 패키징 비용이다. 3회차·6회차의 존재 이유가 이 표 한 장에 들어 있다.

> **스펙시트를 볼 때의 주의:** 벤더 자료는 희소성 적용 값을 병기하는데, 그 값을 쓰면 I^* 가 두 배로 뻥튀기된다. 반드시 dense 기준인지 확인할 것.

실습 2 풀이 — 디코딩 상한

▾ 풀이

$$\text{tokens/s} \lesssim \frac{B}{P \cdot b} \times B_{sz}$$

H100 SXM 한 장($B = 3.35$ TB/s, 용량 80GB), 배치 1 기준으로 계산한다. $P \cdot b$ 는 가중치의 총 바이트다.

모델	정밀도	가중치 용량	필요 장수	총 B	이론 상한
7B	fp16 ($b = 2$)	14 GB	1	3.35 TB/s	**239 tok/s**
7B	INT8	7 GB	1	3.35	479
7B	INT4	3.5 GB	1	3.35	957
70B	fp16	140 GB	**2** (용량 초과)	6.7	**48 tok/s**
70B	INT8	70 GB	1 (빠듯)	3.35	48
70B	INT4	35 GB	1	3.35	96
405B	fp16	810 GB	**11**	36.85	45
405B	INT8	405 GB	6	20.1	50
405B	INT4	202.5 GB	3	10.05	50

계산 예: 70B fp16 2장 → $6.7 \times 10^{12} / (1.4 \times 10^{11}) = 47.9$ tok/s. 총 B 는 장수 × 3.35 TB/s이고, 상한은 반올림하지 않은 값으로 계산했다.

30 tok/s 요구를 만족하는 조합: 위 표에서는 전부다. 다만 **이론 상한**임에 주의한다. 실제로는 대역폭 이용률이 60~80%에 그치고(이 책의 가정값), 텐서 병렬 통신과 커널 실행 오버헤드가 더 붙는다. 이용률만 넣어도 70B fp16 2장은 48 → 약 29~38 tok/s이고, 오버헤드가 더해지면 30 아래로 내려갈 수 있어 아슬아슬해진다.

세 가지 통찰

1. **같은 장비라면 모델 크기와 속도는 반비례한다.** 식에 P 가 분모에 있으므로, 장수와 정밀도를 고정하면 7B에서 70B로 갈 때 상한이 10배 낮아진다. 위 표에서 fp16 7B(1장, 239)와 70B(2장, 48)가 5배 차이인 것은 70B가 장수가 두 배라 총 대역폭도 두 배이기 때문이다. "더 큰 모델을 쓰자"는 결정은 같은 장비에서 그만큼 느려지거나, 같은 속도를 위해 그만큼 칩을 더 사야 한다는 결정이다.
2. **양자화는 두 번 이긴다.** 가중치 바이트가 절반이 될 때마다 같은 장수에서 상한이 두 배가 되거나, 같은 상한을 절반의 장수로 낸다. 70B를 fp16에서 INT4로 바꾸면 이 둘을 함께 얻는다. 2장 48 tok/s가 1장 96 tok/s가 되어 장비 비용이 절반이 된다.
3. **용량을 겨우 채우는 장수로 담으면 배치 1 상한은 모델 크기와 거의 무관해진다.** 70B fp16 2장(48), 70B INT8 1장(48), 405B fp16 11장(45), 405B INT4 3장(50)이 모두 비슷하다. 칩마다 자기 메모리를 거의 꽉 채운 가중치를 토큰마다 한 번씩 다 읽으므로, 상한이 칩 한 장의 「용량 ÷ 대역폭」(H100은 80 GB ÷ 3.35 TB/s ≈ 24 ms, 약 42 tok/s)보다 조금 위에 모이기 때문이다. 같은 모델·정밀도에서 칩을 더 얹으면 이 어림식의 상한은 장수에 비례해 오른다(405B fp16 11장 45 → 12장 50). 실제로는 칩 사이 텐서 병렬 통신이 붙어 그만큼 오르지 않는다. 이 통신 비용은 어림식 밖에 있다(부록 B 질문 1).

실습 3 풀이 — KV 캐시 예산

▾ **풀이**

fp16 KV 캐시 크기는 시퀀스 하나당 $4 \cdot L \cdot S \cdot d$ 바이트다 (K와 V 둘 ×2, fp16 ×2).

설정: 70B급 모델, $L = 80$ 레이어, GQA로 $d = 8$헤드 $\times 128 = 1{,}024$, $S = 4{,}096$, H100 80GB 한 장.

시퀀스 하나당 KV $= 4 \times 80 \times 4096 \times 1024 = 1.34 \times 10^9$ 바이트 = **1.34 GB**

정밀도	가중치	남는 용량	최대 배치	연산강도 $I = 2B_{sz} / b$	그 정밀도의 I^*	이용률 I / I^*
INT8	70 GB	10 GB	**7**	14	590	**2.4%**
INT4	35 GB	45 GB	**33**	132	1,181	**11.2%**

(I^* 는 H100 BF16 dense 값 295에서 연산 피크가 INT8은 2배, INT4는 4배로 는다고 본 값이다. INT8의 2배는 H100 사양표의 INT8 텐서 코어 값과 맞고, INT4의 4배는 같은 비율을 이어 붙인 가정이다. 실무에서는 활성화 버퍼와 단편화로 여유가 더 줄어 위 값은 낙관적 상한이다. 반대로 모든 시퀀스가 S 까지 꽉 찬다고 본 것은 보수적인 가정이다. 쓰는 만큼만 자리를 잡는 PagedAttention 같은 방식에서는 평균 길이가 기준이 된다.)

정밀도 가정이 둘이다. 위 표는 연산도 가중치와 같은 정밀도로 한다고 보는 어림이다. 그러면 연산강도와 리지 포인트가 함께 $2 / b$ 배로 늘어, 이용률은 fp16 때처럼 $B_{sz} / 295$ 가 된다(실습 3의 KV 그림이 쓰는 「B_{sz} / I^*」가 이것이다). 6절의 루프라인 그림(화면본)은 가중치만 저정밀로 읽고 연산은 BF16으로 하는 경우다. 연산강도는 똑같이 $2B_{sz} / b$ 이지만 리지 포인트를 BF16 값 295로 두므로 이용률이 $2 / b$ 배 높게 나온다. 계산하면 위 표의 INT8 배치 7은 약 4.7%, INT4 배치 33은

약 45%가 된다. 두 값이 다른 것은 오류가 아니라 가정이 다르기 때문이다.

드러나는 것 — 이 실습의 핵심

용량 제약이 배치를 누르고, 배치가 눌리면 연산강도가 눌리고, 연산강도가 눌리면 **가장 비싼 부품인 연산 유닛의 97%가 논다.** 대역폭 제약과 용량 제약은 별개의 것 같지만, 용량이 배치를 통해 대역폭 효율을 지배한다.

한 걸음 더 — GQA가 없었다면

Multi-Head Attention이라면 $d = 64 \times 128 = 8{,}192$ 로 8배가 되어 시퀀스당 KV가 **10.7 GB**다. INT8에서는 배치 1조차 들어가지 않는다. **GQA·MQA가 선택이 아니라 필수인 이유**가 이 한 줄 계산에 있다.

처리량을 늘리는 수단의 우선순위 (이 계산에서 자연히 나온다)

1. KV 캐시 양자화(INT8) → KV가 절반, 배치 두 배
2. 가중치 양자화 → 남는 용량이 급증, 배치 급증
3. PagedAttention 등 단편화 제거 → 실효 용량 회복
4. 시퀀스 길이 제한 → S 에 선형 비례

아래 그림은 H100 80GB 한 장에 70B급 모델을 INT8 가중치·fp16 KV 캐시·GQA로 올리고 $S = 4{,}096$ 으로 둘 때 메모리가 어떻게 나뉘는지 보여 준다. 화면본에서는 칩, 시퀀스 길이, 가중치·KV 캐시 정밀도, 어텐션 방식, 예비 비율을 바꿔 가며 최대 배치와 연산 유닛 이용률이 어떻게 움직이는지 볼 수 있다.



그림 2-3. H100 80GB 한 장에 70B급 모델을 INT8 가중치(70 GB)로 올리고 GQA·KV fp16으로 S = 4,096 시퀀스를 받으면 KV 캐시 자리가 시퀀스 7개분뿐이다. 배치가 7이면 연산도 INT8로 한다고 볼 때 연산 유닛 이용률은 약 2.4%다. — 화면본에서는 이 자리에서 값을 바꿔 볼 수 있다(화면본 2회차 실습 3 풀이).

실습 4 풀이 — ASIC 손익분기

▾ 풀이

$$N^* = \frac{C_{\text{NRE}}}{(c_G + o_G) - (c_A + o_A)}$$

가정: NRE 1억 달러, ASIC 단가 3,000달러, GPU 단가 30,000달러. ASIC 한 대가 GPU 한 대를 1:1로 대신한다(5절). 3년 전력비는 다음과 같이 세운다 — GPU 700W, ASIC 300W, PUE 1.3, 가동률 80%, 전기료 $0.08/kWh.

- GPU 전력량 $= 0.7 \times 1.3 \times 8760 \times 3 \times 0.8 = 19{,}132$ kWh → $o_G \approx 1{,}531$ 달러
- ASIC 전력량 $= 0.3 \times 1.3 \times 8760 \times 3 \times 0.8 = 8{,}199$ kWh → $o_A \approx 656$ 달러

$$N^* = \frac{10^8}{(30{,}000 + 1{,}531) - (3{,}000 + 656)} = \frac{10^8}{27{,}875} \approx 3{,}590 \text{ 대}$$

모델 수명이 2년이면? 전력비가 2/3이 되어 분모는 27,583 → $N^* \approx 3{,}630$ 대(나누면 3,625.4). **거의 변하지 않는다.**

전력을 어디서 재는가 — 칩, 서버, 시설. 위 700 W는 NVIDIA H100 사양표가 H100 SXM의 최대 TDP로 적은 값(「최대 700 W, 설정 가능」)이다. 칩 하나가 낼 수 있는 열의 상한이지 작업 중 평균 소비가 아니고, 서버 전체의 전력은 더더욱 아니다. 전력은 경계를 넓힐수록 커진다.

경계	값	근거
칩	최대 700 W (H100 SXM TDP)	H100 사양표
서버	최대 10.2 kW (DGX H100, GPU 8장) → 장당 약 1.28 kW	DGX H100/H200 사용자 안내서 전원 사양. 계산하면 $10.2 \div 8 = 1.275$ kW
시설	서버 전력 × PUE	이 실습의 가정 PUE 1.3

같은 가정(PUE 1.3, 가동률 0.8, $0.08/kWh, 3년)으로 서버 경계에서 다시 세면 다음과 같다.

- GPU 전력량 $= 1.275 \times 1.3 \times 8760 \times 3 \times 0.8 \approx 34{,}847$ kWh → $o_G \approx 2{,}788$ 달러 (칩 경계의 약 1.8배)
- ASIC도 같은 경계로 올려야 공정하다. ASIC 서버의 사양을 정할 근거가 없으므로 GPU와 같은 비율(1.275 / 0.7 ≈ 1.82 배)로 올려 약 0.546 kW($0.3 \times 1.275 / 0.7 = 0.54643\ldots$)로 두면, 전력량 ≈ 14,935 kWh(비율을 반올림하지 않고 계산한 값) → $o_A \approx 1{,}195$ 달러

$$N^* = \frac{10^8}{(30{,}000 + 2{,}788) - (3{,}000 + 1{,}195)} = \frac{10^8}{28{,}593} \approx 3{,}500 \text{ 대}$$

GPU만 서버 경계로 올리고 ASIC은 칩 300 W 그대로 두면 분모가 29,132달러, $N^* \approx 3{,}430$ 대가 되어 ASIC 쪽으로 기운다. 경계를 한쪽만 옮기면 비교가 틀어진다. 어느 쪽으로 세든 N^*는 3,400~3,600대에 머물고, 서버 경계에서 전력비 차이 1,593달러는 분모의 약 6%라 결론(취득원가가 지배한다)은 그대로다. 같은 경계의 예를 하나 더 들면, Blackwell 세대의 DGX B200 데이터시트는 GPU 8장 시스템의 최대 전력을 약 14.3 kW로 적는다(계산하면 장당 약 1.79 kW).

단가·전기료·PUE·가동률은 여전히 예시용 가정값이고, 제조사 사양에서 온 것은 칩과 서버의 전력 상한뿐이다.

이 실습이 진짜 가르치는 것 — 결과의 방향이 직관과 반대다

1. **손익분기 물량이 놀랍도록 작다.** 3,600대는 하이퍼스케일러 기준으로 몇 주치 도입량이다. 순수 회계만 보면 자체 칩은 압도적으로 유리하다.
2. **전력비는 결정적이지 않다.** 분모 27,875달러 중 전력비 차이는 875달러, 겨우 3%다(서버 경계로 세도 약 6%). 지배하는 것은 **취득원가 차이 27,000달러**다. "전기를 아끼려고 자체 칩을 만든다"는 흔한 설명은 이 수치와 맞지 않는다.
3. **그렇다면 왜 대부분은 하지 않는가.** N^* 가 답이 아니기 때문이다. 진짜 제약은 **시간**이다 — 개발에 2~3년이 걸리는데, 완성 시점에도 이 칩이 겨냥한 연산 패턴이 여전히 지배적이어야 한다. 2년 전의 설계 결정이 오늘의 모델 구조(MoE, 긴 컨텍스트, 새로운 어텐션 변형)를 맞힐 확률이 실제 의사결정 변수다.

따라서 하이퍼스케일러의 진짜 이점은 물량이 아니라 「자기 워크로드를 자기가 통제한다」는 데 있다. 칩을 모델에 맞추는 대신 **모델을 칩에 맞출 수 있으면** 위 시간 리스크 가운데 「설계가 모델 구조를 못 맞히는」 몫이 크게 준다. 수요가 예상만큼 오는지, 경쟁 모델이 요구하는 성능이 어디로 움직이는지 같은 위험은 그대로 남는다. 7절 소프트웨어 락인 논의가 여기서 완성된다.

> 4회차 5절에서 c_A 를 수율 식으로 직접 계산해 이 식을 정밀화한다.

아래 그림은 실습 4의 가정값(NRE 1억 달러, ASIC 3,000달러, GPU 30,000달러, PUE 1.3, 가동률 80%, $0.08/kWh, 3년)을 칩 경계(GPU 700 W, ASIC 300 W)에서 센 경우다. 화면본에서는 전력 경계를 칩·서버(양쪽)·서버(GPU만)로 바꾸고, 단가·전력·PUE·가동률·전기료·기간을 움직여 두 총비용 직선의 교차점 N^* 와 분모 가운데 전력비 몫이 어떻게 변하는지 볼 수 있다.

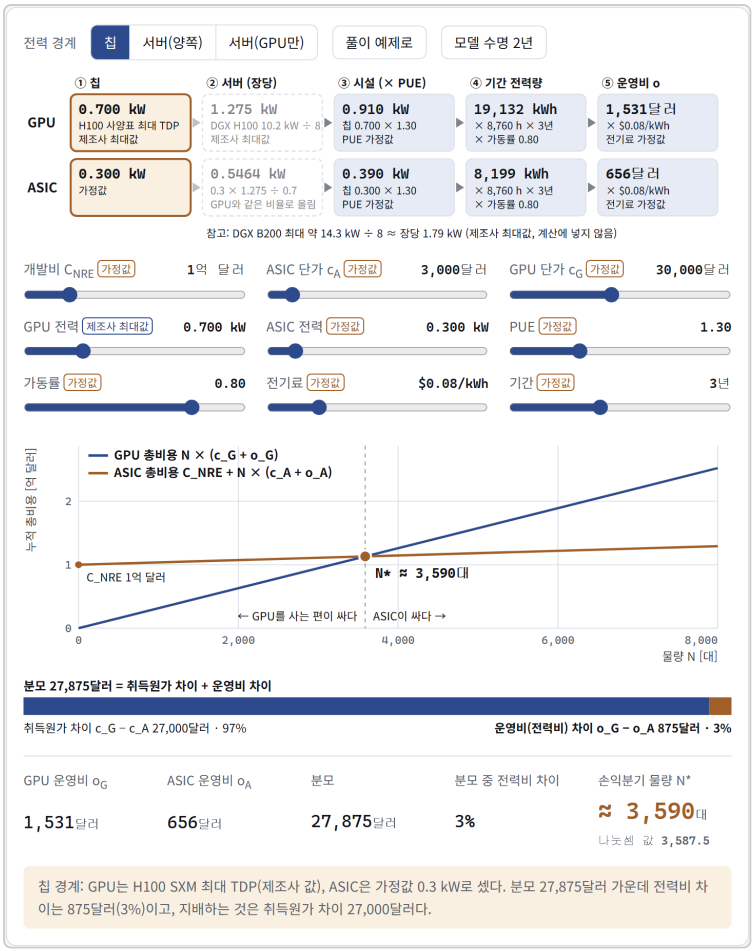


그림 2-4. 칩 경계(GPU 700 W, ASIC 300 W)와 실습 4의 가정값(NRE 1억 달러, 단가 3,000·30,000달러, PUE 1.3, 가동률 80%, $0.08/kWh, 3년)에서 두 총비용 직선은 N* ≈ 3,590대에서 만난다. 분모 27,875달러 가운데 전력비 차이는 875달러(3%)다. — 화면본에서는 이 자리에서 값을 바꿔 볼 수 있다(화면본 2회차 실습 4 풀이).

자습 부록 B — 토론 질문 생각의 실마리

질문	핵심 축	실마리
1. 루프라인의 한계	**정상상태 가정**	루프라인은 "충분히 오래 도는 커널"의 상한이다. 지연시간(첫 토큰까지의 시간), 캐시 계층(단일 B 로 뭉갠다), 노드 간 통신, 커널 실행·동기화 오버헤드, 전력 제한에 의한 클럭 하향은 전혀 담지 못한다. 6절의 디코딩 예는 가중치만 읽는다고 보았으므로 KV 캐시를 읽는 트래픽도 빠져 있고, 문맥이 길면 실제 연산강도는 더 낮다(실습 3). 저정밀 연산기의 피크도 단순화했다. 6절의 루프라인 그림은 INT8·INT4 가중치에도 P_{peak} 을 BF16 dense 값으로 두는데, 연산까지 같은 정밀도로 하면 피크와 연산강도가 함께 늘어 이용률은 fp16 때와 비슷해진다(실습 3 풀이). 실행 오버헤드의 크기는 NVIDIA의 2019년 CUDA Graphs 글이 V100·CUDA 10.1로 잰 예가 감을 준다. 실행에 2.9 µs 걸리는 짧은 커널을 띄울 때마다 동기화하면 커널당 9.6 µs가 들었고, 동기화를 단계마다 한 번으로 줄이면 3.8 µs, 20개 커널을 CUDA Graph 하나로 묶어 띄우면 3.4 µs였다. 현세대 GPU의 수치가 아니라 규모를 보여 주는 예다. 커널이 마이크로초 단위로 짧아지면 실행 시간보다 띄우고 기다리는 시간이 더 클 수 있다. **상한을 주는 모델이지 예측 모델이 아니다**
2. $I^* \approx 300$ 은 우연인가	**설계 의도**	우연이 아니다. 대역폭을 못 따라 올리면 연산 유닛이 노는 칩이 되므로 벤더가 균형을 맞춘다. 위로 벗어나면(연산 과잉) 실사용 효율이 떨어지고, 아래로 벗어나면(대역폭 과잉) HBM 원가가 이익을 잠식한다
3. MoE의 루프라인 효과	**읽는 가중치 vs 가진 가중치**	총 파라미터는 늘어 **용량** 압박이 커지고, 활성 파라미터는 줄어 **디코딩 상한**이 올라간다. 즉 용량 제약과 대역폭 제약을 맞바꾸는 구조다. 다만 라우팅이 불규칙 접근을 만들어 실효 대역폭 이용률이 떨어지고, 배치 안에서 전문가가 흩어지면 이득이 반감된다
4. 소프트웨어 락인의 지속성	**누가 커널을 쓰는가**	락인의 실체는 언어가 아니라 **최적화된 커널의 축적**이다. Triton·XLA가 커널 작성 비용을 낮추면 그 축적의 가치가 줄어든다. 반대로 하드웨어가 특수해질수록(전용 명령, 특수 메모리 계층) 컴파일러가 흡수하기 어려워져 락인이 오히려 강화된다
5. 추론 전용 칩 설계	**리지 포인트를 어디에 둘 것인가**	디코딩은 $I \approx B_{\text{sz}}$ 로 낮으므로, 연산을 줄이고 대역폭·용량에 예산을 몰아 I^* 를 **낮게** 설계하는 것이 합리적이다. 정밀도도 INT8/INT4 위주로 고정하고, 남는 면적은 온칩 SRAM에 쓴다. 실제 추론 특화 칩들이 취한 방향이다

자습 부록 C — 핵심 용어

용어	영문	한 줄 정의
루프라인 모델	roofline model	성능 상한을 연산 한계와 대역폭 한계의 최솟값으로 나타낸 모형
리지 포인트	ridge point	연산 바운드와 대역폭 바운드가 갈리는 연산강도 $I^* = P_{\text{peak}} / B$
연산 바운드	compute-bound	연산 유닛이 한계인 상태 ($I > I^*$)
대역폭 바운드	memory-bound	메모리 전송이 한계인 상태 ($I < I^*$)
GEMM	general matrix multiply	일반 행렬-행렬 곱셈
프리필 / 디코딩	prefill / decode	입력 전체를 한꺼번에 처리하는 단계 / 토큰을 하나씩 생성하는 단계
배치	batch size	동시에 처리하는 시퀀스 수
타일링·블로킹	tiling / blocking	데이터를 온칩에 올릴 크기로 잘라 재사용하는 최적화

용어	영문	한 줄 정의
커널 퓨전	kernel fusion	여러 연산을 합쳐 중간 결과를 메모리에 쓰지 않는 최적화
FlashAttention	—	어텐션을 퓨전해 HBM 왕복을 줄인 대표적 커널
MoE	mixture of experts	토큰마다 일부 전문가만 활성화하는 희소 모델 구조
GQA / MQA	grouped / multi-query attention	KV 헤드 수를 줄여 KV 캐시를 줄이는 어텐션 변형
시스톨릭 어레이	systolic array	데이터를 격자로 흘리며 곱셈-누산을 반복하는 연산 구조
희소성	structured sparsity	가중치 일부를 0으로 만들어 연산을 줄이는 기법. 스펙시트 수치 부풀리기의 주범
NRE	non-recurring engineering	칩 하나를 개발하는 데 드는 일회성 비용
텐서 병렬	tensor parallelism	한 층의 행렬을 여러 장의 가속기에 쪼개 얹는 방식
TTFT	time to first token	요청부터 첫 토큰까지의 시간. 프리필이 좌우한다
TPOT	time per output token	출력 토큰 사이 간격의 평균. 디코딩이 좌우한다
P99	99th percentile latency	요청의 99%가 그 이하로 끝나는 지연. 꼬리 지연의 지표
폴백	fallback	가속기가 지원하지 않는 연산을 CPU로 넘겨 실행하는 것. 장치-호스트 왕복이 붙는다

자습 부록 D — 자가진단

☐ $P_{\text{attainable}} = \min(P_{\text{peak}}, B \times I)$ 를 보지 않고 쓰고, 각 항이 무엇인지 말할 수 있다 → 6절 ★

☐ 임의의 칩 스펙시트를 받아 30초 안에 I^* 를 계산할 수 있다 → 실습 1 ★

☐ 프리필은 연산 바운드, 디코딩은 대역폭 바운드인 **이유**를 각각의 I 로 설명할 수 있다 → 6절

☐ 디코딩 상한식으로 "70B fp16을 H100 한 장에"가 왜 불가능한지 두 가지 이유를 댈 수 있다 → 실습 2

☐ 양자화가 왜 상한과 용량 **둘 다** 개선하는지 설명할 수 있다 → 실습 2

☐ KV 캐시 크기 식을 세우고, 용량이 어떻게 대역폭 효율을 제한하는지 설명할 수 있다 → 실습 3 ★

☐ 연산강도를 올리는 방법과 대역폭을 올리는 방법을 구분해 각각 세 개씩 댈 수 있다 → 6절 마지막 표

☐ ASIC 손익분기에서 **전력비가 아니라 취득원가가 지배한다**는 것을 계산으로 보일 수 있다 → 실습 4

☐ 루프라인이 설명하지 **못하는** 것을 세 가지 댈 수 있다 → 부록 B

★ 세 개를 통과하지 못하면 3회차로 넘어가지 말 것. 3회차 전체가 "그래서 B 를 어떻게 올릴 것인가"에 대한 답이다.

다음 회차로 넘어가기 전에

이 회차의 결론은 한 문장이다. **연산강도가 리지 포인트에 못 미치는 일, 대표적으로 배치가 작은 디코딩에서는 성능을 연산이 아니라 대역폭이 정한다.** 프리필과 학습처럼 연산강도가 높은 일은 여전히 연산 바운드다.

그러면 자연히 다음 질문이 나온다.

이 회차가 남긴 질문	3회차의 답
B 를 올리려면 무엇을 해야 하나?	버스를 넓히거나 핀 속도를 올리거나 — 그리고 왜 전자를 택했는지
그것이 왜 그렇게 비싼가?	적층·TSV·KGD 선별·본딩, 그리고 적층 수율의 잔인한 산수
용량 제약(실습 3)은 어떻게 푸나?	HBF와 메모리 계층화

출처

스터디 커리큘럼이 고른 입문 자료 ("AI 칩 (연산 반도체)" 항목)

커리큘럼 소주제	링크	이 책에서의 활용
AI칩이란	엘리먼트코리아 — 「AI칩이란 무엇인가? 전 세계가 주목하는 이유와 NPU GPU 차이까지」	입문용 개념 정리. 2·5절의 CPU/GPU/NPU 구분에 대응하는 쉬운 설명. 이 책은 여기서 더 나아가 에너지·오버헤드 관점으로 재구성했다
AI칩이란	MSAP — 「AI 가속기란 무엇인가? GPU, NPU, TPU 대세는?」	AI 가속기 유형 개관. 5절 유연성-효율 스펙트럼의 배경
AI칩이란	ScienceON JAKO202412457639314 — 「국내외 인공지능 반도체에 대한 연구 동향」	국내 학술 문헌. AI 반도체 기술 개관
AI칩 시장	Tilnote — 「엔비디아 GPU vs 구글·아마존 AI 칩, 뭐가 다른가?」	시장 동향 요약. 1회차 4절의 "수치 비판적 읽기" 실습 대상으로 활용 권장
AI칩 시장	ScienceON JAKO202504350488400 — 「AI 반도체와 주요 응용 분야에 관한 연구」	국내 학술 문헌. AI 반도체 시장 분석
AI칩 시장	InitNode(처음 eigenstate.dev에 실렸다가 옮겨진 글) — "CPU vs GPU vs TPU vs NPU: AI Hardware Architecture Guide 2026"	네 가지 아키텍처를 비교한 심층 영문 해설. **3·4절의 구조 비교와 가장 직접적으로 대응하는 자료로, 예습·복습 모두에 권장**

검증 완료 (2026-08-02)

- Horowitz(ISSCC 2014) 에너지 수치 — 널리 인용되는 표준 참조값(어느 값이 원 발표에 있는지는 아래 2026-10-08 항목). **45nm 공정 기준이므로 절대값이 아니라 비율의 크기(두 자릿수 배)를 이해하는 용도로 사용**
- NVIDIA H100 SXM: BF16 dense 989 TFLOP/s, HBM3 80GB, 3.35 TB/s — 제조사 공표 사양
- NVIDIA B200: BF16 dense 2,250 TFLOP/s, HBM3e 192GB, 7.7 TB/s — 제조사 공표 사양(Blackwell 아키텍처 기술 개요 V2.1 표 3, HGX B200). **192GB·7.7 TB/s는 칩 공표값(최대)이다.** DGX B200 시스템은 GPU당 180GB, 8.0 TB/s(합계 64 TB/s ÷ 8)다. 시스템 사양은 다르다 — 아래 2026-10-08 항목
- TPU v1 256×256 MAC 어레이 구성 — Jouppi et al. ISCA 2017 원논문

검증 완료 (2026-10-08) — 아래는 모두 1차 출처를 직접 열어 확인했다.

- B200 메모리 용량의 칩·시스템 구분 — NVIDIA Blackwell 아키텍처 기술 개요 V2.1의 HGX B200 GPU당 「최대 192 GB HBM3e」, DGX B200 사용자 안내서·DGX B200 데이터시트의 8장 합계 1,440 GB(계산하면 GPU당 180 GB)
- 전력 경계 — H100 SXM 최대 TDP 700 W(H100 사양표), DGX H100 최대 10.2 kW(DGX H100/H200 사용자 안내서), DGX B200 최대 약 14.3 kW(DGX B200 데이터시트·사용자 안내서). 실습 4의 서버 경계 재계산은 이 값과 실습 4의 가정값으로 다시 했다
- H100 SXM PCIe Gen5 128 GB/s·GPU 메모리 3.35 TB/s — H100 사양표의 공식 표기. H100 백서 V1.04는 128 GB/s를 양방향 합계(한 방향 64 GB/s)로 밝힌다
- 폴백 동작 — ONNX Runtime 실행 제공자 문서의 우선순위 목록 설명, PyTorch 2.14 MPS 환경변수 문서의 `PYTORCH_ENABLE_MPS_FALLBACK`
- `torch.compile` 의 구성(TorchDynamo·TorchInductor·Triton) — PyTorch 2.14 torch.compiler 문서
- 커널 실행 오버헤드 2.9·9.6·3.8·3.4 µs — NVIDIA 기술 블로그 「Getting Started with CUDA Graphs」(2019-09-05) 본문. Tesla V100, CUDA 10.1
- Intel AMX 타일 구성(8개 · 합계 8 KB · 최대 16행 × 64바이트)과 TMUL — Intel AMX 코드 예제 문서
- MLPerf Inference Llama2-70B 과제의 정확도 하한(FP32 기준 99.9%)·TTFT/TPOT 제약, TTFT·TPOT 정의, Server·Interactive 시나리오의 99번째 백분위수 — MLCommons 추론 규칙 master 판(마지막 수정 2026-08-20, 커밋 ff7edba)
- DeepSeek-V3의 H800 2,048장, 2.788M GPU시간(2,664K + 119K + 5K), 시간당 2달러 가정 557.6만 달러와 사전 연구·절제 실험 비용 제외 단서, DualPipe, SM 132개 중 20개 — arXiv:2412.19437 v2 본문
- 시스톨릭 어레이 박자 수와 TPU v1 크기 이용률(0.2%·약 89%) — 위젯과 같은 식 $M / (M + K + N - 2)$로 다시 계산
- 1절 에너지 표 — Horowitz 발표문(ISSCC 2014 논문 1.1) 그림 1.1.9 「Rough energy costs for various operations in 45nm 0.9V」 공개 사본에서 8비트 정수 덧셈 0.03 pJ, 32비트 부동소수 덧셈 0.9 pJ·곱셈 3.7 pJ, 캐시(64비트 접근) 8 KB 10 pJ·32 KB 20 pJ·1 MB 100 pJ, DRAM 1.3~2.6 nJ을 확인. 32비트 SRAM 5 pJ·DRAM 640 pJ은 이 그림에 없고, Han 외(2015) arXiv:1506.02626 v3의 그림 1(「Energy table for 45nm CMOS process」, Horowitz 인용)에 있다

- 메모리 코얼레싱의 32바이트 트랜잭션, 연속 4바이트 × 32스레드 = 128바이트 = 트랜잭션 4회, 최악 32회 — CUDA Programming Guide §2.3.4.1. 같은 장 §2.3.3.2·§2.3.3.6이 L1·L2 캐시와 공유 메모리를 설명한다
- H100 SXM INT8 텐서 코어 3,958 TOPS(희소성 적용, dense는 그 절반 1,979로 BF16 dense 989의 2배) — H100 사양표. 사양표에 INT4 값은 없다(실습 3의 INT4 피크는 가정)

미확인 · 유보

- ASIC NRE·마스크셋 비용의 구체적 금액 — 노드·업체·계약에 따라 편차가 매우 크고 공개되지 않는 경우가 많다. 본문에서는 자릿수 규모로만 언급하고, 실습에서는 **가정값**임을 명시했다
- 업체별 AI 가속기 점유율 — 1회차 과 동일하게 단정하지 않음
- 5절 실습의 단가·전력비 — 전부 예시용 가정값이며 실제 시장가가 아니다

3회차

AI 메모리 — DRAM · HBM · HBF

핵심 TAKEAWAY

HBM의 기괴해 보이는 형태(수직 적층 + 2,048비트 초광폭 버스 + 실리콘 인터포저)는 디자인 선택이 아니라 물리적 제약의 필연적 귀결이다. 대역폭을 올리는 길은 "신호를 빠르게" 또는 "선을 많이" 둘뿐인데, 전자는 신호 무결성과 전력에서 막혔다. 그래서 후자를 극단까지 밀었고, 선을 수천 개 뽑으려면 물리적으로 아주 가까이 붙여야 해서 TSV와 인터포저가 필수가 되었다. **메모리 문제가 곧 패키징 문제가 된 이유가 여기 있다.** HBM이 비싼 이유는 수율이 곱해지기 때문이고(12단 적층 = 다이 수율의 12제곱), HBF가 등장한 이유는 대역폭과 별개로 **용량**이 독립적인 병목이기 때문이다.

원자료 바로가기

반도체네트워크 「AI 시대 메모리 반도체의 진화… 플래시·SRAM·DRAM에서 HBM까지」 · Rambus 「High Bandwidth Memory (HBM): Everything You Need to Know」 · HyperAccel 「HBF 이해하기」 · ETRI 전자통신동향분석 41권 3호 「AI 메모리 반도체 분야 주요 기술 및 시장 동향」 · JEDEC JESD270-4(HBM4) 보도자료 · 삼성전자 HBM4 양산 출하 보도자료(2026-02-12) · SK하이닉스 HBF 첫 표준 사양 공개(2026-08-04) · Micron 16Gb DDR4 데이터시트 Rev. H

학습목표

1. SRAM·DRAM·NAND의 소자 구조 차이가 왜 그런 속도·용량·비용 특성을 낳는지 설명할 수 있다.
2. DRAM의 지연시간이 20년간 거의 개선되지 않은 이유를 설명할 수 있다.
3. HBM의 구조를 "왜 이렇게 생겼나"의 관점에서 재구성할 수 있다.
4. 적층 수율의 곱셈 구조를 계산하고, HBM 원가 구조를 설명할 수 있다.
5. HBF가 겨냥하는 병목이 HBM과 어떻게 다른지 구분할 수 있다.

0. 자습 안내 — 혼자 읽는 사람에게

이 회차는 2회차의 답변이다

2회차가 "배치가 작은 디코딩에서는 성능을 대역폭이 정한다"로 끝났다. 그러면 대역폭을 올리면 되지 않느냐 — 이 회차는 그 "올리면 되지 않느냐"가 왜 산업 전체를 뒤흔드는 문제인지를 보여준다. **적층 수율 산수(실습 1)를 해 보면 HBM 가격이 왜 그런지 숫자로 보인다.**

직전 회차에서 가져오는 것

2회차에서	이 회차에서
$I^* = P_{\text{peak}} / B$	B를 물리적으로 어떻게 만드는가
"디코딩은 대역폭 바운드"	그래서 HBM이 필요하다
실습 3의 용량 제약	그래서 HBF(대용량) 논의가 나온다

선행 지식 점검

개념	한 줄 답
DRAM 셀	커패시터 하나에 전하를 담아 0/1을 저장한다. 새면 사라지므로 주기적으로 다시 써 준다(리프레시)
NAND 플래시	전하를 절연막 안에 가둬 전원이 꺼져도 유지한다. 대신 느리고 쓰기 수명이 있다
버스 폭 / 전송률	한 번에 몇 가닥으로 보내는가 / 한 가닥이 초당 몇 번 보내는가. **대역폭 = 둘의 곱**
TSV	through-silicon via. 실리콘을 수직으로 관통하는 전기 통로
pJ/bit	1비트를 옮기는 데 드는 에너지. 대역폭 시대의 진짜 통화 단위

읽는 순서 (총 110분 권장)

순서	절	시간	비고
1	메모리 계층·DRAM 기초	25분	배경
2	**대역폭을 만드는 두 방법**	20분	실습 2와 함께 읽으면 훨씬 빠르다
3	**HBM 구조와 적층 수율**	30분	**정독.** 실습 1을 여기서 바로 풀 것
4	6절 HBF	20분	용량이라는 다른 축
5	7절 PIM/PNM + 부록	15분	

이 회차를 관통하는 한 문장

> **대역폭과 용량은 서로 다른 제약이고, 서로 다른 기술로 푼다.** 대역폭은 HBM(가깝고 넓게), 용량은 HBF·계층화(멀지만 크게). 둘을 섞어 생각하면 이 회차 내내 헷갈린다.

막히기 쉬운 지점 — 「왜 넓고 느리게」인가

직관적으로는 핀 속도를 올리는 쪽이 쉬워 보인다. 핀을 1,024개 심는 것보다 한 가닥을 두 배 빠르게 하는 게 간단하지 않은가?

아니다. 전송률을 올리면 **비트당 에너지가 비선형으로 늘고**(등화 회로, 클럭 복원, 드라이버 세기), 신호 무결성 문제가 급격히 나빠진다. 대역폭 4 TB/s를 1,024개의 32 Gbps 핀으로 만들면 인터페이스만으로 스택 하나가 100W 가까이(3 pJ/bit로 잡으면 96W) 먹을 수 있다. 가속기 한 장에 스택 8개면 메모리 인터페이스에만 약 770W로, GPU 자체의 전력(700~1,000W)과 맞먹는 열원이 하나 더 생긴다. 실제 온도가 얼마나 오르는지는 그 열이 퍼지는 면적과 냉각 조건에 달렸지만, 메모리 쪽에 이만한 전력 예산을 주기는 어렵다. 그래서 HBM은 **느린 핀을 아주 많이** 쓰는 쪽을 택했고, 그 대가로 핀을 심을 자리가

필요해져 **적층과 TSV, 인터포저**가 따라왔다. 실습 2가 이 논리를 숫자로 확인한다.

---

1. 직관 잡기 — 메모리 계층은 물리적 타협의 지도

컴퓨터 구조를 배운 독자라면 "레지스터 → L1 → L2 → L3 → DRAM → SSD" 계층이 낯익을 것이다. 그런 수업에서는 이 계층이 **왜** 그런 크기·속도로 나뉘는지는 대개 다루지 않는다. 답은 소자 구조에 있다.

계층	소자 구조	셀당 트랜지스터	결과
SRAM (캐시)	교차 결합 인버터 래치	6개 (6T)	빠름, 비휘발성 아님, **면적 큼 → 용량 작고 비쌈**
DRAM (메인메모리)	트랜지스터 1 + 커패시터 1	1개 (1T1C)	셀이 작아 **밀도 높고 쌈**, 대신 느리고 리프레시 필요
NAND (스토리지)	전하 저장층 있는 트랜지스터, 수직 적층	1개. 대신 셀 하나에 3~4비트(TLC·QLC)	**초고밀도·비휘발성**, 대신 블록 단위 소거, 느림, 내구성 유한

한 문장 요약: 한 비트를 담는 데 드는 소자 수가 계층 전체의 성격을 결정한다. 비트당 트랜지스터가 SRAM 6개 → DRAM 1개(커패시터 하나 추가) → NAND 1/3~1/4개(셀 하나에 3~4비트)로 줄수록 싸고 촘촘해지고, 대신 느려지고 다루기 까다로워진다. NAND는 여기에 셀을 수직으로 수백 층 쌓아 같은 면적에 담는 비트를 한 번 더 늘린다. 소자 수를 줄이는 것과 층을 쌓는 것은 서로 다른 밀도 요인이다.

AI 가속기에서 본 계층 (2026년 10월 기준)

같은 소자가 AI 서버 안에서는 아래 순서로 놓인다. 위로 갈수록 지연이 짧고, 아래로 갈수록 크고 싸다. HBF는 스택 대역폭만 보면 호스트 DRAM보다 클 수 있지만, NAND의 읽기 지연이 길어 DRAM 아래 단에 둔다. SK하이닉스는 2026년 8월 FMS 기조연설에서 메모리 계층을 HBM·DRAM·NAND(HBF)·SSD 네 단으로 나눠 제시했다. 이 표는 그 순서에 칩 안의 SRAM을 맨 위에 더한 것이다.

계층	매체	놓이는 곳	맡는 일
온칩 SRAM	SRAM	가속기 다이 안	지금 계산 중인 타일
HBM	적층 DRAM	가속기 바로 옆(인터포저 위)	매 토큰 읽는 가중치·KV 캐시
DRAM	DDR·LPDDR(CXL 확장 메모리 포함)	CPU 쪽 메모리 슬롯·확장 장치	호스트 메모리, 넘치는 KV 캐시
HBF	적층 NAND	프로세서에 UCIe로 연결(6절)	자주 바뀌지 않는 대용량 데이터(제안 단계)
SSD	NAND	서버의 저장 장치	모델 파일·데이터셋

순서는 속도와 용량으로 매긴 것이고 물리적 거리 순서가 아니다. HBF는 DRAM보다 아래 단이지만 HBM처럼 가속기 가까이 붙이는 것을 상정한다. 비유하면 SRAM은 책상 위, HBM은 손 닿는 책꽂이, DRAM은 같은 방의 서가, SSD는 지하 서고다. HBF는 책상 옆에 새로 들인 대형 서고 칸이다 — 가까이 있지만 꺼내는 데 서가보다 오래 걸린다. 이 비유는 거리와 꺼내는 속도, 크기만 담는다. 쓸 때마다 닳는 NAND의 수명이나 블록 단위로만 지울 수 있는 성질은 책 비유에 없다.

2. DRAM의 물리 — 왜 느린가, 왜 안 빨라지는가

읽기 동작의 실제

DRAM 셀은 아주 작은 커패시터에 전하가 있는지 없는지로 1비트를 저장한다. 읽으려면:

1. **워드라인 활성화** — 해당 행의 접근 트랜지스터를 켠다
2. **전하 공유** — 셀 커패시터의 전하가 비트라인으로 흘러나가 비트라인 전압이 아주 조금 변한다. 셀 커패시턴스가 비트라인 커패시턴스보다 훨씬 작아서 변화량이 수십 mV 수준에 불과하다
3. **감지증폭 (sense amplifier)** — 기준 전압과 비교해 그 미세한 차이를 풀 스윙으로 증폭
4. **읽기 = 파괴** — 전하가 빠져나갔으므로 원래 값을 **다시 써넣어야 한다** (restore)
5. **프리차지** — 다음 접근을 위해 비트라인을 중간 전압으로 되돌린다

타이밍 파라미터의 의미

이 절차가 곧 DRAM 타이밍의 정체다.

파라미터	대응하는 물리 과정
tRCD (RAS to CAS delay)	워드라인 활성화 → 감지증폭 완료
tCL (CAS latency)	열 주소 지정 → 데이터 출력
tRP (row precharge)	프리차지 완료 시간
tRAS (row active time)	복원(restore)까지 포함한 행 개방 최소 시간
tREFI / tRFC	리프레시 주기 / 리프레시 소요 시간

핵심 사실: 위 과정들은 전하량과 커패시턴스가 지배하는 아날로그 물리 과정이라, 공정 미세화로 잘 빨라지지 않는다. 오히려 셀이 작아질수록 저장 전하가 줄어 감지가 더 어려워진다.

> **20년의 비대칭:** DRAM의 **대역폭**은 수십 배 늘었지만 **지연시간**은 거의 그대로다. 대역폭은 병렬화와 인터페이스 속도로 늘릴 수 있지만, 지연시간은 한 셀의 물리 과정이 결정하기 때문이다. 이 비대칭이 컴퓨터 구조 전체를 "지연을 감추는" 방향으로 몰아갔다(프리페치, 다중 뱅크, GPU의 대규모 스레드 병렬성).

리프레시 세금

전하는 누설되므로 주기적으로 읽고 다시 써야 한다. 리프레시 중에는 해당 뱅크에 접근할 수 없어 대역폭과 전력을 잠식한다. 용량이 커질수록 리프레시 부담이 커지는 구조라, 고용량 DRAM 설계의 상시 과제다.

앞의 4단계(읽은 뒤 복원)와 헷갈리기 쉽다. 복원은 **접근이** 일으키고, 리프레시는 **시간이** 일으킨다. 아무도 읽지 않는 셀도 정해진 시간 안에 한 번씩 다시 채워야 한다.

계산 예 — 세금이 얼마인가. Micron 16Gb DDR4 데이터시트(Rev. H, 2021)를 보자. 85°C 이하에서는 64ms 안에 REFRESH 명령을 8,192번 내야 하므로 명령 간격 tREFI는 7.8µs(64ms ÷ 8,192 = 7.8125µs)

다. 85~95°C에서는 32ms 안에 8,192번이라 tREFI가 절반인 3.9μs가 된다. 명령 하나가 칩을 붙잡는 시간 tRFC1은 밀도마다 다르다. 칩이 리프레시에 묶여 있는 시간의 비율은 대략 tRFC/tREFI다.

다이 밀도	TRFC1	85°C 이하 (TREFI 7.8MS)	85~95°C (TREFI 3.9MS)
2Gb	160 ns	2.0%	4.1%
4Gb	260 ns	3.3%	6.7%
8Gb	350 ns	4.5%	9.0%

밀도가 네 배가 되는 동안 세금은 두 배 넘게 늘었고, 뜨거워지면 다시 두 배가 된다. 이 비율은 tRFC1을 7.8125μs·3.90625μs로 나눠 계산한 값이고, 리프레시 동안 칩 전체가 다른 일을 못 한다고 본 단순 계산이다.

---

3. 대역폭을 올리는 두 가지 길, 그리고 막다른 골목

메모리 대역폭은 근사적으로

$$B = (\text{버스 폭}) \times (\text{핀당 전송률})$$

이다. 따라서 방법은 둘뿐이다.

도로에 빗대면 버스 폭은 차선 수, 핀당 전송률은 차가 달리는 속도다. 시간당 지나가는 차의 수를 늘리려면 차선을 늘리거나 차를 빨리 달리게 해야 한다. 비유가 놓치는 것도 있다. 도로는 차가 빨라진다고 차 한 대를 보내는 비용이 가파르게 오르지 않지만, 메모리 핀은 빠르게 보낼수록 한 비트를 보내는 에너지(pJ/bit)가 가파르게 는다. 아래 두 길이 갈리는 이유가 이 비용이다.

길 ①: 핀당 전송률을 올린다 (GDDR의 길)

- 신호를 더 빠르게 토글한다. GDDR6X·GDDR7은 다중레벨 신호(PAM3/PAM4)까지 동원한다.
- 얼마나 빨라졌나: JEDEC GDDR7 표준(JESD239, 2024-03-05 공표)은 전압 3단계로 2사이클에 3비트를 보내는 PAM3를 쓰고, 디바이스당 최대 192 GB/s를 낸다. 32비트 디바이스로 계산하면 핀당 48 Gbps로, HBM4 표준 핀 속도 8 Gbps의 6배다. 핀 속도를 못 올려서 HBM이 느린 핀을 쓰는 것이 아니다.
- **한계:** 주파수를 올리면 신호 무결성이 급격히 나빠진다. 기판 배선에서의 손실, 반사, 크로스토크가 주파수에 따라 악화되고, 이를 보상하는 등화(equalization)·직렬화 회로가 전력을 잡아먹는다. **전송 비트당 에너지(pJ/bit)가 올라가는 것이 결정적 문제다.**
- 데이터센터 규모에서는 이 전력이 곧 운영비이자 열이므로, 이 길만으로는 갈 수 없다.

길 ②: 버스 폭을 늘린다 (HBM의 길)

- DDR5 채널 하나가 64비트, GDDR이 32비트/디바이스인 반면 **HBM은 스택 하나가 1,024비트(HBM3까지), HBM4는 2,048비트**다.
- 같은 대역폭을 훨씬 낮은 주파수로 얻을 수 있어 **pJ/bit가 크게 낮아진다.**
- **한계:** 선을 수천 개 뽑으려면 그 선들이 갈 곳이 있어야 한다. 일반 PCB 기판으로는 이런 배선 밀도가 불가능하고, 패키지 핀 수도 부족하다.
- **단서 — 두 길을 섞은 표준도 있다.** JEDEC가 2026-07-13에 공표한 SPHBM4(JESD330-4)는 HBM4와 같은 DRAM 다이를 쓰되 베이스 다이를 바꿔, 데이터 신호를 2,048개에서 512개로 줄이고 4:1 직렬화로 같은 총대역폭을 내게 했다. 핀 수가 줄어 범프 간격이 넓어지므로 실리콘 인터포저 대신 일반 유기 기판에 올릴 수 있고, 메모리를 SoC에서 더 멀리 둘 수 있다. 「광폭 버스에는 인터포저가 필수」는 2,048개를 다 쓰는 경우의 이야기이고, 신호 한 가닥의 속도를 네 배로 올리는 대가를 치르면 길 ① 쪽으로 일부 되돌아갈 수 있다. 그 대가가 pJ/bit로 얼마인지는 보도자료에 없다.

그래서 나온 답

1. 선을 아주 많이 뽑으려면 → 배선을 극도로 촘촘히 할 수 있는 매개체가 필요하다 → **실리콘 인터포저**
2. 스택 안에서 여러 다이를 연결하려면 → **TSV(실리콘 관통 전극)**
3. 거리가 짧아야 전력·지연이 준다 → **GPU 바로 옆에 붙인다 (2.5D 패키징)**

이 순간 메모리 문제는 패키징 문제가 되었다. 6회차가 이 책에서 가장 중요한 회차인 이유다.

---

4. HBM 구조 정밀 해설

물리적 구성



세로 점선 = TSV (실리콘 관통 배선) · 점 = 마이크로범프 / 하이브리드 본딩

요소별 역할

- **DRAM 다이 적층:** 용량을 면적이 아니라 높이로 확보한다. 8단·12단·16단.
- **TSV:** 각 다이를 관통하는 수직 배선. 스택 전체를 하나의 넓은 버스로 묶는다.
- **베이스 다이:** DRAM 공정이 아니라 로직 공정으로 만드는 층. PHY(물리계층 인터페이스), 테스트 회로, 제어 로직이 들어간다. **HBM4에서는 이 베이스 다이를 파운드리 로직 공정으로 만드는 일이 실현됐다.** 삼성전자는 2026년 2월 양산 출하한 HBM4에 1c DRAM 코어와 4nm 로직 공정을 썼고, 2026년 5월 HBM4E 샘플 보도자료에서 이 베이스 다이가 자사 파운드리의 4nm 로직 다이이며 HBM4와 같은 코어·베이스 다이 조합이라고 밝혔다. 메모리 업체와 파운드리의 경계가 흐려지는 지점이다.
- **다이 간 접합 — 마이크로범프와 그다음:** 지금 양산 HBM은 마이크로범프로 다이를 잇고 그 사이를 메운다. SK하이닉스는 HBM4에 MR-MUF(쌓은 칩 사이에 액상 보호재를 흘려 넣어 굳히는 방식)를 썼고(2025-09 발표), 삼성전자는 HBM3E 12단에 TC-NCF(층마다 비전도성 필름을 깔고 열압착하는 방식)를 썼다(2024-02 발표). 범프 없이 구리끼리 직접 붙이는 하이브리드 본딩은 더 높은 단수를 위한 다음 후보로 개발 중이며, 2026년 10월 현재 이를 쓴 HBM의 양산을 알린 제조사 발표는 확인하지 못했다. 하이브리드 본딩 자체는 6회차에서 다룬다.
- **인터포저:** GPU와 HBM 사이의 수천 개 배선을 실제로 그어주는 실리콘 판.

세대별 사양

세대	인터페이스 폭	채널 수	스택 대역폭 — 표준 기준	스택 대역폭 — 양산 제품 (발표일)
HBM2E	1,024비트	8	~0.4~0.5 TB/s (대략)	
HBM3	1,024비트	16	~0.8 TB/s (대략)	
HBM3E	1,024비트	16		핀 최대 9.6 Gbps → 계산하면 약 1.2 TB/s (삼성 HBM4 보도자료의 비교값, 2026-02-12). 삼성 HBM3E 12H 보도자료(2024-02-27)는 최대 1,280 GB/s
HBM4	**2,048비트**	**32** (채널당 의사채널 2)	**8 Gbps → 최대 약 2 TB/s** (JESD270-4, 2025-04-16)	핀 11.7 Gbps(최대 13 Gbps), 스택당 최대 3.3 TB/s (삼성, 2026-02-12 양산 출하)
HBM4E				핀 14 Gbps(최대 16 Gbps), 스택당 최대 3.6 TB/s, 12단 48 GB (삼성, 2026-05-29 샘플 출하)

표준 열은 JEDEC 표준이 정한 핀 속도로 낼 수 있는 값이고, 제품 열은 제조사가 발표한 값이다. 2026년 10월 현재 양산 HBM4는 표준 핀 속도를 넘어섰다.

- **SK하이닉스는** 2025-09-12에 HBM4 개발 완료와 양산 체제 구축을 알렸다. 2,048개 I/O, 핀당 10 Gbps 이상(JEDEC 표준 8 Gbps 초과), 1b DRAM(10나노급 5세대), Advanced MR-MUF 공정이다.
- **삼성전자는** 2026-02-12에 HBM4 양산 출하를 알렸다. 핀 11.7 Gbps로 동작하고 13 Gbps까지 올릴 수 있으며, 12단 24~36 GB다. 11.7 Gbps를 2,048비트에 곱하면 약 3.0 TB/s, 13 Gbps면 약 3.3 TB/s로, 보도자료의 「스택당 최대 3.3 TB/s」는 13 Gbps 쪽 값이다.
- 같은 회사의 HBM4E 12단 샘플(2026-05-29)도 HBM4처럼 2,048비트라고 보고 14 Gbps를 곱하면 약 3.6 TB/s로 보도자료 값과 맞는다(보도자료는 인터페이스 폭을 적지 않았다).
- 표준 열의 HBM2E·HBM3 값과 HBM3E 행은 벤더·제품마다 핀 속도가 달라 대략적 범위다. 빈칸은

이번 판에서 1차 출처로 확인하지 않은 칸이다.

HBM4의 주요 변경점

- 인터페이스 폭 1,024 → **2,048비트** (2배)
- 독립 채널 16 → **32개**, 각 채널에 의사채널(pseudo-channel) 2개
- HBM3 컨트롤러와의 하위 호환 — 컨트롤러 하나로 HBM3와 HBM4를 함께 다룰 수 있게 인터페이스를 정의했다
- DRFM(directed refresh management) 도입 — 로우해머 완화와 신뢰성(RAS) 개선
- 4~16단 적층 지원, 다이 밀도 24Gb 또는 32Gb → 32Gb 16단이면 큐브당 64GB
- 코어 전압 하향(HBM3/3E의 1.1V → 1.0/1.05V), VDDQ는 0.7~0.9V 중 벤더 선택

> **주목:** HBM4가 주파수를 크게 올리는 대신 **폭을 2배로 늘린 것**은 3절에서 본 "길 ②"를 다시 한 번 선택했다는 뜻이다. 전력 효율 때문에 이 길이 계속 선택되고 있다.

왜 GPU마다 HBM 스택을 여러 개 붙이나

스택 하나로는 대역폭도 용량도 부족하기 때문이다. 최신 가속기는 GPU 다이 주변에 HBM 스택 6~8개를 배치한다. 한 예로 NVIDIA는 Rubin GPU가 12단 HBM4 스택으로 GPU당 최대 288 GB, 최대 22 TB/s를 낸다고 밝혔다(2026-07-21 기술 블로그, Blackwell 대비 대역폭 2.8배). 스택 수는 그 글에 없다. 36 GB 스택이라고 가정하면 8개이고, 그때 스택당 2.75 TB/s, 핀당 약 10.7 Gbps가 된다 — 어디까지나 역산한 추정이다. 그런데 스택을 늘리면:

- 인터포저 면적이 커진다 → 인터포저도 노광 필드 제약을 받는다 → CoWoS-L 같은 확장 기법 필요
- 패키지가 커지고 열밀도가 올라간다
- **패키징 캐파가 곧 AI 칩 출하량의 상한이 된다** (1회차 공급망 병목의 실체)

---

5. HBM은 왜 비싼가 — 수율의 곱셈

이 절은 통계 전공자에게 가장 직관적인 부분이다.

기본 계산

DRAM 다이 하나의 수율을 y 라 하고, n 단을 적층한다고 하자. 단순히 무작위로 골라 쌓는다면 스택 전체가 정상일 확률은

$$Y_{\text{stack}} = y^{n} \times \prod_{i=1}^{n-1} y_{\text{bond},\,i}$$

여기서 y_{bond} 는 각 접합 공정의 수율이다.

수치 예시 ($y = 0.99$, 본딩 수율 0.995)

적층 수	다이 항 y^n	본딩 항	스택 수율
4단	0.961	0.985	0.946
8단	0.923	0.966	0.891
12단	0.886	0.946	0.839
16단	0.851	0.928	0.790

스택 수율 열은 반올림하기 전 값끼리 곱한 것이라, 표에 적힌 두 열을 곱하면 끝자리가 하나 다를 수 있다.

개별 다이가 99%로 훌륭해도 16단이면 스택 5개 중 1개가 버려진다. 그리고 버려질 때는 정상 다이도 함께 버려진다. 다이 하나가 불량이면 나머지 정상 다이 15개가, 다이는 16개 모두 정상인데 접합 하나가 불량이면 16개 전부가 버려진다. 이것이 HBM 가격 구조의 핵심이다.

대응책

1. **KGD (Known Good Die) 선별** — 적층 전에 각 다이를 완전히 테스트해서 양품만 쓴다. 위 식의 y 를 1에 가깝게 만드는 방법. 다만 웨이퍼 레벨에서 완전한 테스트를 하려면 별도의 테스트 구조와 시간이 필요하고, 이 자체가 원가다.
2. **여분 회로(redundancy)** — 여분 행/열/TSV를 두고 결함을 대체한다. DRAM은 전통적으로 리페어에 능한 제품이다.
3. **본딩 공정 수율 개선** — 하이브리드 본딩이 다음 후보로 검토되는 동기 중 하나.

원가 구조 요약

HBM이 같은 용량의 일반 DRAM보다 몇 배 비싼 이유는 (a) 낮은 적층 수율, (b) TSV 형성·박막화·본딩 등 추가 공정, (c) 인터포저와 첨단 패키징 비용, (d) 공급이 3사로 제한된 시장 구조가 겹친 결과다.

> **토론거리:** 이 중 어느 항이 시간이 지나면 개선되고, 어느 항이 구조적으로 남는가?

---

6. HBF (High Bandwidth Flash) — 용량이라는 또 다른 병목

왜 나왔나

2회차에서 본 디코딩 상한식을 다시 보자.

$$\text{tokens/s} \lesssim \frac{B}{P \cdot b} \times B_{sz}$$

여기서 배치 B_{sz} 를 키우면 처리량이 올라가는데, 배치를 키우려면 **KV 캐시를 담을 메모리 용량**이 필요하다. 또한 모델 자체가 커지면 가중치를 담을 용량도 필요하다. **대역폭과 용량은 별개의 제약이다.**

HBM은 대역폭에 최적화되어 있지만 용량은 비싸다(다이당 24~32Gb 수준). 반면 NAND는 밀도가 압도적으로 높다. **"NAND를 HBM처럼 적층하고 광폭 버스를 붙이면 어떨까"** 가 HBF의 발상이다.

현황과 사양

- 2025년 8월 6일, **Sandisk와 SK하이닉스가 HBF 표준화 협력(양해각서)을 발표했다.**
- 2026년 2월 25일, 양사가 **OCP(Open Compute Project) 내 전용 워크스트림**으로 표준화 작업을 착수했다.
- 2026년 8월 4일, 양사가 **첫 HBF 표준 사양을 OCP로 공개했다**(SK하이닉스 발표). 8단·16단 두 구성으로 스택당 최대 512 GB, 대역폭은 세 등급(Grade 1~3)으로 약 0.4~3.0 TB/s, 프로세서와의 연결은 칩렛 연결 표준인 UCIe다. Google과 Tenstorrent가 컨소시엄에 참여했다. 중간 등급의 값은 공개 자료에서 확인하지 못했다.
- 이보다 앞선 Sandisk의 1세대 제품 목표(2025년 7월 자료)는 **다이당 256Gb(기가비트), 16단 스택 512GB, 읽기 대역폭 1.6 TB/s였고,** HBM4와 비슷한 풋프린트·전력 프로파일·스택 높이를 지향했다. 256Gb는 32GB이므로 16단이면 512GB가 된다. 1.6 TB/s는 위 표준의 등급 범위 안에 있는 한 제품의 목표치다.
- 일정: Sandisk는 2025년 8월에 2026년 하반기 첫 HBF 샘플, 2027년 초 HBF 탑재 AI 추론 장치 샘플을 목표로 제시했다. 2026년 8월 투자자의 날 발표 자료(Sandisk Investor Day 2026, HBF 발표 PDF)는 첫 HBF 메모리 다이의 테이프아웃과 2027년 첫 HBF 추론 제품 샘플을 로드맵으로 제시했고, 2026년 하반기 메모리 샘플 목표는 다시 적지 않았다.
- 기술 기반: Sandisk의 BiCS NAND와 CBA(CMOS directly Bonded to Array, CMOS 회로 웨이퍼를 셀 어레이 웨이퍼에 직접 붙이는 구조) + SK하이닉스의 HBM 패키징·3D 적층 역량.

거래의 구조 — 무엇을 주고 무엇을 받나

항목	HBM4 (참고)	HBF (2026-08 공개 사양)
스택 용량	수십 GB급 (양산 12단 24~36GB, 표준 최대 64GB)	**최대 512GB (8단·16단)**
읽기 대역폭	표준 약 2 TB/s, 양산품 최대 3.3 TB/s	세 등급, 약 0.4~3.0 TB/s
쓰기 특성	대칭적, 자유로움	**비대칭 — 쓰기 느림, 블록 소거 필요**
내구성	사실상 무제한	**P/E 사이클 유한**
지연시간	낮음	높음
용도	학습·추론 전반	**추론 시 가중치 상주 중심**

핵심 이해: HBF는 HBM을 대체하는 물건이 아니라 **메모리 계층에 한 단을 추가하는** 물건이다. 자주 바뀌지 않는 대용량 데이터(모델 가중치, MoE의 비활성 전문가, 긴 컨텍스트 캐시)를 HBF에 두고, 활발히 갱신되는 데이터는 HBM에 두는 구조를 상정한다.

두 계층을 섞으면 — 단순 모형

한 번의 디코딩 스텝에서 읽는 바이트 가운데 비율 f 가 HBF에서, 나머지 $1-f$ 가 HBM에서 온다고 하자. 두 곳을 차례로 읽는다고 보면 걸리는 시간은 두 몫의 합이므로 실효 대역폭은

$$B_{\text{eff}} = \frac{1}{(1-f)\,/\,B_{\text{HBM}} + f\,/\,B_{\text{HBF}}}$$

이고, 담을 수 있는 용량은 두 계층의 합 $C_{\text{HBM}} + C_{\text{HBF}}$ 다.

계산 예. HBM4 스택 하나(삼성 발표 최대 3.3 TB/s, 12단 36GB)와 HBF 스택 하나(512GB)를 짝지었다고 하자. HBF 대역폭은 공개 사양 범위의 아래 끝 0.4 TB/s를 가정값으로 쓴다. 읽는 바이트의 10%만 HBF에서 와도($f = 0.1$)

$$B_{\text{eff}} = \frac{1}{0.9 / 3.3 + 0.1 / 0.4} \approx 1.91 \text{ TB/s}$$

로, HBM만 읽을 때의 58% 수준이 된다. 같은 f 에서 HBF를 위 끝 3.0 TB/s 등급으로 가정하면 약 3.27 TB/s로 거의 손해가 없다. 용량은 어느 경우든 36 + 512 = 548GB다. 느린 계층에서 읽는 몫이 조금만 커져도 그 계층이 전체 속도를 정한다. 그래서 f 를 작게 유지할 수 있는 데이터 — 예컨대 MoE에서 이번 스텝에 쓰이지 않는 전문가 가중치처럼, 대부분은 담아 두기만 하고 가끔 읽는 것 — 가 HBF의 자리다 (실습 4).

이 모형은 두 계층을 순서대로 읽는다고 본 것이다. 두 곳을 동시에 읽어 겹치거나, 필요한 데이터를 미리 HBM으로 옮겨 두는(프리페치) 경우는 넣지 않았고, NAND의 긴 지연도 대역폭 하나로 뭉갰다. 대역폭 값은 모두 스택 하나 기준의 발표치·가정값이다.

아래 그림은 위 계산 예, 곧 $f = 0.1$, HBM4 3.3 TB/s·36GB, HBF 0.4 TB/s(가정값)·512GB일 때다. 화면본에서는 f 와 HBM 대역폭·용량, HBF 대역폭(공개 사양 범위 0.4~3.0 TB/s 안)을 바꿔 볼 수 있다.



그림 3-1. 읽는 바이트의 10%를 HBF(0.4 TB/s, 가정값)에서 읽으면 HBM4 스택(3.3 TB/s)과 섞은 실효 대역폭은 약 1.91 TB/s로 HBM만 읽을 때의 58%이고, 용량은 548 GB다. 띠는 HBF 공개 사양의 범위 0.4~3.0 TB/s다. — 화면본에서는 이 자리에서 값을 바꿔 볼 수 있다(화면본 3회차 6절).

남은 과제 — 아직 검증되지 않은 부분

- NAND 읽기 지연이 DRAM보다 훨씬 큰데, 이를 대역폭으로 상쇄할 수 있는 워크로드가 실제로 얼마나 되는가
- 쓰기 내구성 관리(웨어 레벨링)를 가속기 수준에서 어떻게 할 것인가
- 소프트웨어 스택(어떤 데이터를 어느 계층에 둘지 결정하는 정책)이 성숙할 것인가
- 표준화 진행 상황과 실제 채택 여부 — **2026년 10월 현재 첫 표준 사양은 공개됐지만 양산 제품은 없으며, 성능 수치는 사양 범위·목표치다. "확정된 사실"이 아니라 "진행 중인 시도"로 다뤄야 한다.** DRAM과 NAND 사이를 메우려던 앞선 시도인 인텔 Optane(3D XPoint)은 2022년 2분기에 사업 정리에 들어갔다. 새 계층 후보가 반드시 자리를 잡는 것은 아니다(부록 B).

> **용량 축의 다른 해법 — CXL**
>
> HBF가 가속기 가까이에 용량을 붙이는 길이라면, CXL(Compute Express Link)은 서버 쪽에서 메모리를 늘리는 길이다. CXL은 PCIe의 물리·전기 인터페이스를 빌려 쓰면서 CPU의 메모리 공간과 연결된 장치의 메모리 사이에 캐시 일관성을 유지하는 개방 표준이다. 서버 한 대에 메모리를 덧붙이거나(확장), 여러 서버가 메모리 묶음을 나눠 쓰는(풀링) 데 쓴다. CXL 컨소시엄은 2025-11-18에 CXL 4.0을 공개하며 전송 속도를 64 GT/s에서 128 GT/s로 두 배 올렸다. SK하이닉스는 FMS 2026에서 CXL 풀드 메모리로 KV 캐시를 여러 서버가 옮기고 재사용하는 시연을 했다. 1절 계층 표에서 CXL 메모리는 DRAM 단에 들어간다. x16 링크 하나를 계산하면 한 방향 약 256 GB/s(128 GT/s × 16 ÷ 8, 부호화 손실 무시)로 HBM4 스택 하나의 10분의 1 안팎이다. 그래서 매 토큰 읽는 데이터가 아니라 넘치는 KV 캐시처럼 가끔 꺼내 쓰는 데이터를 위한 자리다.

---

7. PIM / PNM — 연산을 메모리로 가져가기

루프라인 식 $P = \min(P_{\text{peak}}, B \times I)$에서 지금까지 본 접근은 I를 올리거나(소프트웨어) B를 올리는(HBM) 것이었다. 세 번째 접근은 **애초에 데이터를 옮기지 않는 것**이다.

- **PIM (Processing-in-Memory):** 메모리 다이 안에 간단한 연산기를 넣는다. DRAM 뱅크마다 곱셈누산기를 배치하면 내부 대역폭(외부 핀에 제약받지 않는)을 그대로 쓸 수 있다. 삼성 HBM-PIM, SK하이닉스 AiM 등의 시도가 있다.
- **PNM (Processing-Near-Memory):** 메모리 근처(예: 베이스 다이나 버퍼 칩)에 연산기를 둔다.

왜 아직 주류가 아닌가

- DRAM 공정은 로직 공정이 아니다. 같은 다이에 복잡한 로직을 넣으면 비용과 수율이 나빠진다.
- 프로그래밍 모델이 없다. 어떤 연산을 메모리로 내릴지 결정하는 컴파일러·런타임이 미성숙하다.
- 범용성이 낮아, 특정 연산(예: 임베딩 룩업, GEMV)에서만 이득이 크다.

그럼에도 주목할 이유: 삼성 HBM4처럼 베이스 다이를 파운드리 로직 공정으로 만드는 흐름(4절)은 PNM의 문을 여는 변화다. **메모리와 연산의 경계가 물리적으로 흐려지고 있다.**

---

8. 자주 하는 오해

오해 1: "HBM은 그냥 빠른 DRAM이다" → 셀 자체는 일반 DRAM과 크게 다르지 않다. 다른 것은 **인터페이스 폭과 패키징**이다. HBM은 메모리 기술이라기보다 패키징 기술에 가깝다.

오해 2: "HBM이 GDDR보다 항상 좋다" → 절대 대역폭과 전력효율에서는 우수하지만 원가가 훨씬 높고 용량 확장이 제한적이며 공급이 부족하다. 게임 GPU가 여전히 GDDR을 쓰는 데는 이유가 있다.

오해 3: "HBM 용량을 늘리면 메모리 문제가 해결된다" → 용량과 대역폭은 다른 문제다. 대역폭 쪽도 늘린다고 끝나지 않는다. P_{peak} 가 그대로일 때 B 를 키우면 리지 포인트 $I^{*} = P_{\text{peak}} / B$ 는 내려간다(예: 2회차 H100의 989 TFLOP/s ÷ 3.35 TB/s ≈ 295 FLOP/byte는 B 를 두 배로 하면 약 148 FLOP/byte가 된다). 리지 포인트가 내려오면 더 낮은 연산강도의 워크로드도 연산 바운드 쪽으로 넘어가지만, 디코딩처럼 연산강도가 새 리지 포인트보다도 낮은 워크로드는 여전히 대역폭에 묶인다.

오해 4: "DRAM도 미세화하면 계속 빨라진다" → 대역폭은 늘어도 지연시간은 개선되지 않는다. 오히려 셀 커패시터가 작아져 감지가 어려워지는 방향의 압력이 있다.

오해 5: "HBF는 HBM을 대체할 것이다" → 쓰기 특성과 내구성 때문에 대체가 아니라 계층 추가다. 그리고 2026년 10월 현재 첫 표준 사양만 공개됐을 뿐 양산 제품이 없다.

오해 6: "적층을 더 하면 용량을 선형 비용으로 늘릴 수 있다" → 양품 스택 하나의 용량은 단수에 비례한다(다이 용량 × 단수). 비례하지 않는 것은 그 스택을 얻는 비용이다. 선별 없이 쌓으면 스택 수율이 y^{n} 으로 떨어져, 양품 스택 하나를 얻는 데 드는 다이가 n / y^{n} 개로 단수보다 빨리 는다. 단수를 올릴수록 열 방출과 스택 높이(패키지 규격) 제약도 커진다.

오해 7: "HBM의 채널은 메인프레임의 I/O 채널 같은 것이다" → 이름만 같다. 메인프레임의 I/O 채널은 CPU 대신 입출력을 맡는 전용 처리 장치를 가리킨다. HBM의 채널은 스택의 넓은 버스를 독립적으로 동작하는 몫으로 나눈 단위다. 버스 폭 = 채널 수 × 채널 폭이어서, HBM3는 16채널 × 64비트 = 1,024비트, HBM4는 32채널 × 64비트 = 2,048비트다(채널 폭은 표준이 밝힌 버스 폭과 채널 수로 나눈 값). GDDR에도 같은 구조가 있어, JEDEC는 GDDR7에서 독립 채널을 GDDR6의 2개에서 4개로 늘렸다.

오해 8: "리프레시는 읽을 때 빠져나간 전하를 다시 채우는 것이다" → 그것은 복원(restore)이다. 복원은 읽기라는 접근이 일으키고, 리프레시는 시간이 일으킨다. 읽지 않는 셀도 누설 때문에 정해진 시간 안에 다시 채워야 한다. DDR4 기준으로 85°C 이하에서는 64ms 안에 REFRESH 명령 8,192번을 내야 한다(2절).

오해 9: "핀 속도는 물리적 한계에 막혀 더 올릴 수 없다" → 빨리 못 보내서가 아니라 빨리 보내면 비싸서다. GDDR7 표준은 이미 핀당 48 Gbps(디바이스당 192 GB/s를 32비트로 나눈 값)를 정해 두었다. 막는 것은 절대 한계가 아니라 속도를 올릴수록 커지는 채널 손실과 그것을 메우는 등화 회로가 키우는 pJ/bit, 그리고 짧아지는 신호 도달 거리다(3절).

9. 정량 실습

실습 1 — 적층 수율 다이 수율 y 를 0.95 / 0.98 / 0.995로 바꿔가며 8단·12단·16단 스택 수율을 계산하고, KGD 선별이 없을 때 폐기되는 정상 다이의 개수를 구하라. **"왜 HBM 업체가 테스트에 그렇게 투자하는가"의 답이 숫자로 나온다.**

실습 2 — 대역폭 설계 목표 대역폭 4 TB/s를 (a) 버스 폭 1,024비트, (b) 2,048비트로 각각 달성하려면 핀당 전송률이 얼마여야 하는가? 두 경우의 pJ/bit 차이가 시스템 전력에 어떤 영향을 주는지 정성적으로 논하라.

실습 3 — 메모리 예산 배분 특정 모델을 서빙할 때 가용 메모리를 가중치와 KV 캐시에 어떻게 배분하는 것이 처리량을 최대화하는가? 배치를 늘리면 KV 캐시가 늘고, 양자화하면 가중치가 준다. 최적점을 찾는 문제로 세워보라.

실습 4 — HBF 적합성 판정 어떤 워크로드가 HBF에서 이득을 볼지 판정하는 기준을 세워보라. (힌트: 읽기/쓰기 비율, 재사용 거리, 지연 민감도)

---

10. 토론 질문

1. HBM의 형태를 물리적 제약에서 재구성해 보라. 만약 신호 무결성 문제가 없었다면 HBM은 존재했을까?
2. HBM4에서 베이스 다이를 로직 파운드리 공정으로 만드는 흐름(삼성은 자사 파운드리 4nm 공정으로 실현했다)은 메모리 산업의 경쟁 구도를 어떻게 바꿀 수 있는가?
3. 메모리 계층이 SRAM–HBM–DRAM–HBF–SSD로 더 세분화되면 소프트웨어는 무엇을 해야 하는가? 이 결정을 자동화할 수 있는가?
4. PIM이 실용화되려면 무엇이 먼저 해결되어야 하는가? 하드웨어 문제인가 소프트웨어 문제인가?
5. 대역폭 경쟁이 계속 pJ/bit 개선에 의존한다면, 물리적 한계는 어디인가? 광연결(optical interconnect)이 답이 될 수 있는가? (7회차에서 이어서)

---

11. 후속 공부

- **메모리 시스템 교과서:** Jacob, Ng & Wang, *Memory Systems: Cache, DRAM, Disk* (Morgan Kaufmann, 2007) — DRAM 타이밍과 구조의 표준 참고서
- **DRAM 연구:** Onur Mutlu 교수(ETH Zürich) 홈페이지와 SAFARI 연구실의 강의 자료와 논문 — DRAM 지연시간, 리프레시, PIM에 관한 가장 체계적인 자료군이며 강의 영상이 공개되어 있다. **DRAM을 더 깊이 보려는 독자에게 권한다**
- **HBM 표준:** JEDEC JESD235(HBM), JEDEC JESD270-4(HBM4) — 표준 문서 자체(2026년 10월 현재 HBM4 표준 페이지는 개정판 JESD270-4A를 싣고 있다). 요약본은 Rambus·Synopsys의 기술

블로그가 접근성이 좋다

- **신호 속도 쪽 표준:** JEDEC JESD239(GDDR7) — PAM3로 핀 속도를 올린 길 ①의 현재 모습
- **HBF:** Sandisk·SK하이닉스 표준화 협력 보도자료(2025-08-06), SK하이닉스 HBF 첫 표준 사양 공개(2026-08-04), IEEE Spectrum 「High Bandwidth Flash Unlocks Massive Model Storage」(2026-07-14), OCP 워크스트림 문서
- **PIM:** 삼성 HBM-PIM, SK하이닉스 AiM 관련 ISSCC/HotChips 발표 자료
- **국내 기술동향:** ETRI 전자통신동향분석 41권 3호 「AI 메모리 반도체 분야 주요 기술 및 시장 동향」

자습 부록 A — 정량 실습 풀이

실습 1 풀이 — 적층 수율

▾ 풀이

본딩 수율을 1로 두고 다이 수율만 보면 $Y_{\text{stack}} = y^n$ 이다.

다이 수율 y	8단	12단	16단
0.95	66.3%	**54.0%**	44.0%
0.98	85.1%	78.5%	72.4%
0.995	96.1%	94.2%	92.3%

여기서 멈추면 절반만 본 것이다. 진짜 질문은 "선별 없이 쌓으면 멀쩡한 다이가 몇 개나 함께 버려지는가"다.

불량 스택 하나에 들어 있는 정상 다이의 기댓값은

$$E[\text{정상 다이} \mid \text{스택 불량}] = \frac{ny - ny^n}{1 - y^n}$$

$n = 12$, $y = 0.95$ 를 넣으면

$$\frac{12(0.95) - 12(0.5404)}{1 - 0.5404} = \frac{11.40 - 6.485}{0.4596} = 10.7 \text{ 개}$$

아홉 조합 모두에 같은 식을 쓰면 다음과 같다. 칸마다 앞의 값이 불량 스택 하나에 든 정상 다이의 기댓값이고, 괄호 안이 스택 1,000개를 만들 때 함께 버려지는 정상 다이의 기댓값 1,000 $(ny - ny^n)$ 이다.

다이 수율 y	8단	12단	16단
0.95	6.8개 (2,293개)	**10.7개 (4,916개)**	14.6개 (8,158개)
0.98	6.9개 (1,034개)	10.9개 (2,343개)	14.8개 (4,099개)
0.995	7.0개 (274개)	11.0개 (641개)	15.0개 (1,153개)

다이 수율이 오르면 버려지는 정상 다이의 총량(괄호)은 크게 줄지만, 불량 스택 하나에 든 정상 다이는 오히려 $n-1$ 에 가까워진다. 불량이 드물어질수록 불량 스택은 「다이 하나만 나쁜」 스택이 되기 때문이다.

12단 스택 1,000개를 만든다고 하자.

- 투입 다이 12,000개, 그중 정상 11,400개
- 완성되는 스택 540.4개(약 540개) → 살아남은 정상 다이 6,484.3개(약 6,484개)
- **버려진 정상 다이 4,915.7개(약 4,916개) — 정상 다이의 43%가 낭비된다**

(기댓값이라 소수가 남는다. 스택 수율 0.54036을 반올림하지 않고 곱한 값이다.)

이 숫자가 설명하는 것

1. **왜 KGD(known good die) 선별에 그렇게 투자하는가.** 선별을 완벽히 하면 위 43%가 0이 된다. 스택 1,000개마다 정상 다이 약 4,916개를 건지는 셈이므로, 다이 하나의 값을 c 라 하면 그만한 선별에 드는 추가 테스트 비용이 대략 $4{,}916c$ 보다 작을 때 선별이 이긴다(버려지는 스택에 들인 적층·접합 공정비까지 넣으면 문턱은 더 높아진다). 실제 테스트 비용과 다이 단가는 이 실습에 없다.
2. **왜 HBM 단수를 올리는 것이 그렇게 어려운가.** 8단에서 16단으로 가면 $y = 0.95$ 기준 수율이 66% → 44%로 떨어진다. 단수를 두 배로 올리려면 다이 수율을 0.95에서 0.975 수준까지 끌어올려야 같은 스택 수율이 유지된다.
3. **왜 HBM 공급이 돈을 더 준다고 빨리 늘지 않는가.** 패키징 캐파가 출하량의 상한을 정한다는 것은 4절에서 보았다. 이 계산은 그 위에 수율 학습곡선이라는 다른 제약을 하나 더 보여 준다. 수율이 낮으면 같은 캐파에서 나오는 양품 스택이 줄고, 수율은 설비를 사듯 돈으로 당겨 오기 어렵다. 이 계산에는 생산능력이나 증설 기간이 입력으로 없으므로 둘 중 어느 쪽이 지금 더 세게 묶고 있는지는 가리지 못한다. 1회차에서 본 병목의 이동이 실제로 어떤 모양인지가 이것이다.

> 본딩 수율까지 넣으면 더 나빠진다. $y_{\text{bond}} = 0.999$ 라도 12단이면 11번 접합해 $0.999^{11} = 0.989$ 가 곱해진다.

이 계산이 빼놓은 것

- 다이 결함이 서로 독립이라고 보았다. 실제 결함은 웨이퍼 위에서 뭉쳐 나타나는 경향이 있다.
- 「선별을 완벽히 하면」은 KGD 테스트가 불량을 하나도 놓치지 않는다는 극단이다. 테스트 비용과 놓치는 비율은 넣지 않았다.
- 여분 회로(redundancy)로 결함을 고치는 효과는 다이 수율 y 안에 이미 들어 있다고 본다.

아래 그림은 다이 수율 $y = 0.95$, 12단, 선별 없이 스택 100개를 만든 표본일 때다. 화면본에서는 다이 수율·적층 수·접합 수율과 적층 전 선별 여부를 바꿔 볼 수 있다.

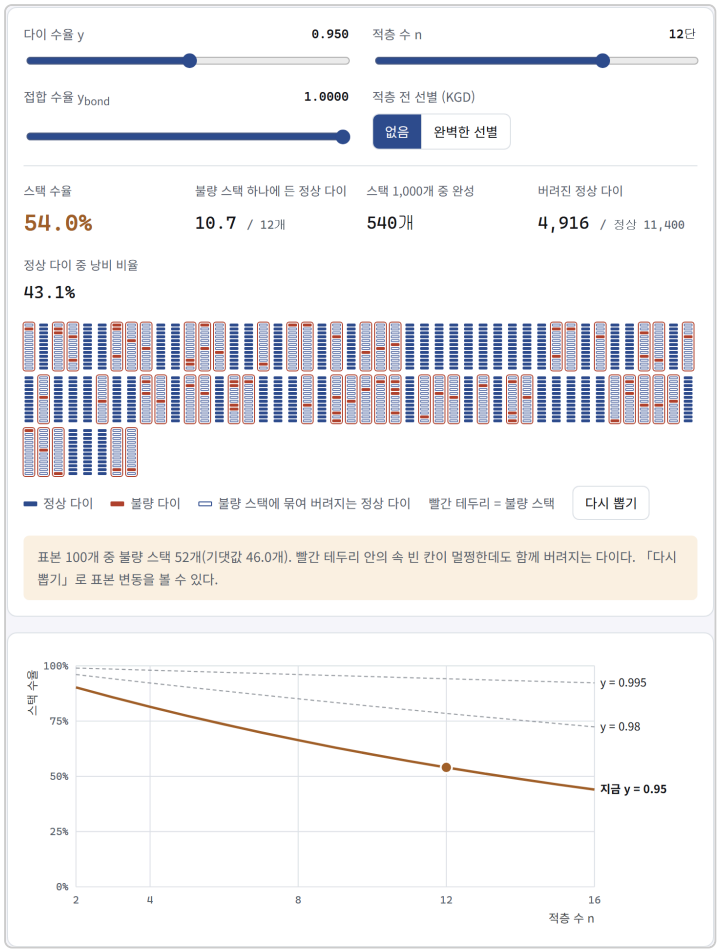


그림 3-2. 다이 수율 0.95로 12단을 선별 없이 쌓으면 스택 수율은 54.0%이고, 불량 스택 하나에 멀쩡한 다이가 평균 10.7개 묶여 버려진다. 막대는 무작위로 뽑은 스택 100개다. — 화면본에서는 이 자리에서 값을 바꿔 볼 수 있다 (화면본 3회차 실습 1 풀이).

실습 2 풀이 — 대역폭 설계

▾ 풀이

B = (버스 폭) × (핀당 전송률). 목표 4 TB/s $= 3.2 \times 10^{13}$ bit/s.

버스 폭	필요한 핀당 전송률	실현 방식
(a) 1,024 bit	$3.2 \times 10^{13} / 1024 =$ **31.25 Gbps**	SerDes 영역. 등화·CDR 회로 필수
(b) 2,048 bit	**15.6 Gbps**	병렬 인터페이스로 가능

pJ/bit 차이와 전력

전송률이 올라가면 비트당 에너지는 선형이 아니라 가파르게 는다. 등화 회로, 클럭·데이터 복원(CDR), 드라이버 세기, 종단 저항 손실이 모두 붙기 때문이다. 대략적인 감으로

방식	대략적 pJ/bit	4 TB/s에서의 전력
고속 SerDes (30 Gbps급)	3~5	**96~160 W**
광폭 병렬 (15 Gbps급)	0.5~1	**16~32 W**

3.2×10^{13} bit/s × 3 pJ/bit = 96 W. 스택 **하나**가 그렇다.

결론: 가속기 한 장에 스택 8개를 붙인다면 (a)는 범위의 아래 끝인 3 pJ/bit로 잡아도 메모리만으로 약 770W(96W × 8)로 800W에 육박하고, 위 끝 5 pJ/bit면 1,280W다. (b)는 0.5~1 pJ/bit 범위에서 약 130~260W다. 가속기 자체가 700~1,000W인 상황에서 이 차이는 시스템 설계를 바꾼다.
HBM이 "느린 핀을 아주 많이" 쓰는 이유가 전력이다.

대가: 2,048개 핀을 심으려면 다이 가장자리로는 불가능하고 **면적 배열**이 필요하다 → 마이크로범프 → 인터포저 → CoWoS. 6회차 실습 5에서 이 배선 수를 직접 세어 본다. 단, 핀을 2,048개 다 쓰는 경우의 이야기다. 3절 끝에서 본 SPHBM4처럼 신호를 512개로 줄이고 한 가닥을 네 배 빠르게 돌리면 유기 기판에 올릴 수 있다. HBM4 표준 8 Gbps의 네 배로 계산하면 한 가닥이 32 Gbps로, 이 실습의 (a)와 비슷한 속도다. 인터포저를 피하는 대신 (a) 쪽의 pJ/bit 부담을 지는 선택이다 (SPHBM4의 실제 전력 값은 JEDEC 보도자료에 없다).

이 계산이 빼놓은 것: pJ/bit 범위는 대략적인 감이지 벤더 실측치가 아니다. 핀 속도와 pJ/bit의 관계도 식으로 만들지 않았다 — 실제 값은 회로와 공정마다 다르다.

아래 그림은 스택당 목표 4 TB/s, 가속기 한 장에 스택 8개, (a) 3 pJ/bit, (b) 0.75 pJ/bit일 때다. 화면본에서는 목표 대역폭과 스택 수, 두 방식의 pJ/bit를 바꿔 볼 수 있다.

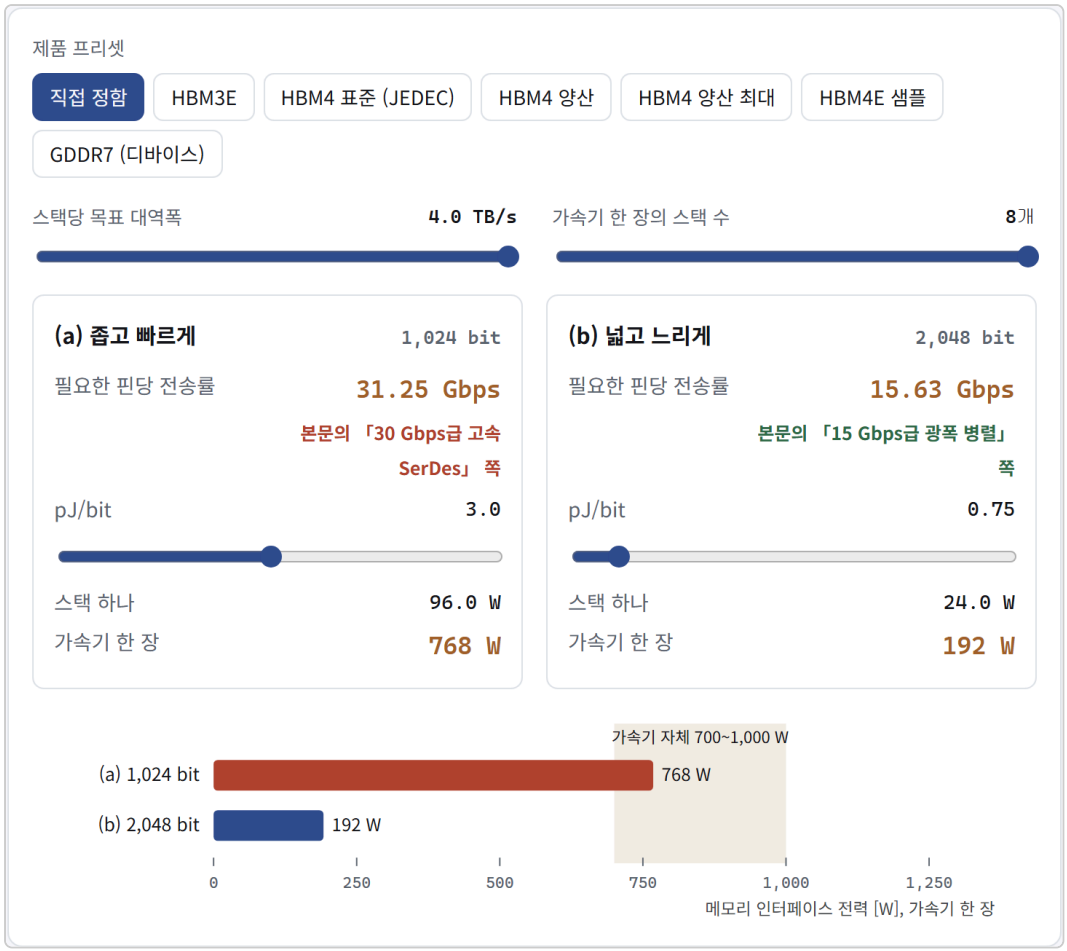


그림 3-3. 스택당 4 TB/s를 1,024비트 버스로 내려면 핀당 31.25 Gbps, 2,048비트로 내면 15.6 Gbps가 필요하다. 비트당 에너지를 3과 0.75 pJ/bit로 두면 스택 8개분 메모리 인터페이스 전력은 (a)가 800 W에 육박하고 (b)는 약 190 W다(0.5~1 pJ/bit 범위면 약 130~260 W). — 화면본에서는 이 자리에서 값을 바꿔 볼 수 있다(화면본 3회차 실습 2 풀이).

실습 3 풀이 — 메모리 예산 배분

▾ 풀이

문제를 식으로 세운다. 총 메모리 M, 가중치 $P \cdot b$, 시퀀스당 KV 캐시 k 일 때, 가중치가 먼저 들어가야 하므로($M \geq P \cdot b$, 아니면 서빙 불가) 배치는 남는 용량에 들어가는 시퀀스 수를 내림한 정수다.

$$B_{sz} = \left\lfloor \frac{M - P \cdot b}{k} \right\rfloor, \qquad \text{처리량} \lesssim \frac{B}{P \cdot b} \times B_{sz} \leq \frac{B\,(M - Pb)}{P\,b\,k}$$

맨 오른쪽은 내림을 뺀 연속 상한이다. 아래 성질은 이 연속 상한으로 읽고, 표의 배치는 내림한 값이다.

b 에 대해 보면 분자는 커지고 분모는 작아진다 — 이중 이득이다. 그래서 양자화의 효과가 직관보다 훨씬 크다.

수치 확인 (70B, $M = 80$ GB, $k = 1.34$ GB, $B = 3.35$ TB/s)

정밀도	가중치	배치	처리량 상한	배수
fp16	140 GB	—	들어가지 않음	—
INT8	70 GB	7	335 tok/s	×1
INT4	35 GB	33	**3,159 tok/s**	**×9.4**

정밀도를 절반으로 줄였는데 처리량이 2배가 아니라 9.4배다. 상한이 2배가 되는 효과에, 남는 용량이 10 GB → 45 GB로 늘어 배치가 4.7배가 되는 효과가 곱해지기 때문이다.

최적화 문제로서의 성질

- b 는 자유변수가 아니다. **정확도 손실이라는 제약** 아래에서 고를 수 있는 최소값이다.
- KV도 따로 양자화하면 k 가 줄어 추가 이득이 있다. 가중치보다 KV 양자화가 정확도에 더 관대한 경우가 많아 실무에서 먼저 적용된다.
- 시퀀스가 길어지면 $k \propto S$ 이므로 배치가 급감한다. **긴 컨텍스트 서비스가 비싼 이유**가 이 항이다.

얻어 갈 것: "메모리를 늘린다"는 말은 대역폭과 용량 중 무엇을 늘린다는 뜻인지 항상 구분해야 한다. 여기서는 **용량**이 처리량을 지배했다.

같은 계산은 2회차 실습 3 풀이 끝의 그림에서 해 볼 수 있다. 화면본에서는 시퀀스 길이와 정밀도를 바꿔 가며 최대 배치가 어떻게 변하는지 본다.

실습 4 풀이 — HBF 적합성 판정

▾ 풀이

NAND 기반 대용량 메모리가 이득을 보려면 그 워크로드가 NAND의 약점을 건드리지 않아야 한다. 판정 기준을 네 축으로 세운다.

축	HBF에 유리	HBF에 불리
읽기/쓰기 비율	읽기 지배적 (쓰기 거의 없음)	쓰기 빈번 — NAND는 쓰기 수명·지연이 취약
재사용 거리	길다 — HBM 캐시로 커버되지 않음	짧다 — 어차피 HBM에 머문다
접근 단위	크고 순차적	작고 무작위
지연 민감도	낮음, 또는 **미리 알 수 있어 프리페치 가능**	높고 예측 불가

적합한 것

- **MoE 전문가 가중치** — 대부분이 비활성이고 읽기 전용이다. 다만 라우터는 그 층에 들어온 숨은 상태를 보고 나서 전문가를 고르므로, 같은 층의 결정을 미리 알 수는 없다. 프리페치가 성립하려면 앞 층의 숨은 상태로 다음 층의 전문가를 예측해 미리 읽거나(예측이 빗나가면 그만큼 늦게 읽는다), 여러 층·여러 토큰의 계산을 겹쳐 읽기 시간을 감추는 스케줄이 있어야 한다. 이런 장치를 전제로 HBF의 가장 유력한 용례로 꼽힌다.
- **초장문 KV 캐시 가운데 덜 읽히는 부분** — 앞의 모든 위치를 보는 dense attention이면 새 토큰마다 이전 위치의 K·V를 전부 읽으므로, 오래된 토큰이라는 이유만으로 덜 읽히지는 않는다. 접근 빈도가 실제로 낮아지는 것은 슬라이딩 윈도·희소(sparse) attention이나 중요한 위치만 골라 읽는 정책을 쓸 때, 또는 지금 디코딩하지 않고 대기 중인 대화의 KV 캐시일 때다. 그런 조건에서 계층화의 대상이 된다.
- **RAG 벡터 인덱스·임베딩 테이블** — 대용량, 읽기 전용, 배치 접근.
- **오프라인 배치 추론** — 지연 요구 자체가 없다.

부적합한 것

- **학습 중 옵티마이저 상태** — 매 스텝 쓰기가 발생한다. NAND 수명이 즉시 문제가 된다.
- **대화형 디코딩의 핫 경로** — 매 토큰 필요한 활성 가중치. 지연이 그대로 사용자에게 보인다.
- **작고 무작위한 접근이 지배적인 것** — 페이지 단위 접근인 NAND에서 증폭이 심하다.

판정 규칙 한 줄: "읽기 전용이고, 크고, 언제 필요한지 미리 알 수 있는가." 셋을 모두 만족하면 HBF, 하나라도 어긋나면 HBM에 남긴다.

자습 부록 B — 토론 질문 생각의 실마리

질문 유형	갈리는 지점	실마리
HBM 과점은 지속되는가	**진입 장벽의 성질**	장벽이 특허라면 만료되고, 캐파라면 돈으로 풀리지만, **수율 학습곡선**이라면 시간으로만 풀린다. 실습 1이 보여 주는 것은 셋째 유형의 무게다 — 수율이 곱해지는 구조에서는 다이 수율의 작은 차이가 스택 수율에서 크게 벌어지고, 후발주자는 선발주자의 누적 학습을 돈으로 건너뛰기 어렵다. 세 장벽 가운데 어느 것이 지금 가장 큰지는 이 계산으로 가리지 못한다
HBF는 HBM을 대체하는가	**대체 vs 계층 추가**	대체가 아니라 **계층이 하나 늘어나는 것**으로 보는 편이 정확하다. 메모리 계층은 역사적으로 사라진 적이 없고 늘어나기만 했다(레지스터→캐시→DRAM→SSD). HBF는 HBM–DRAM–HBF–SSD 순서에서 DRAM과 SSD 사이의 빈칸을 메우는 후보다. 다만 새 단이 늘 자리를 잡은 것은 아니다. 같은 빈칸을 노린 인텔 Optane(3D XPoint)은 2022년 2분기에 사업 정리에 들어갔다. HBF도 같은 시험대에 서 있다
PIM은 왜 아직 안 되나	**생태계 vs 소자**	장벽은 두 겹이다. 하나는 하드웨어로, DRAM 공정에 로직을 넣으면 비용과 수율이 나빠진다(7절). 다른 하나가 여기서 강조하는 **프로그래밍 모델**이다. 연산을 메모리로 보내려면 컴파일러·프레임워크·커널이 전부 바뀌어야 하는데, 2회차 7절의 락인이 정확히 여기서 작동한다. 하이퍼스케일러가 자체 칩에서는 PIM을 시도할 수 있는 이유도 같다
메모리 가격은 정상인가	**사이클 vs 구조 변화**	1회차 토론 5와 같은 질문이다. 다만 여기서는 근거가 하나 더 있다 — HBM은 일반 DRAM 캐파를 잠식하므로, **AI 수요가 일반 메모리 공급을 줄이는 경로가** 존재한다. 이것은 과거 사이클에 없던 구조다

자습 부록 C — 핵심 용어

용어	영문	한 줄 정의
대역폭 / 용량	bandwidth / capacity	초당 옮기는 양 / 담아 두는 양. **이 회차 내내 구분해야 하는 두 축**
HBM	high bandwidth memory	DRAM 다이를 수직 적층하고 광폭 버스로 연결한 메모리
HBF	high bandwidth flash	NAND를 같은 방식으로 적층해 대용량을 노리는 제안. 2026년 8월 첫 표준 사양 공개, 양산 제품은 아직 없다
CXL	Compute Express Link	PCIe의 물리 계층 위에서 CPU와 장치 사이 캐시 일관성을 유지하는 연결 표준. 서버 메모리 확장·풀링에 쓴다
TSV	through-silicon via	실리콘을 관통하는 수직 배선
KGD	known good die	적층 전에 검증을 마친 정상 다이
마이크로범프	microbump	다이 사이를 접합하는 미세 솔더 돌기
하이브리드 본딩	hybrid bonding	범프 없이 구리-구리를 직접 접합하는 방식. 피치와 열 저항 모두에서 유리. HBM에서는 2026년 10월 현재 다음 세대 후보
MR-MUF / TC-NCF	mass reflow molded underfill / thermal compression non-conductive film	양산 HBM의 적층 방식. 쌓은 뒤 액상 보호재를 한꺼번에 채워 굳히는 쪽(SK하이닉스)과 층마다 필름을 깔고 열압착하는 쪽(삼성)
인터포저	interposer	여러 다이를 얹어 서로 연결해 주는 중간 기판

용어	영문	한 줄 정의
리프레시	refresh	DRAM 커패시터의 전하가 새기 전에 정해진 주기로 다시 써 주는 동작. 시간이 일으키며, 읽은 직후의 복원(restore)과 다르다
뱅크	bank	DRAM 다이 안에서 따로 열고 닫을 수 있는 셀 배열 묶음. 내부의 병렬 접근 단위
채널	channel	메모리 인터페이스를 독립적으로 동작하는 몫으로 나눈 단위. 버스 폭 = 채널 수 × 채널 폭(HBM4: 32 × 64비트). 메인프레임의 I/O 채널과는 다른 말이다
pJ/bit	picojoule per bit	1비트 전송 에너지. 대역폭 설계의 실질적 제약
PIM / PNM	processing in / near memory	연산기를 메모리 안에 / 메모리 옆에 두는 구조
적층 수율	stack yield	스택 전체가 정상일 확률. y^n 으로 급락한다
메모리 계층화	memory tiering	접근 빈도에 따라 데이터를 서로 다른 매체에 배치하는 기법

자습 부록 D — 자가진단

☐ 대역폭과 용량이 **다른 제약**임을 예를 들어 설명할 수 있다 → 이 회차 전체 ★

☐ $B =$ 버스 폭 × 전송률에서, 왜 산업이 "넓고 느리게"를 택했는지 **전력으로** 설명할 수 있다 → 실습 2 ★

☐ 8단·12단·16단 적층 수율을 암산 수준으로 어림할 수 있다 → 실습 1

☐ **KGD 선별이 왜 필수인지 숫자로** 말할 수 있다 → 실습 1 ★

☐ HBM 공급이 왜 돈으로 빨리 늘지 않는지 설명할 수 있다 → 실습 1

☐ HBF가 어떤 워크로드에 맞는지 판정 기준 세 가지를 댈 수 있다 → 실습 4

☐ MoE가 왜 HBF의 유력한 용례인지 설명할 수 있다 → 실습 4

☐ PIM을 막는 장벽을 공정(하드웨어)과 프로그래밍 모델(생태계)로 나눠 말할 수 있다 → 7절

다음 회차로 넘어가기 전에

여기까지 세 회차로 **무엇이 문제인지를** 다 봤다. 4회차부터는 **누가, 어떻게 만드는가로** 넘어간다.

이 회차가 남긴 질문	이후 회차의 답
적층 수율이 그렇게 중요하면, 애초에 수율은 무엇이 정하는가	**4회차** — 푸아송 모형과 결함 밀도
인터포저·TSV·본딩은 정확히 어떤 공정인가	**6회차** — 첨단 패키징
스택 8개를 붙일 자리는 어떻게 만드나	**6회차** 실습 3·5
HBM이 그렇게 뜨거우면 어떻게 식히나	**7회차** — 열저항 사슬

출처

스터디 커리큘럼이 고른 입문 자료 ("AI 메모리 반도체" 항목)

커리큘럼 소주제	링크	이 책에서의 활용
메모리의 종류	반도체네트워크 「AI 시대 메모리 반도체의 진화… 플래시·SRAM·DRAM에서 HBM까지」(2026. 6. 29.)	메모리 분류 개관. 1절 SRAM/DRAM/NAND 구분의 기초 자료. 같은 글이 『Semiconductor Network』 2026년 7월호 44쪽에 실렸다
HBM	Rambus 블로그 「High Bandwidth Memory (HBM): Everything You Need to Know」	HBM 구조·세대별 사양의 표준 영문 해설. **4절과 직접 대응하며, 그림 자료가 좋아 발제용으로 권장**
HBF	HyperAccel 기술 블로그 「AI 시대의 필수 소비재, 메모리 이해하기 1편: HBF 이해하기」	HBF 개념 해설. 6절의 기초
HBF	다음뉴스 「[분석] HBM? D램? 진짜 수요는 '낸드'야…AI 시대 '창고 전쟁'」	HBF 국내 보도. 산업 맥락
HBF	그로쓰리서치(Growth Research), 「[HBF 산업보고서] AI의 병목은 GPU가 아니라 메모리였다: HBF가 가져올 AI반도체 혁명」(2025. 11. 17., 한용희·구민준·최혜인) PDF	HBF 산업 전망 보고서. 6절의 시장 측면. **투자 관점 보고서이므로 전망치는 비판적으로 읽을 것.** 공개된 게시처를 찾지 못해 링크를 달지 않았다
AI 메모리 기술동향	ETRI 전자통신동향분석 41권 3호(2026), 홍아름·이재성·연주원, 「AI 메모리 반도체 분야 주요 기술 및 시장 동향」, pp.15–27	국내 학술 리뷰. **AI 메모리 전반을 한국어로 체계적으로 정리한 자료로, 이 챕터 전체의 보조 교재로 적합**

검증 완료 (2026-08-02, 복수 출처 대조)

- HBM4 = JEDEC JESD270-4, **2025년 4월 공표**. 인터페이스 2,048비트, 독립 채널 32개(채널당 의사채널 2개), 스택당 최대 약 2 TB/s, 4~16단, 다이 24Gb/32Gb, 32Gb 16단 시 큐브당 64GB, HBM3 컨트롤러 하위 호환, DRFM, 코어 전압 1.0/1.05V·VDDQ 0.7~0.9V — JEDEC 보도자료, TrendForce(2025-04-21), Tom's Hardware 대조
- HBF 표준화: Sandisk·SK하이닉스 협력 발표 **2025년 8월 6일**, OCP 워크스트림 착수 **2026년 2월 25일**. 목표 사양 다이당 256Gb / 16단 512GB / 읽기 약 1.6 TB/s, HBM4와 비슷한 풋프린트 지향. 첫 샘플 2026년 하반기 목표 — Sandisk·SK하이닉스 보도자료, IEEE Spectrum, Tom's Hardware 대조. 다이당 용량과 풋프린트 표현은 Sandisk HBF 팩트시트 원문(「256Gb per die」, 「closely matching」)으로 확인
- HBM 3사 과점 구조와 매출 기준 선두 SK하이닉스(2026년 1분기 IDC 집계, 1회차 4절) — 검증 완료. 2·3위의 순서는 세대마다 바뀌어 적지 않는다

검증 완료 (2026-10-08, 1차 출처 원문 대조)

- HBM4 양산: SK하이닉스 HBM4 개발 완료·양산 체제 구축(2025-09-12) — 2,048 I/O, 10 Gbps 이상, 1b DRAM, Advanced MR-MUF. 삼성전자 HBM4 양산 출하(2026-02-12) — 11.7 Gbps(최대 13 Gbps), 스택당 최대 3.3 TB/s, 12단 24~36GB, 1c DRAM·4nm 로직. 삼성전자 HBM4E 샘플 출하(2026-05-29) — 14 Gbps(최대 16 Gbps), 스택당 최대 3.6 TB/s, 12단 48GB, 「Samsung Foundry's 4nm logic base die」
- HBM4 표준: JEDEC JESD270-4 보도자료(2025-04-16) — 8 Gb/s, 2,048비트, 최대 2 TB/s, 32채널·채널당 의사채널 2, VDDQ 0.7/0.75/0.8/0.9V·VDDC 1.0/1.05V, HBM3 컨트롤러 하위 호환, DRFM, 4·8·12·16단과 24Gb·32Gb 다이(32Gb 16단 64GB)
- NVIDIA Rubin: NVIDIA 기술 블로그 「Inside NVIDIA Rubin GPU Architecture」(2026-07-21) — HBM4 최대 288GB, 12단 스택, 최대 22 TB/s(Blackwell 대비 2.8배). 스택 수는 글에 없어 본문의 8개는 역산 추정
- HBF: SK하이닉스 HBF 첫 표준 사양 공개(2026-08-04) — OCP 공개, 8단·16단, 최대 512GB, 3등급 약 0.4~3.0 TB/s, UCIe, Google·Tenstorrent 참여. SK하이닉스 FMS 2026 정리(2026-08-07) — HBM·DRAM·NAND(HBF)·SSD 계층, CXL 풀드 메모리 시연. Sandisk 표준화 협력 보도자료(2025-08-06) — 2026년 하반기 첫 샘플·2027년 초 추론 장치 샘플 목표. Sandisk HBF 팩트시트(2025-07) — 1세대 읽기 1.6 TB/s, 다이당 256Gb, 16단 512GB, CBA = CMOS directly Bonded to Array
- 적층 공정: SK하이닉스 HBM4 보도자료(위) — Advanced MR-MUF. 삼성전자 HBM3E 12H 보도자료(2024-02-27) — Advanced TC NCF
- SPHBM4: JEDEC JESD330-4 보도자료(2026-07-13) — HBM4와 같은 DRAM 다이, 데이터 신호 512개·4:1 직렬화, 유기 기판 실장
- GDDR7: JEDEC JESD239 보도자료(2024-03-05) — 디바이스당 최대 192 GB/s, PAM3(2사이클에 3비트), 독립 채널 GDDR6 2개 → GDDR7 4개. 핀당 48 Gbps는 32비트로 나눈 계산값
- DRAM 리프레시: Micron 16Gb DDR4 데이터시트 Rev. H(2021-08), 리프레시 기능 설명과 Table 52 — 0~85°C에서 64ms 안 8,192회(tREFI 7.8125µs), 85~95°C에서 32ms(3.90625µs), tRFC1 2Gb 160ns·4Gb 260ns·8Gb 350ns. 링크는 유통사(Digi-Key)의 사본이다
- CXL: CXL 컨소시엄 CXL 4.0 공개 보도자료(2025-11-18) — 64 GT/s → 128 GT/s. CXL 컨소시엄 개요 발표자료 — PCIe 물리·전기 인터페이스 활용
- Optane: Intel Form 10-Q(2022-07-02 종료 분기) — SEC EDGAR 원문. 2022년 2분기에 Optane 메모리 사업의 정리(wind-down)를 시작했고, 같은 분기에 재고 손상 5억 5,900만 달러를 매출원가에 반영했다고 적었다. 다음 분기 10-Q(2022-10-01 종료 분기, 원문)도 같은 금액을 1~3분기 누계로 적었다(2026-10-08 대조)

미확인 · 유보 (본문에서 단정하지 않음)

- **HBF의 실측 성능** — 2026년 10월 현재 양산 제품이 없다. 본문의 512GB·0.4~3.0 TB/s·1.6 TB/s는 전부 **사양 범위·목표치**이며 검증된 실측값이 아니다. 중간 등급(Grade 2)의 대역폭 값은 공개 자료에서 확인하지 못했다
- **HBF 샘플 일정** — 2025년 8월 목표는 2026년 하반기 첫 샘플이었다. 2026년 8월 Sandisk 투자자의 날 자료는 첫 HBF 메모리 다이 테이프아웃과 2027년 첫 추론 제품 샘플만 제시했다(2026-10-08 확인). 메모리 샘플 출하 시점은 공식 발표 미확인

- **HBM2E/HBM3/HBM3E의 정확한 스택 대역폭** — 벤더·제품별로 핀 속도가 달라 표의 수치는 대략적 범위다. 정확한 값은 각 벤더 데이터시트 확인 필요
- **하이브리드 본딩의 HBM 양산 적용 시점** — 2026년 10월 현재 양산 HBM에 썼다는 제조사 발표를 찾지 못했다. Micron의 적층 방식은 1차 자료로 확인하지 못했다
- **HBM4의 명령 버스 구조 변화** — 명령 버스와 데이터 버스의 분리를 HBM4 변경점으로 드는 서술은 JEDEC 보도자료(2024-07 예고·2025-04 공표) 원문에서 확인되지 않아 4절 변경점 목록에 넣지 않았다. 표준 본문은 열람하지 못했다
- **HBM 업체별 점유율 퍼센트** — 자료별 편차가 커 단정하지 않음
- **5절 수율 계산의 y 값** — 실제 업체 수율은 영업비밀이다. 본문 수치는 **계산 원리를 보이기 위한 가정값**이며 실제 수율이 아니다
- **PIM 제품의 상용화 수준** — 시제품·한정 적용 단계로, 범용 채택 여부는 미정

4회차

팹리스 · 파운드리와 밸류체인

핵심 TAKEAWAY

반도체 산업의 분업 구조(팹리스·파운드리·IDM·OSAT)는 취향이 아니라 비용 구조의 귀결이다. 팹 건설비가 지수적으로 증가하면서 제조는 규모의 경제가 지배하는 소수 과점이 되었고, 설계는 자산이 가벼운 다수 경쟁이 되었다. 그리고 이 산업의 거의 모든 경제적 판단은 하나의 식으로 수렴한다: **수율 $Y = e^{-AD_0}$ — 다이 면적이 커지면 수율이 지수적으로 떨어진다.** 이 식이 칩렛 분할의 근거이고, 파운드리 가격 구조의 근거이며, 6회차 첨단 패키징의 존재 이유다. "3nm 공정"은 물리 치수가 아니라 마케팅 명칭이다. 비교해야 할 것은 트랜지스터 밀도와 셀 구조다.

원자료 바로가기

삼성전자 반도체 뉴스룸 「파운드리? 팹리스? 반도체 생태계 한눈에 보기!」 · ScienceON JAKO202470953293995 「최신 인공지능 반도체 및 컴파일러 지원 기술 동향」 · ScienceON JAKO202369265025811 「초거대 인공지능 프로세서 반도체 기술 개발 동향」 · ScienceON ASTI00673636 「Market Insight 336: AI 주문형 반도체(AI ASIC)」 · 윈스퍽 「반도체 공정 종류 뭐가 있을까? 전공정·후공정 차이 정리」 · TSMC Open Innovation Platform® · Samsung Foundry SAFE™ 프로그램 · Stapper 외(1980), doi:10.1147/rd.243.0398 · TSMC 2nm(N2) 기술 페이지

학습목표

1. 팹리스·파운드리·IDM·OSAT·IP 벤더의 역할과 각자의 이익 구조를 구분할 수 있다.
2. 칩 설계 파이프라인 전체를 소프트웨어 개발 과정에 대응시켜 설명할 수 있다.
3. 푸아송 수율 모형을 세우고 다이 크기별 웨이퍼당 굿다이 수를 계산할 수 있다.
4. "3nm"라는 표현에서 실제로 확인해야 할 지표가 무엇인지 안다.
5. ASIC 개발 여부를 정량적으로 판단할 수 있다.

0. 자습 안내 — 혼자 읽는 사람에게

> **이 회차는 두 번째 공용 도구를 만드는 회차다**
>
> 2회차가 루프라인을 만들었다면, 이 회차는 **수율 식**을 만든다. $Y = e^{-AD_0}$ 이 한 줄이 5회차(왜 공정 하나하나가 six sigma여야 하는가), 6회차(왜 칩렛인가), 7회차(왜 공급이 안 늘어나는가)를 전부 지배한다. **실습 1·2는 반드시 손으로 계산할 것.** 표를 눈으로 읽는 것과 직접 지수함수에 넣어 보는 것은 남는 것이 다르다.

직전 회차에서 가져오는 것

앞 회차에서	이 회차에서
1회차 밸류체인 3층	그 층 구분을 **공정과 원가**로 정밀화한다
2회차 ASIC 손익분기 N^*	c_A(ASIC 단가)를 수율 식으로 실제 계산한다
3회차 적층 수율 y^n	그 y 가 어디서 오는지 — 결함 밀도와 다이 면적

선행 지식 점검

개념	한 줄 답	필요도
푸아송 분포	정해진 구간에 드물게 일어나는 사건의 수. $P(X=k) = e^{-\lambda}\lambda^k / k!$	★ 필수 (쓰는 것은 $k=0$ 하나뿐. 그 식이 어디서 나오는지는 5절)
감마 분포	양의 실수를 값으로 갖는 연속분포. 여기서는 결함 밀도의 변동을 나타낸다	5절 심화에만
음이항 분포	푸아송의 평균이 감마로 흔들릴 때 나오는 분포	5절 심화에만
다이(die)	웨이퍼를 잘라 낸 칩 하나	★ 필수
레티클 한계	노광기가 한 번에 찍을 수 있는 최대 면적(약 858 mm²)	실습 1 풀이·5회차 실습 5

> **통계 배경이 있다면:** 이 회차 5절은 학부 확률론 첫 장이 그대로 산업의 원가 구조가 되는 흔치 않은 사례다. 실습 3(푸아송 vs 음이항)은 젠센 부등식 한 줄로 정리된다 — 부록 풀이 참조.

읽는 순서 (총 110분 권장)

순서	절	시간	비고
1	1~4절 (분업 구조)	20분	1회차의 정밀화. 빠르게
2	**5절 수율의 경제학**	**35분**	**이 회차의 중심.** 실습 1·2를 여기서 바로
3	실습 1·2·3	25분	계산기 필수
4	6절 ASIC·FPGA 선택	15분	
5	실습 4·5 + 자가진단	15분	실습 4는 6회차 예고편이다

절대 건너뛰지 말 것: 5절 「수치로 느끼기」 표. 그리고 그 표를 보고 **왜 면적이 16배 늘 때 수율이 절반이 되는가**를 지수함수의 성질로 설명할 수 있어야 한다.

막히기 쉬운 지점 — GDPW 식이 왜 저렇게 생겼나

$$\text{GDPW} \approx \frac{\pi d^2 / 4}{A} - \frac{\pi d}{\sqrt{2A}}$$

겁먹을 식이 아니다. 두 항을 따로 읽으면 된다.

- **첫 항** = 웨이퍼 넓이 ÷ 다이 넓이. 그냥 "몇 개가 들어가나"다.
- **둘째 항** = 가장자리 손실 보정. 원판에 사각형을 채우면 테두리에 걸려 잘리는 다이가 생긴다. 그 개수를 **웨이퍼 둘레 ÷ 다이 대각선으로** 어림한 것이 둘째 항이다. 정사각 다이의 한 변이 $\sqrt{A}$ 이면 대각선은 $\sqrt{2A}$ 이므로 $\pi d / \sqrt{2A}$ 가 된다. 분모를 한 변 $\sqrt{A}$ 로 잡으면 손실이 $\sqrt{2}$ 배로 커지는데, 이 상수 하나가 결과를 눈에 띄게 바꾼다(아래 표).

즉 "면적으로 나눈 뒤, 테두리에서 버려지는 만큼 뺀다"가 전부다. 다이가 클수록 둘째 항의 비중이 커진다 **— 큰 칩은 수율에서도, 테두리에서도 두 번 손해를 본다.**

이 식도 근사다. 직접 세어 보면 차이가 보인다. 조건은 지름 300 mm 웨이퍼, 정사각 다이, 가장자리 제외 폭 0, 스크라이브 라인 폭 0이다. 다이 격자를 웨이퍼 위에 깔고, 네 모서리가 모두 반지름 150 mm 원 안에 드는 다이만 센다. 격자의 원점을 가로·세로로 한 변의 1/20씩 옮겨 400가지 위치에서 세고 평균과 범위를 냈다. 계산하면 다음과 같다.

다이 면적	직접 센 개수: 평균 (최소~최대)	$\sqrt{2A}$ 식	$\sqrt{A}$ 식
100 mm²	647.8 (640~657)	640.2 (−1.2%)	612.6 (−5.4%)
400 mm²	147.8 (144~151)	143.4 (−3.0%)	129.6 (−12.3%)
800 mm²	68.2 (66~71)	64.8 (−5.0%)	55.0 (−19.3%)

괄호 안 백분율은 직접 센 평균에 대한 차이다. $\sqrt{2A}$ 식은 평균보다 1~5% 적게 나오고, $\sqrt{A}$ 식은 5~19% 적게 나온다. 다이가 클수록 오차가 커지고, 격자를 어디에 놓느냐에 따라서도 개수가 달라진다(800 mm²에서 66개와 71개). 다이 꼭짓점 하나를 웨이퍼 중심에 맞추면 100·400·800 mm² 에서 각각 648·148·68개로, 평균과 거의 같다. 실제 웨이퍼에는 가장자리 제외 폭과 스크라이브 라인이 있으므로 직접 센 값도 이보다 줄어든다. 여러 근사식을 비교한 문헌으로 de Vries(2005), doi:10.1109/TSM.2004.836656이 있다.

---

1. 직관 잡기 — 왜 설계와 제조가 갈라졌나

고정비의 폭발

칩을 만드는 비용은 두 덩어리다.

- **고정비(NRE, Non-Recurring Engineering):** 설계 인건비, EDA 툴 라이선스, IP 라이선스, **마스크셋 제작비**, 검증·시제품. 물량과 무관하게 한 번 든다.

- **변동비:** 웨이퍼 값, 패키징, 테스트. 개당 든다.

여기에 **팹 건설비**라는 훨씬 큰 고정비가 있다. 최신 첨단 팹 하나가 수백억 달러 규모다. 이 정도 고정비를 회수하려면 팹을 24시간 가동하며 엄청난 물량을 흘려야 한다.

결론: 자기 칩만 만들어서는 팹을 채울 수 없다. 그래서 두 가지 길이 생겼다.

1. **IDM (종합반도체회사):** 자기 제품 물량이 충분히 큰 경우(메모리처럼 표준품 대량생산). 삼성전자·SK하이닉스·Micron·Intel.
2. **파운드리-팹리스 분업:** 여러 회사의 설계를 한 팹에 모아 채운다. TSMC가 이 모델을 확립했다.

> **경제학 프레임:** 고정비가 크고 한계비용이 낮은 산업은 자연독점 경향을 갖는다. 반도체 제조는 여기에 **학습곡선(누적 생산량이 늘수록 수율이 오름)** 까지 겹쳐서, 선두가 계속 유리해지는 구조다.
> 후발주자가 따라잡기 어려운 이유는 자본만이 아니라 **경험의 축적**에도 있다.

---

2. 플레이어별 역할

유형	하는 일	대표 기업	특징
팹리스	설계만. 제조는 외주	NVIDIA, AMD, Broadcom, Qualcomm, Apple, MediaTek	자산 가벼움, 이익률 높음, 경쟁 격심
파운드리	제조만. 설계는 안 함	TSMC, Samsung Foundry, Intel Foundry, GlobalFoundries, UMC, SMIC	초고정비, 고객 기밀 유지가 생명
디자인하우스 (설계 서비스)	팹리스의 논리 설계를 특정 파운드리 공정에 맞춰 물리 구현·검증하고, 제조·조립·테스트까지 대행하기도 함	파운드리마다 공인 협력망이 있다. TSMC는 OIP 안의 DCA(Design Center Alliance, 칩 구현 서비스)와 VCA(Value Chain Alliance), 삼성은 SAFE™ 안의 디자인 서비스 얼라이언스(DSP·VDP)	협력하는 파운드리의 공정·IP에 맞춰 일한다. 시스템 회사·신생 기업이 양산까지 가는 통로
IDM	설계+제조	Samsung, SK하이닉스, Micron, Intel, TI	메모리·아날로그에 많음
OSAT	패키징·테스트 외주	ASE, Amkor, JCET, PTI	전통적으로 저부가가치였으나 첨단 패키징으로 위상 상승 중
IP 벤더	설계 블록 라이선스	Arm, Synopsys, Cadence, Imagination	로열티 모델
EDA	설계 자동화 툴	Synopsys, Cadence, Siemens EDA	3사 과점, 진입장벽 극도로 높음
장비	제조 장비	ASML, Applied Materials, Lam Research, TEL, KLA	공정별 사실상 과점 또는 독점
소재	웨이퍼·가스·포토레지스트 등	신에츠, SUMCO, JSR, TOK, 린데	일본 기업 비중 높음

경계가 흐려지는 지점 (2026년)

- **메모리 업체가 로직 공정을 필요로 한다** — HBM4 베이스 다이를 로직 파운드리 공정으로 만든다. SK하이닉스는 2024년 4월 TSMC와 HBM4 협력 양해각서를 맺으면서 TSMC의 첨단 로직 공정을 쓰고 베이스 다이의 성능 개선부터 함께 하겠다고 발표했다. 메모리 업체와 파운드리의 협업이 필수가 된다.
- **파운드리가 패키징으로 확장한다** — TSMC의 CoWoS·SoIC. 전통적 OSAT 영역을 파운드리가 흡수하고 있다. 이유는 명확하다: **첨단 패키징이 성능을 결정하므로 전공정과 통합 최적화해야 한다.**
- **하이퍼스케일러가 팹리스가 된다** — Google·Amazon·Meta·Microsoft가 자체 칩을 설계한다. 다만 대부분 Broadcom·Marvell 같은 설계 파트너와 함께한다.
- **IDM이 파운드리를 연다** — Intel Foundry, Samsung Foundry. 신뢰 문제(경쟁사에 설계를 맡길 수 있는가)가 구조적 난제다.

---

3. 칩 설계 파이프라인 — 소프트웨어 개발과 대응시켜 보기



대응표

칩 설계 단계	소프트웨어 개발의 대응	하는 일	그 단계에서 실물로 있는 것
사양 정의 (아키텍처)	시스템 설계	성능·전력·면적(PPA) 목표 설정	문서
RTL 설계 (Verilog/VHDL)	**소스 코드 작성**	레지스터 전송 수준으로 동작 기술	코드 파일
기능 검증 (시뮬레이션·포멀)	테스트·CI	논리적으로 맞는지 확인	시뮬레이션 결과 파일
논리 합성	**컴파일**	RTL → 표준셀 게이트 네트리스트	네트리스트 파일
배치배선 (P&R)	링킹 + 레이아웃 최적화	셀을 실제 위치에 놓고 배선	레이아웃 데이터
타이밍/전력/신호 검증 (STA 등)	프로파일링·최적화	클럭 주기 안에 신호가 도착하는가	검증 보고서
DRC / LVS	린트 + 정적 분석	제조 규칙 위반, 레이아웃-회로 일치 확인	검증을 통과한 레이아웃 데이터
테이프아웃	**릴리즈 (되돌릴 수 없는)**	마스크 제작으로 데이터 전달	레이아웃 파일 한 벌. 아직 데이터뿐이다
마스크 제작	배포 매체 제작	첨단 노드 마스크셋 한 벌에 수천만 달러 규모, 수 주 소요	**처음 생기는 물건** — 층별 포토마스크 묶음(마스크셋)
웨이퍼 제조	빌드 실행	수개월 소요	회로가 새겨진 웨이퍼
웨이퍼 테스트 (EDS/probe)	통합 테스트	다이 단위 양불 판정, 리페어	다이마다 양불이 기록된 웨이퍼 (웨이퍼 맵)
패키징·최종 테스트	패키징·배포	6회차	잘라 낸 다이 → 패키지에 든 칩

넷째 열을 위에서 아래로 읽으면 테이프아웃까지 칩은 **파일로만 있다.** 테이프아웃 때 파운드리로 넘어가는 것은 레이아웃 파일이고, 오래 쓰여 온 형식은 GDSII 스트림이다. SEMI는 그 후속 형식으로 OASIS(SEMI P39)를 표준화했다. 소프트웨어 비유가 테이프아웃에서 끝나는 이유가 이 열에 있다. 그 아래로는 고치려면 물건을 다시 만들어야 한다.

소프트웨어와 결정적으로 다른 두 가지

1. **핫픽스가 없다.** 테이프아웃 이후 버그가 발견되면 마스크를 다시 만들고(리스핀) 수개월과 수천만 달러를 다시 쓴다. 그래서 검증에 전체 개발 공수의 절반 이상을 쓴다. **"버그 하나의 비용"이 소프트웨어와 자릿수가 다르다.**
2. **물리가 개입한다.** 같은 RTL이라도 배치배선 결과에 따라 성능이 크게 달라진다. 타이밍 클로저(모든 경로가 클럭 주기를 만족하도록 만드는 일)가 개발 후반의 주된 고통이다.

「**RTL = 소스 코드**」 **비유가 어긋나는 곳이 하나 더 있다.** RTL 문장은 위에서 아래로 실행되는 명령이 아니라 **동시에 존재하는 회로를** 적은 것이다. `assign` 문 세 줄은 덧셈을 세 번 하는 것이 아니라 덧셈기 세 개를 나란히 놓는 것이고, `if` 는 분기가 아니라 두 입력 중 하나를 고르는 멀티플렉서가 된다. 소프트웨어에서 온 사람이 가장 자주 틀리는 것은 논블로킹 대입(`<=`)이다.

```
always @(posedge clk) begin
  q1 <= d;
  q2 <= q1;
end
```

클럭 직전에 `d = 1`, `q1 = 0`, `q2 = 0` 이라고 하자. 위에서 아래로 실행된다고 읽으면 클럭 뒤에 `q1` 과 `q2` 가 모두 1이어야 한다. 실제로는 `q1 = 1`, `q2 = 0` 이다. 논블로킹 대입은 클럭 순간에 오른쪽 값을 모두 먼저 읽어 두고 왼쪽을 한꺼번에 바꾸기 때문이다(IEEE Std 1364-2005의 논블로킹 절차 대입 규정. 이 규격은 지금 IEEE Std 1800(SystemVerilog)으로 대체되었다). 두 줄의 순서를 바꿔도 결과는 같다. 이 코드는 명령 두 개가 아니라 플립플롭 두 개를 이어 붙인 2단 시프트 레지스터의 그림이다.

PDK와 표준셀 — 추상화 계층

파운드리는 **PDK(Process Design Kit)** 를 제공한다. 여기에는 설계 규칙(DRC 룰), 소자 모델(SPICE 파라미터), 표준셀 라이브러리(NAND·플립플롭 등 기본 블록의 레이아웃과 타이밍 특성)가 들어 있다.

이것이 왜 결정적인가: PDK 덕분에 팹리스는 반도체 물리를 몰라도 설계할 수 있다. 소프트웨어의 API/ABI와 정확히 같은 역할이다. 그리고 API가 그렇듯, PDK에 맞춰 설계한 것은 다른 파운드리로 그대로 옮겨지지 않는다. 표준셀과 설계 규칙이 바뀌므로 합성부터 배치배선·타이밍 클로저·DRC/LVS까지 다시 돌려야 하고, 마스크셋도 새로 만든다. 이 이식 비용이 파운드리 전환을 막는 실제 장벽이고, TSMC의 생태계 우위도 여기서 나온다.

이 생태계는 이름이 붙은 제도다. TSMC는 협력사를 OIP(Open Innovation Platform®) 아래 얼라이언스로 묶는다. 2018년 10월 클라우드 얼라이언스가 다섯째로 더해졌을 때 구성은 EDA·IP·디자인센터·VCA(당시 이름 Value Chain Aggregator, 지금은 Value Chain Alliance)·클라우드였고, 2022년 10월 여섯째로 3DFabric 얼라이언스(첨단 패키징·3D IC)가 생겼다. 삼성 파운드리는 같은 일을 SAFE™ 프로그램으로 한다. 인증한 PDK·설계 방법론·IP를 묶어 내놓고, IP·EDA·클라우드·디자인 서비스·멀티다이 통합(MDI)·OSAT 얼라이언스를 둔다(2026년 10월 확인). 고객이 보기에 파운드리를 고르는 일은 공정 하나를 고르는 일이 아니라 이 협력망 전체를 고르는 일이다.

---

4. "3nm"의 진실 — 노드 이름 읽는 법

무엇이 아닌가

과거에는 노드 이름이 게이트 길이 같은 실제 물리 치수를 가리켰다. FinFET 이후로는 **어떤 물리 치수와도 대응하지 않는 마케팅 명칭**이 되었다. 서로 다른 회사의 "3nm"는 밀도와 성능이 다르고, 같은 회사의 세대 간 비교에서도 이름이 성능 향상을 그대로 반영하지 않는다.

그러면 무엇을 봐야 하나

지표	의미
트랜지스터 밀도 (MTr/mm²)	제곱밀리미터당 백만 트랜지스터. 가장 직접적인 비교 지표. 다만 로직/SRAM 비율에 따라 달라진다
CPP (Contacted Poly Pitch)	게이트 간 간격
MMP (Minimum Metal Pitch)	최소 배선 간격
셀 높이 (track 수)	표준셀의 높이를 그 안에 지나가는 가로 배선 트랙 수(최소 배선 피치 단위)로 센 값. 6T·5T의 T는 track이다. 3회차 SRAM 6T의 T(트랜지스터 6개)와 다른 말이다
PPA 개선율	이전 노드 대비 성능·전력·면적 개선. 벤더 발표치는 조건을 반드시 확인
SRAM 셀 면적	**최근 특히 중요.** SRAM은 로직만큼 미세화되지 않아, 캐시가 큰 칩은 노드 이득이 작다

> **AI 가속기 맥락에서 더 중요한 사실:** AI 칩 성능은 노드보다 **메모리 대역폭과 패키징**에 더 좌우된다. 노드가 한 세대 앞선 것보다 HBM 스택을 하나 더 붙이는 것이 성능에 더 큰 영향을 줄 수 있다.

2026년 현재 2nm급 상황

각 회사의 공식 발표로 2026년 10월 8일에 다시 확인한 상태다.

파운드리	노드	트랜지스터 구조	후면전력공급	상태
TSMC	N2	나노시트 GAA	없음 (A16에서 Super Power Rail 도입. 2024년 4월 발표 때는 2026년 생산 계획이었으나, 2026년 4월 기술 심포지엄에서 보도된 TSMC 임원 발언에 따르면 양산은 2027년 시작 예상. 2026년 10월 22일 확인)	2025년 4분기 양산 시작. 개량판 N2P는 2026년 하반기 양산 예정
Intel	18A	RibbonFET (GAA)	**PowerVia (적용)**	첫 제품 Core Ultra 시리즈 3를 2026년 1월 발표, 첫 서버 제품 Xeon 6+를 2026년 2분기 출시. 개량판 18A-P는 2026년 6월 리스크 생산 진입
Samsung	SF2	MBCFET (GAA)	—	2025년 양산 시작. Exynos 2600에 적용

주목할 구도: Intel은 GAA와 후면전력공급을 동시에 도입하는 공격적 전략을, TSMC는 GAA를 먼저 안정화하고 후면전력공급을 다음 노드로 미루는 보수적 전략을 택했다. **이것은 기술 우열이 아니라 위험 배분(risk budgeting) 전략의 차이다.** 토론하기 좋은 주제다.

---

5. 수율의 경제학 — 이 책의 두 번째 핵심 도구

푸아송 모형

웨이퍼에 결함이 무작위로(공간적으로 독립·균일하게) 떨어진다고 가정하자. 단위 면적당 평균 치명결함 밀도를 D_0 [결함/cm²], 다이 면적을 A [cm²]라 하면, 한 다이에 떨어지는 결함 수는 평균 $\lambda = AD_0$ 인 푸아송 분포를 따른다. 결함이 하나도 없을 확률, 즉 수율은

$$Y = P(X = 0) = e^{-AD_0}$$

통계 전공자에게: 이것은 그냥 푸아송 분포의 $P(X=0)$ 이다. 반도체 수율 이론의 출발점이 학부 확률론 첫 장이라는 것이 이 분야의 재미있는 점이다.

식이 어디서 오나 — 다이를 칸으로 쪼개 보기. 다이를 같은 넓이의 작은 칸 n 개로 나누고, 칸마다 결함이 떨어질 확률을 $p = \lambda / n$ 으로 둔다. 칸끼리 독립이면 다이가 깨끗할 확률은 모든 칸이 깨끗할 확률이고, 칸을 한없이 잘게 쪼개면 고교 극한 $\lim_{n\to\infty}(1 + x/n)^n = e^x$ 에 $x = -\lambda$ 를 넣은 것이 된다.

$$P(\text{결함 } 0) = \left(1 - \frac{\lambda}{n}\right)^n \longrightarrow e^{-\lambda} \quad (n \to \infty)$$

계산하면 $\lambda = 0.8$ 일 때 $n = 10$ 에서 0.4344, $n = 100$ 에서 0.4479, $n = 1{,}000$ 에서 0.4492이고, 극한값 $e^{-0.8}$ 은 0.4493이다. 칸 열 개로만 나눠도 답의 97%에 이른다.

▾ **풀이**

일반 k — 이항분포의 극한. 칸 n 개 가운데 정확히 k 칸에 결함이 떨어질 확률은 이항분포다. 이것을 네 인수로 나눠 적는다.

$$P(X=k) = \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} = \frac{\lambda^k}{k!}\cdot\frac{n(n-1)\cdots(n-k+1)}{n^k}\cdot\left(1-\frac{\lambda}{n}\right)^n\cdot\left(1-\frac{\lambda}{n}\right)^{-k}$$

k 를 고정하고 $n \to \infty$ 로 보내면 뒤의 세 인수가 각각 1, $e^{-\lambda}$, 1로 간다. 남는 것이 $P(X=k) = e^{-\lambda}\lambda^k / k!$, 곧 푸아송 분포다. 드문 사건을 아주 많은 기회에 나눠 놓으면 그 횟수가 푸아송을 따른다는 이 결과를 희귀 사건의 법칙(law of rare events)이라 부른다. $k = 0$ 이면 첫째·둘째·넷째 인수가 모두 1이 되고 셋째만 남아 본문의 식이 된다.

수치로 보면 $\lambda = 0.8$, $k = 1$ 일 때 이항 확률이 $n = 10$ 에서 0.3777, $n = 100$ 에서 0.3612, $n = 1{,}000$ 에서 0.3596이고, 푸아송 값은 0.3595다(계산하면).

수치로 느끼기

$D_0 = 0.1$ 결함/cm² (성숙 공정 수준의 예시값)일 때:

다이 면적	$\lambda = AD_0$	수율 $e^{-\lambda}$
50 mm² (0.5 cm²)	0.05	95.1%
100 mm²	0.10	90.5%
200 mm²	0.20	81.9%
400 mm²	0.40	67.0%
600 mm²	0.60	54.9%
800 mm² (레티클 한계 근처)	0.80	**44.9%**

면적이 16배 늘 때 수율은 절반 이하로 떨어진다. 그리고 큰 다이는 웨이퍼당 개수도 적으므로, 굿다이 수는 훨씬 급격히 준다.

웨이퍼당 굿다이 수

300mm 웨이퍼 면적은 약 $\pi \times 15^2 = 706$ cm²다. 가장자리 손실을 고려한 근사식(Gross Die Per Wafer)은 여러 형태가 있는데, 흔히 쓰는 것은

$$\text{GDPW} \approx \frac{\pi d^2 / 4}{A} - \frac{\pi d}{\sqrt{2A}}$$

(d = 웨이퍼 지름, A = 다이 면적. 두 번째 항이 가장자리 손실 보정으로, 웨이퍼 둘레를 다이 대각선 $\sqrt{2A}$로 나눈 것이다. 100~800 mm² 다이에서 이 식이 직접 센 개수보다 1~5% 적게 나온다는 비교는 0절 「막히기 쉬운 지점」에 있다)

굿다이 수 = GDPW × Y 이고, **다이당 원가 = 웨이퍼 가격 / 굿다이 수** 다.

> **실습:** 웨이퍼 한 장 값을 W 라 할 때, 다이 면적 100 mm²와 800 mm²의 다이당 원가 비율을 구해보라. 면적비는 8배지만 원가비는 훨씬 크게 나온다.

푸아송 모형이 틀리는 이유 — 결함 클러스터링

실제 결함은 독립적으로 흩어지지 않고 **뭉친다**(파티클 오염, 장비 이상, 웨이퍼 가장자리 효과). 클러스터링이 있으면 푸아송 모형은 수율을 **과소평가**한다. 결함이 몇 개 다이에 몰리면 나머지 다이는 깨끗하기 때문이다.

대안 모형들

모형	식	가정
푸아송	$Y = e^{-AD_0}$	완전 무작위
머피(Murphy)	$Y = \left(\frac{1 - e^{-AD_0}}{AD_0}\right)^2$	결함밀도가 삼각분포
지수(Seeds)	$Y = \frac{1}{1 + AD_0}$	결함밀도가 지수분포
음이항 (Negative Binomial)	$Y = \left(1 + \frac{AD_0}{\alpha}\right)^{-\alpha}$	**결함밀도가 감마분포 → 클러스터링 반영**

음이항 모형에서 α 는 클러스터링 파라미터로, $\alpha \to \infty$ 이면 푸아송으로 수렴하고 α 가 작을수록 클러스터링이 심하다. **감마-푸아송 혼합이 음이항이 되는 것은 통계학에서 익숙한 결과이고, 반도체 수율 모형이 정확히 그 구조다.** 결함 밀도 D 가 평균 D_0 인 감마 분포를 따르면 그 분산은 D_0^2 / α 다. $\alpha = 1$ 이면 결함 밀도의 표준편차가 평균과 같다는 뜻이다(적분은 실습 3 풀이). IBM이 1980년에 낸 메모리 수율 모형도 푸아송의 평균을 흔드는 혼합 모형이었다(Stapper 외, 1980).

> **더 파고들기:** 이 절은 통계 전공자가 가장 편하게 심화할 수 있는 지점이다. 실제 팹에서는 웨이퍼 맵(어느 위치의 다이가 불량인지) 데이터에 공간 통계와 머신러닝을 적용해 결함 패턴을 분류하고 원인 장비를 역추적한다. 5회차 검사·계측 절과 이어진다.

A 는 다이 면적 전체가 아니다 — 임계 면적. 결함이 떨어졌다고 모두 불량이 되지는 않는다. 배선 두 줄 사이에 떨어진 파티클이 두 줄에 함께 닿을 만큼 크지 않으면 단락은 생기지 않고, 회로가 없는 자리에 떨어지면 아무 일도 없다. 그래서 $\lambda = AD_0$ 의 A 는 엄밀히는 다이 면적이 아니라 **임계 면적(critical area)**, 곧 그 크기의 결함이 떨어졌을 때 회로가 망가지는 자리의 넓이다. 같은 다이라도 결함이 클수록 임계 면적이 넓어지고, 배선이 촘촘한 블록일수록 넓다. IBM의 메모리 수율 모형(Stapper 외, 1980)은 제품을 이 임계 면적으로, 공정을 결함 밀도로 따로 기술했다. 이 책이 D_0 를 「치명결함 밀도」라 부르며 다이 면적을 그대로 쓰는 것은 임계 면적의 효과를 밀도 쪽에 흡수한 단순화다.

수율이 설명하는 것들

1. **왜 칩렛인가** — 800 mm² 다이 1개(수율 45%)보다 200 mm² 다이 4개(각 82%)를 따로 만들어 테스트로 양품만 골라(KGD) 묶는 편이 훨씬 유리하다. 고르지 않고 네 개를 그대로 묶으면 완성품 수율은 $(e^{-0.2})^4 = e^{-0.8} \approx 45\%$ 로 단일 다이와 같다(3회차의 곱셈 수율). 이득은 다이를 작게 나눈 데서가 아니라 불량을 조립 전에 버려 양품 실리콘만 쓰는 데서 나온다. 다만 붙이는 비용과 조립 수율이 추가된다. 이 손익 계산이 6회차의 핵심이다.
2. **왜 파운드리는 웨이퍼 단위로 파는가** — 수율 위험을 누가 지느냐의 문제다. 웨이퍼 단위 계약이면 팹리스가, 굿다이 단위 계약이면 파운드리가 진다. 신규 노드일수록 파운드리는 웨이퍼 단위를 선호한다.
3. **왜 신규 노드가 비싼가** — 초기에는 D_0 가 높다. 시간이 지나며 학습곡선을 따라 내려간다(yield learning). 이것이 선두 주자의 구조적 이점이다.
4. **왜 비닝(binning)을 하는가** — 비닝은 두 종류다.
   - **결함 비닝.** 일부 결함은 해당 코어만 비활성화하면 쓸 수 있다. 8코어 설계를 6코어 제품으로 파는 것은 수율 회수 전략이다. 결함을 고쳐 쓰는 길도 있다. 메모리는 예비 행·열을 두었다가 결함 난 줄을 예비로 바꿔 끼우고(리페어), 일부만 동작하는 칩을 따로 팔기도 한다. IBM은 64K비트 RAM의

수율 추이를 보고하면서 예비 회로와 부분 동작 제품으로 얻은 수율 향상을 함께 제시했다(Stapper 외, 1982). 3회차 HBM의 여분 회로가 같은 원리다.

- **속도·전력 비닝.** 결함이 없는 다이끼리도 공정 변동 때문에 낼 수 있는 최대 동작 주파수와 전력이 다르다. 그래서 다이마다 최대 주파수를 시험해 등급을 나눈다. 이것은 결함 개수를 세는 e^{-AD_0} 로는 설명되지 않는 **파라메트릭 수율의** 문제다. 결함 비닝이 「몇 개가 살아 있나」를 다룬다면 속도 비닝은 「살아 있는 것이 얼마나 빠른가」를 다룬다.

---

6. ASIC · FPGA · 범용 프로세서 — 정량적 선택

손익분기 (2회차에서 세운 식의 정밀화)

$$N^* = \frac{C_{\text{NRE}}}{(c_G + o_G) - (c_A + o_A)}$$

- c : 단위 취득원가, o : 수명기간 운영비(주로 전력비)의 현재가치
- 2회차와 같이 ASIC 한 대가 GPU 한 대의 일을 1:1로 대신한다고 둔다. 처리량이 다르면 N 을 같은 수요를 처리하는 대수로 환산해야 한다
- c_A 는 위 5절의 다이당 원가 + 패키징 + 테스트로 계산한다

다섯 축 트레이드오프

	개발기간	NRE	단가	전력효율	유연성
범용 CPU/GPU	0 (구매)	0	높음	낮음	최고
FPGA	수개월	낮음	매우 높음	낮음~중간	재구성 가능
구조화 ASIC	1년 내외	중간	중간	중간	제한적
풀커스텀 ASIC	2~3년	매우 높음	낮음	최고	없음

시간 축이라는 함정

N^* 를 넘길 물량이 있어도, 그 물량을 그 칩으로 처리하는 기간이 짧으면 소용없다. 모델 아키텍처가 2년마다 크게 바뀌는데 칩 개발에 2~3년이 걸린다면, 완성 시점에 그 칩이 겨냥한 연산 패턴이 여전히 지배적이어야 한다.

하이퍼스케일러가 이 위험을 감수할 수 있는 이유

1. 물량 N 이 압도적이다.
2. **자기 워크로드를 자기가 통제한다.** 모델을 칩에 맞춰 설계할 수 있다는 것이 결정적이다.
3. 전력비 절감액의 총액이 크다(데이터센터 규모). 장당으로 보면 손익분기 대수를 주로 가르는 것은 취득가 차이이고 전력비 차이는 몇 %에서 십몇 %에 그치지만(2회차 실습 4 풀이, 이 회차 실습 5 풀이), 수만 장을 몇 년 돌리면 그 몫도 큰 금액이 된다.
4. 벤더 협상력이라는 옵션 가치가 있다.
5. 실패해도 감당 가능한 자본 여력이 있다.

대부분의 회사가 안 하는 이유: 위 다섯 가지 중 어느 것도 성립하지 않기 때문이다.

---

7. 자주 하는 오해

오해 1: "TSMC는 설계도 한다" → 순수 파운드리 모델의 핵심은 고객과 경쟁하지 않는다는 신뢰다. IP 블록·패키징 서비스는 제공하지만 완제품 칩 설계는 하지 않는다. 이 중립성이 TSMC 성공의 큰 부분이다.

오해 2: "노드가 앞서면 칩이 빠르다" → AI 가속기에서는 메모리와 패키징이 더 결정적일 수 있다. 또한 노드 이름 자체가 비교 가능한 지표가 아니다.

오해 3: "수율은 공정이 성숙하면 100%에 수렴한다" → 푸아송 식 안에서는 맞다. 면적이 얼마든 $D_0 \to 0$ 이면 $e^{-AD_0} \to 1$ 이다. 문제는 D_0 가 0까지 내려가지 않는다는 것이다. 같은 D_0 에서는 다이가 클수록 수율이 지수적으로 낮으므로, 공정이 성숙해도 큰 다이는 작은 다이보다 낮은 수율에 머문다(800 mm² 다이는 $D_0 = 0.05$ 에서도 67.0%). 여기에 결함 개수로 세지 않는 파라메트릭 손실(5절 속도·전력 비닝)이 식 밖에 따로 남는다. 그리고 신규 노드는 항상 다시 낮은 수율에서 시작한다.

오해 4: "팹리스는 제조를 몰라도 된다" → PDK가 추상화해 주지만, 고성능 설계는 공정 특성(변동성, 전자이동, 열)을 깊이 고려해야 한다. DFM(제조용이성 설계)이 별도 분야로 존재하는 이유다.

오해 5: "OSAT는 저부가가치 하청이다" → 첨단 패키징이 성능을 결정하게 되면서 위상이 크게 올라갔다. 다만 최첨단 영역은 파운드리가 흡수하고 있어, OSAT의 위치는 유동적이다.

오해 6: "EDA는 그냥 CAD 소프트웨어다" → 배치배선과 타이밍 최적화는 대규모 조합최적화 문제이고, 최근에는 강화학습·그래프신경망이 실제로 투입되는 분야다. **CS·AI 전공자에게 가장 직접적인 진입 지점이 여기다.**

---

8. 정량 실습

실습 1 — 수율 곡선 $D_0 = 0.05, 0.1, 0.2$ 각각에 대해 다이 면적 50~800 mm² 구간의 수율 곡선을 그려라. 어느 면적부터 수율이 급락하는가?

실습 2 — 다이당 원가 300mm 웨이퍼 가격을 W 라 두고, 다이 면적 100 / 400 / 800 mm²의 다이당 원가를 W 의 배수로 구하라. GDPW 근사식과 수율을 모두 적용할 것.

실습 3 — 푸아송 vs 음이항 같은 AD_0 에 대해 푸아송과 음이항($\alpha = 1, 2, 5$)의 수율을 비교하라. 클러스터링이 심할수록 실제 수율이 푸아송 예측보다 높은 것을 확인하고, 그 이유를 확률적으로 설명하라.

실습 4 — 칩렛 손익 (6회차 예고) 800 mm² 단일 다이 vs 200 mm² 칩렛 4개를 비교하라. 칩렛 쪽에는 (a) 칩렛 간 인터페이스 회로 면적 오버헤드 10%, (b) 조립 수율 95%, (c) 인터포저·조립 추가 비용을 넣어라. (c)는 완성품 한 개에 얹히는 비용(실패한 조립 시도의 몫까지 포함)으로 잡는다. 어느 쪽이 유리한가? D_0 를 바꾸면 결론이 뒤집히는 지점이 있는가?

실습 5 — ASIC 판단 가정: NRE 1.5억 달러, ASIC 단가 4천 달러, 상용 GPU 단가 3만 달러, 3년 전력비는 ASIC 2천 달러·GPU 6천 달러. N^* 를 구하고, 이 물량이 현실적인 조직이 어디인지 논하라.

---

9. 토론 질문

1. 파운드리-팹리스 분업은 되돌릴 수 있는가? Intel의 IDM 2.0 전략과 하이퍼스케일러의 수직통합은 각각 어떤 방향인가?
2. Intel(GAA+후면전력 동시)과 TSMC(GAA 먼저, 후면전력 나중)의 위험 배분 전략 중 어느 쪽이 합리적인가? 어떤 조건에서 각각 옳은가?
3. 수율 데이터는 영업비밀이다. 외부에서 수율을 추정할 수 있는 관찰 가능한 대리변수는 무엇인가? (가격, 리드타임, 제품 라인업, 비닝 구성 등)
4. EDA 3사 과점은 왜 깨지지 않는가? 오픈소스 EDA(OpenROAD 등)와 오픈 PDK가 이를 바꿀 수 있는가?
5. 노드 이름이 물리적 의미를 잃은 상황에서, 산업은 왜 그 명명법을 계속 쓰는가? 대안이 있는가?

---

10. 후속 공부

- **VLSI 설계 표준 교과서:** Weste & Harris, *CMOS VLSI Design: A Circuits and Systems Perspective* — 수율·면적·전력 모형이 정리되어 있다
- **수율 이론:** Stapper의 IBM 수율 모형 논문들 — Stapper, McLaren, Dreckmann, "Yield Model for Productivity Optimization of VLSI Memory Chips with Redundancy and Partially Good Product", *IBM J. Res. Dev.* 24(3):398–409 (1980), doi:10.1147/rd.243.0398 · Stapper 외, "Evolution and Accomplishments of VLSI Yield Management at IBM", *IBM J. Res. Dev.* 26(5):532–545 (1982), doi:10.1147/rd.265.0532. 여러 수율 모형의 비교는 Cunningham, "The Use and Evaluation of Yield Models in Integrated Circuit Manufacturing", *IEEE Trans. Semiconductor Manufacturing* 3(2):60–71 (1990), doi:10.1109/66.53188
- **GDPW 근사식:** de Vries, "Investigation of Gross Die Per Wafer Formulas", *IEEE Trans. Semiconductor Manufacturing* 18(1):136–139 (2005), doi:10.1109/TSM.2004.836656 — 0절 표를 직접 계산해 본 뒤 읽으면 좋다
- **직접 해보기:** Tiny Tapeout — 소액으로 실제 칩에 자기 설계를 올려볼 수 있는 프로젝트. **CS 전공자가 테이프아웃 파이프라인을 체감하기에 가장 좋은 방법**
- **오픈 툴체인:** OpenLane / OpenROAD + SkyWater 130nm 오픈 PDK — RTL부터 GDS까지 무료로 돌려볼 수 있다
- **AI × EDA:** Mirhoseini et al., "A graph placement methodology for fast chip design", *Nature* 594 (2021), doi:10.1038/s41586-021-03544-w — 강화학습 기반 배치. 이 논문은 재현성 논쟁이 있었으므로 후속 반박·재반박까지 함께 읽으면 좋다
- **산업 구조:** TSMC·Intel·Samsung의 IR 자료와 기술 심포지엄 발표 — TSMC 분기 실적 · Intel

보도자료

자습 부록 A — 정량 실습 풀이

8절 실습의 풀이다. 면적은 모두 cm²로 환산해서 넣는다(100 mm² = 1 cm²).

실습 1 풀이 — 수율 곡선

▾ **풀이**

$Y = e^{-AD_0}$.

다이 면적	$D_0 = 0.05$	$D_0 = 0.1$	$D_0 = 0.2$
50 mm²	97.5%	95.1%	90.5%
100 mm²	95.1%	90.5%	81.9%
200 mm²	90.5%	81.9%	67.0%
400 mm²	81.9%	67.0%	44.9%
600 mm²	74.1%	54.9%	30.1%
800 mm²	67.0%	**44.9%**	**20.2%**

"어느 면적부터 급락하는가"에 답하는 법

지수함수에는 무릎이 따로 없다. 대신 기준선을 하나 정하면 깔끔한 식이 나온다. **수율 50%가 되는 면적**은

$$Y = 0.5 \;\Rightarrow\; A_{50} = \frac{\ln 2}{D_0} = \frac{0.693}{D_0}\ \text{cm}^2$$

D_0	A_{50}
0.05	1,386 mm² (레티클 한계보다 큼 — 50% 기준으로는 노광 면적이 먼저 막는다. 레티클 한계 858 mm²에서도 수율은 65.1%라 원가 영향은 남는다)
0.1	**693 mm²**
0.2	**347 mm²**

읽는 법: D_0 가 두 배가 되면 같은 목표 수율(여기서는 50%)에서 허용되는 다이 면적이 **절반이** 된다. 50%는 설명을 위해 고른 기준선일 뿐이고, 이 면적을 넘는 칩도 만들 수는 있다. 다만 수율이 낮아진 만큼 원가가 오른다. 신규 노드 초기에 D_0 가 높다는 것은 큰 칩을 감당할 만한 원가로 만들기 어렵다는 뜻이고, 이것이 선두 파운드리의 구조적 이점이자 대형 AI 가속기가 최신 노드를 바로 쓰기 어려운 이유의 하나다. 실제로 만들 수 있는 최대 크기는 수율이 아니라 레티클 한계가 정한다.

실습 2 풀이 — 다이당 원가

▾ 풀이

$d = 30$ cm, $D_0 = 0.1$ 로 두고 GDPW와 수율을 모두 적용한다.

$$\text{GDPW} \approx \frac{\pi \cdot 900 / 4}{A} - \frac{\pi \cdot 30}{\sqrt{2A}} = \frac{706.9}{A} - \frac{94.25}{\sqrt{2A}}$$

다이 면적	GDPW	수율	굿다이	다이당 원가
100 mm² (1 cm²)	640.2	0.905	579.3	0.00173 W
400 mm² (4 cm²)	143.4	0.670	96.1	0.01040 W
800 mm² (8 cm²)	64.8	0.449	29.1	0.03435 W

계산 예 (800 mm²): $706.9 / 8 = 88.36$, $94.25 / \sqrt{16} = 23.56$, GDPW $= 64.80$. 굿다이 $= 64.80 \times 0.4493 = 29.1$. 원가 $= W / 29.1$.

핵심 결과

$$\frac{0.03435}{0.00173} \approx 19.9\text{배}$$

면적은 8배인데 원가는 20배다. 세 가지가 곱해진 결과다 — 면적으로 8배, 수율로 2.0배, 가장자리 손실로 1.2배.

함의

- 800 mm²급 AI 가속기 한 개의 실리콘 원가는 100 mm²급 칩 20개와 맞먹는다. **"큰 칩은 비싸다"가 아니라 "면적에 대해 초선형으로 비싸다"** 가 정확한 진술이다.
- 이 초선형성이 **칩렛의 경제적 동인**이다. 같은 총 면적을 잘게 나눌 수 있으면 원가가 급락한다. 실습 4에서 계산한다.
- 2회차 실습 4의 c_A 를 여기서 실제로 채울 수 있다. 웨이퍼 한 장을 15,000달러로 놓으면 800 mm² 다이의 실리콘 원가만 약 515달러이고, 여기에 HBM·패키징·테스트가 얹힌다.

아래 그림은 다이 면적 800 mm², 결함 밀도 $D_0 = 0.1$, 다이 꼭짓점을 웨이퍼 중심에 맞춘 배치일 때다. 화면본에서는 다이 면적, 격자 위치, D_0, 웨이퍼 값, 결함 모형을 바꿔 가며 직접 센 개수와 두 근사식, 수율, 다이당 원가가 어떻게 움직이는지 볼 수 있다. 그림의 결함 점은 무작위 표본 하나이고, 판독부의 수율과 원가는 식으로 계산한 값이다.



그림 4-1. 300 mm 웨이퍼에 800 mm² 다이를 깔면 직접 센 개수는 68개(격자 위치에 따라 66~71개)이고, √(2A) 근사식은 64.8개다. $D_0 = 0.1$이면 수율 44.9%, 굿다이 29.1개, 다이당 원가는 웨이퍼 값의 0.03435배로 100 mm² 다이의 약 19.9배다. 결함 점은 무작위 표본 하나다. — 화면본에서는 이 자리에서 값을 바꿔 볼 수 있다(화면본 4회차 실습 2 풀이).

실습 3 풀이 — 푸아송 vs 음이항

▾ 풀이

$\lambda = AD_0 = 0.8$ (800 mm², $D_0 = 0.1$) 로 고정하고 비교한다.

모형	식	수율
푸아송	$e^{-0.8}$	44.93%
음이항 $\alpha = 5$	$(1 + 0.16)^{-5}$	47.61%
음이항 $\alpha = 2$	$(1 + 0.40)^{-2}$	51.02%
음이항 $\alpha = 1$	$(1 + 0.80)^{-1}$	**55.56%**

클러스터링이 심할수록(α 가 작을수록) 수율이 높다. 결함의 총 기댓값은 똑같은데도 그렇다.

확률적 설명 — 젠센 부등식 한 줄

결함 밀도 자체가 웨이퍼마다 흔들린다고 보자. 즉 Λ 가 확률변수이고 $E[\Lambda] = \lambda$ 다. 수율은 조건부 푸아송의 $P(X = 0)$ 를 평균한 것이므로

$$Y = E\left[e^{-\Lambda}\right] \;\geq\; e^{-E[\Lambda]} = e^{-\lambda}$$

e^{-x} 가 **볼록함수**이므로 젠센 부등식이 항상 이 방향으로 성립한다. **결함이 뭉칠수록(분산이 클수록) 수율은 푸아송 예측보다 반드시 높다.**

$\Lambda \sim \text{Gamma}(\alpha, \lambda / \alpha)$ 로 두면 위 기댓값이 정확히 감마 분포의 적률생성함수이고, 그 값이 $(1 + \lambda / \alpha)^{-\alpha}$ — 음이항 수율식이다. $\alpha \to \infty$ 이면 분산이 0으로 가서 푸아송으로 수렴한다.

적분을 직접 적으면. 결함 밀도 D 가 평균 D_0, 모양 모수 α 인 감마 분포를 따른다고 두고, 조건부 수율 e^{-AD} 를 그 밀도함수로 평균한다.

$$Y = \int_0^\infty e^{-AD}\,\frac{(\alpha / D_0)^\alpha}{\Gamma(\alpha)}\,D^{\alpha-1}e^{-\alpha D / D_0}\,dD = \frac{(\alpha / D_0)^\alpha}{(A + \alpha / D_0)^\alpha} = \left(1 + \frac{AD_0}{\alpha}\right)^{-\alpha}$$

가운데 단계는 감마 함수의 정의 $\int_0^\infty D^{\alpha-1}e^{-cD}\,dD = \Gamma(\alpha) / c^\alpha$ 에 $c = A + \alpha / D_0$ 를 넣은 것이다. 이 감마 분포의 분산은 다음과 같다.

$$\text{Var}\,(D) = \frac{D_0^2}{\alpha}$$

이 식이 있으면 α 를 「클러스터링 파라미터」라는 이름 대신 숫자로 읽을 수 있다. $\alpha = 1$ 이면 결함 밀도의 표준편차가 평균과 같고, $\alpha = 5$ 이면 평균의 약 45%다$(1 / \sqrt{5})$.

수치로 확인하면(계산하면), $D_0 = 0.1$, $A = 8$ 에서 위 적분을 수치로 구한 값과 닫힌 식이 $\alpha = 1, 2, 5$ 각각 0.555556, 0.510204, 0.476113으로 일치하고, 분산은 0.01, 0.005, 0.002로 D_0^2 / α 와 같다. 위 표의 음이항 수율이 이 값들이다.

직관: 결함 100개가 웨이퍼 전체에 고르게 흩어지면 거의 모든 다이가 하나씩 맞는다. 같은 100개가 다이 10개에 몰리면 나머지는 전부 깨끗하다. **수율은 "결함이 몇 개인가"가 아니라 "몇 개의 다이가 오염됐는가"의 문제다.**

실무적 함의: 푸아송을 쓰면 수율을 보수적으로(낮게) 추정한다. 팹이 실제 웨이퍼 맵 데이터로 α 를 추정해 쓰는 이유가 이것이고, 그 웨이퍼 맵 패턴 분류가 5회차 검사·계측 절의 머신러닝 응용 지점이다.

아래 그림은 $\lambda = 0.8$ 에서 다이를 칸 10개로 쪼갠 경우와, 800 mm² 다이·$D_0 = 0.1$ 에서 푸아송과 음이항($\alpha = 2$)을 비교한 경우다. 화면본에서는 칸 수, λ, 다이 면적, D_0, α 를 바꿔 볼 수 있다. 오른쪽 작은 격자 둘은 같은 수의 결함을 고르게 뿌린 표본과 감마로 흔든 밀도에 따라 뿌린 표본이다.

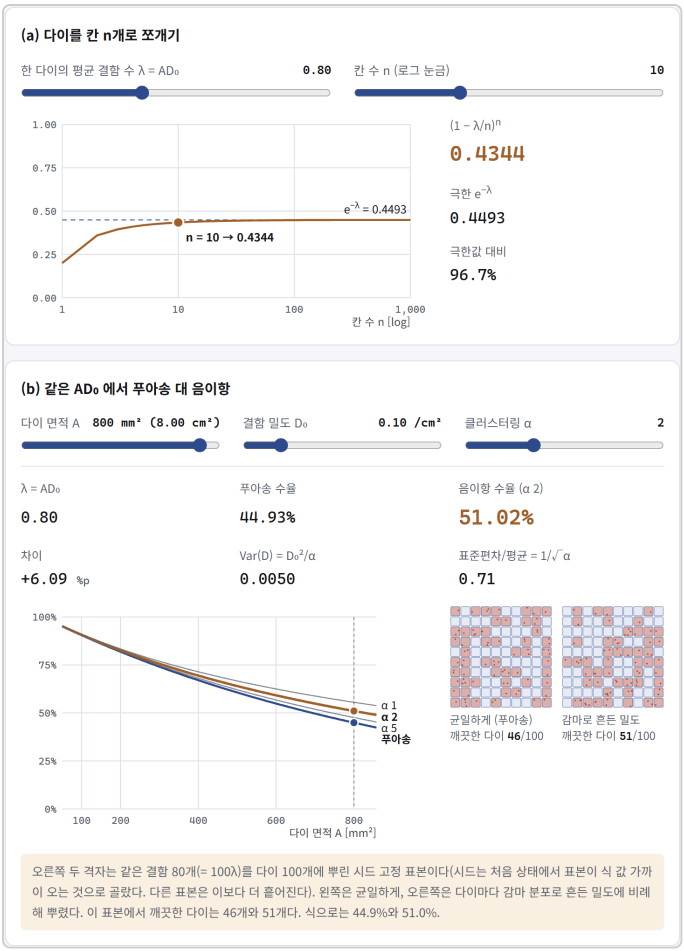


그림 4-2. λ = 0.8에서 다이를 칸 10개로 쪼개면 $(1-\lambda/n)^n = 0.4344$로 극한 $\exp(-0.8) = 0.4493$의 97%에 이른다. 같은 λ에서 음이항(α = 2) 수율은 51.0%로 푸아송 44.9%보다 높고, 결함 밀도의 분산은 $D_0^2/\alpha = 0.005$다. — 화면본에서는 이 자리에서 값을 바꿔 볼 수 있다(화면본 4회차 실습 3 풀이).

실습 4 풀이 — 칩렛 손익 (6회차 예고)

▾ 풀이

800 mm² 단일 다이 vs 200 mm² 칩렛 4개. $D_0 = 0.1$, 인터페이스 오버헤드 10%, 조립 수율 95%.

단일 다이: 위 실습 2에서 $0.03435\,W$

칩렛: 각 200 mm²에 오버헤드 10% → 220 mm² = 2.2 cm²

- GDPW $= 706.9 / 2.2 - 94.25 / \sqrt{4.4} = 321.3 - 44.93 = 276.4$
- 수율 $= e^{-0.22} = 0.8025$ → 굿다이 $= 221.8$
- 칩렛 하나당 $W / 221.8 = 0.004509\,W$, 네 개면 $0.018035\,W$ (양품 칩렛만 골라 쓴다는 KGD 가정. 굿다이 수로 나눈 원가가 곧 그 뜻이다)
- 조립 수율 0.95로 나누면 $0.018984\,W$

절감액 $= 0.03435 - 0.018984 = 0.01536\,W$

판정: 인터포저·조립 추가비용이 완성품 하나당 **웨이퍼 한 장 값의 약 1.5%(0.01536 W)보다 작으면** 칩렛이 이긴다. 웨이퍼를 15,000달러로 보면 시스템당 약 230달러의 여유가 있다는 뜻이다.

추가비가 조립을 시도할 때마다 든다고 잡으면 실패한 시도의 몫도 0.95로 나뉘어 완성품 원가가 $(4c + C) / 0.95$ 가 된다. 이때 문턱은 $C < 0.95 \times 0.03435 - 0.018035 = 0.01459\,W$, 시도당 약 219달러로 조금 내려간다. 어느 기준으로 잡든 결론의 방향은 같다.

결론이 뒤집히는 지점 — D_0 를 바꿔 보면

D_0	단일	칩렛 4분할	절감액	판정
0.02	$0.01811\,W$	$0.01592\,W$	$0.0022\,W$	조립비가 웨이퍼값의 **0.22%** 만 넘어도 역전
0.1	$0.03435\,W$	$0.01898\,W$	$0.0154\,W$	칩렛 우세

이 예제에서 읽을 것: 칩렛의 원가 이점은 수율이 나쁠수록 크다. 같은 웨이퍼 값·같은 D_0·4분할·오버헤드 10%·조립 수율 95%라는 조건에서는, 공정이 성숙해 D_0 가 떨어지면 그 이점이 빠르게 줄어든다. 최신 노드·대면적 칩에서 칩렛이 먼저 채택된 것과 맞는 방향이다. 다만 칩렛을 쓰는 이유에는 수율 말고도 레티클 한계를 넘는 면적, 블록마다 다른 공정을 쓰는 이종 집적, 설계 재사용이 있으므로(6회차), 이 계산 하나로 칩렛이 어디에 쓰이고 어디에 안 쓰이는지를 정하지는 못한다.

> 6회차 실습 1에서 분할 수 n 을 자유변수로 놓고 최적점을 찾는다. 결과를 미리 말하면 — **최적 분할 수는 D_0 에 따라 이동한다.**

같은 계산은 6회차 실습 1 풀이 끝의 그림에서 이 실습의 가정으로 바꿔 해 볼 수 있다(화면본).

실습 5 풀이 — ASIC 판단

▾ 풀이

NRE 1.5억 달러, ASIC 4,000달러, GPU 30,000달러, 3년 전력비 각각 2,000 / 6,000달러.

$$N^* = \frac{1.5 \times 10^8}{(30{,}000 + 6{,}000) - (4{,}000 + 2{,}000)} = \frac{1.5 \times 10^8}{30{,}000} = 5{,}000 \text{ 대}$$

이 물량이 현실적인 조직

조직 유형	3년 도입 규모	판정
하이퍼스케일러	수십만 대	압도적으로 초과 — 자체 칩이 합리적
대형 AI 랩	수만 대	초과 — 다만 워크로드 변동 위험이 크다
대기업 사내 AI 인프라	수백~수천 대	**경계선.** 대개 미달
일반 기업·연구기관	수십 대	명백히 미달

여기서 나오는 산업 구조적 함의

5,000대라는 문턱은 생각보다 낮다. 그래서 하이퍼스케일러가 아니어도 **"거의 하이퍼스케일러급" 사업자들이 커스텀 실리콘을 고려할 수 있고**, 이 수요를 받아 주는 것이 Broadcom·Marvell 같은 커스텀 ASIC 설계 서비스 사업이다. 이들이 최근 급부상한 구조적 이유가 이 계산에 있다.

단, 2회차 실습 4의 경고가 그대로 적용된다. N^* 는 필요조건일 뿐이다. 충분조건은 **개발 2~3년 후에도 그 워크로드가 살아 있을 것**이며, 이것을 보장할 수 있는 조직은 자기 모델을 자기가 통제하는 곳뿐이다.

아래 그림은 실습 5의 가정값(NRE 1.5억 달러, ASIC 4,000달러, GPU 30,000달러, 3년 전력비 2,000 / 6,000달러)일 때다. 화면본에서는 다섯 값을 바꿔 N^* 가 위 표의 어느 조직 규모에 놓이는지 볼 수 있다. 6절은 o 를 운영비의 현재가치라고 했지만 할인율을 정하지 않았으므로, 그림은 3년 전력비를 할인하지 않고 그대로 넣는다. 조직 띠의 경계는 표의 「수십·수백~수천·수만·수십만 대」를 자릿수로 옮긴 것이다.

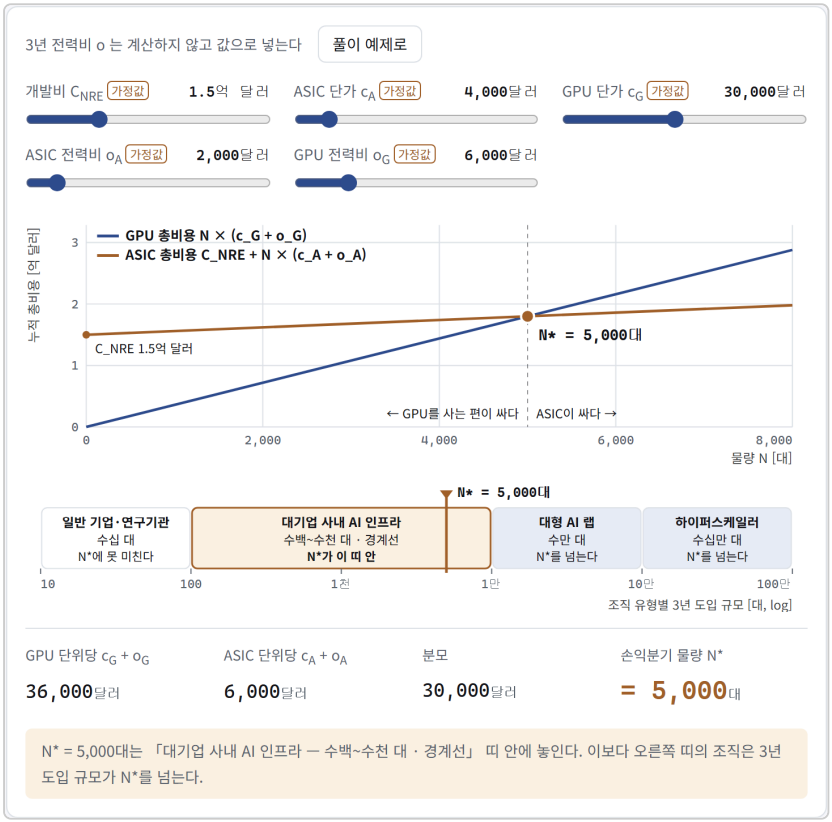


그림 4-3. 실습 5의 가정값(NRE 1.5억 달러, ASIC 4,000·GPU 30,000달러, 3년 전력비 2,000·6,000달러)이면 N* = 5,000대로, 「대기업 사내 AI 인프라 — 수백~수천 대」 띠 안에 놓인다. — 화면본에서는 이 자리에서 값을 바꿔 볼 수 있다(화면본 4회차 실습 5 풀이).

자습 부록 B — 토론 질문 생각의 실마리

질문	핵심 축	실마리
1. 분업은 되돌릴 수 있는가	**자본 집약도**	분업이 일어난 이유는 팹 투자 규모가 단일 기업의 수요를 초과했기 때문이다. 되돌리려면 한 기업의 수요가 팹 하나를 채울 만큼 커야 하는데 — **하이퍼스케일러의 AI 수요는 실제로 그 규모에 접근하고 있다.** 이것이 수직통합 재부상의 근거이자, Intel IDM 2.0과는 방향이 다른 논리다
2. GAA와 후면전력의 위험 배분	**직렬 vs 병렬 리스크**	관찰할 수 있는 것: 두 변화를 동시에 들이면 문제가 생겼을 때 어느 쪽 원인인지 가리기 어려워지고, 둘 중 하나만 늦어져도 전체 일정이 밀린다. 순차로 하면 그 위험은 작지만 시간이 더 든다. 토론용 가설: 뒤처진 쪽은 따라잡으려고 동시에, 앞선 쪽은 자리를 지키려고 순차로 갈 동기가 있다. 이 가설이 맞는지는 각 선택의 성공 확률, 지연이 낳는 손실, 선두 자리의 보상을 놓고 따져 봐야 하며, 여기서는 그 값을 정하지 않았다
3. 수율의 관찰 가능한 대리변수	**가격이 정보를 담는다**	리드타임, 비닝 구성(하위 SKU 비중이 크면 수율이 낮다는 신호), 가격 인하 속도, 신제품의 초기 물량, 파운드리 계약 형태(웨이퍼 단위인가 굿다이 단위인가 — 5절 참조). **계약 형태가 가장 정직한 신호다**
4. EDA 과점이 안 깨지는 이유	**검증된 것의 가치**	툴을 바꾸는 비용이 툴 가격이 아니라 **재검증 비용**이다. 수천만 달러짜리 마스크셋을 새 툴의 검증 결과만 믿고 찍을 수 있는가. 오픈소스 EDA는 이 신뢰 축적을 건너뛸 수 없다
5. 노드 이름을 왜 계속 쓰나	**조정 장치로서의 이름**	물리적 의미는 없어도 **세대를 지칭하는 공통 좌표**로는 여전히 작동한다. 대안(밀도 지표, 성능·전력 지표)은 더 정확하지만 단일 숫자로 압축되지 않아 마케팅과 계약 협상에서 쓰기 어렵다

자습 부록 C — 핵심 용어

용어	영문	한 줄 정의
결함 밀도	defect density D_0	단위 면적당 치명 결함 수 [결함/cm²]
임계 면적	critical area	그 크기의 결함이 떨어지면 회로가 망가지는 자리의 넓이. 결함이 클수록 넓다
다이	die	웨이퍼에서 잘라 낸 칩 하나
GDPW	gross die per wafer	웨이퍼 한 장에서 나오는 다이의 총 개수(수율 적용 전)
굿다이	good die	정상 동작하는 다이. GDPW × 수율
푸아송 수율 모형	Poisson yield model	$Y = e^{-AD_0}$. 결함이 완전 무작위라는 가정
음이항 수율 모형	negative binomial	$Y = (1 + AD_0 / \alpha)^{-\alpha}$. 클러스터링을 반영
클러스터링 파라미터	α	작을수록 결함이 뭉친다. ∞ 이면 푸아송
수율 학습곡선	yield learning	시간이 지나며 D_0 가 내려가는 과정

용어	영문	한 줄 정의
비닝	binning	다이를 등급별로 나눠 파는 일. 불량 코어를 끄고 하위 제품으로 파는 결함 비닝과, 최대 주파수·전력으로 등급을 나누는 속도·전력 비닝이 있다
칩렛	chiplet	큰 칩을 여러 작은 다이로 나눠 만든 뒤 붙이는 방식
레티클 한계	reticle limit	노광기가 한 번에 찍는 최대 면적(약 26×33 mm)
NRE	non-recurring engineering	개발에 드는 일회성 비용
구조화 ASIC	structured ASIC	일부 층만 맞춤 제작해 NRE를 낮춘 중간 형태
IDM	integrated device manufacturer	설계·제조를 함께 하는 종합 반도체 회사
PDK	process design kit	파운드리가 제공하는 공정별 설계 규칙·모델 묶음
트랙	track	표준셀 안을 지나는 가로 배선 한 줄의 자리. 셀 높이를 트랙 수로 센다(6T·5T 셀). SRAM 6T의 T(트랜지스터)와 다르다

자습 부록 D — 자가진단

☐ $Y = e^{-AD_0}$ 를 쓰고, 왜 지수함수인지(푸아송의 $P(X=0)$) 설명할 수 있다 → 5절 ★

☐ 면적이 8배일 때 원가가 왜 20배가 되는지 **세 가지 요인으로 분해**할 수 있다 → 실습 2 ★

☐ $A_{50} = \ln 2 \,/\, D_0$ 를 유도하고, D_0 가 두 배면 같은 목표 수율에서 허용되는 다이 면적이 절반임을 말할 수 있다 → 실습 1

☐ GDPW 식의 두 항이 각각 무엇인지 설명할 수 있다 → 자습 안내

☐ 클러스터링이 수율을 **올린다**는 것과 그 이유(젠센 부등식)를 설명할 수 있다 → 실습 3

☐ 칩렛의 원가 이점이 D_0 **가 높을수록** 커지고 낮아지면 줄어드는 것을 계산으로 보일 수 있다 → 실습 4 ★

☐ N^* 를 계산하고, 그것이 필요조건일 뿐인 이유를 말할 수 있다 → 실습 5

☐ 파운드리 계약 형태(웨이퍼 단위 vs 굿다이 단위)가 무엇을 배분하는지 설명할 수 있다 → 5절

다음 회차로 넘어가기 전에

여기까지가 **바깥에서 본 제조**다. 5회차부터 실제로 안으로 들어간다.

이 회차가 남긴 질문	이후 회차의 답
D_0 는 어디서 오는가	**5회차** — 1,000단계 안팎 공정 각각의 결함
왜 공정 하나하나가 그렇게 정밀해야 하나	**5회차 실습 4** — 스텝 수율의 지수적 누적
레티클 한계가 왜 칩 크기를 정하나	**5회차 실습 5** → **6회차** 칩렛
칩렛을 실제로 어떻게 붙이나	**6회차** — 인터포저·TSV·하이브리드 본딩

출처

스터디 커리큘럼이 고른 입문 자료 ("팹리스" 항목)

커리큘럼 소주제	링크	이 책에서의 활용
용어설명	삼성전자 반도체 뉴스룸, 「파운드리? 팹리스? 반도체 생태계 한눈에 보기!」	분업 구조 입문. 2절의 기초. **발제 도입부로 쓰기 좋은 자료**
팹리스업체동향	ScienceON JAKO202470953293995, 「최신 인공지능 반도체 및 컴파일러 지원 기술 동향」	국내외 AI 반도체 설계사(사피온·리벨리온·퓨리오사AI, 구글·메타·테슬라)와 컴파일러(MLIR·TVM·XLA) 동향을 정리한 논문. 2절 팹리스 층의 국내 사례
팹리스업체동향	ScienceON JAKO202369265025811, 「초거대 인공지능 프로세서 반도체 기술 개발 동향」	초거대 AI용 프로세서를 개발하는 국내외 기업(NVIDIA·Graphcore·Tesla·Google·Meta와 국내 3사)의 기술 동향 보고. 2절 팹리스 층의 사례 배경
ASIC	ScienceON ASTI00673636, 「Market Insight 336: AI 주문형 반도체(AI ASIC)」 (보고서)	AI ASIC 기술·산업 보고서. 6절의 배경
전체 공정	윈스펙, 「반도체 공정 종류 뭐가 있을까? 전공정·후공정 차이 정리」	반도체 전체 공정 개괄. **3절 파이프라인 및 5회차 전공정의 예습 자료로 적합**

이번 개정에서 더한 1차 출처 (2026-10-08 확인)

- TSMC, Open Innovation Platform® · TSMC, Value Chain Alliance · TSMC 3DFabric® Alliance — 2절 디자인하우스, 3절 OIP
- TSMC 보도자료(2018-10-03), 「TSMC Announces 5th Open Innovation Platform® Alliance, the OIP Cloud Alliance」 · TSMC 보도자료(2022-10-27), 「TSMC Launches OIP 3DFabric Alliance ...」 — 3절 얼라이언스 구성과 연도
- Samsung Foundry, SAFE™ 프로그램 — 2절 디자인 서비스 얼라이언스(DSP·VDP), 3절 SAFE 구성
- SEMI P39, Specification for OASIS® – Open Artwork System Interchange Standard — 3절 레이아웃 파일 형식
- IEEE Std 1364-2005 (Verilog HDL) — 3절 논블로킹 대입(대체 규격 IEEE Std 1800)
- TSMC 2nm(N2) 기술 페이지 · TSMC 보도자료(2024-04-24), A16 발표 · Tom's Hardware(2026-04-22), 기술 심포지엄 기사 — A16 양산 2027년(TSMC 임원 발언 인용) — 4절 표
- Intel 2025년 4분기·연간 실적 보도자료(2026-01-22) · Intel 2026년 2분기 실적 보도자료(2026-07-23) · Intel Foundry VLSI 심포지엄 보도자료(2026-06-16) — 4절 표
- Samsung Foundry 로직 노드 페이지 · Exynos 2600 제품 페이지 — 4절 표
- SK하이닉스 뉴스룸(2024-04-19), 「SK hynix Partners with TSMC to Strengthen HBM Technological Leadership」 — 2절 HBM4 베이스 다이
- Stapper, McLaren, Dreckmann (1980), doi:10.1147/rd.243.0398 · Stapper, Castrucci, Maeder, Rowe, Verhelst (1982), doi:10.1147/rd.265.0532 — 5절 임계 면적·혼합 모형·리페어와 부분 양품 (서지는 Crossref, 내용은 출판사 초록으로 확인)

- Zhang, Ren, Su, "On-Chip Structures for Fmax Binning and Optimization", *Sensors* 22(4):1382 (2022), PMC8963108 · Datta 외, "Speed binning aware design methodology to improve profit under parameter variations", ASP-DAC 2006, pp. 712–717 (서지만 Crossref로 확인) — 5절 속도 비닝
- de Vries (2005), doi:10.1109/TSM.2004.836656 — 0절 GDPW 근사식 비교 문헌(서지만 확인, 본문 미열람)

검증 상태

- 2nm급 노드 현황(4절 표) — 각 사 공식 페이지·보도자료로 2026-10-08 재확인. TSMC N2 2025년 4분기 양산 시작·N2P 2026년 하반기 양산 예정(TSMC 2nm 기술 페이지), Intel 18A 첫 제품 Core Ultra 시리즈 3 발표(2026-01-22 실적 보도자료)·Xeon 6+ 출시(2026년 2분기 실적 보도자료)·18A-P 리스크 생산(2026-06-16), Samsung SF2 「2025년 양산 시작」(로직 노드 페이지)과 Exynos 2600 「업계 첫 2nm GAA 공정 기반」(제품 페이지). 트랜지스터 구조 이름(나노시트·RibbonFET·MBCFET)과 PowerVia 적용은 2026-08-02 대조 결과를 유지
- TSMC A16의 Super Power Rail(후면전력공급)과 생산 일정 — 2024-04-24 보도자료(2026년 생산 계획) 원문 확인(2026-10-08). 양산 2027년 시작 예상은 2026-04-22 기술 심포지엄의 TSMC 임원 발언을 옮긴 보도(Tom's Hardware 2026-04-22)로만 확인되고, 같은 날 TSMC 보도자료에는 A16 언급이 없다. 2026-10-15 3분기 실적 보도자료에도 A16 언급이 없고, 양산 개시 공식 발표는 아직 없다(2026-10-22 확인)
- TSMC OIP 다섯 얼라이언스(2018)와 3DFabric 얼라이언스(2022년 10월, 여섯째) — TSMC 보도자료 원문 확인(2026-10-08)
- 삼성 SAFE™의 구성(PDK·설계 방법론·IP 인증, 디자인 서비스 얼라이언스의 DSP·VDP) — 삼성 공식 페이지 원문 확인(2026-10-08)
- 디자인하우스·TSMC VCA의 성격 — TSMC 공식 페이지로 확인(2026-10-08). 회원 명단은 시점에 따라 바뀌므로 본문에 쓰지 않음
- SK하이닉스–TSMC HBM4 협력 양해각서(2024-04-19) — SK하이닉스 뉴스룸 원문 확인(2026-10-08)
- 푸아송·머피·Seeds·음이항 수율 모형의 수식 — 반도체 수율 이론의 표준 정식화
- 이항→푸아송 극한 수치, 감마 혼합 적분과 $\mathrm{Var}(D) = D_0^2 / \alpha$ — 파이썬으로 재계산(2026-10-08)
- GDPW 근사식 둘째 항의 뜻(다이 대각선 $\sqrt{2A}$)과 직접 센 개수 비교 — 300 mm 웨이퍼에 다이 격자 400가지 위치를 깔아 직접 계산(2026-10-08)
- 임계 면적, IBM 메모리 수율 모형의 혼합 구조, 64K비트 RAM의 리페어·부분 양품 수율 향상 — Stapper 외(1980·1982) 출판사 초록으로 확인(2026-10-08). 본문 전문은 열람하지 못해 그림·표 수치는 옮기지 않음
- 속도 비닝이 공정 변동에서 생긴다는 점 — Zhang 외(2022) 원문 확인(2026-10-08)
- OASIS가 GDSII 스트림의 후속 레이아웃 교환 형식이라는 점 — SEMI P39 규격 페이지 확인(2026-10-08)
- 논블로킹 대입의 의미 — IEEE Std 1364-2005 규격 페이지만 확인했고 규격 본문은 열람하지 못함. 의미 자체는 Verilog의 표준 동작
- 300mm 웨이퍼 규격 — 표준

미확인 · 유보 (본문에서 단정하지 않음)

- **팹 건설비, 마스크셋 비용, NRE 규모** — 노드·업체·계약에 따라 편차가 크고 공식 공개가 드물다. 본문은 자릿수 규모로만 언급
- **결함밀도 D_0 값** — 업체 영업비밀. 본문의 0.05~0.2는 **계산 예시용 가정값**이며 특정 업체의 실제 값이 아니다
- **각 사 노드별 수율** — 언론 보도(예: SF2 수율 30~60%대)는 출처가 불명확한 경우가 많아 본문에 수치로 반영하지 않았다
- **트랜지스터 밀도(MTr/mm²) 구체 수치** — 벤더별 측정 기준이 달라 직접 비교가 어렵다. 지표의 존재와 읽는 법만 서술
- **8절 실습의 단가·NRE·전력비** — 전부 예시용 가정값이며 실제 시장가가 아니다

5회차

전공정

핵심 TAKEAWAY

전공정 8대 공정은 외울 목록이 아니라 하나의 반복 루프다: "패턴을 그린다 → 재료를 더하거나 뺀다 → 평탄화한다 → 검사한다." 이 루프가 층마다 반복되어 수백 단계가 된다. 그 중심에 노광이 있고, 노광은 단 하나의 식으로 요약된다 — 레일리 판정식 $\mathrm{CD} = k_1\lambda / \mathrm{NA}$. 파장을 줄이거나, 개구수를 키우거나, 공정계수를 줄이거나. **반도체 미세화의 역사 전체가 이 세 항을 공격한 역사다.** 트랜지스터가 Planar → FinFET → GAA로 바뀐 이유도 한 문장이다: 채널이 짧아지자 게이트의 제어력이 약해졌고, 게이트가 채널을 감싸는 면적을 늘리는 것이 유일한 해법이었다.

원자료 바로가기

삼성전자 반도체 뉴스룸 「반도체 8대 공정 1탄: 웨이퍼」 · 삼성전자 반도체 「반도체 8대 공정 2탄, 웨이퍼 표면을 보호하는 산화공정」 · 삼성전자 반도체 뉴스룸 「반도체 8대 공정 6탄: 증착과 이온주입」 · ETRI 전자통신동향분석 제38권 제6호(2023) 「Complementary FET로 열어가는 반도체 미래 기술」 · ScienceON JAKO202509150403114 「EUV 리소그래피 공정 기술 동향 및 연구 전망」 · ScienceON JAKO201326940561717 「원자층증착 기술: 개요 및 응용분야」 · ASML 「5 things you should know about High NA EUV lithography」(2024) · van Schoot 외(ASML·ZEISS), 「High-NA EUV lithography enabling Moore's law in the next decade」, EUVL Workshop 2017 · TSMC 보도자료(2019-10-07), N7+ EUV 양산 · Deal & Grove(1965), doi:10.1063/1.1713945

학습목표

1. 전공정 8대 공정을 개별 나열이 아니라 하나의 반복 루프로 재구성할 수 있다.
2. 레일리 판정식의 세 항을 설명하고, 산업이 각 항을 어떻게 공격했는지 서술할 수 있다.
3. FinFET·GAA 전환의 물리적 필연성을 설명할 수 있다.
4. EUV의 물리적 난점(반사광학·광원·확률적 결함)을 설명할 수 있다.
5. 각 공정이 어떤 물리량을 제어하는지 구분할 수 있다.

0. 자습 안내 — 혼자 읽는 사람에게

> **이 회차는 가장 길고 가장 물리적인 회차다**
>
> 앞 네 회차가 "얼마나"의 문제였다면 이 회차는 "어떻게"의 문제다. 분량에 겁먹지 말 것 — **실제로 어려운 것은 4절(노광) 하나뿐이고, 나머지는 4절을 이해하기 위한 준비이거나 그 결과다.** 시간이 부족하면 3절과 4절만 읽어도 이 회차의 8할을 가져간다.

직전 회차에서 가져오는 것

앞 회차에서	이 회차에서
4회차 $Y = e^{-AD_0}$	그 D_0 를 만드는 1,000개 안팎의 공정 스텝
4회차 레티클 한계	그 한계가 어디서 오는지 — 노광 필드
3회차 "HBM은 왜 비싼가"	여기서는 로직 칩이 왜 비싼가

선행 지식 점검 — 몰라도 되는 것과 꼭 알아야 하는 것

개념	한 줄 답	필요도
MOSFET	게이트에 전압을 걸면 소스-드레인 사이에 전류 통로(채널)가 열리는 스위치	★ 필수
도핑	Si는 바깥 전자가 4개라 이웃 넷과 빈틈없이 짝을 짓는다. 전자 5개인 인(P)·비소(As)를 섞으면 전자 하나가 남아 n형, 3개인 붕소(B)를 섞으면 빈자리(정공)가 생겨 p형이 된다. 4인석이 꽉 찬 식당에 5인 일행이 오면 한 명이 서서 돌아다니고, 3인 일행이 오면 빈자리를 따라 손님들이 한 칸씩 옮겨 앉는 셈이다. 순수한 Si도 부도체가 아니라 전도도가 낮은 반도체다	☆ 개념만 (7절)
파장·주파수	빛의 파장이 짧을수록 광자 에너지가 크다. $E = hc / \lambda$	★ 필수 (실습 3)
개구수 NA	렌즈가 빛을 얼마나 넓은 각도로 모으는가. 클수록 해상도가 좋다	★ 필수
회절 한계	빛이 그릴 수 있는 가장 작은 선폭은 파장에 비례하고 NA에 반비례한다($k_1\lambda$ / NA). 공정계수 k_1 을 줄이는 기법으로 파장보다 작은 선폭도 그린다 — 실습 1에서 193 nm 빛으로 약 43 nm	★ 필수
플라즈마	전자와 이온으로 분리된 기체. 식각에서 반응성 이온을 만든다	☆ 개념만
열처리·확산	이온주입으로 부서진 결정을 열로 회복시키고 도펀트를 격자 자리에 앉히는 과정(활성화). 열을 오래 주면 도펀트가 퍼진다(확산)	☆ 개념만 (7절)

신호처리·통신 배경이 있다면 지름길이 있다. 4절의 「신호처리로 이해하기」 절을 먼저 읽을 것. 노광은 결국 **저역통과 필터를 통과시킨 뒤 원래 신호를 복원하는 문제**이고, OPC(광학근접보정)는 **전치보상(pre-distortion)** 이다. 이 대응을 잡으면 4절 전체가 익숙한 언어로 읽힌다.

읽는 순서 (총 160분 권장 — 두 번에 나눠 읽어도 좋다)

순서	절	시간	비고
1일차			
1	1·2절 직관·웨이퍼	20분	
2	**3절 트랜지스터 구조 진화**	30분	개념적 하이라이트
3	**4절 노광**	**45분**	**이 회차의 핵심.** 실습 1·2·3을 여기서
2일차			
4	5~7절 식각·증착·이온주입	30분	4절보다 훨씬 쉽다
5	8절 검사·계측	15분	CS·AI 배경이면 가장 재미있는 절
6	9절 통합 + 실습 4·5	20분	

절대 건너뛰지 말 것

1. **4절 레일리 판정식** — 실습 1·2·5가 전부 여기서 나온다.
2. **4절 「EUV의 잔인한 물리」** — 실습 3의 배경이고, EUV가 왜 그렇게 비싼지의 진짜 이유다.
3. **실습 4 스텝 수율** — 5분이면 풀리는데, 반도체 제조의 난이도를 한 숫자로 요약한다.

막히기 쉬운 지점 — 「NA를 키우면 왜 초점심도가 제곱으로 나빠지나」

두 식을 나란히 놓으면 긴장 구조가 보인다.

$$\mathrm{CD} = k_1 \frac{\lambda}{\mathrm{NA}}, \qquad \mathrm{DOF} = k_2 \frac{\lambda}{\mathrm{NA}^2}$$

직관적으로 이렇다. NA가 크다는 것은 **빛을 아주 넓은 각도로 모아 온다**는 뜻이다. 넓은 각도로 모이면 초점에서는 더 뾰족해지지만(해상도 ↑), 초점에서 조금만 벗어나도 각 방향에서 온 빛들이 서로 크게 어긋난다(초점심도 ↓). 초점에서 z 만큼 벗어나면 비스듬히 들어온 가장자리 빛과 수직으로 들어온 빛의 경로차가 대략 $z(1-\cos\theta) \approx z\sin^2\theta\,/\,2$ 가 된다. 공기 중이면 $\sin\theta = \mathrm{NA}$ 이므로 경로차는 NA의 제곱에 비례하고, 이 경로차가 파장의 일정 비율을 넘으면 상이 흐려지므로 허용되는 z 가 $\lambda\,/\,\mathrm{NA}^2$ 에 비례한다. 같은 매질에서 NA를 두 배로 하면 CD는 절반, 초점심도는 4분의 1이 된다.

사진으로 비유하면: 조리개를 활짝 열면(NA ↑) 선명하지만 **초점이 맞는 깊이가 종잇장처럼 얇아진다.** High-NA EUV의 DOF가 수십 나노미터라는 것은, 웨이퍼 표면의 요철이 그 안에 들어와야 한다는 뜻이고, 그래서 **CMP(평탄화)가 선택이 아니라 필수 공정**이 된다.

---

1. 직관 잡기 — 반도체 제조는 "조각"이 아니라 "인쇄와 도장"

전체 그림

칩을 만드는 일을 조각(깎아내기)으로 생각하면 틀린다. 실제로는 **다층 인쇄**에 가깝다.

1. 실리콘 판(웨이퍼) 위에 얇은 막을 씌운다 (**증착**)
2. 감광물질을 바르고 빛으로 패턴을 그린다 (**노광**)
3. 빛을 받은 곳(또는 안 받은 곳)을 현상해 없앤다
4. 노출된 부분만 깎아낸다 (**식각**)
5. 남은 감광물질을 제거한다
6. 표면을 평평하게 다듬는다 (**CMP**)
7. 검사한다 (**계측·검사**)
8. → 1번으로 돌아가 다음 층

이 루프가 **수십 번 반복**되고, 각 반복 안에 세부 단계가 많아 전체 공정 스텝은 **1,000단계**를 넘는다. 팹에 웨이퍼를 넣고 완성품이 나오기까지 **수개월**이 걸린다.

왜 층이 이렇게 많나

칩은 크게 두 부분이다.

- **FEOL (Front-End of Line):** 실리콘 표면에 트랜지스터를 만드는 부분. 소자층.
- **BEOL (Back-End of Line):** 그 위에 배선을 여러 층 쌓아 트랜지스터들을 연결하는 부분. 최신 칩은 배선층이 15~20층에 이른다.

> 도시로 비유하면 FEOL은 건물이고 BEOL은 도로망이다. 그리고 **최근에는 도로망이 더 큰 문제다** — 배선이 가늘어지면 저항이 급증해 신호 지연과 전력 소모를 지배하게 되었다. 이것이 후면전력공급(BSPDN) 같은 기법이 나온 배경이다.

---

2. 웨이퍼 — 출발점

만드는 법

1. **다결정 실리콘 정제** — 규석에서 금속급 실리콘을 얻고, 지멘스 공정 등으로 초고순도로 정제. **11N(99.999999999%) 수준**의 순도가 요구된다. 불순물 하나가 소자 특성을 바꾸기 때문이다.
2. **단결정 성장 (초크랄스키법)** — 용융 실리콘에 씨결정을 담가 회전시키며 천천히 끌어올리면, 원자가 씨결정의 결정 방향을 그대로 이어받아 거대한 단결정 잉곳이 자란다.
3. **슬라이싱 → 래핑 → 에칭 → 폴리싱** — 잉곳을 얇게 썰고 표면을 원자 수준으로 평탄하게 연마한다.
4. **에피택시(선택적)** — 웨이퍼 위에 결정 방향을 이어받는 고품질 실리콘 층을 추가 성장시킨다.

왜 300mm에서 멈췄나

웨이퍼가 클수록 한 번에 더 많은 다이를 뽑아 다이당 원가가 내려간다. 그래서 150 → 200 → 300mm로 커졌다. 그런데 **450mm 전환은 좌초했다.**

이유는 순수하게 경제적이다. 웨이퍼 지름이 1.5배가 되면 면적은 2.25배가 되지만, 그 큰 웨이퍼를 다루는 장비를 전부 새로 만들어야 하고(노광기·식각기·이송장치 전부), 웨이퍼 전체에서 균일한 공정 결과를 얻기가 훨씬 어려워진다. **개발비가 절감액을 넘어섰다.** 산업이 대신 선택한 것은 웨이퍼 확대가 아니라

미세화와 3차원화(적층)였다.

> **일반 교훈:** 반도체 산업의 기술 선택은 물리적 가능성이 아니라 **비용 대비 효과**로 결정된다. 450mm 좌초는 그 대표 사례다.

웨이퍼 위의 이름들

공정을 마친 웨이퍼를 내려다보면 같은 네모가 바둑판처럼 반복된다. 부위마다 이름이 있다.

- **칩(다이):** 바둑판의 칸 하나. 노광기가 같은 패턴을 칸마다 되풀이해 찍은 결과다.
- **스크라이브 라인:** 다이 사이의 좁은 통로. 공정이 끝나면 이 선을 따라 잘라 낸다(6회차 후공정).
- **TEG(test element group):** 스크라이브 라인 같은 빈자리에 넣어 두는 시험용 소자와 패턴. 선폭·저항·문턱전압이 목표대로 나왔는지 재는 데 쓰고, 제품에는 들어가지 않는다.
- **가장자리 다이:** 둥근 판에 네모를 깔면 테두리에 온전하지 못한 칸이 생긴다. 4회차 웨이퍼당 다이 수 식의 둘째 항이 바로 이 손실이다. 웨이퍼 사양서에 적힌 가장자리 제외 폭(300 mm 사양서 예: 3 mm)도 쓸 수 없는 띠다.
- **방향 표지:** 결정 방향과 웨이퍼가 놓인 각도를 장비에 알려 주는 표시. 공급사 사양서를 보면 150 mm 이하 웨이퍼는 대개 가장자리를 직선으로 잘라 낸 **플랫(flat)을**, 200·300 mm 웨이퍼는 V자 홈인 **노치(notch)를** 쓴다. 300 mm 사양서의 노치는 깊이 1.0 mm, 각 90°, 방향 <110> 이다(같은 사양서의 지름 300±0.1 mm, 두께 775±25 μm).

열산화 — 실리콘이 스스로 기르는 첫 막

1절의 루프에는 빠져 있지만 많은 공정이 여기서 출발한다. 실리콘 표면을 산화시켜 SiO_2 막을 기르는 일이다. 증착(6절)이 위에 다른 물질을 쌓는 것이라면 열산화는 **실리콘 자신을 산화막으로 바꾸는 것이다.** 막이 자라는 만큼 아래 실리콘이 소모된다.

- **조건:** 보통 800~1,200 ℃의 고온에서 한다. 산소(O_2)만 쓰면 **건식**, 수증기(H_2O)를 함께 쓰면 **습식이다.** 건식은 느리게 자라지만 막이 치밀하고 전기적 특성이 좋아 얇은 막에 쓴다. 습식은 같은 온도와 시간에서 몇 배 두껍게 자라는 대신 밀도가 낮다.
- **쓰임:** 소자 사이의 절연막, 이온주입 때 원하지 않는 곳을 가리는 막, 식각 때 아래층을 지키는 막.

두께와 시간 — Deal–Grove 식. 산화제는 이미 자란 산화막을 뚫고 확산해 들어가 실리콘과의 경계면에서 반응한다. 두 과정 중 느린 쪽이 속도를 정한다. Deal과 Grove(1965)는 이 둘을 한 식으로 묶었다.

$$x^2 + Ax = B(t + \tau)$$

x 는 산화막 두께, t 는 산화 시간, τ 는 처음부터 있던 막을 시간으로 환산한 보정값이다. A 와 B 는 온도와 산화제(O_2·H_2O)에 따라 정해지는 계수다. 원논문은 700~1,300 ℃에서 산소와 수증기 양쪽의 산화 데이터가 이 식에 잘 맞음을 보였다.

식의 두 극한이 성장의 두 국면이다.

- **얇을 때** — x^2 를 무시하면 $x \approx (B / A)(t + \tau)$. 두께가 시간에 비례한다(선형 구간). 경계면 반응이 속도를 정한다.
- **두꺼울 때** — 충분히 오래 키워 x^2 가 지배하면 $x^2 \approx Bt$, 곧 $x \approx \sqrt{Bt}$. 막을 뚫는 확산이 속도를 정하는 포물선 구간이다. 두께를 두 배로 하려면 시간이 네 배 든다.

A·B 의 실제 값은 온도와 산화제마다 따로 정해 표로 쓴다. 이 책은 수치를 싣지 않는다.

---

3. 트랜지스터 구조의 진화 — 평면에서 FinFET·GAA로

MOSFET이 하는 일 (30초 복습)

소스와 드레인 사이에 채널이 있고, 게이트에 전압을 걸면 채널에 전하가 유도되어 전류가 흐른다. 게이트는 **밸브**다. 좋은 트랜지스터란 밸브가 확실히 잠기고(off 누설 작음) 확실히 열리는(on 전류 큼) 것이다.

문제 — 단채널 효과(Short-Channel Effect)

미세화로 채널 길이가 짧아지면 소스와 드레인이 가까워진다. 그러면 **드레인 전압이 채널에 직접 영향을 주기 시작한다.** 게이트가 아니라 드레인이 채널을 제어하는 상황이 생기고, 게이트를 껐는데도 전류가 새는(DIBL, 문턱전압 저하) 현상이 나타난다.

밸브 비유: 파이프가 짧아지자 밸브를 잠가도 반대편 압력이 그대로 밀고 들어오는 상황이다.

해법 — 게이트가 채널을 더 많이 감싸기

게이트의 제어력은 **게이트가 채널을 마주하는 면적**에 비례한다. 그래서 구조가 이렇게 진화했다.

구조	게이트가 감싸는 면	시기	특징
Planar (평면)	1면 (위)	~2011	단순, 단채널 효과에 취약
FinFET	**3면** (위+양옆)	2011~	채널을 지느러미처럼 세워 게이트가 감쌈. 핀 개수로 구동력 조절(양자화됨)
GAA / 나노시트	**4면 (전체)**	2022~	채널을 수평 판(시트)으로 만들고 게이트가 완전히 둘러쌈. **시트 폭을 연속적으로 조절해 구동력 미세 조정 가능**
(차세대) CFET	4면 + n/p 수직 적층	연구 단계	nFET과 pFET을 위아래로 겹쳐 면적 절감

GAA의 실질적 이점 두 가지

1. **더 나은 정전 제어** — 4면을 감싸므로 짧은 채널에서도 게이트가 이긴다.
2. **설계 유연성** — FinFET은 구동력을 핀 개수(정수)로만 조절할 수 있어 양자화 손실이 있었다. GAA는 나노시트의 **폭을 연속적으로** 바꿀 수 있어 회로 설계자가 원하는 구동력을 정확히 얻는다.

대가: 나노시트를 여러 장 겹쳐 만들려면 Si/SiGe 초격자를 성장시킨 뒤 SiGe만 선택적으로 제거해 시트를 띄우고, 그 틈으로 게이트 절연막과 금속을 넣어야 한다. **좁은 틈을 고르게 덮는 원자층 증착(ALD)이 이 공정의 열쇠다** (6절 참조).

후면전력공급 (BSPDN) — 다음 분기점

전통적으로 전력과 신호가 모두 웨이퍼 앞면 배선층을 공유했다. 그런데 전력 배선은 굵어야 하고(IR drop 방지), 신호 배선은 촘촘해야 해서 서로 자리를 다툰다.

해법: 전력 배선을 웨이퍼 **뒷면**으로 옮긴다. 웨이퍼를 뒤집어 얇게 갈아내고 뒷면에 전력망을 만든 뒤 나노

TSV로 트랜지스터에 직접 연결한다.

이득: 앞면 배선 혼잡 해소(면적·성능 개선) + 전력 전달 효율 개선(IR drop 감소). **대가:** 웨이퍼 박막화, 정렬, 열 방출 경로 변화 등 난이도가 크게 오른다.

2026년 10월 기준으로 Intel이 PowerVia라는 이름으로 18A에 먼저 적용했다. 18A 첫 제품(Core Ultra 시리즈 3)은 2026년 1월에 발표됐다. TSMC는 A16에 Super Power Rail로 도입한다. 2024년 4월 A16을 발표할 때는 2026년 생산을 계획했다. 2026년 4월 기술 심포지엄에서 보도된 TSMC 임원 발언에 따르면, A16은 2026년에 생산 준비를 마치지만 실제 양산은 고객 제품 일정에 달려 있어 2027년에 시작할 것으로 예상한다. 이 일정은 TSMC 보도자료나 실적 자료에는 나오지 않고, 2026년 10월 22일 현재 양산 개시 공식 발표도 없다. **7회차의 데이터센터 전력 이야기가 트랜지스터 수준까지 내려오는 지점이 여기다.**

---

4. 노광 (리소그래피) — 가장 깊이 들어갈 부분

레일리 판정식

해상 가능한 최소 선폭(CD, Critical Dimension)은

$$\mathrm{CD} = k_1 \frac{\lambda}{\mathrm{NA}}$$

- λ : 광원 파장
- NA : 개구수 (numerical aperture) = $n\sin\theta$. 렌즈가 얼마나 넓은 각도의 빛을 모으는가
- k_1 : 공정계수. 이론적 하한은 단일 노광에서 약 0.25

그리고 초점심도(DOF, Depth of Focus)는

$$\mathrm{DOF} = k_2 \frac{\lambda}{\mathrm{NA}^2}$$

핵심 긴장: NA를 키우면 해상도는 선형으로 좋아지지만 **초점심도는 제곱으로 나빠진다.** 표면이 조금만 울퉁불퉁해도 패턴이 안 맺힌다. 이것이 CMP(평탄화)가 필수 공정인 이유다(7절).

신호처리로 이해하기 — 통계·전산 배경 독자를 위한 지름길

노광계를 **선형 시불변 시스템의 저역통과 필터**로 보면 모든 것이 정리된다.

- 마스크 패턴은 공간 영역의 신호다. 이를 푸리에 변환하면 공간주파수 성분들이 나온다.
- 렌즈(투영계)는 **차단 주파수** $f_c \propto \mathrm{NA}/\lambda$ **인 저역통과 필터**다. 그보다 높은 공간주파수는 렌즈 동공(pupil)을 통과하지 못해 사라진다.
- 따라서 웨이퍼에 맺히는 상은 마스크 패턴의 **저역통과 필터링 결과**다. 날카로운 모서리(고주파 성분)는 뭉개진다.

이 프레임에서 보면:

- **OPC (Optical Proximity Correction):** 원하는 출력을 얻기 위해 입력(마스크)을 미리 왜곡시키는 것. **전형적인 역문제(inverse problem)/디컨볼루션**이다. 모서리에 세리프를 붙이고 선폭을 미세

조정한다.

- **ILT (Inverse Lithography Technology):** 아예 최적화 문제로 정식화해서 마스크 패턴을 푼다. 목적함수는 "웨이퍼 상 패턴과 목표 패턴의 차이", 변수는 마스크 픽셀. **대규모 비볼록 최적화 문제이며, 오늘날 GPU 클러스터로 푼다.**
- **위상반전마스크(PSM):** 인접 개구의 위상을 180° 반전시켜 경계에서 상쇄간섭을 만들어 대비를 높인다. 즉 진폭뿐 아니라 위상까지 제어하는 방식.
- **광원-마스크 동시최적화(SMO):** 조명 형태(동공에서의 조명 분포)까지 함께 최적화한다.

> **더 파고들기:** "계산 리소그래피(computational lithography)"는 이름 그대로 대규모 계산 문제다. **머신러닝이 실제 산업에 투입된 대표 영역**이며, CS·AI 전공자에게 반도체 산업의 가장 자연스러운 진입점 중 하나다.

파장을 줄여온 역사

광원	파장	시기
g-line (수은등)	436 nm	~1990년대 초
i-line	365 nm	1990년대
KrF 엑시머	248 nm	1990년대 후반
ArF 엑시머	193 nm	2000년대
ArF 액침 (immersion)	193 nm (물속 파장 약 134 nm)	2007~
EUV	**13.5 nm**	2019~

액침(immersion)의 트릭: NA $= n\sin\theta$ 에서 n 은 매질의 굴절률이다. 렌즈와 웨이퍼 사이를 공기(n≈1) 대신 물(n≈1.44)로 채우면 NA를 1을 넘겨 최대 1.35까지 올릴 수 있다. **광원을 바꾸지 않고 NA를 키운 영리한 해법이었고, 이 덕분에 193nm로 10여 년을 더 버텼다.** 표의 134 nm는 물속에서 줄어든 파장 $\lambda_0 / n = 193 / 1.44$ 이다. 이 효과는 NA $= n\sin\theta$ 의 n 에 이미 들어 있으므로 레일리 식에는 진공 파장 193 nm와 NA 1.35를 넣는다(실습 1). 134 nm를 넣으면 굴절률을 두 번 셈하게 된다.

축소 투영 — 마스크는 웨이퍼보다 크다

노광기는 마스크(레티클)의 패턴을 같은 크기로 찍지 않고 **줄여서** 찍는다. 0.33 NA EUV 노광기(ASML NXE)는 가로세로 모두 4배로 줄이고, 오늘날 ArF 스캐너도 4배가 표준이다. 마스크 위의 패턴은 웨이퍼에 그려질 패턴보다 4배 크게 그려져 있다는 뜻이다.

이 구조에서 두 가지가 따라 나온다. 첫째, 마스크의 패턴 오차도 웨이퍼에서는 대략 4분의 1 크기로 옮겨진다(아주 미세한 패턴에서는 마스크 오차가 그보다 증폭되어 옮겨지기도 한다. 마스크 오차 증폭계수 MEEF라 부른다). 마스크 패턴면에 붙은 먼지도 마찬가지로 작게 찍힌다. 둘째, 그래도 마스크의 결함은 그 마스크로 찍는 **모든 필드에 똑같이 되풀이된다.** 웨이퍼 위에 무작위로 떨어지는 파티클(4회차의 푸아송 결함)과 달리, 마스크 결함 하나는 웨이퍼의 모든 필드에서 같은 자리에 결함을 만든다.

한 번에 찍는 필드는 웨이퍼 위에서 최대 26×33 mm다. 계산하면 4배 축소에서 마스크 위의 패턴 영역은 104×132 mm이고, 이것이 한 변 6인치(152.4 mm) 마스크 안에 들어간다.

High-NA는 두 방향의 배율이 다르다. NA를 0.55로 키우려면 거울이 커지고, 그러면 빛이 마스크에

들어가는 각도가 커져 반사 마스크가 패턴을 제대로 반사하지 못한다. 양 방향을 8배로 줄이면 각도 문제는 풀리지만 마스크를 더 크게 만들어야 한다. ASML의 답은 **한 방향은 4배, 다른 방향은 8배로 줄이는 아나모픽 광학이다.** 계산하면 마스크 위 영역 104×132 mm는 그대로이고, 8배로 줄이는 방향의 필드만 132 ÷ 8 = 16.5 mm가 되어 웨이퍼 위 필드가 26×16.5 mm로 절반이 된다. **필드가 절반인 이유는 6인치 마스크를 그대로 쓰기 위해서다.**

멀티패터닝 — k_1 을 우회하는 법

단일 노광의 k_1 하한(약 0.25)에 도달하면, 한 층을 여러 번 나눠 찍는다.

- **LELE (Litho-Etch-Litho-Etch):** 패턴을 두 마스크로 나눠 두 번 노광·식각
- **SADP / SAQP (자기정렬 이중/사중 패터닝):** 기존 패턴 측벽에 스페이서를 증착하고 원래 패턴을 제거하면 **간격이 절반인 패턴**이 남는다. 이를 반복해 1/4까지 간다

대가가 크다: 공정 스텝 수가 배로 늘어 원가와 사이클타임이 급증하고, **정렬 오차(overlay error)** 가 새로운 수율 변수가 된다. **EUV 도입의 진짜 동기는 "더 작게"가 아니라 "멀티패터닝 스텝 수를 줄여 원가를 낮추기"였다.**

EUV의 잔인한 물리

파장 13.5nm는 자외선을 넘어 연X선에 가깝다. 여기서 모든 것이 어려워진다.

① **렌즈를 쓸 수 없다** 13.5nm는 거의 모든 물질에 흡수된다. 굴절 렌즈가 불가능하므로 **반사 광학계**만 가능하다. 반사경도 일반 거울이 아니라 몰리브덴과 실리콘을 몇 nm 두께로 번갈아 최대 100층가량 쌓은 **다층 브래그 반사경**이며, 그래도 반사율은 최대 70% 정도다. 반사경을 여러 장 거치면 손실이 곱해져 **광원 출력의 극히 일부만 웨이퍼에 도달한다.**

② **진공이 필수** 공기도 EUV를 흡수하므로 광 경로 전체가 진공이어야 한다.

③ **광원이 극도로 어렵다** 주석(Sn) 액적을 초당 수만 번 떨어뜨리고, 각각에 고출력 CO_2 레이저를 두 번 (예비 펄스로 납작하게 만든 뒤 본 펄스로) 쏘아 플라즈마를 만들어 13.5nm를 방출시킨다. **전기에서 EUV로의 변환 효율이 매우 낮아 장비 전력 소모가 크다.** 게다가 튄 주석이 집광경을 오염시킨다. 양산의 문턱이 된 숫자 하나가 광원 출력이다. TSMC는 2019년 첫 EUV 양산 공정을 알리면서 자사 EUV 장비의 일상 운영 광원 출력이 250 W를 넘었다고 밝혔다(실습 3의 처리량 계산이 이 출력에서 출발한다).

④ **마스크도 반사식** 투과 마스크를 쓸 수 없어 반사 마스크를 쓰고, 빛이 비스듬히 입사하므로 **그림자 효과(shadowing)** 라는 새로운 왜곡이 생긴다. 또한 마스크 방어막(pellicle)이 EUV를 흡수하므로 펠리클 개발이 오랜 난제였다.

⑤ **확률적 결함 (stochastic defect) — 통계 전공자 주목**

이것이 EUV의 가장 흥미로운 문제다.

- 광자 에너지 $E = hc / \lambda$ 는 파장에 반비례하므로, 13.5nm 광자는 193nm 광자보다 약 **14배 에너지가 크다.**
- 같은 노광 에너지(dose)를 주려면 **광자 개수는 그만큼 적어진다.**
- 광자 도달은 푸아송 과정이므로, 개수 N 의 상대적 요동은 $\sigma / \mu = 1 / \sqrt{N}$ 이다. N 이 작으면 요동이 커진다.

- 결과: 어떤 지점은 우연히 광자를 덜 받아 현상이 덜 되고, 어떤 지점은 더 받는다. **선폭 거칠기(LER/LWR), 홀 미형성, 브릿지 결함**이 확률적으로 발생한다.

이것은 공정을 개선해서 없앨 수 있는 결함이 아니라 통계적 필연이다. 대응책은 dose를 올리는 것인데(광자 수를 늘림) 그러면 처리량(throughput)이 떨어진다. **해상도-처리량-확률적 결함의 삼각 트레이드오프**가 EUV 공정의 근본 제약이다. 새로운 레지스트(금속 함유 레지스트 등)가 이 균형을 개선하려는 시도다.

> **통계 전공자를 위한 심화:** 결함 확률의 꼬리(tail)가 문제다. 칩 하나에 수십억 개의 피처가 있으므로, 개별 피처의 실패 확률이 10^{-12} 수준이어도 칩 단위 수율에 영향을 준다. **극단값 이론(extreme value theory)이 실제로 쓰이는 산업 문제다.**

High-NA EUV — 현재 진행형

- NA를 0.33에서 **0.55**로 올린 차세대 EUV. ASML TWINSCAN **EXE:5200**.
- 2026년 10월 기준: Intel이 2세대 EXE:5200B를 처음 설치해 인수 시험을 통과했고, 2026년 7월에는 18A의 특정 층을 High-NA로 찍은 Core Ultra 시리즈 3 일부 제품을 대량 생산해 출하하고 있다고 ASML과 Intel이 발표했다. 수율은 기존 NXE로 찍은 것과 같은 수준이라고 한다. SK하이닉스는 2025년 9월 이천 M16 팹에 양산용 EXE:5200B를 들였고, imec은 2026년 3월 EXE:5200을 받아 2026년 4분기까지 검증을 마칠 계획이다.
- **아직은 일부 층에 쓰는 단계다.** Intel은 18A에서 검증한 뒤 14A 생산으로 넓히겠다고 밝혀 왔다. 다른 회사가 어느 노드부터 High-NA를 쓸지, 전 세계에 몇 대가 설치됐는지는 2026년 10월 현재 공식 발표로 확인하지 못했다.

High-NA의 대가 두 가지

1. **초점심도 급감** — DOF $\propto \lambda / \mathrm{NA}^2$ 이므로 NA를 0.33→0.55로 올리면 DOF가 약 $(0.33/0.55)^2 \approx 0.36$배로 준다. 평탄도 요구가 극심해진다.
2. **아나모픽 광학과 필드 축소** — High-NA는 한 방향으로 배율이 다른 아나모픽 광학계(4배·8배)를 쓰는데, 그 결과 **한 번에 노광할 수 있는 필드 크기가 절반(26×16.5mm)으로 줄어든다.** 기존 레티클 한계 858 mm^2가 429 mm^2가 되는 셈이다. 왜 절반이 되는지는 위 「축소 투영」에서 계산했다.

> **★ 여기가 5회차와 6회차를 잇는 고리다:** 노광 필드가 줄면 필드보다 큰 칩은 한 번에 찍을 수 없다. 길은 둘이다. 두 필드를 이어 찍는 스티칭과, 칩을 나눠 만든 뒤 패키지에서 붙이는 칩렛이다(실습 5). 스티칭은 이음매의 정렬 오차와 설계 제약이라는 비용을 치르므로, High-NA로 찍는 층이 늘수록 큰 칩을 칩렛으로 나눌 유인이 커진다. 6회차에서 이어진다.

---

5. 식각 (Etching) — 빼는 공정

동판화를 떠올리면 된다. 동판에 산에 녹지 않는 막을 바르고 바늘로 긁어 그림을 그린 다음 부식액에 담그면, 긁혀 드러난 곳만 녹아 홈이 파인다. 반도체에서는 감광막이 그 막이고, 노광·현상이 바늘이며, 식각이 부식액이다.

습식 vs 건식

- **습식 식각:** 화학 용액에 담근다. 등방성(모든 방향으로 똑같이 파고듦)이라 미세 패턴에는 부적합. 세정·희생층 제거 등에 쓰인다.
- **건식 식각 (플라즈마):** 반응성 이온을 전기장으로 가속해 수직으로 때린다. **이방성(anisotropic)** 을 얻을 수 있어 미세 패턴의 표준이다.

두 가지 핵심 지표

1. **이방성(anisotropy):** 수직으로만 파야 한다. 옆으로 파고들면(undercut) 선폭이 무너진다. 플라즈마 식각은 이온 충격(물리적)과 화학 반응을 결합해 수직성을 얻고, 측벽에 보호막(passivation)을 형성해 옆으로 파이는 것을 막는다.
2. **선택비(selectivity):** 파야 할 재료와 파지 말아야 할 재료의 식각 속도 비. 선택비가 낮으면 아래층까지 뚫어버린다. GAA 공정에서 **SiGe만 골라 제거하고 Si 나노시트는 남기는 것**이 극단적 선택비를 요구하는 대표 사례다.

관리 인자로는 이 밖에 **식각 속도와 균일도가** 있다. 식각 속도는 단위 시간에 깎이는 깊이로, 장비 한 대가 시간당 처리하는 웨이퍼 수를 정한다. 균일도는 웨이퍼 중심과 가장자리가 같은 깊이로 깎이는가다. 균일도가 나쁘면 덜 깎인 곳을 마저 뚫는 동안 먼저 뚫린 곳은 아래층을 계속 깎이게 되고, 그만큼 선택비에 기대야 할 몫이 커진다.

고종횡비 식각 (HAR) — 난제 중의 난제

깊고 좁은 구멍을 뚫는 일이다.

- **3D NAND:** 수백 층을 관통하는 채널 홀. 종횡비가 100:1을 넘는다.
- **TSV:** 실리콘 웨이퍼를 관통하는 비아 (6회차).
- **DRAM 커패시터:** 면적은 줄이면서 용량을 유지하려면 깊게 파야 한다.

어려운 이유: 깊이 들어갈수록 반응 생성물이 빠져나오기 어렵고, 이온이 측벽에 부딪혀 방향이 흐트러지며, 구멍 입구가 먼저 넓어지는 보잉(bowing)과 아래로 갈수록 좁아지는 테이퍼가 생긴다. 그래서 **극저온 식각, 펄스 플라즈마** 등 특수 기법이 동원된다.

---

6. 증착 (Deposition) — 더하는 공정

방식	원리	특징	용도
PVD (물리기상증착)	타깃 물질을 스퍼터링/증발시켜 날려 붙임	빠름, 단차 피복성 나쁨	금속 배선, 배리어
CVD (화학기상증착)	기체 전구체를 반응시켜 표면에 막 형성	피복성 좋음, 고온 필요할 수 있음	절연막, 폴리실리콘
ALD (원자층증착)	**자기제한 반응**을 교대로 반복	**원자층 단위 두께 제어, 우수한 단차 피복**	게이트 절연막, 배리어, 나노시트 게이트

ALD가 왜 미세화의 구원자인가

ALD는 전구체 A를 넣어 표면이 포화될 때까지만 반응시키고(자기제한, self-limiting), 퍼지한 뒤 전구체 B를 넣어 반응을 완성한다. 표면이 포화되면 반응이 멈추므로 한 사이클에 쌓이는 두께(사이클당 성장량)가 거의 일정하다. 그 값은 전구체 분자의 크기, 표면 상태, 온도에 따라 정해지고, 보통 원자층 하나에 못 미친다.

결과적으로 얻는 두 가지 성질

1. **두께를 사이클 수로 정밀하게 제어**한다. 사이클당 성장량이 일정하므로 반복 횟수로 옹스트롬 단위까지 맞출 수 있다.
2. **복잡한 모양도 고르게 덮는다.** 전구체가 구멍 안쪽과 틈 사이까지 닿고 남은 기체가 다 빠져나갈(퍼지) 만큼 사이클 시간을 주면, 깊은 구멍 안쪽도 나노시트 사이의 좁은 틈도 거의 같은 두께로 코팅된다. 모양이 복잡할수록 사이클 하나에 드는 시간이 길어진다.

GAA는 ALD에 기댄다. 떠 있는 나노시트 사이의 수 나노미터 틈에 게이트 절연막과 금속을 균일하게 채워 넣어야 하는데, 표면 반응이 스스로 멈추며 한 겹씩 쌓는 ALD가 이런 좁은 틈을 고르게 덮는 데 가장 잘 맞는다.

High-k 게이트 절연막 — 양자역학이 산업을 바꾼 사례

게이트 절연막(전통적으로 SiO_2)이 얇아질수록 게이트 제어력은 좋아지지만, **1~2nm 이하가 되면 전자가 양자 터널링으로 절연막을 통과해 누설전류가 폭증한다.**

해법: 정전용량 $C = \varepsilon A / d$ 에서, d 를 줄이는 대신 ε (유전율)이 큰 물질을 쓴다. 유전율이 높으면 물리적으로 두꺼워도 같은 정전용량을 얻어, 터널링은 막으면서 제어력은 유지한다. 하프늄 산화물(HfO_2) 계열이 이 역할을 하며, 폴리실리콘 게이트도 금속 게이트로 바뀌었다(**HKMG: High-k Metal Gate**, 2007년경 도입).

> **더 파고들기:** "터널링 확률이 장벽 두께에 지수적으로 의존한다"는 양자역학 교과서의 결과가 수십억 달러짜리 산업 전환을 강제한 사례다.

---

7. 이온주입 · 열처리 · CMP

이온주입 (Ion Implantation)

도핑의 실체는 우아하지 않다. **도펀트 이온(붕소, 인, 비소 등)을 수십~수백 keV로 가속해 실리콘에 물리적으로 때려박는다.**

- **깊이 제어:** 가속 에너지로 조절한다. 에너지가 클수록 깊이 들어간다.
- **농도 제어:** 이온 빔 전류와 시간(dose)으로 조절한다.
- **분포:** 정지 깊이는 확률적이라, 농도 프로파일이 대략 가우시안 형태(정확히는 왜도를 가진 분포)를 이룬다. **몬테카를로 시뮬레이션으로 예측한다** — 통계 전공자에게 익숙한 도구가 그대로 쓰인다.
- **부작용:** 이온이 결정 격자를 부순다. 그래서 **열처리(어닐링)** 로 결정을 회복시키고 도펀트를 격자

자리에 앉혀 전기적으로 활성화해야 한다.

- **긴장 관계:** 열처리를 오래 하면 결정은 잘 회복되지만 도펀트가 확산해 프로파일이 뭉개진다. 미세 소자에서는 이게 치명적이라 **급속열처리(RTA), 밀리초 어닐링, 레이저 어닐링** 처럼 점점 더 짧고 강한 열처리로 간다.

CMP (화학기계연마)

왜 필수인가: 4절에서 봤듯 초점심도가 얕다. 표면이 울퉁불퉁하면 다음 층 노광에서 패턴이 맺히지 않는다. 층을 쌓을수록 요철이 누적되므로, **매 층마다 평탄화**해야 한다.

어떻게: 화학적으로 표면을 무르게 만드는 슬러리를 흘리면서 패드로 기계적으로 문지른다. 화학과 기계의 결합이라 이름이 CMP다.

어려운 점: 재료마다 연마 속도가 달라 **디싱(dishing, 넓은 금속 영역이 오목해짐)** 과 **에로전(erosion)** 이 생긴다. 그래서 설계 단계에서 패턴 밀도를 균일하게 맞추는 **더미 패턴(dummy fill)** 을 넣는다. **설계가 제조 공정에 맞춰지는 대표적 사례(DFM)** 다.

세정

언급하기 쉽지만 실제로는 전체 공정 스텝의 상당 부분이 세정이다. 파티클 하나가 곧 결함이고, 결함 하나가 곧 수율 손실이므로($Y = e^{-AD_0}$ 의 D_0!) 각 공정 사이마다 세정이 들어간다.

---

8. 검사 · 계측 — CS/AI 전공자의 직접적 접점

검사(Inspection) vs 계측(Metrology)

- **검사:** 결함이 있는가? 어디에? (파티클, 브릿지, 미형성)
- **계측:** 치수가 목표대로인가? (선폭, 막 두께, 정렬 오차)

주요 기법

기법	원리	특징
광학 검사	산란광·명암 대비	빠름, 파장 이하 결함 검출 한계
e-beam 검사	전자빔 주사	고분해능, **매우 느림**
OCD / 스캐터로메트리	회절광 스펙트럼을 모델에 피팅해 형상 역산	비파괴, **역문제 풀이**
오버레이 계측	정렬 마크 측정	층간 정렬 오차 관리
CD-SEM	전자현미경으로 선폭 직접 측정	정확, 느림, 국소적

왜 여기가 AI의 주 무대인가

1. **데이터가 압도적으로 많다.** 웨이퍼 한 장에 수천 개 다이, 팹에서 하루 수만 장. 웨이퍼 맵, 센서 로그, 계측값이 방대하게 쌓인다.
2. **결함 분류가 전형적 분류 문제다.** SEM 이미지에서 결함 유형을 판별하는 일은 이미지 분류 그 자체다.
3. **웨이퍼 맵 패턴이 원인을 가리킨다.** 링 모양, 스크래치, 가장자리 집중 등 공간 패턴이 특정 장비·공정 문제와 대응한다. **공간 통계 + 패턴 인식 문제.**
4. **가상 계측(virtual metrology).** 모든 웨이퍼를 다 측정할 수는 없다(느리고 비쌈). 장비 센서 데이터로 계측값을 예측해 샘플링을 줄인다. **회귀 문제이며, 불확실성 정량화가 실제로 중요하다.**
5. **예지 정비(predictive maintenance).** 장비 고장을 미리 예측한다.
6. **APC (Advanced Process Control).** 계측 결과를 다음 웨이퍼 공정 조건에 피드백하는 제어 루프. **run-to-run 제어는 전형적인 시계열 제어 문제다.**

> **더 파고들기:** 반도체 산업에서 AI가 가장 많이, 가장 실제로 쓰이는 곳은 칩 안이 아니라 **팹 운영**이다.

"측정할 수 없으면 만들 수 없다"

계측 자체가 미세화의 병목이 되고 있다. 수 나노미터 구조를 비파괴로, 빠르게, 정확히 측정하는 것이 점점 어려워진다. 3D 구조(GAA 나노시트, 3D NAND 채널)는 표면에서 보이지 않아 더 어렵다.

---

9. 전체 통합 — 층 하나가 만들어지는 과정

배선층(BEOL) 하나를 예로 들면(다마신 공정):

```
1. 절연막 증착 (CVD)
2. 감광막 도포 → 노광 → 현상          ← 패턴 정의
3. 식각으로 트렌치/비아 형성 (플라즈마)
4. 감광막 제거 + 세정
5. 배리어/시드층 증착 (ALD/PVD)       ← Cu 확산 방지
6. 구리 전해도금으로 채움
7. CMP로 여분 구리 제거 + 평탄화      ← 이래야 다음 층 노광이 가능
8. 캡층 증착
9. 검사·계측
   ↓ 다음 층으로 반복 (15~20회)
```

배선 재료는 한 번 크게 바뀌었다. 1990년대까지 배선은 알루미늄(Al)이었다. 알루미늄 막을 깔고 식각으로 배선 모양만 남기는 방식이고, 트랜지스터와 배선을 잇는 좁은 콘택트 구멍은 텅스텐(W)으로 메웠다(W 플러그). 알루미늄이 실리콘에 바로 닿으면 경계에서 서로 섞여 들어가 접합을 망가뜨리므로, 그 사이에는 배리어 메탈을 깔았다. 배선이 가늘어지자 알루미늄은 두 벽에 부딪혔다. 하나는 저항이고, 다른 하나는 일렉트로마이그레이션(전류가 금속 원자를 조금씩 밀어 배선을 끊거나 뭉치게 하는 현상)이다. 구리는 저항이 낮고 이 현상에 강하지만, 실리콘으로 원자가 새어 들어가 소자를 망가뜨리고 식각으로 깔끔하게 깎기 어렵다. 그래서 구리는 깎는 대신 **절연막에 홈을 파고 채운 뒤 CMP로 덜어 내는** 다마신 방식으로 들어왔다. 비아와 배선 홈을 한 번에 채우는 것이 듀얼 다마신이고, 위 5단계의 배리어가 구리가 새어 나가는 것을 막는다. IBM은 1997년 9월 22일 구리 배선 칩을 본격 양산하겠다고 발표했고(첫 출하는 1998년 9월), 구리 배선의 저항이 알루미늄보다 약 40% 낮다고 밝혔다. 2000년대에 구리가 배선의 업계

표준이 됐다.

여기서 얻어야 할 감각

- 하나의 층에도 이만큼 단계가 있고, 이것이 수십 번 반복된다.
- 어느 한 단계라도 실패하면 그 웨이퍼는 (부분적으로) 손실이다. **1,000단계 각각이 99.99%로 성공해도 전체는** $0.9999^{1000} \approx 90\%$ **다.** 각 스텝의 요구 수율이 얼마나 극단적인지 알 수 있다.
- 사이클타임이 수개월이므로, 문제를 발견해도 피드백이 느리다. 이것이 **가상 계측과 인라인 제어가 중요한 이유**다.

마스크 수와 원가

노드가 미세해질수록 마스크 수(층 수 × 멀티패터닝 배수)가 늘고, 공정 스텝이 늘고, 장비 시간이 늘어 웨이퍼 가격이 오른다. **EUV는 멀티패터닝을 줄여 이 증가를 완화하는 도구다.** "EUV가 비싸다"는 말과 "EUV가 원가를 줄인다"는 말이 동시에 참인 이유가 여기 있다.

---

10. 자주 하는 오해

오해 1: "노광은 그냥 사진 찍듯 인쇄하는 것" → 파장보다 작은 구조를 만드는 일이라 물리적으로는 회절 한계와의 싸움이다. 마스크 패턴과 웨이퍼 패턴이 전혀 다르게 생겼다(OPC/ILT).

오해 2: "EUV가 있으면 멀티패터닝이 필요 없다" → EUV도 단일 노광 한계가 있어 최선단에서는 EUV 멀티패터닝을 한다. High-NA는 이를 늦추려는 시도다.

오해 3: "미세화는 물리 법칙 때문에 곧 끝난다" → 물리적 한계보다 **경제적 한계**가 먼저 온다. 만들 수 있어도 그 원가로는 팔 수 없는 지점. 450mm 웨이퍼 좌초가 그 예다.

오해 4: "GAA는 FinFET보다 그냥 좋다" → 공정 복잡도와 원가가 크게 오른다. 모든 제품이 GAA로 갈 이유는 없고, 성숙 노드는 계속 남는다.

오해 5: "결함은 청정도를 높이면 없앨 수 있다" → 파티클성 결함은 그렇지만, EUV의 확률적 결함처럼 **통계적 필연인 결함**은 청정도로 해결되지 않는다.

오해 6: "웨이퍼가 클수록 좋다" → 450mm 좌초 사례가 보여주듯, 장비 재개발비와 균일도 확보 난이도가 절감액을 넘을 수 있다.

오해 7: "이온주입은 화학적으로 도핑하는 것" → 물리적으로 때려박는다. 그래서 격자가 손상되고 열처리가 반드시 뒤따른다.

오해 8: "EUV는 3 nm 이하 최선단에서만 쓴다" → EUV 양산은 7 nm급에서 시작됐다. TSMC는 N7+를 「업계 첫 상용 EUV 리소그래피 기술」이라 내세우며 2019년 2분기에 양산을 시작했고(삼성전자는 2018년 10월 EUV를 쓴 7LPP 공정의 웨이퍼 생산을 시작했다고 발표했다), 그해 10월 고객 제품이 대량으로 시장에 나가고 있다고 발표했다. 같은 발표에서 후속 N6은 EUV를 더 확대 적용해 N7보다 로직 밀도를 18% 높인다고 했고, ASML은 0.33 NA 장비가 7·5·3 nm 노드의 핵심 층을 찍는 데 쓰인다고 설명한다. 4절에서 본 대로 EUV를 들이는 동기는 「더 작게」보다 멀티패터닝 스텝을 줄이는 데 있었다.

11. 정량 실습

실습 1 — 레일리 판정식 $k_1 = 0.3$ 일 때 (a) ArF 액침(λ=193nm, NA=1.35), (b) EUV(λ=13.5nm, NA=0.33), (c) High-NA EUV(λ=13.5nm, NA=0.55)의 CD를 각각 구하라. 세대 간 개선 배율은?

실습 2 — 초점심도 대가 같은 조건에서 DOF 비율을 구하라(k_2 는 동일 가정). High-NA로 가면서 평탄도 요구가 얼마나 엄격해지는가?

실습 3 — 확률적 결함 같은 노광 에너지에서 193nm와 13.5nm의 광자 수 비율을 구하고($E = hc / \lambda$), 상대 요동 $1/\sqrt{N}$ 이 몇 배 커지는지 계산하라. dose를 몇 배 올려야 같은 요동 수준이 되는가? 그때 처리량은?

실습 4 — 스텝 수율 전체 공정 스텝이 800단계일 때, 목표 라인 수율 90%를 달성하려면 각 스텝의 평균 수율이 얼마여야 하는가? 1,200단계면?

실습 5 — 노광 필드와 다이 크기 기존 레티클 한계(약 26×33mm)와 High-NA 필드(약 26×16.5mm)에서 각각 만들 수 있는 최대 다이 면적을 구하고, 4회차의 수율 식을 적용해 최대 다이의 수율을 비교하라. **6회차 칩렛 논의의 출발점이 된다.**

12. 토론 질문

1. 레일리 판정식의 세 항 중 앞으로 가장 개선 여지가 큰 것은 무엇인가? Hyper-NA나 새로운 파장은 현실적인가?
2. High-NA EUV의 필드 축소는 설계 방법론을 어떻게 바꾸는가? 칩렛이 선택이 아니라 필수가 되는가?
3. 확률적 결함처럼 통계적으로 필연인 결함에 대해, 산업은 어떤 전략을 취해야 하는가? 설계 차원에서 여유(margin)를 두는 것과 공정에서 dose를 올리는 것 중 어느 쪽이 효율적인가?
4. 계측이 미세화의 병목이 되고 있다면, "측정 없이 제어하는" 방법(모델 기반 제어, 가상 계측)은 어디까지 가능한가? 신뢰성은 어떻게 담보하는가?
5. 미세화가 경제적 한계에 부딪힌다면, 성능 향상의 다음 축은 무엇인가? (6회차 예고)

13. 후속 공부

- **표준 교과서:** Plummer, Deal & Griffin, *Silicon VLSI Technology* — 공정 물리의 정석. May & Spanos, *Fundamentals of Semiconductor Manufacturing and Process Control* (Wiley, 2006, doi:10.1002/0471790281) — **통계적 공정 관리를 다루므로 통계 배경 독자에게 특히 적합**
- **열산화의 원논문:** Deal & Grove, "General Relationship for the Thermal Oxidation of Silicon", *J. Appl. Phys.* 36(12):3770–3778 (1965), doi:10.1063/1.1713945 — 2절 식이 어떤 실험 데이터에서 나왔는지 볼 수 있다

- **소자 물리:** Sze & Ng, *Physics of Semiconductor Devices* (Wiley, 3판 2006, doi:10.1002/0470068329, 본격 교재), Taur & Ning, *Fundamentals of Modern VLSI Devices* (Cambridge, 2판 2009, doi:10.1017/CBO9781139195065, 단채널 효과·스케일링 이론)
- **리소그래피:** Chris Mack, *Fundamental Principles of Optical Lithography* (Wiley, 2007, doi:10.1002/9780470723876) — **푸리에 광학 기반 서술이라 신호처리 배경이 있으면 매우 잘 읽힌다.** 저자의 웹사이트 lithoguru.com에 온라인 강의와 기초 해설이 있다
- **EUV:** Bakshi (ed.), *EUV Lithography* (SPIE, 2판 2018, doi:10.1117/3.2305675). High-NA의 아나모픽 설계는 van Schoot 외, EUVL Workshop 2017 발표와 ASML 「5 things you should know about High NA EUV lithography」
- **최신 동향:** IEDM, VLSI Symposium, SPIE Advanced Lithography의 발표 자료. imec의 로드맵 공개 자료
- **AI × 반도체 제조:** 웨이퍼 맵 결함 분류의 공개 데이터셋 WM-811K(Wu, Jang, Chen, *IEEE Trans. Semiconductor Manufacturing* 28(1) (2015), doi:10.1109/TSM.2014.2364237)로 직접 모델을 돌려보면 8절이 체감된다

자습 부록 A — 정량 실습 풀이

실습 1 풀이 — 레일리 판정식

▾ 풀이

CD $= k_1\lambda$ / NA, $k_1 = 0.3$ 으로 고정.

세대	λ	NA	CD
(a) ArF 액침	193 nm	1.35	$0.3 \times 193 / 1.35 =$ **42.9 nm**
(b) EUV	13.5 nm	0.33	$0.3 \times 13.5 / 0.33 =$ **12.3 nm**
(c) High-NA EUV	13.5 nm	0.55	$0.3 \times 13.5 / 0.55 =$ **7.4 nm**

개선 배율

구간	배율	무엇이 기여했나
(a) → (b)	**3.49배**	거의 전부 파장 (193 → 13.5, 14.3배). NA는 오히려 1.35 → 0.33으로 **나빠졌다**
(b) → (c)	**1.67배**	전부 NA (0.33 → 0.55)

읽는 법 — 이 표가 EUV의 성격을 다 말해 준다

EUV로 넘어가면서 NA를 **포기했다.** 13.5 nm는 모든 물질에 흡수되므로 굴절 렌즈를 쓸 수 없고 반사 광학계만 가능한데, 반사경으로는 큰 NA를 만들기가 어렵다. 파장에서 14.3배를 벌고 NA에서 4.1배를 잃어 순이득이 3.5배가 된 것이다.

그리고 **High-NA는 바로 그 잃었던 NA를 되찾는 작업**이다. 0.33 → 0.55는 겨우 1.67배지만, 그것을

위해 광학계 전체를 다시 설계하고 필드를 절반으로 줄이는(실습 5) 대가를 치른다.

> **주의:** $k_1 = 0.3$ 은 단일 노광의 실용 하한 근처다(이론 하한 0.25). 실제 42.9 nm는 하프피치이고, ArF 액침으로 그보다 미세한 패턴을 만들 때는 **멀티패터닝**으로 여러 번 나눠 찍는다 — 즉 k_1 을 우회하지 낮추지 않는다.

실제 장비는? 장비사가 밝힌 해상도에서 k_1 을 거꾸로 구할 수 있다. ASML은 High-NA 장비(EXE)의 해상도를 8 nm로 밝힌다. 계산하면 $k_1 = 8 \times 0.55 / 13.5 = 0.326$ 이다. 0.33 NA 장비(NXE)의 해상도 13 nm로 하면 $k_1 = 13 \times 0.33 / 13.5 = 0.318$ 이다. 위 표에서 가정한 $k_1 = 0.3$ 이 실제 장비 사양과 크게 다르지 않다는 뜻이다.

실습 2 풀이 — 초점심도의 대가

▾ 풀이

DOF $= k_2\lambda / \mathrm{NA}^2$. k_2 는 동일하게 두고 λ / NA^2 만 비교한다.

세대	λ / NA^2	상대 DOF	$k_2 = 0.5$ 일 때
(a) ArF 액침	193 / 1.8225 = 105.9	1.00	53 nm
(b) EUV	13.5 / 0.1089 = 124.0	**1.17**	62 nm
(c) High-NA EUV	13.5 / 0.3025 = 44.6	**0.42**	**22 nm**

뜻밖의 결과 하나: EUV(b)는 ArF 액침(a)보다 초점심도가 **오히려 17% 여유롭다.** NA를 낮춘 덕이다. EUV 도입이 평탄도 요구를 갑자기 올린 것은 아니라는 뜻이다.

진짜 문제는 High-NA다. DOF가 (b) 대비 0.36배, (a) 대비 0.42배로 줄어 **약 22 nm**가 된다.

22 nm 안에 무엇이 들어와야 하는가

- 웨이퍼 표면의 국부 요철 (CMP가 담당)
- 웨이퍼 척의 평탄도와 진동
- 레지스트 막 두께 자체 — 레지스트가 50 nm 두께면 이미 DOF를 초과한다

그래서 연쇄적으로 따라오는 것

1. **CMP 요구가 극단적으로 엄격해진다.** 평탄화가 노광의 전제조건이 된다 (7절).
2. **레지스트를 더 얇게 해야 한다.** 그런데 얇으면 **식각 내성이 떨어진다** — 마스크 역할을 하다가 먼저 없어진다.
3. 그래서 **하드마스크·다층 레지스트**를 쓰거나, **금속산화물 레지스트**처럼 얇아도 버티는 새 재료로 간다.

얻어 갈 것: 해상도 개선 하나가 평탄화·재료·식각까지 연쇄적으로 요구를 밀어 올린다. **미세화가 "노광기만 바꾸면 되는 일"이 아닌 이유**가 이 한 실습에 들어 있다.

아래 그림은 EUV(13.5 nm, NA 0.33)에 $k_1 = 0.3$, $k_2 = 0.5$ 를 넣은 경우다. 화면본에서는 세

세대의 광원과 NA, k_1, k_2 를 바꿔 볼 수 있고, 장비 공칭 해상도에서 k_1 을 거꾸로 구해 볼 수도 있다.

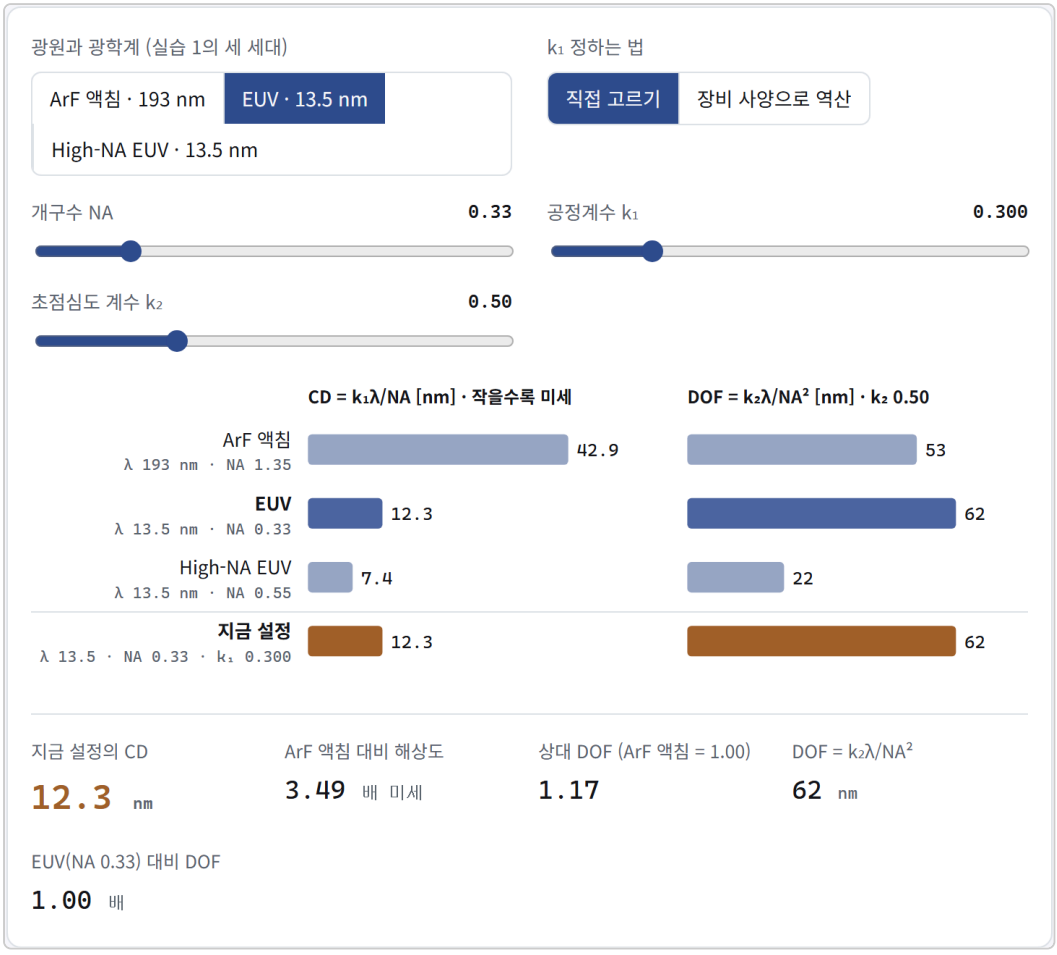


그림 5-1. $k_1 = 0.3$에서 CD는 ArF 액침 42.9 nm, EUV 12.3 nm, High-NA EUV 7.4 nm다. $k_2 = 0.5$로 둔 초점심도는 53·62·22 nm로, High-NA에서 EUV의 0.36배가 된다. — 화면본에서는 이 자리에서 값을 바꿔 볼 수 있다(화면본 5회차 실습 2 풀이).

실습 3 풀이 — 확률적 결함

▾ 풀이

광자 하나의 에너지 $E = hc / \lambda$ 이므로 파장에 반비례한다.

$$\frac{E_{\text{EUV}}}{E_{\text{ArF}}} = \frac{193}{13.5} = 14.3$$

EUV 광자는 ArF 광자보다 **14.3배 에너지가 크다** (약 92 eV vs 6.4 eV).

같은 dose(단위 면적당 에너지)에서의 광자 수

에너지가 같은데 광자 하나가 14.3배 크면, 광자 **개수**는 1/14.3 = **7%** 뿐이다.

$$N_{\text{EUV}} = \frac{N_{\text{ArF}}}{14.3}$$

상대 요동

푸아송 통계에서 개수 N 의 상대 요동은 $1/\sqrt{N}$ 이므로

$$\frac{\sigma_{\text{EUV}}/N_{\text{EUV}}}{\sigma_{\text{ArF}}/N_{\text{ArF}}} = \sqrt{14.3} = 3.78\text{배}$$

같은 요동 수준을 맞추려면 dose를 14.3배 올려야 한다. (N 을 14.3배로 만들어야 하므로.)

그때의 처리량

EUV 소스 출력은 웨이퍼당 처리 시간을 직접 결정한다. 웨이퍼를 옮기고 정렬하는 시간 같은 다른 조건이 같다면, 한 장을 노광하는 데 드는 시간은 dose를 광원 출력으로 나눈 값에 비례한다.

$$t_{\text{노광}} \propto \frac{\text{dose}}{\text{광원 출력}}, \qquad \text{처리량} = \frac{1}{T_0 + t_{\text{노광}}}$$

여기서 T_0 는 웨이퍼 이송·정렬처럼 dose와 무관한 고정 시간이다. 그래서 처리량은 광원 출력 ÷ dose에 그대로 비례하지 않고, 노광 시간이 짧아질수록 $1/T_0$ 에 붙어 포화한다.

기준점은 4절에서 본 일상 운영 광원 출력 250 W 이상(2019년 TSMC 발표)이다. 이 출력을 그대로 두고 dose를 14.3배 올리면 노광 시간이 14.3배가 된다. 노광 시간이 전체를 지배하는 극한에서는 **시간당 웨이퍼 처리량이 1/14.3로 붕괴한다.** 고정 시간 T_0 가 있으면 떨어지는 폭은 그보다 작고, 얼마나 작은지는 T_0 가 원래 노광 시간 t 에 비해 얼마나 큰지에 달렸다. 처리량은 원래의 $(T_0 + t)/(T_0 + 14.3t)$ 배이므로 $T_0 = t$ 이면 13%, $T_0 = 10t$ 이면 45%로 떨어지고, T_0 가 노광 시간보다 훨씬 길면 거의 떨어지지 않는다. 이 실습은 T_0 를 주지 않으므로 1/14.3은 가장 나쁜 경우로 읽는다. 그 극한이면 수억 달러짜리 장비의 생산성이 7%로 떨어지니 경제적으로 불가능하다. 반대로 처리량을 지키려면 광원 출력을 14.3배, 계산하면 250 W에서 약 3.6 kW로 올려야 한다.

그래서 실제로 하는 일

수단	내용
dose를 **일부만** 올린다	ASML은 EUV 장비 처리량을 30 mJ/cm²(NXE:3600D·3800E)와 50 mJ/cm²(High-NA EXE:5000) 기준으로 공표한다. 광자 수 차이만큼 14.3배로 올리지 않고 수십 mJ/cm² 자릿수에 머문다
레지스트 감도·화학 개선	광자 하나가 더 많은 반응을 유발하도록 (화학증폭, 금속산화물 레지스트)
확률적 결함을 **설계로 회피**	규칙적 1차원 패턴 위주 설계, 금지 피치 회피
계측으로 잡아낸다	확률적 결함은 드물게 산재하므로 대면적 고속 검사가 필요 → 8절

이것이 "EUV의 잔인한 물리"의 정체다. 파장을 줄여 해상도를 얻었는데, 그 대가로 **광자 수 자체가 부족해져 통계적 불확실성이 커졌다.** 해상도와 확률적 결함 사이에는 원리적 상충이 있고, 이것이 dose·처리량·수율을 동시에 묶는다.

연결: 이 확률적 결함이 4회차의 D_0 에 그대로 더해진다. 그리고 D_0 가 올라가면 같은 목표 수율에서 허용되는 다이 면적이 줄어든다($A_{50} = \ln 2 / D_0$). **노광의 물리가 칩의 크기를 정한다.**

아래 그림은 EUV의 dose를 193/13.5배, 곧 약 14.3배로 올려 ArF와 같은 요동을 맞춘 경우다. 화면본에서는 dose 배수와 기준 광원 출력을 바꿔 가며 상대 요동, 노광 시간, 같은 노광 시간을 지키는 데 필요한 광원 출력을 볼 수 있다. 처리량은 고정 시간 T_0 를 모르므로 노광 시간이 전부를 지배하는 극한값만 보인다.

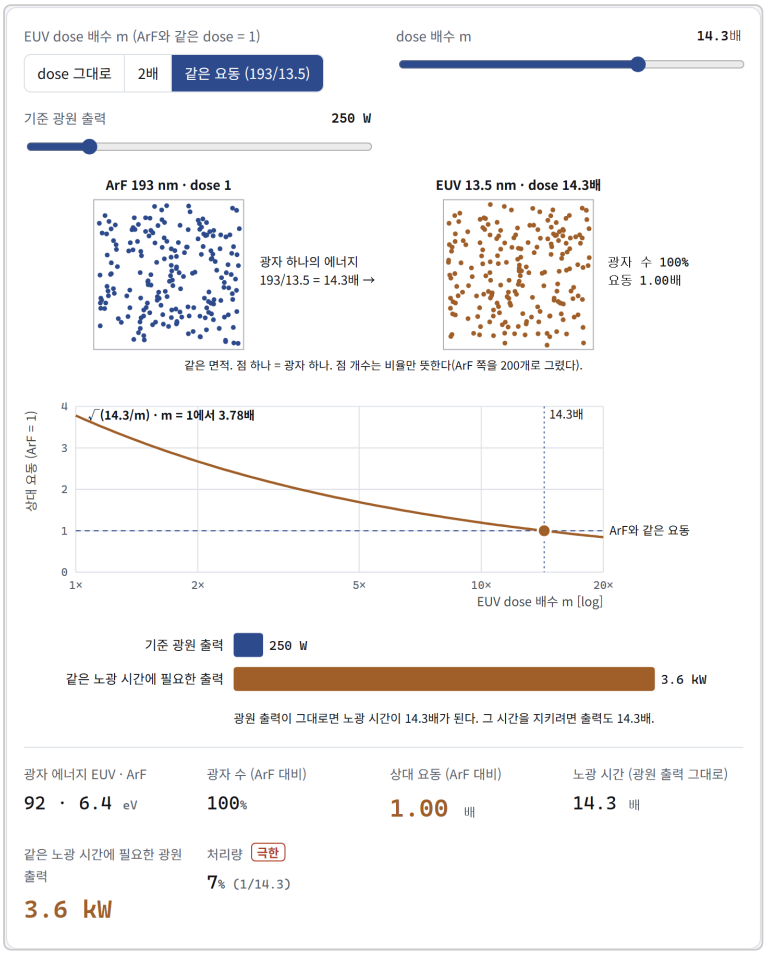


그림 5-2. 같은 dose에서 EUV 광자 수는 ArF의 7%, 상대 요동은 3.78배다. 요동을 맞추려고 dose를 14.3배 올리면 같은 광원 출력에서 노광 시간이 14.3배가 되고, 노광 시간을 지키려면 광원 출력이 250 W에서 약 3.6 kW로 올라가야 한다. — 화면본에서는 이 자리에서 값을 바꿔 볼 수 있다(화면본 5회차 실습 3 풀이).

실습 4 풀이 — 스텝 수율

▾ 풀이

전체 수율이 스텝 수율의 곱이라면 $y^n = Y$ 이므로 $y = Y^{1/n}$.

800단계, 목표 90%

$$y = 0.9^{1/800} = \exp\left(\frac{\ln 0.9}{800}\right) = \exp\left(\frac{-0.10536}{800}\right) = 0.9998683$$

$1 - 0.9998683 = 1.317 \times 10^{-4}$ 이므로 **스텝당 허용 불량률은 131.7 ppm**이다. 스텝 하나가 100만

번 중 132번만 실수해도 목표를 놓친다.

1,200단계, 같은 목표

$$y = \exp\left(\frac{-0.10536}{1200}\right) = 0.9999122 \;\Rightarrow\; \text{스텝당 } 87.8 \text{ ppm}$$

스텝이 1.5배 늘면 허용 불량률은 2/3로 줄어든다.

역방향으로 보면 훨씬 무섭다

스텝당 수율	스텝당 불량률	800단계 라인 수율	1,200단계
99.99%	100 ppm	$0.9999^{800} = 92.3\%$	88.7%
99.9%	1,000 ppm	$0.999^{800} = 44.9\%$	30.1%
99%	10,000 ppm	$0.99^{800} = 0.03\%$	거의 0

표의 라인 수율은 y^n 을 그대로 계산한 값이다. 불량률 $p = 1 - y$ 가 작으면 $y^n = (1-p)^n \approx e^{-pn}$ 이므로 800단계에서 99.99%는 $e^{-0.08}$, 99.9%는 $e^{-0.8}$, 99%는 e^{-8} 로 어림할 수도 있다. 표에 적은 자릿수에서는 두 계산이 같다.

스텝당 99%는 상식적으로 훌륭한 수치인데, 800단계를 거치면 사실상 아무것도 나오지 않는다.

이 실습이 요약하는 것

반도체 제조가 다른 제조업과 근본적으로 다른 이유가 여기 있다. **수백 단계가 직렬로 곱해지므로, 각 단계는 "거의 완벽"이 아니라 "ppm 단위로 완벽"해야 한다.** 그래서

- 공정 하나하나에 six sigma 수준의 관리가 요구되고
- 매 단계마다 검사·계측이 들어가며(8절)
- 스텝 수를 늘리는 결정(멀티패터닝!)이 단순한 비용 증가가 아니라 **수율 위험의 증가**가 된다

> **멀티패터닝의 진짜 비용:** 4절에서 멀티패터닝이 k_1 을 우회한다고 했다. 그 대가가 이것이다 — 한 층을 4번 나눠 찍으면 스텝이 늘고, 위 표에 따라 라인 수율이 곱셈으로 깎인다. **EUV가 비싸도 도입하는 이유는 해상도가 아니라 스텝 수 절감일 때가 많다.**

아래 그림은 800단계에서 라인 수율 90%를 목표로 둔 경우다. 화면본에서는 스텝 수와 목표 수율을 바꿔 볼 수 있고, 거꾸로 스텝당 수율에서 라인 수율을 구해 볼 수도 있다.

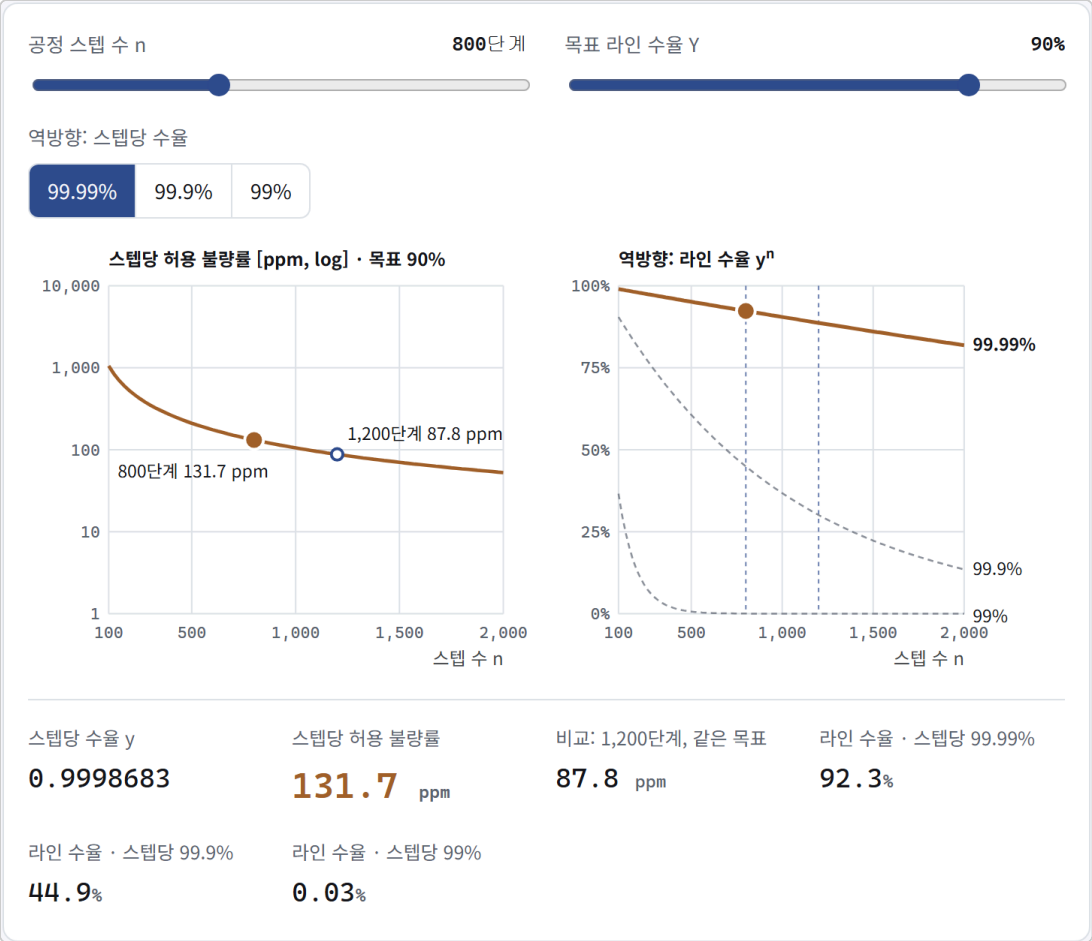


그림 5-3. 800단계에서 라인 수율 90%를 내려면 스텝당 허용 불량률이 131.7 ppm, 1,200단계면 87.8 ppm이다. — 화면본에서는 이 자리에서 값을 바꿔 볼 수 있다(화면본 5회차 실습 4 풀이).

실습 5 풀이 — 노광 필드와 다이 크기

▾ 풀이

노광 방식	필드 크기	최대 다이 면적
기존 (0.33 NA 포함)	26 × 33 mm	**858 mm²**
High-NA EUV	26 × 16.5 mm	**429 mm²**

High-NA는 아나모픽 광학계(한 방향으로만 축소 배율이 다름)를 쓰기 때문에 필드가 절반이 된다.

왜 하필 절반인가. 4절 「축소 투영」의 계산이 답이다. 기존 장비는 가로세로 모두 4배로 줄이므로 마스크 위 104×132 mm가 웨이퍼 위 26×33 mm가 된다. High-NA는 한 방향을 8배로 줄이는데 마스크는 같은 6인치를 쓰므로, 그 방향의 필드가 132 ÷ 8 = 16.5 mm로 줄어든다. 배율이 두 배가 된 방향만 정확히 절반이 되는 것이다.

4회차 수율 식 적용 ($D_0 = 0.1$)

필드	면적	$\lambda = AD_0$	수율
기존 최대 다이	8.58 cm^2	0.858	**42.4%**
High-NA 최대 다이	4.29 cm^2	0.429	**65.1%**

딜레마가 정확히 드러난다

High-NA는 **더 미세한 패턴을 그릴 수 있게 해 주는 대신, 한 번에 그릴 수 있는 면적을 절반으로 줄인다.** 수율은 좋아지지만(42% → 65%) 만들 수 있는 칩이 작아진다.

그런데 AI 가속기는 계속 커지고 있다. 800 mm^2급 다이에 HBM 8스택을 붙이는 방향으로 가는데, High-NA 필드로는 그런 다이를 한 번에 못 찍는다.

해법은 두 가지다

해법	내용	문제
스티칭 (stitching)	두 필드를 이어 붙여 한 다이를 만든다	이음매에서 정렬 오차·결함, 스텝 증가
칩렛	애초에 작은 다이로 나눠 만들고 패키지에서 붙인다	인터페이스 오버헤드, 조립 수율

결론: High-NA 필드보다 큰 다이는 스티칭 비용을 치르거나 칩렛으로 나눠야 한다. 둘 중 어느 쪽이 싼지는 이음매의 정렬·수율 비용과 칩렛의 인터페이스·조립 비용을 견줘 정해지지만, **High-NA로 찍는 층이 늘수록 칩렛 쪽으로 기울 유인이 커진다.** 4회차 실습 4에서 계산한 칩렛 손익에 High-NA의 필드 제약이 비용 항목 하나로 더해지는 셈이다. 6회차가 그 손익을 이어서 따진다.

아래 그림은 6인치 마스크 위의 같은 패턴 영역(104 × 132 mm)이 0.33 NA 장비와 High-NA 장비에서 각각 어떤 필드가 되는지, $D_0 = 0.1$ 일 때 최대 다이의 수율이 얼마인지 보여 준다. 화면본에서는 광학계와 D_0 를 바꿔 볼 수 있다.



그림 5-4. 마스크 위 패턴 영역 104 × 132 mm를 4배씩 줄이면 웨이퍼 필드는 26 × 33 mm(858 mm²), 한 방향만 8배로 줄이는 High-NA는 26 × 16.5 mm(429 mm²)다. D_0 = 0.1이면 최대 다이 수율은 42.4%와 65.1%다. — 화면본에서는 이 자리에서 값을 바꿔 볼 수 있다(화면본 5회차 실습 5 풀이).

자습 부록 B — 토론 질문 생각의 실마리

번호는 12절 토론 질문의 번호다. + 표시는 함께 생각해 볼 만한 질문을 덧붙인 것이다.

질문	핵심 축	실마리
1. 세 항 중 개선 여지가 큰 것	**남은 여지 대 대가**	k_1 은 단일 노광의 이론 하한 0.25에 가깝다(실습 1의 가정 0.3, 장비 사양에서 역산한 0.32~0.33). NA를 0.55보다 더 키우면 초점심도가 NA^2에 반비례해 줄고(실습 2) 필드·마스크 문제가 다시 커진다. 파장을 더 줄이면 광자 하나의 에너지가 커져 같은 dose의 광자 수가 더 줄어든다(실습 3). 어느 항이든 남은 이득보다 따라오는 대가를 먼저 계산해 보는 것이 출발점이다
2. 필드 축소와 설계 방법론	**스티칭 대 칩렛**	High-NA 필드(429 mm^2)보다 큰 다이는 스티칭으로 이어 찍거나 칩렛으로 나눠야 한다(실습 5). 「필수인가」는 이음매의 정렬·수율 비용과 칩렛의 인터페이스·조립 비용을 견주는 문제로 바뀐다. 한 층이라도 High-NA로 찍으면 그 층에서 필드 제약이 걸리므로, High-NA로 찍는 층이 몇 개인지가 답을 바꾼다
3. 확률적 결함 — dose 대 설계 여유	**처리량 대 면적**	dose를 올리면 상대 요동이 $1/\sqrt{\text{dose}}$ 에 비례해 줄지만 노광 시간이 dose에 비례해 늘어난다(실습 3). 설계 여유는 피처를 키우거나 규칙적 패턴을 써서 실패 확률의 꼬리를 줄이는 대신 면적을 쓴다. 둘을 「웨이퍼 한 장에서 나오는 정상 다이의 원가」라는 같은 단위로 바꾸면 4회차 다이당 원가 식으로 견줄 수 있다
4. 측정 없이 제어하기	**예측 대 검증**	가상 계측은 장비 센서로 계측값을 예측해 측정 표본을 줄인다(8절). 예측이 틀렸는지 알아내는 길도 결국 측정이라, 측정을 없애기보다 예측의 불확실성이 큰 웨이퍼에 측정을 몰아주는 쪽으로 쓴다. 신뢰성은 예측 오차의 분포를 계속 실측으로 점검하는 데서 나온다
5. 미세화 다음의 축	**무엇이 먼저 한계인가**	후보가 셋이다 — (i) 물리(원자 수 단위 채널, 확률적 결함), (ii) 경제(마스크·장비 비용 대비 이득), (iii) 열(전력밀도). 현재 가장 먼저 닿은 것은 (ii)와 (iii)이고, (i)은 아직 여유가 있다. **그래서 산업이 미세화에서 통합(패키징)으로 축을 옮기고 있다**
+ EUV 없이 어디까지 가능한가	**스텝 수 vs 해상도**	멀티패터닝으로 해상도 자체는 따라갈 수 있다. 무너지는 것은 **경제성**이다 — 실습 4가 보여 준 대로 스텝이 늘면 수율이 곱셈으로 깎이고, 마스크 수와 사이클 타임이 늘어 원가가 급증한다. "기술적으로 불가능"이 아니라 "경제적으로 불가능"이 정확한 진술이다
+ 후면전력공급의 의미	**배선과 소자의 분리**	앞면을 신호에, 뒷면을 전력에 전담시키는 것은 단순한 개선이 아니라 **레이어 배치 철학의 변경**이다. 7회차 실습 4(IR drop)에서 이것이 왜 GAA만큼 중요한지 정량적으로 확인한다
+ 검사·계측에서 AI의 역할	**탐지 vs 원인 규명**	결함 이미지 분류는 이미 성숙했다. 어려운 것은 **웨이퍼 맵 패턴으로부터 원인 장비·스텝을 역추적**하는 문제이고, 이것은 4회차 클러스터링(α)과 직결된다. 공간 통계 + 인과 추론 문제로 보면 접근이 열린다
+ 국산화는 어디부터인가	**대체 가능성의 위계**	1회차 3층 구조에서 본 대로 대체 불가능성이 높은 곳일수록 국산화 난도가 높다. 다만 난도와 효과가 비례하지 않는다 — **소재·부품 중 대체 가능하지만 공급이 편중된 것**부터가 비용 대비 효과가 크다

자습 부록 C — 핵심 용어

용어	영문	한 줄 정의
전공정	front-end process (FEOL + BEOL)	웨이퍼 위에 트랜지스터(FEOL)와 그 위 배선층(BEOL)을 만드는 웨이퍼 공정 전체. 웨이퍼를 자르고 붙이는 후공정과 구분된다
열산화	thermal oxidation	실리콘을 고온에서 산화시켜 SiO_2 막을 기르는 공정. 두께는 Deal–Grove 식을 따른다
축소 투영	reduction projection	마스크 패턴을 줄여 웨이퍼에 찍는 방식. 4배가 표준, High-NA는 4배·8배
듀얼 다마신	dual damascene	절연막에 비아와 배선 홈을 파고 구리를 한 번에 채운 뒤 CMP로 덜어 내는 배선 공정
CD	critical dimension	만들어야 하는 최소 선폭
레일리 판정식	Rayleigh criterion	$\mathrm{CD} = k_1\lambda$ / NA. 해상도의 근본 한계
개구수	numerical aperture (NA)	렌즈가 빛을 모으는 각도의 척도
초점심도	depth of focus (DOF)	초점이 유지되는 깊이 범위. $\propto \lambda$ / NA^2
EUV	extreme ultraviolet	파장 13.5 nm의 극자외선 노광
High-NA EUV	—	NA를 0.33에서 0.55로 올린 차세대 EUV. 필드가 절반
멀티패터닝	multi-patterning	한 층을 여러 번 나눠 찍어 해상도 한계를 우회
확률적 결함	stochastic defect	광자 수 부족에서 오는 무작위 결함
OPC	optical proximity correction	회절 왜곡을 미리 상쇄하도록 마스크를 변형
레지스트	photoresist	빛에 반응해 성질이 바뀌는 감광막
식각	etching	필요 없는 부분을 제거하는 공정
이방성 / 선택비	anisotropy / selectivity	수직으로만 깎이는 정도 / 원하는 물질만 깎이는 정도
HAR 식각	high aspect ratio etch	깊고 좁은 구멍을 뚫는 식각. 3D NAND·TSV의 난제
증착	deposition	필요한 막을 쌓는 공정
ALD	atomic layer deposition	원자층 단위로 쌓아 두께를 정밀 제어하는 증착
High-k	—	유전율이 높은 게이트 절연막. 누설전류를 막는다
이온주입	ion implantation	불순물 이온을 때려 넣어 전기적 성질을 바꾸는 공정
CMP	chemical mechanical polishing	화학·기계적으로 표면을 평탄화하는 공정
GAA	gate-all-around	게이트가 채널을 사방에서 감싸는 트랜지스터 구조
CFET	complementary FET	n형과 p형을 수직으로 쌓는 차세대 구조
BSPDN	backside power delivery network	전력 배선을 웨이퍼 뒷면으로 옮기는 방식
검사 / 계측	inspection / metrology	결함이 있는지 찾는 것 / 치수·두께를 재는 것

자습 부록 D — 자가진단

☐ 레일리 판정식 두 개를 쓰고, **해상도와 초점심도가 상충한다**는 것을 설명할 수 있다 → 4절 ★

☐ EUV가 파장에서 14배를 벌고 NA에서 4배를 잃었다는 것을 계산으로 보일 수 있다 → 실습 1

☐ High-NA의 DOF가 왜 22 nm 수준인지, 그것이 CMP·레지스트에 무엇을 요구하는지 말할 수 있다 → 실습 2 ★

☐ **왜 EUV에서 확률적 결함이 문제가 되는지** 광자 수와 $1/\sqrt{N}$ 으로 설명할 수 있다 → 실습 3 ★

☐ 스텝당 99%가 왜 재앙인지 숫자로 말할 수 있다 → 실습 4 ★

☐ 멀티패터닝의 진짜 비용이 해상도가 아니라 **수율**임을 설명할 수 있다 → 실습 4

☐ High-NA의 필드 축소가 왜 칩렛 유인을 키우는지, 스티칭과 견줘 설명할 수 있다 → 실습 5 ★

☐ 평면 → FinFET → GAA → CFET의 동기를 한 줄로 꿰어 말할 수 있다 → 3절

☐ 검사와 계측의 차이를 말하고, AI가 주로 어디에 쓰이는지 설명할 수 있다 → 8절

다음 회차로 넘어가기 전에

이 회차는 **하나의 칩을 만드는 이야기**였다. 6회차는 **여러 칩을 붙이는 이야기**다.

이 회차가 남긴 질문	6회차의 답
High-NA 필드보다 큰 칩은 어떻게 만드나	스티칭으로 이어 찍거나, 칩렛으로 나눠 패키지에서 붙인다
나눈 칩들을 무엇으로 연결하나	인터포저·TSV·하이브리드 본딩
나누면 배선이 늘 텐데 전력은?	실습 2 — 배선 길이가 에너지를 정한다
HBM 스택을 붙일 자리는?	실습 3 — 인터포저 면적이 레티클 한계를 넘는다

출처

스터디 커리큘럼이 고른 입문 자료 ("전공정" 항목)

커리큘럼 소주제	링크	이 책에서의 활용
웨이퍼	삼성전자 반도체 뉴스룸, 「반도체 8대 공정 1탄. 반도체 집적회로의 핵심재료, 웨이퍼란 무엇일까요?」	2절 웨이퍼 제조의 기초. **8대 공정 시리즈는 국문 입문 자료로 가장 접근성이 좋다**
트랜지스터	ETRI 전자통신동향분석 제38권 제6호(2023), 「Complementary FET로 열어가는 반도체 미래 기술」	CFET(n/p 수직 적층) 기술 동향 리뷰. 3절 구조 진화 표의 마지막 행(CFET)의 배경
트랜지스터	ScienceON 특허 KOR1020250044769, 「다층 수직 적층 게이트-올-어라운드(GAA) 트랜지스터 및 그 제조 방법」	적층 GAA 구조의 특허 문헌. 3절 GAA·CFET의 참고
노광	하나금융투자 반도체 산업 보고서 「반도체: 아는 것이 힘이다 — EUV」(2019-04-23, PDF)	EUV 노광장비 산업 리포트. **투자 리포트지만 EUV 기술 설명이 상세하고 그림이 좋아 4절 발제 자료로 유용.** 다만 2019년 자료이므로 High-NA 부분은 최신 자료로 보완 필요
노광	무역안보관리원 학술지 제8호, 전영민, 「첨단 반도체 EUV 리소그래피 기술과 국제안보 정세」(2024, PDF)	EUV 리소그래피의 기술적 특수성과 수출통제(한일·미중) 정세를 함께 다룬 논문. 4절 EUV와 1회차 공급망 논의를 잇는다
노광	ScienceON JAKO202509150403114, 「EUV 리소그래피(Lithography) 공정 기술 동향 및 연구 전망」	EUV 공정 기술 동향 학술 문헌. 4절
식각	ScienceON JAKO198965349552573, 「건식 식각기술의 동향분석」	건식 식각 기술 학술 문헌 (1989년 자료 — 원리 이해용이며 최신 공정은 별도 자료 필요)
식각	ScienceON 특허 KOR1019990043077, 「반도체 식각 장치」	식각 장치 특허
증착	삼성전자 반도체 뉴스룸, 「반도체 8대 공정 6탄. 반도체가 원하는 전기적 특성을 갖게 하려면?」	6·7절 증착·이온주입의 기초
증착	ScienceON JAKO201326940561717, 「원자층증착 기술: 개요 및 응용분야」	ALD 개요 학술 문헌. 6절 ALD
CMP	ScienceON 보고서 KAR2006021584, 「화학기계연마(CMP)의 메커니즘과 마찰특성」	CMP 메커니즘 보고서. 7절 배경
검사측정	ScienceON DIKO0015747253, 「반도체 계측 품질관리와 Particle 검사설비 검출력 향상에 대한 연구」	계측 품질관리·파티클 검사 학위논문. 8절
이온주입	삼성전자 반도체 공식 사이트, 「반도체 8대 공정 6탄, 반도체에 전기적 특성을 입히다! 증착& 이온주입 공정」	위 뉴스룸 6탄과 같은 시리즈의 공식 사이트판(국문). 같은 시리즈 2탄이 산화공정이다

이번 개정에서 더한 1차 출처 (2026-10-08 확인)

- 삼성전자 반도체, 「반도체 8대 공정 2탄, 웨이퍼 표면을 보호하는 산화공정」 — 2절 열산화 (800~1,200 ℃, 건식·습식)
- Deal, B. E. & Grove, A. S., "General Relationship for the Thermal Oxidation of Silicon", *J. Appl. Phys.* 36(12):3770–3778 (1965), doi:10.1063/1.1713945 — 2절 Deal–Grove 식(서지는 Crossref, 식·온도 범위는 초록으로 확인. 계수 표는 대조하지 않음)
- Silicon Valley Microelectronics, 300 mm 웨이퍼 사양서 SV025 · SV027 — 2절 노치·지름·두께·가장자리 제외. 같은 공급사의 150 mm 이하 사양서(SV001~SV013)는 플랫, 200 mm 사양서(SV017~SV020)는 노치
- van Schoot 외(ASML·ZEISS), 「High-NA EUV lithography enabling Moore's law in the next decade」, EUVL Workshop, 2017-06-15 — 4절 축소 배율(0.33 NA 4×/4×, 0.55 NA 4×/8×), 필드 26×33 mm·26×16.5 mm, 6인치 마스크
- ASML, 「5 things you should know about High NA EUV lithography」(2024) — 4절 아나모픽 광학의 이유, 필드 절반, EXE 해상도 8 nm
- ASML, EUV 리소그래피 시스템 페이지 — NXE 해상도 13 nm, 7·5·3 nm 노드 사용
- ZEISS SMT, 「EUV lithography: The light behind the digital age」 — 4절 반사경(몰리브덴·실리콘 교대 다층 최대 100층가량, 반사율 최대 70%). 2026-10-22 확인
- ASML TWINSCAN NXE:3600D · NXE:3800E · EXE:5000 제품 페이지 — 11절 실습 3 풀이(처리량 사양의 기준 dose 30·30·50 mJ/cm²). 2026-10-22 확인
- ASML 보도자료(2026-07-15), 「High NA EUV reaches new readiness milestone with first high-volume Logic product」 · Intel 2026년 2분기 실적 보도자료(2026-07-23) — 4절 High-NA 양산 적용 현황
- imec 보도자료(2026-03-18), 「Imec receives the world's most advanced High NA EUV system」 · SK하이닉스 뉴스룸(2025-09-03), 「SK hynix Introduces Industry's First Commercial High NA EUV」 · Intel 뉴스룸(2024-04-18), 「With High NA EUV, Intel Foundry Opens New Frontier in Chipmaking」 — 4절 High-NA 현황
- TSMC 보도자료(2019-10-07), 「TSMC's N7+ Technology is First EUV Process Delivering Customer Products to Market in High Volume」 — 4절 광원 출력 250 W 이상, 10절 오해 8
- IBM History, 「Copper interconnects」 — 9절 구리 배선(1997-09-22 발표, 알루미늄 대비 저항 약 40% 낮음, 듀얼 다마신)
- TSMC 보도자료(2024-04-24), A16 발표 · Tom's Hardware(2026-04-22), 기술 심포지엄 기사 — A16 양산 2027년(TSMC 임원 발언 인용) · TSMC 2nm(N2) 기술 페이지 · Intel Foundry VLSI 심포지엄 보도자료(2026-06-16) · Intel 2025년 4분기·연간 실적 보도자료(2026-01-22) — 3절 후면전력공급 현황

검증 상태

- High-NA EUV 현황(4절) — 2026-10-08 재확인. Intel의 EXE:5200B 첫 인수 시험 통과와 2026년 7월 18A 특정 층 High-NA 양산 출하(ASML 보도자료 2026-07-15, Intel 2분기 실적 보도자료), SK하이닉스 M16 EXE:5200B 도입(2025-09-03 뉴스룸), imec EXE:5200 도입·2026년 4분기 검증 계획(2026-03-18 보도자료), Intel의 14A 생산 적용 계획(2024-04-18 뉴스룸). 「전 세계 설치 십수 대」「양산 적용 2027~2028년 전망」「TSMC A14·Samsung SF1.4가 대상 노드」 같은 서술은 1차 출처로 확인되지 않아 싣지 않았다
- 축소 배율과 노광 필드(26×33 mm, High-NA 26×16.5 mm, 4×·4×8×) — ASML·ZEISS 2017 발표와 ASML 2024 글로 확인(2026-10-08). 마스크 위 영역 104×132 mm는 이 값에서 계산한 것
- EXE 해상도 8 nm, NXE 13 nm — ASML 공식 페이지 확인(2026-10-08). 실습 1의 k_1 역산(0.326, 0.318)은 이 값으로 계산
- TSMC N7+(TSMC가 내세운 「업계 첫 상용 EUV 리소그래피 기술」, 2019년 2분기 양산), N6의 EUV 확대 적용·로직 밀도 18%, 광원 출력 일상 운영 250 W 이상 — TSMC 보도자료 원문 확인(2026-10-08). 삼성전자 7LPP EUV 웨이퍼 생산 시작은 삼성전자 보도자료(2018-10-18)로 확인
- IBM 구리 배선 1997-09-22 발표, 알루미늄 대비 저항 약 40% 낮음 — IBM 공식 연혁 페이지 확인(2026-10-08). W 플러그·배리어 메탈·일렉트로마이그레이션의 설명은 공정 교과서의 표준 내용
- 열산화 온도(800~1,200 ℃)와 건식·습식의 차이 — 삼성전자 반도체 공식 사이트 확인. Deal–Grove 식과 적용 범위(700~1,300 ℃, O_2·H_2O) — 원논문 초록 확인(2026-10-08). 두 극한은 식에서 바로 나온다
- 300 mm 웨이퍼의 노치(깊이 1.0 mm, 90°, <110>), 지름·두께·가장자리 제외 3 mm — 공급사 사양서(SVMI SV025·SV027) 확인(2026-10-08). **SEMI M1 원문(유료)은 대조하지 못했다(SEMI M1 원문 미대조)**
- 후면전력공급: Intel 18A PowerVia 적용·첫 제품 2026년 1월 발표, TSMC A16 Super Power Rail(2024년 발표 때 2026년 생산 계획 → 2026년 4월 기술 심포지엄의 임원 발언 보도로는 양산 2027년 시작 예상) — 각 사 보도자료로 2026-10-08 재확인, A16 일정은 2026-10-22 재확인(Tom's Hardware 2026-04-22, TSMC 3분기 실적 보도자료 2026-10-15에는 A16 언급 없음). 4회차 4절 표와 같은 출처
- EUV 반사경의 구성(몰리브덴·실리콘 교대 다층, 최대 100층가량, 층마다 몇 nm)과 반사율(최대 70%) — ZEISS SMT 기술 소개 페이지 확인(2026-10-22)
- EUV 처리량 사양의 기준 dose(NXE:3600D·3800E 30 mJ/cm², EXE:5000 50 mJ/cm²) — ASML 제품 페이지 확인(2026-10-22). 실제 공정 dose는 층·레지스트에 따라 다르다
- EUV 파장 13.5nm, 반사 다층경 방식, 주석 액적 플라즈마 광원 — 표준 기술 사실
- 레일리 판정식 및 초점심도 식 — 광학 표준
- 트랜지스터 구조 진화(Planar→FinFET→GAA) — 2026-08-02 확인
- 도핑(5가·3가 불순물, 진성 실리콘의 반도체 성질) — 교과서 표준(Sze & Ng)

미확인 · 유보 (본문에서 단정하지 않음)

- **TSMC A16 양산 시점** — 2027년 양산이라는 일정은 2026년 4월 TSMC 임원 발언 보도로만 확인되고 TSMC 보도자료·실적 자료에는 없다. 양산 개시 공식 발표는 아직 없다(2026-10-22 확인)
- **High-NA 전 세계 설치 대수, Intel 외 회사의 적용 노드** — 2026년 10월 현재 공식 발표로 확인하지 못함
- **Deal–Grove 계수 A·B 의 수치** — 원논문 표를 대조하지 않아 본문에 싣지 않음
- **EUV 광학계 전체 효율** — 반사경 장수·세대에 따라 다르다. 본문은 "손실이 곱해진다"는 정성적 요지만 확정
- **웨이퍼 순도 11N** — 널리 인용되나 용도·등급별 편차가 있다
- **공정 스텝 수(1,000단계 이상)·사이클타임(수개월)** — 제품·노드에 따라 크게 다르다. 자릿수 규모로만 서술
- **k_1 이론 하한 0.25** — 단일 노광 결상 이론상의 값이며, 실제 공정계수는 조명·마스크 기법에 따라 달라진다
- 11절 실습의 k_1, k_2, 스텝 수 등 — **계산 연습용 가정값**

6회차

후공정과 첨단 패키징

핵심 TAKEAWAY

AI 반도체를 이해하려면 이 회차가 가장 중요하다. 한때 "자르고 붙이고 포장하는" 저부가가치 마무리였던 후공정이, 오늘날 성능 향상의 주된 무대가 되었다. 이유는 세 가지 압력이 동시에 작동했기 때문이다: ① **수율** — 다이가 크면 $Y = e^{-AD_0}$ 로 수율이 무너진다, ② **노광 필드 한계** — 한 번에 찍을 수 있는 면적에 상한이 있고 High-NA에서 그 절반이 된다, ③ **이종 최적화** — 로직·SRAM·아날로그·I/O는 최적 공정이 서로 다르다. 셋 다 "쪼개서 붙여라"를 가리킨다. 그리고 3회차의 HBM 문제(광폭 버스를 뽑으려면 아주 가까이 붙여야 한다)가 여기서 TSV·인터포저·적층 본딩으로 답을 얻는다. **메모리 문제와 패키징 문제는 같은 문제였다.**

원자료 바로가기

SK하이닉스 뉴스룸 「반도체 후공정」 6편 컨벤셔널 패키지 공정 · ScienceON 「인공지능 반도체 및 패키징 기술 동향」(2023) · ScienceON 「AI 하드웨어를 위한 3D 패키징 고대역폭/대용량 메모리 기술 발전과 연구 동향」(2026) · Siemens EDA 「Comprehensive 3D IC guide to heterogeneous integration」 · TSMC 3DFabric(CoWoS·InFO·SoIC) · JEDEC JESD47 표준 페이지 · SK하이닉스 HBM4E 12단 샘플 공급(2026-06-18) · UCIe 사양 페이지

학습목표

1. 전통적 후공정과 첨단 패키징의 목적 차이를 설명할 수 있다.
2. 칩렛 분할의 손익을 수율 모형으로 계산할 수 있다.
3. TSV·인터포저·하이브리드 본딩이 각각 어떤 물리적 문제를 푸는지 설명할 수 있다.
4. 2.5D와 3D를 정확히 구분하고, 설계 목표에 맞는 방식을 고를 수 있다.
5. CoWoS 캐파가 왜 AI 칩 공급의 병목인지 설명할 수 있다.

0. 자습 안내 — 혼자 읽는 사람에게

> **이 회차는 앞선 다섯 회차가 모이는 지점이다**
>
> 2회차의 대역폭, 3회차의 HBM, 4회차의 수율, 5회차의 레티클 한계가 여기서 하나의 문제로 합쳐진다 — **"큰 칩을 못 만드는데 큰 성능이 필요하면 어떻게 하는가."** 실습 1(칩렛 최적 분할)이 그 답을 숫자로 준다. 이 회차의 실습은 다섯 개 중 넷이 앞 회차의 도구를 재사용하므로, 앞을 이해했다면 의외로 빨리 풀린다.

직전 회차들에서 가져오는 것 — 이 회차는 특히 의존도가 높다

어디서	무엇을
4회차	$Y = e^{-AD_0}$, GDPW 식 — **실습 1에 그대로 쓴다**
5회차 실습 5	레티클 한계 858 mm², High-NA는 429 mm²
3회차 실습 2	pJ/bit — **실습 2에 그대로 쓴다**
3회차	HBM 스택 구조, TSV, 적층 수율 y^n, 양산 적층 방식(TSV + 마이크로범프, MR-MUF·TC-NCF)과 차세대 후보 하이브리드 본딩
2회차	왜 대역폭이 전부인가

> 4회차 실습 4(칩렛 손익)를 풀지 않았다면 **지금 돌아가서 풀고 올 것.** 이 회차 실습 1은 그것의 일반화다.

선행 지식 점검

개념	한 줄 답
기생 커패시턴스	배선이 의도치 않게 갖는 전기 용량. 길이에 대략 비례하고, 이것을 채우고 비우는 데 에너지가 든다
$E = \frac{1}{2}CV^2$	커패시턴스 C 를 전압 V 로 채우는 데 드는 에너지
열저항 $R = t\,/\,(kA)$	두께 t, 열전도율 k, 면적 A 인 층을 열이 통과할 때의 저항 [K/W]. 패키지 전체를 한 값으로 묶은 θJA·θJC는 7절에서 다룬다
피치	인접한 배선·범프 사이의 중심 간 거리
신호 무결성	전송된 신호가 왜곡·잡음 없이 도착하는가의 문제

읽는 순서 (총 130분 권장)

순서	절	시간	비고
1	1~3절 (후공정 개관, 왜 쪼개나)	25분	4회차의 연장
2	4~6절 (2.5D·3D, TSV, 본딩)	35분	HBM 구조가 여기서 완성된다
3	**실습 1**	**25분**	**이 회차의 중심.** 표를 직접 채울 것
4	7절 열·신뢰성 + 실습 4	20분	7회차로 이어지는 다리
5	8·9절 UCIe·신호 무결성 + 실습 2·3·5	25분	

이 회차를 관통하는 한 문장

> **패키징은 "칩을 담는 상자"가 아니라 「배선의 연장」이다.** 다이를 얼마나 가까이 붙이느냐가 대역폭·전력·열을 동시에 결정한다. 실습 2가 이것을 200배라는 숫자로 보여 준다.

막히기 쉬운 지점 — 2.5D와 3D가 헷갈릴 때

용어가 혼란스러운 것은 이름이 기하학이 아니라 **연결 방식**을 가리키기 때문이다.

	배치	연결	대표 예
2D	나란히, 기판 위	기판 배선 (길다)	전통적 멀티칩 모듈
2.5D	나란히, **인터포저 위**	인터포저 배선 (짧다)	GPU + HBM (CoWoS)
3D	**위아래로 적층**	TSV + 마이크로범프 또는 하이브리드 본딩 (가장 짧다)	HBM 스택 내부(지금은 마이크로범프), 로직 온 로직

기억법: 2.5D는 "옆에, 그러나 아주 가까이", 3D는 "위에". 그리고 **오늘의 AI 가속기는 둘을 동시에 쓴다** — HBM 스택 내부는 3D, GPU와 HBM 사이는 2.5D다.

---

1. 직관 잡기 — 후공정의 목적이 바뀌었다

옛날 후공정

```
웨이퍼 테스트 (전기 특성 검사 → 웨이퍼 번인 → 기능 테스트 → 리페어 → 재테스트)
  → 백그라인딩(뒷면 연삭) → 다이싱(자르기) → 다이 부착 → 와이어 본딩
  → 몰딩(수지 봉지) → 볼 부착·개별 분리(BGA의 경우) → 패키지 테스트 → 출하
```

테스트가 맨 끝에 한 번 있는 것이 아니다. **자르기 전에 웨이퍼 상태에서 먼저 걸러 내고, 싸고 난 뒤에 다시 검사한다.** 앞쪽 테스트가 2절의 KGD를 만든다.

패키지가 하는 일은 넷이다. ① **칩을 물리적·화학적으로 보호하고,** ② **전력을 칩까지 나눠 보내고,** ③ **신호를 바깥세상과 주고받게 하고,** ④ **칩에서 난 열을 빼낸다.** 이 분야의 표준 참고서로 꼽히는 Tummala 등이 엮은 『Microelectronics Packaging Handbook』 2판(Chapman & Hall, 1997)의

『Semiconductor Packaging』 권도 출판사 소개문에서 전자 패키징을 상호연결·전력 공급·냉각·보호로 정의한다(서지는 출판사 목록으로 확인했고, 책 본문은 대조하지 못했다). 넷째가 이 회차 7절과 7회차의 주제다. 옛날에는 이 작업이 상대적으로 단순해서 인건비가 싼 지역의 OSAT에 외주를 주었고, 부가가치도 낮았다.

패키지는 층위로도 나눈다. 다이 자체, 다이를 담은 단품 패키지(1차 패키지), 그 패키지를 실장한 보드(2차), 보드를 모은 시스템 순이다. 다이를 0차로 세는지, 시스템을 몇 차로 부르는지는 문헌마다 다르다. 번호보다 **어느 층에서 어느 층으로 잇는가를** 보면 된다. 첨단 패키징은 1차 패키지 안에 다이 여러 개와 인터포저를 넣어, 원래 보드가 하던 연결을 패키지 안으로 끌어들인 것이다.

오늘날 첨단 패키징

목적이 완전히 다르다. **여러 개의 칩을 초고대역폭으로 서로 연결해서, 마치 하나의 거대한 칩처럼 동작하게 만드는 것**이다.

이제 패키징이 성능을 결정한다. 같은 GPU 다이라도 HBM을 몇 개 붙였는지, 인터포저가 어떤 종류인지에 따라 시스템 성능이 크게 달라진다. 그래서 **파운드리가 패키징을 직접 흡수하기 시작했다** (TSMC의 CoWoS·SoIC, Intel의 Foveros·EMIB, Samsung의 I-Cube·X-Cube).

> **한 문장:** 전공정이 "트랜지스터를 작게 만드는 경쟁"이라면, 후공정은 "칩들을 가깝게 붙이는 경쟁"이다. 그리고 지금은 후자가 더 빠르게 성과를 낸다.

---

2. 왜 무게중심이 옮겨갔나 — 세 가지 압력

압력 ① 수율 — 큰 다이는 만들 수 없다

4회차의 식을 다시 쓴다. $Y = e^{-AD_0}$.

계산 비교 ($D_0 = 0.1/\text{cm}^2$ 가정)

방식	다이 구성	개별 수율	시스템 수율
단일 대형 다이	800 mm² × 1	44.9%	**44.9%**
칩렛 분할	200 mm² × 4	81.9% (각)	$0.819^4 = 45.0\%$ ← 어?

잠깐, 같지 않은가? 순진하게 계산하면 그렇다. e^{-AD_0} 에서 면적을 4등분해도 전부 필요하면 곱해져서, 면적을 cm²로 바꿔 넣으면 $\left(e^{-2\,\text{cm}^2 \times D_0}\right)^4 = e^{-8\,\text{cm}^2 \times D_0} = e^{-0.8} = 44.9\%$ 로 같아진다(800 mm² = 8 cm²).

핵심은 여기서부터다. 칩렛이 이기는 이유는 다음과 같다.

1. **KGD 선별이 가능하다.** 작은 다이는 개별 테스트가 쉽고 싸다. 양품만 골라 조립하면 시스템 수율이 조립 수율에 가까워진다. 큰 다이는 결함이 있으면 통째로 버린다.
2. **버릴 때 버리는 양이 다르다.** 800mm² 다이가 불량이면 800mm²를 버린다. 200mm² 칩렛 하나가 불량이면 200mm²만 버리고 나머지 셋은 재고로 쓴다.

3. **재사용된다.** 같은 I/O 칩렛을 여러 제품에 쓸 수 있다. NRE가 분산된다.
4. **혼합 노드가 가능하다.** 압력 ③ 참조.
5. **웨이퍼당 다이 수가 늘어 가장자리 손실 비율이 준다.**

KGD는 어디서 나오나. 1절 흐름도의 앞쪽, 웨이퍼 테스트가 만든다. 웨이퍼 상태에서 전기 특성을 재고, 높은 온도와 전압을 걸어 일찍 죽을 칩을 미리 드러내고(번인), 기능을 검사한다. 메모리는 여기서 끝나지 않는다. 불량 셀이 나온 행·열을 여분 행·열로 바꿔 넣는 리페어를 하고 다시 테스트한다. 초기 고장을 얼마나 오래 걸러야 하는지는 표준이 실마리를 준다. JEDEC JESD47I(2012)는 초기 고장률(ELFR)을 재는 인증 시험을 접합 온도 125 °C 이상, 48~168시간으로 둔다. 이것은 제품 인증 시험의 조건이고 양산 번인 조건 자체는 아니다.

정확한 비교식: 칩렛 n 개, 각 면적 A/n, KGD 선별 후 조립 수율 y_{asm} 일 때

$$Y_{\text{chiplet-system}} \approx y_{\text{asm}}, \qquad \text{양품 시스템 하나당 비용} \approx \frac{1}{y_{\text{asm}}}\left[n \times \frac{c_{\text{wafer}}}{\text{GDPW}(A/n) \times e^{-(A/n)D_0}} + c_{\text{assembly}}\right]$$

대괄호 안은 조립을 한 번 시도하는 데 드는 비용이다. 조립에 실패한 시스템의 비용은 살아남은 양품이 떠안으므로 y_{asm} 으로 나눈다. 실습 1 풀이는 c_{assembly} 를 0으로 두고 이 식으로 셈한다.

대가: 칩렛 간 통신을 위한 인터페이스 회로(PHY)가 면적과 전력을 잡아먹고, 다이를 건너는 통신은 온다이 통신보다 느리고 비싸다. **분할이 항상 이기는 것은 아니다** — 4회차 실습 4에서 손익분기를 계산했다.

압력 ② 노광 필드 한계 — 물리적으로 못 만든다

노광기는 한 번에 일정 면적만 찍는다(레티클 필드). ASML은 0.33 NA EUV 노광기(NXE)의 최대 노광 필드를 26 mm × 33 mm로 적는다. 계산하면 858 mm²다. **High-NA EUV(EXE)는 아나모픽 광학 때문에 필드가 NXE의 절반이라고 ASML이 밝히고, 계산하면 약 429 mm²다.** 왜 절반인지는 5회차에서 계산했다.

이보다 큰 칩은 **원리적으로 한 번에 만들 수 없다.** 스티칭(필드를 이어붙이는 기법)이 있지만 정렬 오차와 수율 문제로 일반적이지 않다.

> ★ **5회차에서 예고한 고리:** High-NA EUV 도입은 해상도를 얻는 대신 필드를 잃는다. 그래서 **429 mm²보다 큰 다이가 필요한 설계는 High-NA로 넘어가면서 스티칭과 칩렛 가운데 하나를 골라야 하고, 칩렛을 택할 유인이 커진다.** 필요한 다이 면적이 429 mm² 안에 들어가거나, 미세화로 같은 회로가 그 안에 들어가거나, 스티칭의 정렬·수율 부담을 감수한다면 분할하지 않아도 된다. 미세화의 다음 단계가 패키징 쪽 해법의 유인을 키운다는 이 고리가 5회차와 이 회차를 잇는다.

압력 ③ 이종 최적화 — 모든 회로가 미세화의 이득을 보지는 않는다

회로 유형	미세화 이득	최적 선택
로직 (연산)	큼	최신 노드
SRAM (캐시)	**작아지고 있음** — SRAM 셀 면적 축소가 로직보다 훨씬 느려졌다	최신 노드의 이득이 제한적
아날로그·I/O	**거의 없음** — 전압 스윙과 물리적 크기가 필요	성숙 노드가 훨씬 저렴
전력 관리	없음	성숙 노드

최신 노드의 웨이퍼는 성숙 노드보다 훨씬 비싸다. **미세화 이득이 없는 회로를 최신 노드에 태우는 것은 낭비다.** 그래서 로직은 최신 노드, I/O는 성숙 노드로 나눠 만들고 붙인다. AMD가 이 전략을 대규모로 실증했고, 지금은 업계 표준이 되었다.

---

3. 다이싱 — 자르기도 어려워졌다

방식	원리	특징
블레이드	다이아몬드 날로 절단	전통적, 저렴. 칩핑(가장자리 깨짐), 커프(절단 폭) 손실
레이저 스크라이빙	레이저로 홈을 내고 분리	커프 작음, 열영향부 발생
스텔스 다이싱	레이저 초점을 실리콘 **내부**에 맺어 개질층을 만든 뒤 테이프 확장으로 갈라냄	표면 손상·파티클 거의 없음, **얇은 웨이퍼에 적합**
플라즈마 다이싱	식각으로 분리 (DBG와 조합)	임의 형상 가능, 칩핑 없음, 다이 강도 우수

왜 어려워졌나

1. **웨이퍼가 얇아졌다.** TSV용·적층용 웨이퍼는 수십 μm까지 갈아낸다. 종잇장처럼 얇은 것을 깨지 않고 자르고 다뤄야 한다.
2. **커프를 줄여야 한다.** 절단 폭이 곧 버리는 면적이다. 다이가 작아질수록 커프 비중이 커진다.
3. **다이 가장자리 강도가 중요해졌다.** 적층하면 응력이 걸리므로, 절단면에 미세 크랙이 있으면 나중에 파손된다.
4. **Low-k 절연막이 약하다.** 배선층의 저유전율 물질은 기계적으로 취약해서 블레이드 절단 시 박리된다.

---

4. TSV — 실리콘을 관통하는 수직 배선

왜 필요한가

3회차에서 본 문제: **대역폭을 얻으려면 선을 아주 많이 뽑아야 하고, 그러려면 아주 가까워야 한다.** 칩을 위아래로 쌓으면 거리가 μm 수준으로 줄지만, 위아래 칩을 연결할 통로가 있어야 한다. 그 통로가 TSV다.

전기적으로 왜 좋은가

배선의 지연은 대략 RC에 비례하고, 구동 에너지는 $E = \frac{1}{2}CV^2$ 에 비례한다. 배선 길이가 짧아지면 R 과 C 가 함께 줄어든다.

- 패키지 기판을 돌아가는 배선: **수 mm ~ 수십 mm**
- 인터포저 배선: **수백 μm ~ 수 mm**
- TSV(칩 두께 관통): **수십 μm**

길이가 100배 줄면 기생 커패시턴스도 대략 그만큼 줄고, 비트당 전송 에너지가 급감한다. 이것이 HBM이

GDDR보다 pJ/bit가 훨씬 좋은 이유의 핵심이다.

공정 흐름

```
1. 고종횡비 식각 (Bosch 공정 등) — 깊고 좁은 구멍
2. 절연막 증착 (ALD/CVD) — 실리콘과 전기적 격리
3. 배리어/시드층 증착 (ALD/PVD) — Cu 확산 방지 + 도금 씨앗
4. 구리 전해도금 — 보이드 없이 채우기 (bottom-up fill)
5. CMP — 표면 여분 구리 제거
6. 웨이퍼 박막화(백그라인딩) — 뒷면을 갈아 TSV를 노출
7. 뒷면 처리 — 절연·패드 형성
```

난제

- **보이드 없는 충전:** 깊은 구멍을 구리로 빈틈없이 채워야 한다. 입구가 먼저 막히면 안쪽에 공극이 생긴다. 첨가제를 써서 바닥부터 채우는 도금 화학이 핵심 기술이다.
- **열팽창 불일치(CTE mismatch):** 구리와 실리콘의 열팽창계수가 다르다. 온도가 변하면 TSV 주변에 응력이 생기고, 이 응력이 근처 트랜지스터 특성을 바꾼다. 그래서 TSV 주변에 소자를 두지 않는 **금지영역(keep-out zone)** 을 설정한다. **면적 오버헤드가 발생하는 지점.**
- **언제 뚫나:** 트랜지스터를 만들기 전에 뚫으면 via-first, 트랜지스터를 만든 뒤 배선층을 쌓기 전에 뚫으면 via-middle, 배선층까지 끝낸 뒤 뚫으면 via-last다. 순서에 따라 TSV가 견뎌야 할 고온 공정과 금지영역 설계가 달라진다. HBM 같은 적층 메모리는 via-first·via-middle 쪽의 대표 예로 꼽힌다.
- **박막화된 웨이퍼 취급(WSS, wafer support system):** 수십 μm 두께의 웨이퍼는 스스로 형태를 유지하지 못하고 휜다. 그래서 단단한 캐리어에 잠시 붙여 놓고 뒷면 공정을 한다. 순서대로 보면, 캐리어에 붙이기 전에 웨이퍼 앞면 가장자리를 깎아 낸다(에지 트리밍). 웨이퍼 가장자리는 둥글게 다듬어져 있어서, 그대로 붙여 수십 μm까지 갈면 가장자리가 칼날처럼 얇게 남고 뒤 공정에서 그 끝부터 깨진다. 그다음 임시 접착제로 캐리어에 붙이고, 뒷면을 갈아 TSV를 드러내고, 뒷면 범프를 만든 뒤 캐리어를 떼어 낸다(디본딩). 떼어 낼 때는 얇아진 웨이퍼를 깨뜨리지 않고 범프를 상하지 않게 하면서 접착제를 남기지 않아야 해서 공정 조건이 까다롭다. 디본딩 방식은 접착층을 무엇으로 무력화하느냐로 나눌 수 있다. 힘으로 들어 올려 떼는 기계적 방식, 열로 접착제를 무르게 하는 방식, 용제로 녹이는 화학적 방식, 레이저로 접착층을 분해하는 광학적 방식이다. 레이저 방식은 빛이 캐리어를 통과해야 하므로 유리 캐리어를 쓴다. (공정 순서와 디본딩 방식은 SK하이닉스 뉴스룸 「반도체 후공정」 8편을 바탕으로 정리했다.)
- **스택째 검사해 넘긴다:** HBM은 베이스 다이와 DRAM 다이를 쌓아 몰딩한 뒤 스택 단위로 시험해 KGSD(known good stacked die) 형태로 출하한다. GPU 패키지를 만드는 쪽이 받는 것은 낱장 DRAM이 아니라 검증된 스택이다.

TGV와 유리기판

TGV (Through-Glass Via): 유리를 관통하는 비아. 유리기판이 주목받는 이유:

유리의 장점	의미
낮은 유전손실	고주파 신호 손실이 적다 → 고속 인터커넥트에 유리
우수한 평탄도·강성	대면적에서도 휘지 않는다 → 큰 패키지에 유리
CTE 조절 가능	실리콘에 맞출 수 있다 → 응력 감소
대면적 패널 공정 가능	원형 웨이퍼가 아니라 사각 패널로 처리 → 생산성

남은 과제: 유리는 취성 재료라 크랙이 잘 전파된다. 비아 형성(레이저 유도 식각 등), 금속화, 취급 중 파손이 실용화의 관문이다. 일정은 아직 목표 단계다. Intel은 2023년 9월 유리기판을 발표하며 2020년대 후반에 완전한 유리기판 솔루션을 내놓겠다고 했고, 2026년 7월에는 유리 가공 업체 Lens Technology와 유리기판 패키징 개발을 앞당기려는 협력을 발표했다. 2026년 10월 현재 유리기판 패키지의 대량 양산을 알린 공식 발표는 확인하지 못했다.

---

5. 인터포저 — 칩들을 잇는 판

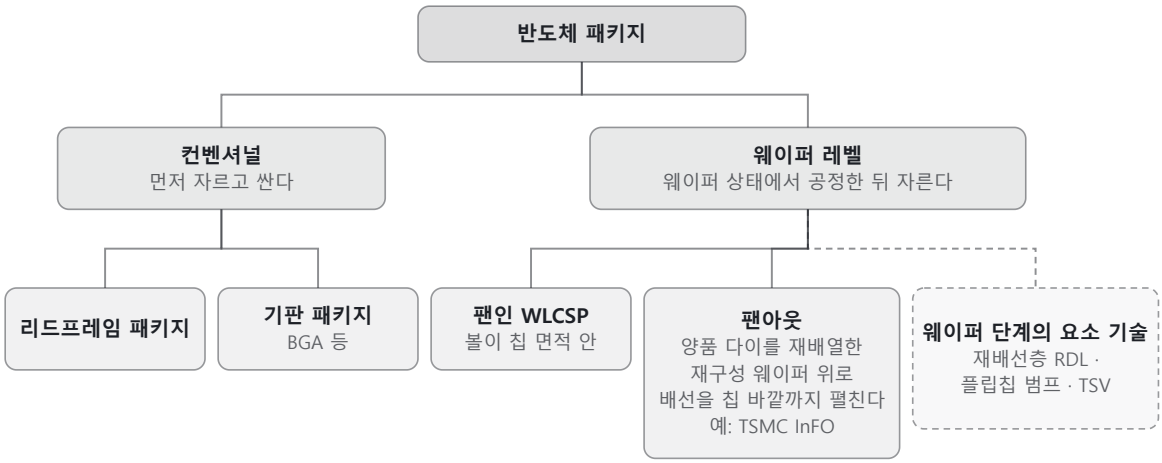
종류와 트레이드오프

유형	배선 밀도	비용	용도
유기 기판	낮음 (수 µm 피치 배선 어려움)	저	일반 패키지, 저대역 연결
RDL 인터포저 (팬아웃)	중간	중	InFO 등. 유기물 위에 재배선층 형성
실리콘 인터포저	**높음** (미세 다마신 배선 가능)	**고**	**HBM 연결의 표준.** CoWoS-S
실리콘 브리지 (EMIB 등)	높음 (필요한 곳만)	중	인터포저 전체를 실리콘으로 하지 않고 연결 부위에만 실리콘 조각 삽입
유리 인터포저	높음 (잠재)	?	개발 중

패키지 분류 — 팬인과 팬아웃

위 표의 「RDL 인터포저(팬아웃)」가 어디서 왔는지 보려면 패키지 분류부터 잡아야 한다. 큰 갈래는 **언제 자르는가다.** 컨벤셔널 패키지는 웨이퍼를 먼저 잘라 낱개 다이를 리드프레임이나 기판(서브스트레이트)에 붙이고 싼다. 웨이퍼 레벨 패키지는 웨이퍼 상태에서 재배선층·범프 같은 패키지 공정을 마친 뒤에 자른다.

웨이퍼 레벨 안에서 다시 둘로 갈린다. **팬인(WLCSP)은** 솔더 볼이 칩 면적 안에만 놓인다. 칩이 작으면 볼 수가 칩 면적에 묶인다. **팬아웃은** 양품 다이만 골라 간격을 두고 다시 늘어놓은 뒤 몰딩 수지(EMC)로 메워 「재구성 웨이퍼」를 만들고, 칩 바깥 몰딩 위까지 재배선층을 펼쳐 볼을 놓는다. 양품만 골라 늘어놓는다는 점에서 2절의 KGD 논리와 뿌리가 같다. TSMC의 InFO(Integrated Fan-Out)가 이 계열의 웨이퍼 레벨 패키지이고, TSMC는 2016년부터 InFO를 대량 출하해 왔다고 밝힌다. InFO에는 로직 다이와 HBM을 재배선층으로 묶는 변형(InFO_MS)도 있어서, TSMC는 이를 레티클 1배 미만 크기에서 CoWoS의 비용·성능 대안으로 소개한다.



나누는 기준은 「언제 자르는가」와 「볼이 칩 면적 밖으로 나가는가」 둘이다. 재배선층·범프·TSV는 패키지 종류라기보다 여러 패키지가 함께 쓰는 웨이퍼 단계의 요소 기술이라 점선으로 따로 묶었다.

왜 HBM에는 실리콘 인터포저가 필요한가

HBM4 스택 하나가 2,048비트다. 스택 8개면 **16,384개의 데이터 배선**이고, 여기에 제어·전원·접지가 더해진다. 이 배선들을 GPU 다이 가장자리에서 HBM까지 수 mm 거리에 밀집시켜야 한다.

유기 기판의 배선 피치로는 이 밀도가 불가능하다. **실리콘 인터포저는 반도체 공정(다마신 배선)으로 만들기 때문에 μm 이하 피치가 가능하다.** 즉 인터포저는 "배선만 있는 칩"이다.

그래서 생기는 문제: 실리콘 인터포저도 반도체 공정으로 만들므로 **노광 필드 제약을 받는다.** GPU + HBM 8개를 올리려면 인터포저가 레티클보다 커야 한다. 그래서:

- **CoWoS-S:** 실리콘 인터포저. 레티클 여러 배 크기를 만들기 위해 스티칭 사용. TSMC는 2012년부터 양산했고 인터포저가 레티클의 3.3배(TSMC 표기 약 2,700 mm^2. 이 책의 858 mm^2로 곱하면 약 2,830 mm^2)까지 커졌다고 밝힌다
- **CoWoS-R:** RDL(재배선층) 기반. 더 크게 만들 수 있고 저렴하나 밀도가 낮음. TSMC는 레티클 3.3배보다 큰 인터포저에는 CoWoS-L이나 CoWoS-R을 권하며, -R은 2023년부터 양산 중이라고 적는다
- **CoWoS-L:** 로컬 실리콘 브리지(LSI)를 RDL에 삽입. **고밀도가 필요한 곳만 실리콘, 나머지는 RDL** — 두 방식의 장점을 결합. 최신 대형 AI 패키지의 주류. 레티클 3.5배 크기의 첫 CoWoS-L이 2024년부터 양산 중이다(TSMC, 2026년 10월 확인)

CoWoS가 왜 AI 칩 공급의 병목인가

1. **공정이 복잡하고 느리다.** 인터포저 제작 + 칩 접합 + 기판 조립의 다단계이며, 각 단계에서 수율 손실이 있다.
2. **캐파 증설에 시간이 걸린다.** 장비 발주·설치·검증에 수년.
3. **수요가 폭증했다.** AI 가속기 하나당 필요한 인터포저 면적이 매우 크고, 출하량이 급증했다.
4. **사실상 특정 파운드리에 집중되어 있다.**

TSMC는 CoWoS 소개 페이지에서 2022년 말 생성형 AI가 등장한 뒤 CoWoS 수요가 더 커졌고 첨단

패키징 캐파를 계속 늘리고 있다고 적는다(2026년 10월 확인). 캐파 규모와 업체별 할당량은 보도마다 편차가 커서 이 책에서는 수치로 단정하지 않는다.

왜 다른 경로로 돌리기가 쉽지 않은가 — 조립 순서. 2.5D 패키지를 조립하는 순서는 크게 둘이다. **CoW-first는** 인터포저를 웨이퍼 상태로 둔 채 그 위에 칩을 먼저 붙이고, 웨이퍼 단위로 몰딩한 다음, 임시 캐리어에 붙여 뒷면을 갈아 TSV를 드러내고 뒷면 재배선층과 범프를 만든 뒤 잘라서 기판에 올린다. 300 mm 실리콘 인터포저로 이 순서를 실제로 거친 공정 보고가 있다(Ho 외, EPTC 2014). CoWoS라는 이름(chip-on-wafer-on-substrate)부터가 이 순서를 가리킨다. **CoS-first는** 인터포저를 먼저 잘라 유기 기판에 붙이고 그 위에 칩을 올린다. 둘을 가르는 것은 휨이다. 실리콘의 열팽창계수는 유기 기판보다 훨씬 작다. 큰 실리콘 인터포저를 먼저 기판에 붙이면 접합 온도에서 식는 동안 두 층이 다르게 줄어 휘고, 칩을 올릴 때는 그 휜 면 위에서 수만 개의 마이크로범프를 맞춰야 한다. CoW-first는 평평하고 단단한 웨이퍼 위에서 미세 접합을 먼저 끝내므로 이 문제를 줄인다(몰딩한 인터포저의 휨은 CoW-first에서도 관리해야 한다). 인터포저가 클수록 차이가 벌어진다. 그런데 CoW-first는 웨이퍼 단위의 몰딩·박막화·재배선 공정이 필요해서, 그런 설비와 경험을 가진 곳이 아니면 쉽게 대신할 수 없다. 병목이 한 곳에 몰리는 이유 가운데 하나다.

> **1회차의 "병목은 계속 이동한다"가 여기서 구체화된다.** 로직 웨이퍼 캐파는 여유가 있어도, CoWoS 캐파와 HBM 공급이 부족하면 AI 가속기는 그만큼밖에 못 만든다. **AI 칩 출하량의 상한을 결정하는 것은 최신 노드가 아니라 패키징과 메모리다.**

---

6. 본딩 — 붙이는 기술의 세대 교체

진화 경로

세대	방식	피치	특징
와이어 본딩	금·구리 와이어로 칩 패드와 기판을 연결	수십 µm	가장자리에만 연결 가능, I/O 수 제한, 인덕턴스 큼
플립칩 (C4 범프)	칩을 뒤집어 솔더 범프로 직접 접합	~100 µm	**면 전체 사용 가능** → I/O 수 급증
마이크로 범프	작은 솔더 범프	~40 µm 이하	2.5D/3D 적층의 주력
마이크로 범프 + 틈 메움 (MR-MUF / TC-NCF)	MR-MUF: 칩을 다 쌓은 뒤 한꺼번에 리플로우해 범프를 붙이고, 칩 사이에 액체 보호재를 흘려 넣어 굳힌다. TC-NCF: 층마다 비전도성 필름을 깔고 열과 압력으로 한 장씩 눌러 붙인다	마이크로 범프와 같음	**2026년 10월 현재 양산 HBM의 적층 방식.** 범프는 그대로이고 틈을 메우는 방식이 다르다
하이브리드 본딩 (Cu-Cu)	**범프 없이 구리 패드끼리 직접 접합** + 주변 절연막도 접합	**10 µm 미만**	최고 밀도·최저 저항·최저 인덕턴스

적층 형태와 수율 — W2W와 D2W. 다이를 쌓는 형태는 칩 위에 칩(C2C), 웨이퍼 위에 다이(D2W, C2W라고도 한다), 웨이퍼 위에 웨이퍼(W2W)로 나뉜다. W2W는 웨이퍼 두 장을 한 번에 붙여 생산성이 높지만 위아래 다이의 크기와 배열이 같아야 하고, 무엇보다 **양품을 골라 붙일 수 없다.** 위 웨이퍼의 불량 다이가 아래 웨이퍼의 양품 다이 위에 그대로 얹힌다. 그래서 n 층을 쌓으면 3회차 5절의 적층 수율 식이

그대로 적용된다.

$$Y_{\mathrm{W2W}} = y^{n} \times y_{\mathrm{bond}}^{\,n-1}$$

D2W는 다이를 잘라 시험한 뒤 양품(KGD)만 골라 웨이퍼 위에 붙인다. 바닥 웨이퍼도 미리 시험해 양품 다이 자리에만 쌓는다고 하자. 선별이 완벽하다고 두면 다이 수율 항이 빠지고 본딩 수율만 남는다.

$$Y_{\mathrm{D2W}} \approx y_{\mathrm{bond}}^{\,n-1}$$

이 식의 분모는 바닥 웨이퍼의 양품 자리다. 바닥 웨이퍼의 불량 자리에도 쌓거나 그 자리까지 분모에 넣으면 바닥 다이 수율이 한 번 곱해져 $y \cdot y_{\mathrm{bond}}^{\,n-1}$ 이 된다. 아래 가정값으로 $n = 8$ 이면 0.869이고, W2W 대비 2.32배가 아니라 2.09배다. 버린 바닥 자리의 값은 비용으로 남는다.

다이 수율 $y = 0.9$, 본딩 한 번의 수율 $y_{\mathrm{bond}} = 0.995$(둘 다 가정값, 본딩 수율은 3회차 5절 예시와 같은 값)로 계산하면 다음과 같다.

적층 수 n	W2W $y^{n} y_{\mathrm{BOND}}^{\,n-1}$	D2W $y_{\mathrm{BOND}}^{\,n-1}$	D2W ÷ W2W
2	0.806	0.995	1.23배
8	0.416	0.966	2.32배
12	0.267	0.946	3.54배
16	0.172	0.928	5.40배

두 층이면 차이가 20%대지만 층이 늘면 몇 배로 벌어진다. W2W가 층수가 적고 다이 크기가 같으며 수율이 높은 공정에 맞는 이유다. TSMC도 웨이퍼 대 웨이퍼 방식의 SoIC-X를 수율이 높은 노드, 같은 다이 크기의 설계에 맞는 방식으로 소개한다. 다만 D2W에서도 버린 불량 다이의 값은 비용으로 남는다. 수율 손실이 사라지는 것이 아니라 **적층하기 전으로 옮겨 가서, 정상 다이를 함께 버리지 않게 되는 것이다.** 2절 압력 ①의 KGD 논리를 적층 방향으로 늘린 셈이다.

아래 그림은 위 표의 가정값($y = 0.9$, $y_{\mathrm{bond}} = 0.995$)으로 8층을 쌓은 경우다. 화면본에서는 다이 수율, 본딩 수율, 적층 수를 바꿔 볼 수 있다.

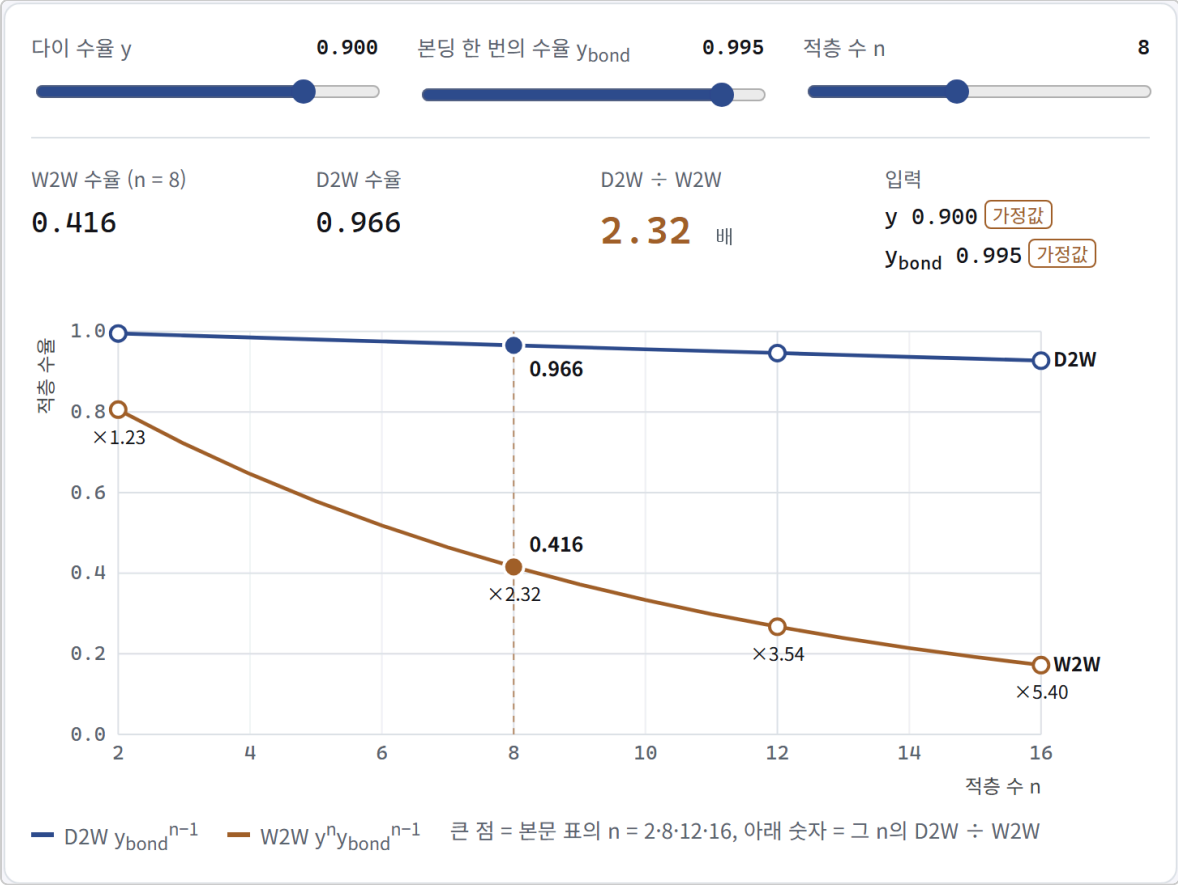


그림 6-1. 다이 수율 0.9, 본딩 수율 0.995(가정값)로 8층을 쌓으면 W2W 수율은 0.416, 양품만 골라 붙이는 D2W는 0.966으로 2.32배 차이가 난다. — 화면본에서는 이 자리에서 값을 바꿔 볼 수 있다(화면본 6회차 6절).

하이브리드 본딩이 왜 중요한가

범프의 문제: 범프는 물리적 크기가 있다. 피치를 줄이려면 범프를 작게 해야 하는데, 작은 솔더는 접합 신뢰성이 떨어지고 전기적 저항이 커진다. 또 범프 사이를 채우는 언더필 재료도 좁아지면 주입이 어렵다. **범프 방식은 대략 수십 μm 피치에서 벽에 부딪힌다.**

하이브리드 본딩의 원리: 두 웨이퍼(또는 다이)의 표면을 원자 수준으로 평탄하게 만들고, 구리 패드와 절연막(주로 SiO_2)이 함께 노출된 표면끼리 맞댄다. 상온에서 절연막끼리 먼저 결합하고(반데르발스 → 공유결합), 이후 어닐링하면 구리가 팽창하며 서로 확산 접합된다.

얻는 것: 피치가 10 μm 아래로 내려가 면적당 연결 수가 마이크로범프(약 40 μm)의 16배를 넘고(피치 4 μm면 100배), 접합부 저항·기생 성분이 극적으로 줄며, 스택 높이가 낮아진다.

요구되는 것 (극단적)

- **표면 평탄도:** 옹스트롬 단위. CMP 품질이 전부를 좌우한다
- **청정도:** 파티클 하나가 그 자리의 접합을 망친다
- **정렬 정밀도:** 수백 nm 이하. 다이 간 정렬 장비가 새로운 병목
- **구리 리세스 제어:** 구리가 절연막보다 살짝 낮아야 하는데, 그 양이 나노미터 단위로 관리되어야 한다

적용처: CMOS 이미지센서는 여러 해 전부터 이 방식으로 대량 생산되어 왔다. 3D 로직 적층도 이 방향이다(TSMC SoIC 등. TSMC는 SoIC의 결합 피치가 10 μm 미만에서 시작하고, 3 nm 칩 적층이

2025년 양산에 들어갔다고 밝힌다). HBM에서는 **차세대 고단 적층의 후보다.** 단수가 늘면 범프와 틈 메움층의 높이가 스택 높이 제한에 걸리기 때문이다. SK하이닉스는 2022년에 HBM2E 8단을 하이브리드 본딩으로 쌓아 전기 테스트까지 마쳤다고 2023년에 밝혔지만, 2026년 6월 고객사에 공급한 HBM4E 12단 샘플에는 어드밴스드 MR-MUF를 썼다. 2026년 10월 현재 하이브리드 본딩을 쓴 HBM의 양산을 알린 제조사 발표는 확인하지 못했다(3회차 4절과 같은 판단).

> **짚어 둘 점:** 하이브리드 본딩은 **후공정에 전공정 수준의 정밀도를 요구한다.** OSAT가 아니라 파운드리가 이 영역을 가져가는 이유가 여기 있다.

---

7. 2.5D와 3D — 정확한 구분

정의

- **2.5D:** 칩들을 인터포저 위에 **나란히(수평으로)** 배치. 칩끼리 위아래로 겹치지 않는다. **대표: GPU + HBM 스택 (CoWoS)**
- **3D:** 칩을 **수직으로 겹친다.** 대표: HBM 스택 내부의 DRAM 적층, 로직 위에 로직/캐시 적층(SoIC, Foveros)

"2.5D"라는 이름의 유래: 완전한 3D는 아니지만 평면 배치보다는 통합적이라는 뜻. HBM 스택 자체는 3D이므로, GPU+HBM 시스템은 **2.5D 안에 3D가 들어있는** 구조다.

비교

	2.5D	3D
연결 거리	mm 수준 (인터포저 경유)	μm 수준 (직접 적층)
대역폭 밀도	높음	**매우 높음**
열 방출	각 칩이 방열판을 볼 수 있음	**위 칩이 아래 칩을 덮어 열이 갇힘**
면적	인터포저 면적 큼	풋프린트 작음
공정 난도	중간	높음
설계 복잡도	중간	높음 (열·전력·EDA 전부)

3D의 최대 적 — 열

이것이 3D 적층의 근본 난제다.

- 칩을 겹치면 단위 면적당 발열이 층수만큼 곱해진다.
- **아래 칩의 열이 위 칩을 통과해서 나가야 하는데**, 실리콘과 접합층의 열저항이 있다. 위에 얹힌 칩이 사실상 담요 역할을 한다.
- 온도가 오르면 누설전류가 늘고, 누설이 늘면 열이 더 나는 **양의 되먹임**이 생긴다.
- 그래서 3D 로직 적층은 대개 **발열이 적은 층을 위에 올린다**(예: 캐시 SRAM을 로직 위에). 고발열 로직끼리 겹치는 것은 아직 어렵다.

HBM에서도 같은 문제: 12~16단 적층에서 가장 아래 다이의 열이 잘 빠지지 않는다. 스택 높이 제한과

함께 적층 수를 제약하는 요인이다. 제조사도 이 값을 내세운다. SK하이닉스는 2026년 6월 HBM4E 12단 샘플을 공급하며 열저항을 HBM4보다 약 17% 낮췄다고 밝혔다.

패키지 밖으로 나가는 열 — θ와 ψ. 지금까지 본 $R = t / (kA)$ 는 층 하나의 열저항이다. 데이터시트는 패키지 전체를 몇 개의 숫자로 묶어 적는다. JEDEC JESD51 계열이 정한 **θJA**(접합–주변 공기), **θJC**(접합–케이스), **θJB**(접합–보드)는 접합 온도 T_j 와 기준점 온도의 차를 소비전력 P 로 나눈 값이다. 예컨대 $\theta_{JA} = (T_j - T_a) / P$ 이므로 $T_j = T_a + P\,\theta_{JA}$ 다. 그런데 θJA는 표준 시험 보드 위에서 정지 공기나 정해진 풍속으로 잰 값이라, 보드와 바람이 바뀌면 값도 바뀐다. Renesas 기술 자료는 구리 평면 두 장을 넣은 시험 보드(1S2P)가 한 장도 없는 보드(1S0P)보다 θJA를 대략 20~40% 이상 낮춘다고 적고, TI는 θJA를 같은 조건에서 패키지끼리 비교하는 값이지 실제 설계에서 T_j 를 계산하는 값이 아니라고 적는다. 실제 보드 위에서 패키지 윗면이나 보드 온도를 재 T_j 를 거꾸로 추정할 때는 **ψJT·ψJB를** 쓴다. 식의 꼴은 θ와 같지만 열이 전부 그 두 점 사이로 흐르지는 않으므로, 진짜 열저항이 아니라 특성 파라미터다.

직렬 사슬로 보기. 방열판을 붙인 패키지에서 열은 접합 → 케이스 → TIM → 방열판(또는 콜드플레이트) → 주변(공기나 냉각수)으로 차례로 흐른다. 직렬이므로 더한다.

$$T_j = T_a + P\,(\theta_{JC} + \theta_{\mathrm{TIM}} + \theta_{SA})$$

θ_{SA} 는 방열판에서 주변까지의 열저항이다. 계산 예를 하나 두자. 아래 값은 모두 **가정값이고** 특정 제품의 값이 아니다. P = 1,000 W, T_a = 35 °C(냉각수 입구 온도로 읽어도 된다), θ_{JC} = 0.015 K/W, θ_{TIM} = 0.025 K/W, θ_{SA} = 0.020 K/W로 두면 합은 0.060 K/W다. 단계별 온도 차는 15 K, 25 K, 20 K이고 합 60 K이므로 T_j = 95 °C다. 이 값이 제품마다 데이터시트가 정한 최대 접합 온도 $T_{j,\ \max}$ 보다 낮아야 한다. 가정값에서 가장 큰 몫을 TIM에 준 것은 7회차 3절의 논점, 곧 칩과 냉각판 사이의 얇은 층이 사슬 전체의 병목이 되기 쉽다는 점을 미리 보이기 위해서다.

> **7회차로 이어짐:** 열 문제가 칩 안에서 랙 수준으로 확장된 것이 데이터센터 냉각 문제다.

아래 그림은 위 계산 예의 가정값으로 주변에서 접합까지 온도가 쌓이는 모양이다. 화면본에서는 소비전력, 주변 온도, 세 열저항을 바꿔 볼 수 있고, 데이터시트의 최대 접합 온도를 넣으면 그 값을 넘는지 볼 수 있다.

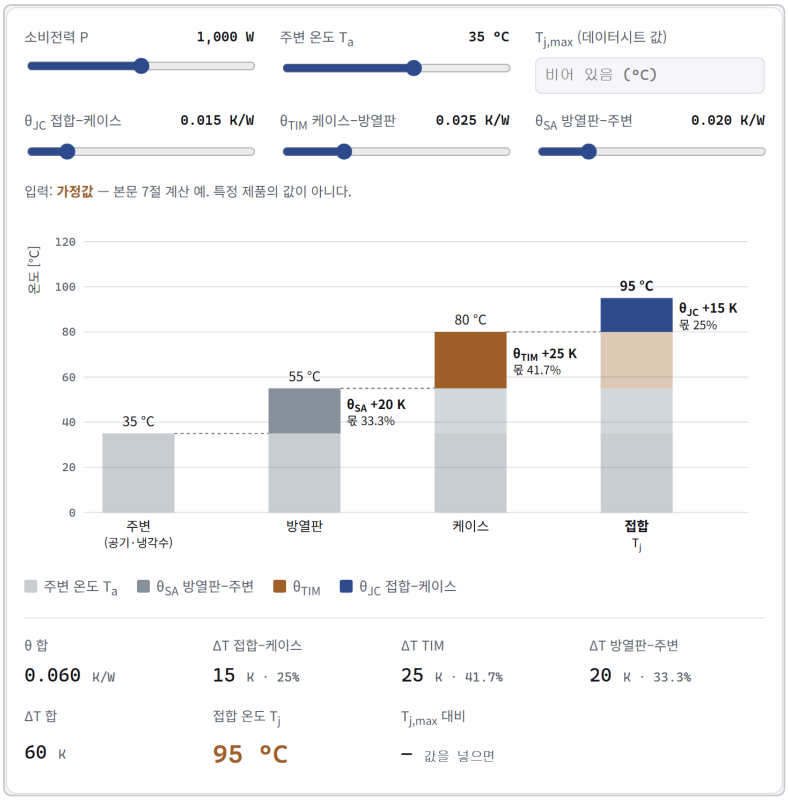
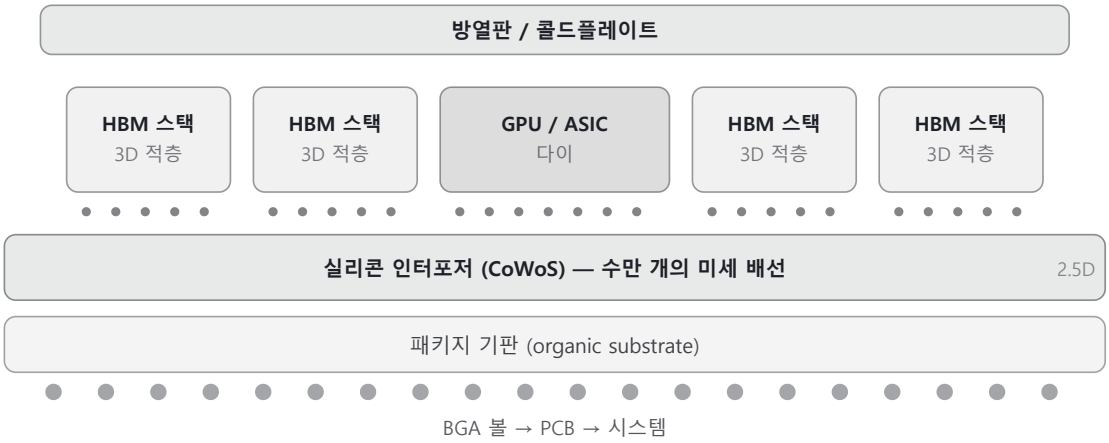


그림 6-2. 1,000 W 칩을 θJC 0.015, θTIM 0.025, θSA 0.020 K/W(가정값) 사슬로 35 °C 주변에 내보내면 단계별 온도 차가 15·25·20 K이고 접합 온도는 95 °C다. — 화면본에서는 이 자리에서 값을 바꿔 볼 수 있다(화면본 6회차 7절).

실제 AI 가속기 패키지의 구조



다이 아래의 점 = 마이크로범프(차세대는 하이브리드 본딩)

여기서 확인할 것

- HBM 스택 내부 = 3D, GPU와 HBM의 관계 = 2.5D
- 패키지 전체 면적이 레티클 여러 배 → CoWoS-L 같은 확장 기법 필요
- 이 패키지 하나의 소비전력이 킬로와트급에 이르며, 그 열이 콜드플레이트로 빠져나가야 한다 (7회차)

패키지 신뢰성 — 열과 습기를 얼마나 견디나

열은 온도를 올리는 데서 끝나지 않는다. 오래 뜨거우면 재료가 열화하고, 켜고 끌 때마다 재료가 서로 다르게 늘었다 줄어 접합부가 피로해진다. 습기는 부식을 부른다. 그래서 새 제품은 출하 전에 가혹 조건으로 미리 망가뜨려 보는 인증 시험을 거친다. 아래는 JEDEC JESD47I(2012년 7월판)의 표 1(소자 시험)과 표 2(비밀봉 패키지 시험)에서 주요 항목을 옮긴 것이다. JEDEC은 2026년 8월 JESD47N을 냈다. 최신판 본문과는 대조하지 않았다.

시험	표준 번호	조건 (JESD47I)	표본·합격 기준	겨누는 고장
HTOL (고온 동작 수명)	JESD22-A108	접합 온도 125 °C 이상, 최대 전원 전압 이상, 1,000시간	3로트 × 77개, 불량 0	동작 중 열·전압에 따른 고장
ELFR (초기 고장률)	JESD22-A108, JESD74	접합 온도 125 °C 이상, 48~168시간	ELFR 표본표	출하 직후 몇 달의 초기 고장
HTSL (고온 보관)	JESD22-A103	150 °C, 1,000시간	3로트 × 25개, 불량 0	확산·산화·금속간 화합물 성장 같은 재료 열화
THB (온습도 바이어스)	JESD22-A101	85 °C, 상대습도 85%, 전압 인가, 1,000시간	3로트 × 25개, 불량 0	습기에 의한 부식, 이온 이동
HAST (고가속 온습도)	JESD22-A110	130 °C 또는 110 °C, 상대습도 85%, 전압 인가, 96시간 또는 264시간	3로트 × 25개, 불량 0	THB를 짧게 가속한 것
UHAST (무바이어스 HAST)	JESD22-A118	130 °C·85%에서 96시간 또는 110 °C·85%에서 264시간	3로트 × 25개, 불량 0	부식
TC (온도 사이클)	JESD22-A104	예: 조건 B는 −55~+125 °C, 700사이클	3로트 × 25개, 불량 0	열팽창 차이에서 오는 열기계적 손상

출처: JEDEC JESD47I(2012-07) 표 1·2에서 골라 옮겼다. 표의 짜임과 「겨누는 고장」 열은 이 책의 정리다.

표 밖에도 규칙이 있다. THB·HAST·TC·UHAST(와 오토클레이브) 앞에는 **프리컨디셔닝을** 한다. JESD22-A113에 따라 J-STD-020의 습기 민감도 등급(MSL)대로 습기를 먹이고 리플로우를 거치게 해서, 보드에 납땜될 때 받는 스트레스를 미리 준다. 포화 수증기로 찌는 오토클레이브는 실제 사용 환경에 없는 손상을 낼 수 있어 인증 시험으로 권하지 않고, 유기 기판 패키지에는 그 대신 HAST를 요구한다. TC 조건 가운데 영하로 내려가는 G·B·C는 유기 기판 플립칩 패키지에 맞지 않을 수 있다는 주석도 있다. **3D 적층과 대형 인터포저에서 문제가 되는 열팽창 차이와 휨은 TC가 겨누는 고장이다.** 4절의 keep-out zone, 5절의 CoW-first가 모두 같은 물리에서 나왔다.

1,000시간은 몇 년인가 — 아레니우스 환산. 온도로 가속되는 고장은 흔히 아레니우스 식으로 시험 시간을 사용 시간으로 바꾼다.

$$AF = \exp\left[\frac{E_a}{k}\left(\frac{1}{T_{\text{use}}} - \frac{1}{T_{\text{stress}}}\right)\right]$$

E_a 는 겉보기 활성화 에너지, $k = 8.617 \times 10^{-5}$ eV/K는 볼츠만 상수, 온도는 절대온도(K)다. 시험 125 °C(398.15 K), 사용 55 °C(328.15 K), E_a = 0.7 eV로 계산하면 $AF \approx 77.7$ 이다. HTOL 1,000시간은 약 77,700시간, 1년을 8,760시간으로 나누면 **약 8.87년에** 해당한다. JESD47I도 표 1 주석에서 같은 조건을 예로 들어 가속계수 78.6, 약 9년이라고 적는다(상수와 반올림 차이로 끝자리가 다르다).

같은 조건에서 E_a 만 바꾸면 결론이 크게 흔들린다.

E_a	가속계수 AF	HTOL 1,000시간의 등가 사용 기간
0.5 eV	22.4	약 2.56년
0.7 eV	77.7	약 8.87년
0.9 eV	269.3	약 30.7년

0.5 eV와 0.9 eV 사이에서 같은 시험의 등가 사용 기간이 약 12배 갈린다. E_a 는 고장 기구마다 다르다. 하나의 시험으로 「이 칩은 9년 간다」고 말할 수 없는 이유다. JESD47I도 이 기간이 실제 사용 요구보다 짧을 수 있고 고장 기구와 사용 환경에 따라 달라진다고 단서를 달며, 장기 마모 기구는 웨이퍼 수준 신뢰성 자료로 보완하라고 한다. 고장 기구별 가속 모형은 JEDEC JEP122가 모아 둔다.

아래 그림은 E_a = 0.7 eV, 125 °C → 55 °C, 1,000시간일 때다. 화면본에서는 E_a, 시험·사용 온도, 시험 시간을 바꿔 볼 수 있다.



그림 6-3. 활성화 에너지 0.7 eV, 시험 125 °C, 사용 55 °C에서 가속계수는 약 77.7이고 HTOL 1,000시간은 약 8.87년에 해당한다. 0.5 eV면 약 2.56년, 0.9 eV면 약 30.7년이다. — 화면본에서는 이 자리에서 값을 바꿔 볼 수 있다(화면본 6회차 7절).

8. 이종집적과 칩렛 표준 (UCIe)

칩렛 생태계의 꿈

칩렛을 표준 인터페이스로 연결할 수 있다면, **서로 다른 회사가 만든 칩렛을 조합해 시스템을 만들 수 있다.** 컴퓨터를 조립하듯 칩을 조립하는 것이다.

UCIe (Universal Chiplet Interconnect Express) 가 이를 위한 표준이다. PCIe/CXL 프로토콜을 다이-투-다이 물리계층 위에 올리는 구조로, 패키지 유형별(표준 패키지 / 첨단 패키지) 사양을 규정한다. 업계 연합이 2022년 3월 출범을 알렸고, 2024년 8월의 2.0은 하이브리드 본딩을 겨냥한 3D 패키징(UCIe-3D)과 칩렛을 가로지르는 관리·테스트 구조를 더했다. 2026년 10월 현재 공식 사양 페이지의 최신판은 2025년 8월 5일에 공개된 3.0이다. 3.0은 데이터 전송률 48 GT/s와 64 GT/s를 더해, 최고 속도가 2.0의 32 GT/s의 두 배가 되었다.

소프트웨어 비유로 이해하기

칩렛 개념	소프트웨어 대응
칩렛	마이크로서비스
UCIe	표준 API / RPC 프로토콜
인터포저	서비스 메시 / 네트워크
이종집적	폴리글랏 아키텍처 (서비스마다 최적 언어)
다이-투-다이 지연	네트워크 홉 지연

그리고 트레이드오프도 같다. 모놀리식은 내부 통신이 공짜지만 통째로 배포해야 하고, 분할하면 재사용·독립 개발이 가능하지만 경계를 넘는 비용이 든다.

왜 아직 어려운가

1. **테스트와 품질보증.** 칩렛 하나가 불량이면 이미 조립된 패키지 전체가 손실이다. 조립 전 완전한 KGD 보장이 필수인데, 미세한 고속 인터페이스를 조립 전에 완전히 테스트하기가 어렵다.
2. **책임 소재.** 여러 회사의 칩렛을 조합했는데 시스템이 안 돌면 누구 책임인가? 상용 생태계의 실질적 장벽이다.
3. **열·전력 통합 설계.** 칩렛마다 발열이 다르고, 전력 공급을 패키지 수준에서 설계해야 한다.
4. **물리적 호환성.** 같은 프로토콜을 써도 범프 피치·다이 두께·패키지 공정이 맞아야 한다. **표준이 프로토콜만 정해서는 부족하다.**
5. **EDA 도구.** 여러 다이를 함께 설계·검증하는 도구(멀티다이 STA, 열-전기 공동 시뮬레이션)가 아직 성숙 중이다.

현실: 2026년 현재 칩렛은 주로 **같은 회사 안에서** 쓰인다(사내 칩렛). 진정한 개방형 칩렛 시장은 아직 형성되지 않았다. 다만 표준화가 진행되며 방향은 명확하다. 2.0이 테스트·관리 구조를 표준에 넣은 것은 위 1번(테스트와 품질보증) 문제를 겨냥한 것이다.

---

9. 신호 무결성 — 통신 전공 배경이 있으면 익숙한 영역

배선이 수만 개로 늘고 속도가 오르면 전기 신호는 더 이상 "0과 1"이 아니라 **고주파 전송선 문제**가 된다.

현상	원인	대응
삽입손실 (insertion loss)	도체 표피효과, 유전체 손실. 주파수↑일수록 심함	저손실 재료, 등화(EQ)
반사 (reflection)	임피던스 불연속	임피던스 정합, 종단저항
크로스토크	인접 배선 간 용량·유도 결합	배선 간격, 접지 차폐, 배선 순서 최적화
동시 스위칭 잡음(SSN)	많은 출력이 동시에 전환되며 전원망이 흔들림	디커플링 커패시터, 전원망 설계
지터 (jitter)	클럭·데이터 타이밍 요동	CDR, 이퀄라이저
스큐 (skew)	배선 길이 차이로 도착 시각 어긋남	길이 매칭

평가 도구: 아이 다이어그램(eye diagram)은 받은 파형을 비트 하나 길이씩 잘라 겹쳐 그린 것이다. 가운데 열린 「눈」이 클수록 0과 1을 잘못 읽을 여유(마진)가 크고, 손실·반사·잡음·지터가 늘면 눈이 닫힌다. 대응에는 디지털 통신의 개념이 그대로 쓰인다. 위 표의 CDR(clock and data recovery)은 별도 클럭선 없이 받은 데이터 파형의 전환 시점에서 클럭 타이밍을 되찾는 회로다. 등화 가운데 CTLE(연속시간 선형 등화기)는 채널이 깎아 먹은 고주파 성분을 수신단의 아날로그 필터로 다시 키우고, DFE(판정 궤환 등화기)는 이미 0·1로 판정한 앞 비트들이 남긴 꼬리(심볼 간 간섭)를 추정해 빼 준다. 송신단에서 전환 순간의 신호를 미리 키워 보내는 프리엠퍼시스, 신호에 규칙을 심는 부호화도 같은 목적으로 쓴다.

> **여기서 3회차의 논리가 완성된다.** 왜 HBM은 주파수 대신 폭을 택했는가? 위 표의 모든 문제가 주파수에 따라 악화되기 때문이다. 폭을 늘리면 낮은 주파수로 같은 대역폭을 얻어, 신호 무결성과 전력 문제를 동시에 완화한다. **HBM은 신호 무결성 문제에 대한 패키징적 해답이다.**

---

10. 자주 하는 오해

오해 1: "후공정은 전공정보다 쉽다" → 하이브리드 본딩은 옹스트롬 단위 평탄도와 나노미터 정렬을 요구한다. 첨단 패키징은 전공정 수준의 정밀도를 요구한다.

오해 2: "2.5D도 3D의 일종이니 대충 같다" → 열 특성이 완전히 다르다. 이 차이가 설계 가능 범위를 결정한다.

오해 3: "칩렛으로 나누면 무조건 싸진다" → 인터페이스 회로의 면적·전력 오버헤드, 조립 비용, 조립 수율 손실이 있다. 다이가 작으면 나누는 게 손해다.

오해 4: "HBM은 메모리 회사가 다 만든다" → HBM 스택은 메모리 회사가 만들지만, GPU와 붙이는 2.5D 패키징은 파운드리(주로 TSMC)가 한다. **공급망이 갈라져 있고, 양쪽 다 병목일 수 있다.**

오해 5: "UCIe가 나왔으니 곧 칩렛을 사다 조립하는 시대가 온다" → 프로토콜 표준만으로는 부족하다.

테스트·품질보증·책임소재·물리적 호환성이 남아 있다.

오해 6: "TSV는 그냥 구멍에 구리를 채운 것" → 열팽창 불일치로 인한 응력이 주변 트랜지스터 특성을 바꿔서, 금지영역이라는 설계 제약을 만든다. 전기 문제가 아니라 기계 문제가 설계를 제약한다.

오해 7: "패키징 캐파는 팹보다 늘리기 쉽다" → 첨단 패키징 장비도 리드타임이 길고, 수율 램프업이 필요하다. 실제로 AI 칩 공급의 병목이 되어 왔다.

오해 8: "지금의 HBM은 하이브리드 본딩으로 쌓는다" → 2026년 10월 현재 양산 HBM과 제조사가 발표한 고객용 샘플은 TSV와 마이크로범프로 쌓고, 칩 사이 틈을 MR-MUF나 TC-NCF로 메운다. 하이브리드 본딩은 더 높은 단수를 위한 차세대 후보다(6절).

오해 9: "HTOL 1,000시간을 통과했으니 수명이 9년이다" → 그 환산은 활성화 에너지 E_a 하나에 달려 있고, E_a 는 고장 기구마다 다르다. 같은 조건에서 0.5 eV로 두면 약 2.6년, 0.9 eV로 두면 약 31년이다(7절 신뢰성).

---

11. 정량 실습

실습 1 — 칩렛 손익 완전판 $D_0 = 0.1/\text{cm}^2$, 총 면적 800 mm^2를 (a) 1개, (b) 2개, (c) 4개, (d) 8개로 나눌 때 각각 다음을 구하라. 비용은 웨이퍼 한 장 값을 W 로 두고 그 배수로 적는다(4회차와 같은 기호다. 전력 단위 와트 W와 다르다).

- 다이당 수율과 웨이퍼당 굿다이 수
- 인터페이스 오버헤드 (분할 개수당 면적 +5%)
- 조립 수율 (다이 하나 추가마다 ×0.99)
- 총 비용

을 계산해 최적 분할 수를 찾아라. D_0 를 0.05, 0.2로 바꾸면 최적점이 어떻게 이동하는가?

실습 2 — 배선 에너지 배선 길이 10mm(기판)와 50μm(TSV)의 기생 커패시턴스 비를 길이 비례로 가정하고, $E = \frac{1}{2}CV^2$ 로 비트당 에너지 비를 구하라. 초당 10^{12} 비트를 전송할 때 전력 차이는?

실습 3 — 인터포저 면적 GPU 다이 800 mm^2, HBM 스택 8개(각 약 100 mm^2), 스택 간 여유를 고려해 필요한 인터포저 면적을 추정하라. 레티클 한계의 몇 배인가? 어떤 CoWoS 변형이 필요한가?

실습 4 — 3D 열 문제 칩 하나의 전력밀도가 1 W/mm^2일 때 2층·4층으로 쌓으면 단위 풋프린트당 열유속은? 실리콘 두께 50μm를 통과하는 열저항을 고려해 층간 온도차를 개략 추정하라. (실리콘 열전도율 약 150 W/m·K)

실습 5 — 배선 수 세기 HBM4 스택 8개를 붙일 때 데이터 배선만 몇 개인가? 여기에 전원·접지·제어를 데이터의 2배로 가정하면 총 몇 개인가? HBM 스택을 GPU 다이의 마주 보는 두 변에 4개씩 붙이고(각 변 약 30 mm), 이 배선들이 그 두 변의 가장자리에서 한 줄로 나온다면 필요한 피치는?

---

12. 토론 질문

1. High-NA EUV의 필드 축소가 칩렛을 택할 유인을 키운다면, 미세화와 통합은 대립하는 전략인가 보완하는 전략인가?
2. 파운드리가 패키징을 흡수하는 흐름은 OSAT 산업을 어떻게 바꾸는가? OSAT에게 남는 영역은 무엇인가?
3. 개방형 칩렛 시장이 형성되려면 무엇이 먼저 해결되어야 하는가? 소프트웨어 생태계의 역사에서 배울 것이 있는가?
4. 3D 로직 적층의 열 문제는 냉각으로 풀 수 있는 문제인가, 아니면 설계로 회피해야 하는 문제인가?
5. 유리기판이 실리콘 인터포저를 대체할 수 있는가? 대체된다면 산업 구조에 어떤 변화가 생기는가?
6. AI 칩 공급의 병목이 로직 → CoWoS → HBM으로 옮겨온 것을 볼 때, 다음 병목은 어디일 것 같은가?

---

13. 후속 공부

- **표준 교과서:** Lau, *Semiconductor Advanced Packaging* (Springer, 2021) 및 동 저자의 3D IC·팬아웃 관련 저서 시리즈 — 이 분야의 가장 체계적인 자료군
- **국문 입문 연재:** SK하이닉스 뉴스룸 「반도체 후공정」 11편(2022-09~2023-09) — 1편 반도체 테스트의 이해부터 11편 반도체 패키지 신뢰성까지. 테스트·패키지 종류·컨벤셔널과 웨이퍼 레벨 공정·재료·신뢰성을 차례로 다룬다. 수치와 용어는 이 책의 1차 출처와 대조해 읽을 것
- **하이브리드 본딩:** imec·CEA-Leti의 공개 발표 자료, ECTC(Electronic Components and Technology Conference) 논문
- **업계 기술 자료:** TSMC 3DFabric(CoWoS/InFO/SoIC) 기술 소개, Intel 첨단 패키징(Foveros/EMIB) 소개, Siemens EDA의 3D IC 이종집적 가이드
- **칩렛 표준:** UCIe 사양서(1.0~3.0, 신청하면 받을 수 있다)
- **신뢰성·열:** JEDEC JESD47(인증 시험의 뼈대. 본문은 2012년 7월판 JESD47I를 기준으로 했고, 현행판은 JESD47N이다), JEDEC JESD51-12(열 정보 보고·사용 지침), Renesas TB379 「Thermal Characterization of Packaged Semiconductor Devices」 — θ와 ψ를 시험 보드 그림과 함께 설명한다
- **신호 무결성:** Bogatin, *Signal and Power Integrity — Simplified* — 직관적 서술로 정평. Hall & Heck, *Advanced Signal Integrity for High-Speed Digital Designs* (심화)
- **학회:** ECTC, IEEE 3D-IC Conference, Hot Chips(제품 수준 아키텍처 공개)

자습 부록 A — 정량 실습 풀이

실습 1 풀이 — 칩렛 손익 완전판

▾ 풀이

설정: 총 실리콘 면적 800 mm²를 n 개로 분할. 분할마다 인터페이스 면적 오버헤드 5%씩 누적(총 면적 $= 800 \times [1 + 0.05(n-1)]$), 조립 수율 0.99^{n-1}, $D_0 = 0.1/\text{cm}^2$. 비용은 웨이퍼 한 장 값 W 의 배수로 적은 양품 시스템 하나의 실리콘 원가다(2절 정확한 비교식에서 $c_{\text{assembly}} = 0$). 아래 표의 W는 와트가 아니다.

$$\text{GDPW}(A) = \frac{\pi d^2/4}{A} - \frac{\pi d}{\sqrt{2A}} \approx \frac{706.9}{A} - \frac{94.25}{\sqrt{2A}} \quad (d = 30\ \text{cm},\ A\ \text{in cm}^2)$$

표의 값은 반올림하지 않은 π 로 셈했다. 반올림한 계수 706.9·94.25로 셈하면 몇 칸에서 마지막 자리가 하나 달라진다(예: $n = 8$ 의 GDPW 466.3, $D_0 = 0.1$ 에서 $n = 1$ 의 비용 $0.03434\,W$). 결론은 바뀌지 않는다.

$D_0 = 0.1$ 일 때

n	칩렛 면적	GDPW	수율	굿다이	시스템/웨이퍼	조립수율	유효	비용
1	8.00 cm²	64.8	0.449	29.1	29.1	1.000	29.1	$0.03435\,W$
2	4.20	135.8	0.657	89.2	44.6	0.990	44.2	$0.02264\,W$
4	**2.30**	263.4	0.795	209.3	52.3	0.970	50.8	$0.01970\,W$
8	1.35	466.2	0.874	407.4	50.9	0.932	47.5	$0.02107\,W$

최적은 $n = 4$, 단일 다이 대비 42.6% 절감. $n = 8$ 에서 다시 나빠지는 것이 중요하다 — 인터페이스 오버헤드와 조립 수율 손실이 수율 이득을 잡아먹기 시작한다.

D_0 를 바꾸면 최적점이 이동한다 — 이것이 이 실습의 핵심 발견

n	$D_0 = 0.05$	$D_0 = 0.1$	$D_0 = 0.2$
1	$0.02302\,W$	$0.03435\,W$	$0.07644\,W$
2	0.01836	0.02264	0.03446
4	0.01756	0.01970	0.02479
8	0.01969	0.02107	0.02412
최적	$n = 4$	$n = 4$	$n = 8$
절감률	23.7%	42.6%	68.5%

이 표는 실습이 고른 네 값만 비교했다. 분할 수를 1부터 16까지 모든 정수로 넓혀 계산하면 최적은 $D_0 = 0.05$ 에서 $n = 3$($0.01754\,W$, 절감률 23.8%), $D_0 = 0.1$ 에서 $n = 4$(그대로. $n = 5$ 의 $0.01971\,W$ 와 거의 같다), $D_0 = 0.2$ 에서 $n = 6$($0.02364\,W$, 절감률 69.1%)이다. 결함이 많을수록 더

잘게 쪼갠다는 방향은 같다. 다만 최적점 근처에서는 곡선이 평평하다. $D_0 = 0.1$ 에서 $n = 4$~6 은 모두 최적과 2% 안에 든다.

세 가지 결론

1. **결함이 많을수록 잘게 쪼개는 것이 유리하다.** $D_0 = 0.2$ 에서는 최적 분할이 8로 이동하고 절감률이 68.5%에 이른다.
2. **결함이 적으면 칩렛의 이점이 사라진다.** $D_0 = 0.05$ 에서는 절감률이 23.7%로 줄고, 조립·인터포저 비용을 넣으면 역전될 수 있다.
3. **최적 분할 수는 존재한다.** 무한정 쪼개면 오버헤드가 이긴다. 실제 제품이 2~8개 칩렛 범위에 몰려 있는 이유가 이 곡선의 형태다.

함의 — 산업 전략으로 읽기

원가로 보면 칩렛은 **최신 노드의 수율이 나쁠 때 그 노드를 일찍 쓸 수 있게 해 주는 도구다.** 최신 노드에 먼저 진입하려는 쪽과 큰 다이가 필요한 AI 가속기가 칩렛을 일찍 쓴 것은 이 방향과 맞는다. 반대로 성숙 노드의 작은 칩에서는 수율 이득이 작아, 원가만 보면 칩렛이 손해이기 쉽다. 다만 이 실습이 보여 주는 것은 수율에서 오는 원가 동기 하나다. 레티클 한계를 넘는 면적(압력 ②), 블록마다 다른 공정을 쓰는 이종 집적(압력 ③), 설계 재사용도 칩렛을 쓰는 이유이므로(4회차 실습 4 풀이와 같은 범위), 이 계산만으로 칩렛이 어디에 쓰이고 어디에 안 쓰이는지를 정하지는 못한다.

> 5회차 실습 5와 합치면 그림이 완성된다. High-NA는 한 번에 찍을 수 있는 최대 다이를 429 mm^2로 줄인다. 800 mm^2 면적의 회로가 필요하면 스티칭과 분할 가운데 하나를 골라야 하고, 이 실습은 이 가정에서 **분할이 경제적으로도 유리하다는 것을** 보여 준다. 칩렛을 택할 유인이 그만큼 커진다.

아래 그림은 이 실습의 가정으로 $D_0 = 0.1$, $n = 4$ 일 때다. 화면본에서는 D_0, 분할 수, 오버헤드, 조립 수율을 바꿔 볼 수 있고, 4회차 실습 4의 가정(칩렛마다 오버헤드 10%, 조립 수율 0.95를 한 번)으로 바꿔 볼 수도 있다.



n	칩렛 면적	GDPW	수율	굿다이	시스템/웨이퍼	조립수율	유효	비용 [W]
1	8.00 cm²	64.8	0.449	29.1	29.1	1.000	29.1	0.03435
2	4.20 cm²	135.8	0.657	89.2	44.6	0.990	44.2	0.02264
4	**2.30 cm²**	**263.4**	**0.795**	**209.3**	**52.3**	**0.970**	**50.8**	**0.01970**
8	1.35 cm²	466.2	0.874	407.4	50.9	0.932	47.5	0.02107

그림 6-4. 800 mm² 실리콘을 n개로 나눌 때 시스템당 비용(웨이퍼 값의 배수). $D_0 = 0.1$, 분할마다 오버헤드 5%, 다이 하나 추가마다 조립 수율 0.99면 n = 4에서 0.01970으로 가장 싸고, 단일 다이보다 42.6% 싸다. — 화면본에서는 이 자리에서 값을 바꿔 볼 수 있다(화면본 6회차 실습 1 풀이).

실습 2 풀이 — 배선 에너지

▾ **풀이**

기생 커패시턴스를 길이에 비례한다고 두면

$$\frac{C_{\text{기판}}}{C_{\text{TSV}}} = \frac{10\ \text{mm}}{50\ \mu\text{m}} = 200\text{배}$$

$E = \frac{1}{2}CV^2$ 에서 같은 전압이면 **비트당 에너지도 200배**다.

초당 10^{12} 비트 전송 시

TSV 쪽 비트당 에너지를 0.05 pJ/bit로 두면

경로	PJ/BIT	1 TBIT/S에서의 전력
TSV (50 µm)	0.05	**0.05 W**
기판 배선 (10 mm)	10	**10 W**

차이 약 10 W. 이것만 보면 크지 않아 보인다. 그런데 실제 규모로 확장하면 이야기가 달라진다.

HBM 한 스택이 4 TB/s $= 3.2 \times 10^{13}$ bit/s 일 때

경로	전력
인터포저·TSV 수준	1.6 W
기판 배선 수준	**320 W**

스택 8개면 **12.8 W vs 2,560 W**다. 7절에서 본 것처럼 이런 패키지 하나의 소비전력은 킬로와트급이다(7절 계산 예도 1,000 W로 두었다). 메모리 배선에만 2.56 kW를 쓰면 패키지 전력 예산 전체를 넘으므로, 이 가정에서는 설계가 성립하지 않는다.

결론: HBM을 인터포저 위에 GPU 바로 옆에 붙이는 것은 편의가 아니라 **필수다.** 거리를 못 줄이면 그 대역폭은 전력 예산 안에서 낼 수 없다.

여기서 3회차와 이어진다. 3회차 실습 2는 "느린 핀을 많이 쓰는" 선택이 전력 때문이라고 했다. 그런데 핀을 많이 쓰려면 붙일 자리가 필요하고, 붙일 자리를 만들려면 인터포저가 필요하다. **대역폭 → 전력 → 거리 → 패키징**이 한 줄로 이어진다.

> **주의:** 200배는 단위 길이당 커패시턴스와 전압이 같다는 가정에서 나온 값이다. 실제 배선은 단면·절연체·기생 성분·드라이버 회로가 달라 이 비가 그대로 나오지 않는다. 한 방향의 차이만 짚으면, 기판 I/O는 신호 무결성 때문에 더 높은 전압 스윙을 쓰는 경우가 많아 V^2 항에서 차이가 더 벌어질 수 있다.

아래 그림은 기판 배선 10 mm 대 TSV 50 µm, TSV 쪽 0.05 pJ/bit, 초당 10^{12} 비트일 때다. 화면본에서는 두 길이와 비트당 에너지, 비트율(HBM 한 스택 4 TB/s 포함)을 바꿔 볼 수 있다.

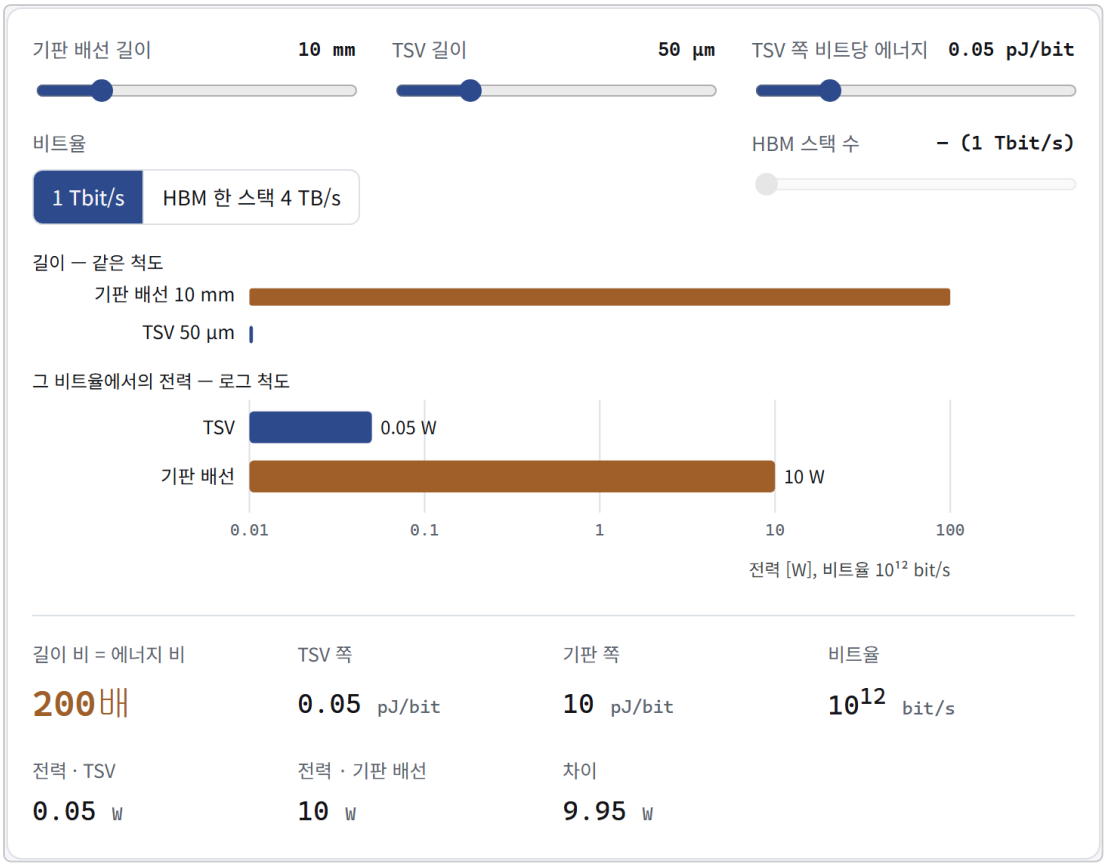


그림 6-5. 배선 10 mm와 TSV 50 µm는 길이 비 200배이므로 비트당 에너지도 200배다. TSV 쪽 0.05 pJ/bit이면 초당 10^{12} 비트에서 0.05 W 대 10 W다. — 화면본에서는 이 자리에서 값을 바꿔 볼 수 있다(화면본 6회차 실습 2 풀이).

실습 3 풀이 — 인터포저 면적

▾ 풀이

필요 면적 추정

구성요소	면적
GPU 다이	800 mm²
HBM 스택 8개 × 100 mm²	800 mm²
소계	1,600 mm²
스택 간 간격·몰딩·주변 배선 여유 (25%)	400 mm²
총계	**약 2,000 mm²**

레티클 한계 대비

$$\frac{2{,}000}{858} \approx 2.3\text{배}$$

단일 노광으로는 만들 수 없다. 인터포저 자체가 실리콘이고 실리콘 인터포저의 배선도 노광으로 만들기 때문에, 같은 858 mm² 제약을 받는다.

해법 세 가지

방식	내용	성격
스티칭	여러 필드를 이어 붙여 큰 인터포저를 만든다	정렬 정밀도와 결함이 문제
국소 브릿지 (CoWoS-L / EMIB 계열)	전면 실리콘 인터포저 대신 **연결이 필요한 곳에만** 작은 실리콘 브릿지를 넣는다. 브릿지를 둘러싼 층은 구조마다 다르다. CoWoS-L은 브릿지(LSI)를 RDL 인터포저 안에 넣고, 그 인터포저를 다시 패키지 기판에 올린다. Intel EMIB는 인터포저 없이 브릿지를 패키지 기판 안에 묻고, 두 다이의 맞닿은 가장자리끼리 잇는다	면적 제약을 우회. 현재 주류 방향
대면적 RDL 기판	유기 재배선층으로 전체를 덮는다	배선 밀도가 실리콘보다 낮다

실리콘 인터포저를 통째로 키우면 레티클 한계에 더해 수율 부담이 커진다. 얼마나 커질 수 있는지 가늠만 해 보자. 로직 다이의 $D_0 = 0.1/\text{cm}^2$를 2,000 mm^2(20 cm^2) 인터포저에 그대로 적용하면 4회차 식으로 $e^{-2.0} = 13.5\%$ 다. 배선만 있는 수동 인터포저의 실제 결함밀도는 로직 다이와 같다는 근거가 없으므로, 13.5%는 실제 제조 수율이 아니라 같은 결함밀도를 가정했을 때의 민감도 예시다. 방향은 **"전면을 실리콘으로"에서 "필요한 곳만 실리콘으로"다.** TSMC가 실리콘 인터포저(CoWoS-S)는 레티클 3.3배까지로 두고, 그보다 큰 패키지에는 RDL 기반 CoWoS-R과 국소 브릿지를 넣은 CoWoS-L(3.5배, 2024년 양산)을 내놓은 것도 이 방향이다.

AI 가속기 공급량이 로직 웨이퍼가 아니라 CoWoS 캐파에 먼저 걸리는 이유가 여기 있다 — 1회차 5절에서 예고한 그 병목의 실체다.

실습 4 풀이 — 3D 열 문제

▾ 풀이

단위 풋프린트당 열유속

전력밀도 1 W/mm^2인 칩을 쌓으면 그대로 곱해진다.

적층	풋프린트당 열유속
1층	1 W/mm^2
2층	2 W/mm^2
4층	**4 W/mm^2**

참고로 4 W/mm^2 = 400 W/cm^2다. 감을 잡자면, 지름 20 cm 화구에서 2 kW를 내는 전기레인지라고 두고 전력을 화구 면적으로 나누면 약 6.4 W/cm^2다(가정값으로 한 계산). 4층 적층의 풋프린트당 열유속은 그 60배를 넘는다.

실리콘 50 μm의 열저항 (1 mm^2 기준)

$$R = \frac{t}{kA} = \frac{50 \times 10^{-6}}{150 \times 10^{-6}} = 0.333\ \text{K/W}$$

4층 적층에서 층간 온도차 (위쪽으로만 방열, 각 층 1 W/mm² 발열)

아래층의 열은 위층들을 모두 통과해야 한다.

구간	통과 열량	ΔT
1층 → 2층	1 W	0.33 K
2층 → 3층	2 W	0.67 K
3층 → 4층	3 W	1.00 K
4층 → 방열면	4 W	1.33 K
합계		**3.33 K**

뜻밖의 결과: 실리콘만 보면 3.3 K에 불과하다. 실리콘은 열전도가 매우 좋은 재료다.

그렇다면 3D 적층의 열 문제는 어디서 오는가 — 계면이다

층 사이에는 실리콘만 있는 것이 아니다. 마이크로범프·언더필·접착층이 있고, 이들의 유효 열전도율은 약 1~5 W/(m·K)로 **실리콘의 1/50 수준**이다. 두께 20 μm, $k = 3$ W/(m·K)로 두면

$$R_{\text{계면}} = \frac{20 \times 10^{-6}}{3 \times 10^{-6}} = 6.67\ \text{K/W}$$

계면은 몇 곳인가. 4층이면 층과 층 사이가 세 곳이고, 맨 위 다이와 방열면 사이에도 접합층이 하나 있다. 여기서는 그 네 곳을 모두 같은 재료(20 μm, $k = 3$)로 둔다. 각 계면을 지나는 열량은 실리콘 구간과 같이 1·2·3·4 W이므로 $(1 + 2 + 3 + 4) \times 6.67 = 66.7$ K 다. 맨 위 접합을 TIM 같은 다른 재료로 따로 셈하고 내부 세 곳만 넣으면 $(1 + 2 + 3) \times 6.67 = 40$ K이고, 실리콘과 합하면 43.3 K다.

실리콘 3.3 K vs 계면 66.7 K — 20배 차이다. 둘을 더하면 맨 아래층과 방열면의 온도 차는 약 70 K다(네 계면 모형).

결론 두 가지

1. **3D 적층의 열 병목은 실리콘이 아니라 접합 계면이다.** 따라서 해법은 실리콘을 얇게 하는 것보다 계면의 열저항을 줄이는 쪽에 있다.
2. 하이브리드 본딩은 그 유력한 후보다. 범프와 언더필 층 대신 구리와 절연막을 직접 맞붙이므로 접합층이 훨씬 얇아진다. 하이브리드 본딩은 흔히 "피치를 줄이는 기술"로 소개되지만 **열 관점에서도 의미가 크다.** 다만 이 계산이 보여 주는 것은 거기까지다. 아래 그림에서 계면을 빼면 70 K가 3.3 K로 떨어지는데, 이것은 맨 위 방열 접합까지 네 계면의 열저항을 모두 0으로 둔 이상화다. 하이브리드 접합층에도 구리와 절연막이라는 두 접합면이 남고(6절), 그 유효 열전도율과 구리 면적 비율은 여기서 셈하지 않았다. 또 양산 HBM은 틈 메움 재료를 개선하는 길로도 열저항을 낮추고 있다(6절, SK하이닉스 HBM4E 12단 샘플의 약 17%). 실제 개선량을 말하려면 접합층의 재료값으로 다시 셈해야 한다.

> 이 열저항 사슬 논리가 7회차에서 그대로 확장된다. 칩에서 냉각수까지의 경로도 같은 구조이고, 거기서도 병목은 가장 열전도가 나쁜 층(TIM)이다.

아래 그림은 1 W/mm² 칩 4층, 실리콘 50 µm, 계면 20 µm·3 W/(m·K)일 때 구간마다 쌓이는 온도 차다. 화면본에서는 층 수, 전력밀도, 두께와 열전도율을 바꾸고 계면을 빼 볼 수 있다.

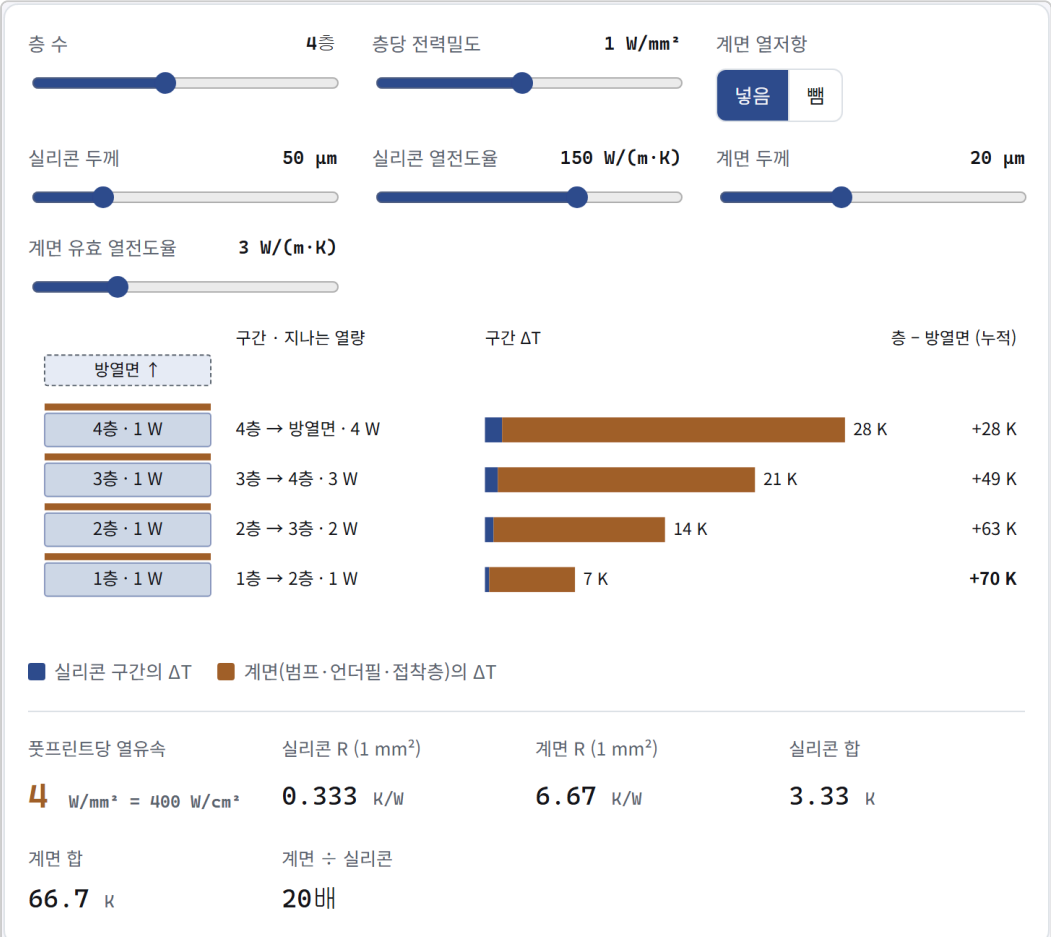


그림 6-6. 1 W/mm² 칩 4층을 쌓으면 실리콘(50 µm)에서 생기는 온도 차는 합 3.33 K, 범프·언더필 같은 계면(20 µm, 3 W/(m·K))에서는 66.7 K로 20배다. — 화면본에서는 이 자리에서 값을 바꿔 볼 수 있다(화면본 6회차 실습 4 풀이).

실습 5 풀이 — 배선 수 세기

▾ 풀이

데이터 배선

HBM4 스택 하나의 데이터 버스를 2,048비트로 두면

$$8\text{ 스택} \times 2{,}048 = 16{,}384\text{ 개}$$

전원·접지·제어를 데이터의 2배로 가정

$$16{,}384 \times 2 = 32{,}768\text{ 개}$$

$$\text{총계} = 16{,}384 + 32{,}768 = 49{,}152\text{ 개}$$

두 변(각 30 mm)의 가장자리에서 한 줄로 나온다면?

문제의 배치대로 마주 보는 두 변에 4스택씩 붙이면 한 변당 49,152 ÷ 2 = 24,576개다.

$$\text{피치} = \frac{30\text{ mm}}{24{,}576} = 1.22\ \mu\text{m}$$

(여덟 스택을 모두 한 변에 붙인다면 49,152개가 30 mm에 들어가야 하므로 0.61 μm다.)

지금의 연결 방식으로는 비현실적이다. GPU와 HBM을 인터포저에 붙이는 마이크로범프의 피치는 약 40 μm이고(6절), 1.22 μm는 그 33분의 1이다. 피치 숫자만 보면 하이브리드 본딩이 내세우는 10 μm 미만 영역에 들기는 하지만, 6절에서 본 것처럼 하이브리드 본딩은 2026년 10월 현재 HBM 연결의 양산 방식이 아니고, 그 피치도 면적 배열의 패드 피치이지 가장자리 한 줄을 전제한 값이 아니다.

그래서 실제로 하는 일 — 2차원 면적 배열

배선을 가장자리 한 줄이 아니라 **다이 아랫면 전체에 격자로** 깐다. 필요한 면적은

피치	필요 면적	GPU 다이(800 mm²) 대비
40 μm (마이크로범프)	$49{,}152 \times (0.04)^2 = 78.6\text{ mm}^2$	9.8%
9 μm (하이브리드 본딩)	4.0 mm^2	0.5%

두 가지 결론

1. **가장자리 배선(peripheral I/O)에서 면적 배열(area array)로의 전환은 선택이 아니었다.** 이 배선 수를 마이크로범프 피치로 한 줄에 뽑을 수는 없다. 플립칩·마이크로범프가 표준이 된 근본 이유가 이것이다.
2. **범프 피치가 대역폭의 상한을 누른다.** 이 실습의 가정(배선 수가 데이터의 3배)에서 40 µm 피치 면적 배열은 다이 면적의 9.8%를 쓴다. 배선 수가 몇 배로 늘면(HBM 세대 교체, 스택 수 증가) 이 비율도 같은 배수로 늘어 다이를 잠식한다. 피치를 줄이는 하이브리드 본딩은 그때의 유력한 경로다. 다만 실제 한계는 피치만이 아니라 인터포저 배선 밀도, HBM 인터페이스 규격, 접합 공정의 수율이 함께 정한다.

> 3회차 실습 2에서 "넓고 느리게"를 택한 대가가 여기 있다. 넓다는 것은 곧 배선 수이고, 배선 수는 곧 피치이고, 피치는 곧 본딩 기술이다. **메모리 대역폭 로드맵과 패키징 로드맵은 같은 곡선이다.**

자습 부록 B — 토론 질문 생각의 실마리

질문 유형	핵심 축	실마리
후공정이 왜 갑자기 중요해졌나	**미세화의 수익 체감**	전공정에서 얻는 이득이 줄고 비용이 급증하는 동안, 패키징에서 얻을 수 있는 이득(대역폭·통합)은 아직 초기 구간이다. 한계수익이 역전된 것이지 패키징이 갑자기 좋아진 것이 아니다
칩렛 표준(UCIe)은 성공할까	**누가 이득을 보는가**	표준화는 **후발주자와 소규모 설계사에 유리**하고 선두 통합업체에는 불리하다. 선두가 자체 인터페이스를 고수할 유인이 있으므로, 표준은 "모두가 쓰는 것"이 아니라 "선두가 아닌 진영의 연합"으로 자리 잡을 가능성이 높다. PC·서버 표준의 역사와 비교해 볼 것
OSAT vs 파운드리	**패키징이 공정인가 조립인가**	첨단 패키징이 노광·식각을 쓰는 순간 그것은 조립이 아니라 **전공정의 연장**이 된다. 그래서 파운드리가 이 영역을 흡수하고 있고, OSAT는 성숙 패키지 쪽으로 밀린다. 부가가치가 어디로 이동하는지의 문제
열이 3D 적층의 상한인가	**어느 계층의 열저항인가**	실습 4가 답을 준다 — 실리콘은 문제가 아니고 **계면**이 문제다. 하이브리드 본딩 등으로 계면의 열저항을 줄이면 상한이 올라간다. 얼마나 오르는지는 실제 접합층의 재료값으로 셈해야 한다. 다만 로직 위에 로직을 쌓는 경우(둘 다 고발열)는 여전히 어렵고, **발열이 적은 것을 위에 쌓는** 비대칭 설계가 현실적 답이다
공급 병목은 언제 풀리나	**장비와 학습곡선**	CoWoS 캐파는 자본으로 늘릴 수 있지만, 하이브리드 본딩 장비는 공급자가 소수이고 수율 학습이 필요하다. 3회차 실습 1의 논리가 여기에도 적용된다 — **캐파 문제와 학습 문제를 구분해야 예측이 가능하다**

자습 부록 C — 핵심 용어

용어	영문	한 줄 정의
후공정	back-end (assembly & test), packaging	웨이퍼를 잘라 칩을 패키지로 만들고 검사하는 단계. 웨이퍼 공정 안의 배선층 형성인 BEOL(5회차)과 다르다
다이싱	dicing	웨이퍼를 잘라 개별 다이로 분리하는 공정
플립칩	flip chip	다이를 뒤집어 범프로 기판에 직접 붙이는 방식
마이크로범프	microbump	다이 사이를 잇는 미세 솔더 돌기. 피치 약 40 µm
하이브리드 본딩	hybrid bonding	범프 없이 구리-구리를 직접 접합. 피치 10 µm 미만 (TSMC SoIC 기준). CIS·3D 로직에 쓰이고, HBM에서는 2026년 10월 현재 차세대 후보
MR-MUF / TC-NCF	mass reflow molded underfill / thermal compression non-conductive film	양산 HBM의 적층 방식. 다 쌓은 뒤 한꺼번에 리플로우하고 액체 보호재로 틈을 메우는 쪽과, 층마다 필름을 깔고 열압착하는 쪽
팬인 / 팬아웃	fan-in WLCSP / fan-out	볼이 칩 면적 안에만 있으면 팬인, 양품 다이를 재배열한 재구성 웨이퍼 위로 배선을 칩 바깥까지 펼치면 팬아웃 (예: TSMC InFO)
via-middle	via-middle TSV	트랜지스터를 만든 뒤, 배선층을 쌓기 전에 뚫는 TSV. 앞이면 via-first, 뒤면 via-last
WSS	wafer support system	얇게 갈 웨이퍼를 임시 접착제로 캐리어에 붙여 다루고, 뒷면 공정 뒤 떼어 내는(디본딩) 방식
KGSD	known good stacked die	스택 단위로 시험을 마친 HBM. 메모리 업체가 이 형태로 출하한다
CoW-first / CoS-first	chip-on-wafer first / chip-on-substrate first	2.5D 조립 순서. 인터포저 웨이퍼 위에 칩을 먼저 붙이느냐, 인터포저를 잘라 기판에 먼저 붙이느냐
W2W / D2W	wafer-to-wafer / die-to-wafer	웨이퍼끼리 붙이면 양품을 고를 수 없어 수율이 y^n 으로 곱해지고, 양품 다이만 골라 붙이면 본딩 수율만 남는다
θJA · θJC · ψJT	junction-to-ambient / junction-to-case thermal resistance, junction-to-top characterization parameter	접합 온도와 기준점 온도의 차를 전력으로 나눈 값. θJA는 표준 시험 조건의 비교값, ψ는 진짜 열저항이 아닌 특성 파라미터
HTOL	high temperature operating life	접합 온도 125 °C 이상에서 1,000시간 동작시키는 인증 시험(JESD47I)
가속계수	acceleration factor	시험 조건 1시간이 사용 조건 몇 시간에 해당하는지. 아레니우스 식에서는 E_a 에 크게 좌우된다
인터포저	interposer	여러 다이를 얹어 서로 연결하는 중간 기판
TSV	through-silicon via	실리콘을 수직 관통하는 배선
TGV	through-glass via	유리 기판을 관통하는 배선. 실리콘 대체 후보
2.5D	—	다이를 인터포저 위에 나란히 놓는 방식
3D 적층	3D stacking	다이를 수직으로 쌓는 방식
CoWoS	chip-on-wafer-on-substrate	TSMC의 2.5D 패키징 계열

용어	영문	한 줄 정의
EMIB / 브릿지	embedded bridge	필요한 구간만 작은 실리콘으로 잇는 방식
RDL	redistribution layer	패드 위치를 재배치하는 재배선층
이종집적	heterogeneous integration	서로 다른 공정·기능의 다이를 한 패키지에 통합
UCIe	universal chiplet interconnect express	칩렛 간 인터페이스 표준
KGD	known good die	조립 전 검증을 마친 정상 다이
언더필	underfill	범프 사이를 채워 기계적으로 보강하는 수지
TIM	thermal interface material	칩과 방열판 사이의 열전달 매개 물질
신호 무결성	signal integrity	신호가 왜곡·잡음 없이 전달되는 정도
워피지	warpage	열팽창 차이로 패키지가 휘는 현상

자습 부록 D — 자가진단

☐ 2D·2.5D·3D를 **연결 방식**으로 구분해 설명할 수 있다 → 자습 안내 ★

☐ 칩렛 최적 분할 수가 D_0 에 따라 이동한다는 것을 계산으로 보일 수 있다 → 실습 1 ★

☐ 무한정 쪼개면 왜 손해인지 두 가지 이유를 댈 수 있다 → 실습 1

☐ **HBM을 인터포저 위에 붙이는 것이 왜 필수인지** 전력으로 설명할 수 있다 → 실습 2 ★

☐ 인터포저 면적이 레티클 한계를 넘는 문제와 그 세 가지 해법을 말할 수 있다 → 실습 3

☐ 3D 적층의 열 병목이 **실리콘이 아니라 계면**임을 계산으로 보일 수 있다 → 실습 4 ★

☐ 하이브리드 본딩이 피치와 열 **둘 다에서** 후보가 되는 이유와, 이 책의 계산이 거기서 멈추는 지점을 설명할 수 있다 → 실습 4·5

☐ 면적 배열이 왜 불가피했는지 배선 수로 설명할 수 있다 → 실습 5

☐ "AI 가속기 공급이 CoWoS에 걸린다"는 말의 물리적 의미를 말할 수 있다 → 실습 3

☐ W2W와 D2W의 적층 수율이 층 수에 따라 얼마나 벌어지는지 계산할 수 있다 → 6절

☐ θJA와 ψJT의 차이, 열저항 직렬 사슬로 T_j 를 구하는 법을 말할 수 있다 → 7절

☐ HTOL 1,000시간의 등가 사용 기간이 E_a 에 따라 얼마나 흔들리는지 보일 수 있다 → 7절 신뢰성

다음 회차로 넘어가기 전에

여기까지가 **칩과 패키지**다. 7회차는 그것을 수만 개 모아 놓았을 때 벌어지는 일이다.

이 회차가 남긴 질문	7회차의 답
열저항 사슬은 패키지 밖에서 어떻게 이어지나	칩 → TIM → 콜드플레이트 → 냉각수 → 냉각탑
랙 하나에 120 kW가 들어가면 어떻게 식히나	실습 1 — 공랭은 왜 불가능한가
전력을 칩까지 어떻게 가져오나	실습 2·4 — 변환 손실과 IR drop
이 모든 것이 결국 무엇에 걸리나	전력·부지·계통 — **물리 세계의 상한**

출처

입문 자료 (주제별)

주제	링크	이 책에서의 활용
후공정 개념	SK하이닉스 뉴스룸 「반도체 후공정」 6편 컨벤셔널 패키지 공정 (연재 전체 11편: 1편 · 2편 · 3편 · 4편 · 5편 · 7편 · 8편 · 9편 · 10편 · 11편)	1절 후공정 개념의 기초. **국문 연재물로 접근성이 좋아 예습 자료로 권장.** 본문의 사실 서술은 되도록 아래 1차 출처에 기댔다. 웨이퍼 테스트 순서(1편), 패키지 층위(2편), WSS·디본딩·에지 트리밍·팬아웃 재구성 웨이퍼(8편)는 이 연재가 근거다
패키징기술동향	ScienceON — 김희주·정재필, 「인공지능 반도체 및 패키징 기술 동향」, 마이크로전자 및 패키징 학회지 30(3), 2023	국내 패키징 기술 동향 학술 리뷰
패키징기술동향	ScienceON — 정학순·권지민, 「AI 하드웨어를 위한 3D 패키징 고대역폭/대용량 메모리 기술 발전과 연구 동향」, 같은 학회지 33(2), 2026	최신 3D 패키징·메모리 기술 동향
다이싱	ScienceON 보고서 — 임무생, 「반도체 제조장치의 레이저-다이싱 머신」, ReSEAT 분석리포트, 2004	레이저 다이싱 장비 보고서. 3절 배경
TSV TGV 인터포저	ScienceON — 서일웅 외, 「TSV 인터포저 기술을 이용한 3D 패키지의 방열 해석」, 같은 학회지 21(2), 2014	TSV 인터포저 3D 패키지의 열 해석. 4·5절과 7절 열 논의의 배경
TSV TGV 인터포저	ScienceON — 선신희 외, 「AI 반도체 패키징을 위한 TGV 및 Cu 충전 연구의 최근 동향」, 같은 학회지 31(4), 2024	TGV·유리기판 관련 문헌
본딩	ScienceON — 주승환 외, 「HBM 반도체 초고집적 하이브리드본딩 스택장비 기술 동향」, 같은 학회지 32(1), 2025	HBM용 하이브리드 본딩 장비 문헌. 6절 배경

주제	링크	이 책에서의 활용
2.5D·3D 패키징과 HBM 인터포저	Ovaga — "Advanced Packaging & Heterogeneous Integration"	첨단 패키징·이종집적 영문 개관. 7·8절과 대응. 2026-10-08 확인 때 페이지가 자바스크립트 관문만 돌려줘 제목을 확인하지 못해 링크를 달지 않았다
2.5D·3D 패키징과 HBM 인터포저	Granite River Labs — 「DDR 메모리의 신호 무결성 및 프로토콜 테스트」	**9절 신호 무결성의 실무적 배경.** 측정 관점 자료
2.5D·3D 패키징과 HBM 인터포저	Siemens EDA — "Comprehensive 3D IC guide to heterogeneous integration" (2025.05.19)	**EDA 벤더의 3D IC 설계 가이드. 8절 이종집적과 설계 도구 문제의 직접 배경**

검증 완료 (2026-08-02)

- CoWoS 변형(CoWoS-S / -R / -L)의 구조적 차이 — TSMC 기술 자료 및 복수 산업분석 대조
- 하이브리드 본딩의 원리(구리-구리 직접 접합 + 절연막 접합, 어닐링 확산) — 표준 기술 사실
- 2.5D / 3D 구분 및 HBM 시스템이 "2.5D 안의 3D" 구조라는 점 — 표준 기술 사실
- TSV 공정 흐름 및 CTE 불일치로 인한 keep-out zone — 표준 기술 사실
- NVIDIA가 CoWoS 캐파의 상당 부분을 확보하고 있고, 첨단 패키징 캐파가 AI 가속기 공급의 제약이라는 구조 — 복수 산업분석 일치. **다만 구체적 캐파 수치·점유율은 자료별 편차가 커 본문에서 수치로 단정하지 않음**
- 노광 레티클 필드 한계와 High-NA 필드 축소 — 5회차와 동일 (2026-10-08 ASML 공식 자료로 확인, 아래)

검증 완료 (2026-10-08)

- 레티클 필드 — ASML TWINSCAN NXE:3600D 제품 페이지(NXE 플랫폼 최대 노광 필드 26 mm × 33 mm), ASML 「5 things you should know about High NA EUV lithography」(EXE의 노광 필드는 NXE의 절반). 858·429 mm^2는 계산값
- 패키지 흐름·기능 — Intel 「How Silicon Die Become Chip Packages」(assembly and test, 패키지가 열·외부로부터 보호하고 연결한다, TIM·리드 부착, 고전압·고온 시험). 패키지의 네 요구(상호연결·전력 공급·냉각·보호)는 Tummala·Rymaszewski·Klopfenstein (eds.), *Microelectronics Packaging Handbook: Semiconductor Packaging*, 2판(Chapman & Hall, 1997; 양장본 ISBN 978-0-412-08441-6)의 Springer 도서 페이지 소개문("Electronic packaging is defined as interconnection, powering, cooling, and protecting semiconductor chips for reliable systems.")으로 확인했다. 판(2판)·연도(양장본 1997-01-31 출간, 전자책은 2013-11-27에 따로 나옴)·편자·ISBN은 이 출판사 목록으로 확인했고, 책 본문은 대조하지 못했다
- EMIB 구조 — Intel Foundry 「Advanced Packaging Innovations | Chip Packages」(「Silicon bridge embedded in package substrate for shoreline-to-shoreline connection」, 2017년부터 양산)
- 인증 시험·아레니우스 — JEDEC JESD47I(2012-07) 표 1(HTOL·ELFR·HTSL), 표 2(PC·HTSL·THB·HAST·TC·UHAST·AC)와 주석, 표 1 주석 a(0.7 eV·125→55 °C에서 가속계수 78.6, 약 9년). 본문 표의 숫자는 이 판과 대조했다. 가속계수는 $k = 8.617 \times 10^{-5}$ eV/K로 다시 계산했다(77.66). JEDEC JESD47 페이지에 따르면 현재판은 JESD47N(2026-08)이다. **최신판 조건은 미대조**

- 열저항 정의 — TI 「One-Page Thermal Reference Sheet」 SLRR012(θJA는 JEDEC 51-2A 정의, 비교용이지 T_j 계산용이 아님, ψJT로 T_j 추정), Renesas TB379(시험 보드 1S0P·1S2P, 1S2P가 θJA를 대략 20~40% 이상 낮춤, ψ는 진짜 열저항이 아님). JEDEC JESD51-12 원문은 보지 못했다
- HBM 적층 방식 — SK하이닉스 「HBM4E 12단 샘플 공급」(2026-06-18)(어드밴스드 MR-MUF, 12단 48GB, 열저항 HBM4 대비 약 17% 개선), SK하이닉스 「어드밴스드 패키징을 견인하는 인터커넥션 기술의 가치와 SK하이닉스 패키징 기술 혁신」(2023-08-18)(TC NCF·MR-MUF 정의, CIS의 하이브리드 본딩 양산, 2022년 HBM2E 8단 하이브리드 본딩 시험). 삼성의 TC-NCF는 3회차 출처(삼성전자 2024-02-27 보도자료)에 있다. 마이크론의 방식과 하이브리드 본딩 도입 일정은 1차 출처가 없어 쓰지 않았다
- KGSD — SK하이닉스 Hot Chips 28 튜토리얼 「The Era of High Bandwidth Memory」(2016) (Known Good Stacked Die, molded KGSD)
- TSV 순서 — US 11,011,499 B2 명세서(via first/middle/last 정의, HBM·HMC가 via first·via middle의 대표 예). 메모리 TSV가 via-middle이라는 것은 이 출처로는 「via-first 또는 via-middle」까지만 확인된다
- CoW-first 순서 — Ho et al., 「2.5D through silicon interposer package fabrication by chip-on-wafer (CoW) approach」, EPTC 2014 초록. CoS-first의 휨은 열팽창계수 차이의 원리로 설명했고 분석기관 자료는 인용하지 않았다
- TSMC 3DFabric — 3DFabric, InFO(Integrated Fan-Out, 2016년부터 대량 출하, InFO_MS는 CoWoS의 비용·성능 대안), CoWoS(2012년 양산, -S 3.3배·~2,700 mm², -R 2023년 양산, -L 3.5배 2024년 양산, 생성형 AI 이후 수요 증가·캐파 확대), SoIC(결합 피치 10 μm 미만부터, 3 nm 칩 적층 2025년 양산, SoIC-X 웨이퍼 대 웨이퍼는 수율 높은 노드·같은 다이 크기에 적합) — 모두 2026-10-08 확인
- UCIe — UCIe 사양 페이지(1.0~3.0, 3.0의 48·64 GT/s, 2.0의 UCIe-3D·DFx), UCIe 보도자료 목록(2022-03-02 출범, 2024-08-06 2.0, 2025-08-05 3.0)
- 유리기판 — Intel 유리기판 발표(2023-09-18)(2020년대 후반 목표), Intel·Lens Technology 협력(2026-07-24)(개발 가속 협력)
- WSS 공정 순서 — SK하이닉스 뉴스룸 「반도체 후공정」 8편 원문과 대조했다(에지 트리밍은 캐리어 본딩 전, 캐리어 본딩 → 뒷면 연삭 → 뒷면 범프 → 캐리어 디본딩, 디본딩 방식 넷). 4절 문단은 이 내용을 이 책의 순서와 분류로 다시 정리한 것이다
- 계산 — W2W·D2W 수율 표(바닥 웨이퍼를 분모에 넣은 경우 포함), 열저항 사슬 예, 가속계수 표, 2절 수율 비교의 단위 변환, 실습 1의 GDPW(정밀 π 식과 반올림 계수 식), 실습 4의 계면 수(네 곳과 세 곳), 실습 5의 피치(두 변과 한 변)는 파이썬으로 다시 계산했다. 입력값(y, y_{bond}, 단계별 θ, P, T_a)은 모두 가정값이다

미확인 · 유보 (본문에서 단정하지 않음)

- **CoWoS 캐파 규모, 업체별 할당량** — 보도마다 편차가 크고 공식 확인이 어렵다. 구조적 사실(병목이라는 점)과 TSMC가 캐파를 늘리고 있다는 공식 서술만 쓴다
- **유리기판 상용화 시점** — Intel의 2020년대 후반 목표 외에 양산 일정을 밝힌 공식 발표를 2026년 10월 현재 확인하지 못했다
- **하이브리드 본딩의 HBM 양산 적용 시점** — 2026년 10월 현재 양산 HBM에 썼다는 제조사 발표를 찾지

못했다(3회차와 같다)

- **JESD47 최신판(JESD47N, 2026-08)의 시험 조건** — 본문 표는 2012년 JESD47I 기준이며 최신판과 대조하지 못했다
- **메모리 리페어의 퓨즈 방식, 양산 번인 조건** — 업체별 공정이라 1차 출처를 찾지 못해 본문에 구체적으로 쓰지 않았다
- **UCIe 실제 채택 현황** — 표준은 공개되었으나 이종 벤더 간 상용 조합 사례는 제한적이다. "방향은 명확하나 시장은 미형성"으로 서술
- **HBM 스택 다이 면적, GPU 다이 면적** — 11절 실습의 수치는 **계산 연습용 가정값**이며 특정 제품의 실제 치수가 아니다
- **실리콘 열전도율 150 W/m·K** — 온도·도핑에 따라 달라지는 근사값

7회차

AI 데이터센터

핵심 TAKEAWAY

칩 수준에서 시작한 열과 전력 문제가 건물 수준에서 최종 상한이 된다. AI 랙의 전력밀도가 전통적 데이터센터의 열 배로 뛰면서 냉각·배전·건물 구조가 전부 바뀌었고, 이제 AI 인프라의 제약은 칩 공급이 아니라 **전력 계통 접속과 부지**가 되어가고 있다. 데이터센터 효율 지표(PUE·WUE)는 측정 경계를 확인하지 않으면 비교가 무의미하다. **PUE는 IT 장비가 비효율적일수록 오히려 좋아 보이는 구조적 함정을 갖고 있다.** 마지막으로 이 회차는 1회차에서 던진 "병목은 어디인가"로 돌아가, 일곱 회차의 답을 하나로 모은다.

원자료 바로가기

IBM Think 「AI 데이터 센터」 · Bloom Energy 「Data Center Cooling for Hyperscale and AI Workloads」 · Bloom Energy 2026 Data Center Power Report: Mid-Year Pulse · Counterpoint 「Global DRAM and HBM Market Share」 · KISTEP 「AI반도체를 활용한 K-클라우드 기술개발사업」(ScienceON TRKO202500001876) · IEA, Key Questions on Energy and AI (2026-04) · LBNL 2024 미국 데이터센터 전력 보고서 발표 · Uptime Institute Global Data Center Survey 2025 · NVIDIA 800 VDC 기술 블로그

학습목표

1. AI 데이터센터가 기존 데이터센터와 무엇이 근본적으로 다른지 설명할 수 있다.
2. 계통에서 트랜지스터까지의 전력 체인을 설명하고, 각 단계의 손실을 이해한다.
3. 왜 특정 밀도를 넘으면 액체 냉각이 불가피한지 열전달로 설명할 수 있다.
4. PUE·WUE 등 효율 지표를 비판적으로 읽을 수 있다.
5. 랙 사양에서 전력·냉각·메모리 수요를 개략 추정할 수 있다.

0. 자습 안내 — 혼자 읽는 사람에게

이 회차는 모든 것이 착지하는 회차다

여섯 회차 동안 쌓은 도구가 여기서 하나의 시설로 합쳐진다. 특히 **실습 5는 일곱 회차의 내용을 한 문제에 담고 있으므로, 시간이 없다면 다른 실습을 건너뛰더라도 실습 5만은 반드시 풀 것.** 그것 하나가 이 책의 최종 점검이다.

앞 회차들에서 가져오는 것 — 이 회차는 전부를 쓴다

어디서	무엇을	어디에
2회차	디코딩 상한식	실습 6 (토큰당 전력 원가)
3회차	HBM 용량·스택 구조	실습 5 (c)
4회차	수율·공급 제약	6절 거시 수치
5회차	BSPDN(후면전력공급)	실습 4 (IR drop)
6회차	열저항 사슬 $R = t\,/\,(kA)$	실습 1·이 회차 3절

선행 지식 점검

개념	한 줄 답
비열 c_p	물질 1 kg의 온도를 1 K 올리는 데 드는 에너지. 공기 약 1,005 / 물 약 4,186 J/(kg·K)
질량유량 $\dot{m}$	초당 흐르는 유체의 질량 [kg/s]
$Q = \dot{m} c_p \Delta T$	유체가 단위 시간에 실어 나를 수 있는 열, 곧 열전달률 [W]. **이 회차에서 가장 많이 쓰는 식**
옴의 법칙 $V = IR$	실습 4에서 그대로 쓴다
kW와 kWh	순간 전력과 누적 전력량. 1 kW로 1시간 = 1 kWh
PUE	시설 총 전력 ÷ IT 장비 전력. 1.0이 이상값

읽는 순서 (총 135분 권장)

순서	절	시간	비고
1	1~2절 (데이터센터 구조, 전력 체인)	25분	실습 2를 여기서
2	**3절 냉각과 열저항 사슬**	25분	6회차 실습 4의 연장. 실습 1을 여기서
3	**4절 PUE와 그 함정**	20분	**실습 3을 꼭 풀 것.** 지표 비판의 훈련이다
4	5~7절 네트워크·거시 수치·메모리 시장	20분	1회차 4절의 수미상관
5	**실습 5**	**30분**	**최종 종합.** 종이 한 장에 전부 써 볼 것
6	실습 4·6 + 자가진단	15분	

이 회차를 관통하는 한 문장

> **모든 전력은 결국 열이 되고, 모든 열은 어딘가로 빠져나가야 한다.** 데이터센터의 상한은 연산 능력이 아니라 이 두 문장이 정한다.

막히기 쉬운 지점 — 「왜 갑자기 수냉인가」

랙 전력이 10 kW에서 120 kW로 올라간 것이 왜 냉각 방식의 교체를 강제하는지는, 물과 공기의 물성 차이 하나로 설명된다.

	밀도	비열	부피당 열용량
공기	1.18 kg/m³	1,005 J/(kg·K)	1,186 J/(m³·K)
물	997 kg/m³	4,186 J/(kg·K)	4,173,000 J/(m³·K)

$$\frac{4{,}173{,}000}{1{,}186} \approx 3{,}500\text{배}$$

같은 열을 나르는 데 물은 공기의 1/3,500 부피면 된다. 공기로 120 kW를 빼려면 초당 6.75 m³(작은 방 하나 분량, 약 14,300 CFM)를 랙 정면 1.2 m²로 밀어 넣어야 하고, 그러면 정면 평균 풍속이 약 5.6 m/s가 된다. 이 풍량은 3절에서 볼 최고 성능 바닥 타일 하나(1,900 CFM)의 약 7.5배다. 그만큼의 공기를 서버 안으로 밀어 넣는 팬 동력과 소음, 압력손실이 함께 커진다. 실습 1이 이것을 숫자로 확인한다.

> **한 줄 요약:** 수냉은 더 좋은 기술이라서 채택된 것이 아니라, **이 밀도에서는 공랭에 드는 풍량·팬 동력·소음이 실용 범위를 넘어서** 채택된 것이다.

---

1. 직관 잡기 — 밀도가 바뀌면 모든 것이 바뀐다

무엇이 달라졌나

전통적 데이터센터는 랙 하나에 수 킬로와트를 소비하는 서버들을 넣고 공기로 식혔다. AI 랙은 그보다 한 자릿수 이상 높은 밀도를 갖는다.

항목	전통적 데이터센터	AI 데이터센터
랙 전력밀도	5~10 kW급	**수십~100 kW 이상** (최신 시스템은 그 이상)
냉각	공랭 (CRAC/CRAH)	**액체 냉각 필수** (직접칩냉각·침지)
부하 특성	비교적 완만	**학습 잡의 동기화된 급변** — 수천 대가 동시에 부하를 올렸다 내린다
네트워크	남북 트래픽 중심	**동서 트래픽 폭발** (노드 간 집합통신)
배치 제약	지연시간(사용자 근접)	**전력 확보 가능 여부**
병목	부지·건물	**전력 계통 접속**

> **한 문장:** AI 데이터센터의 제약 조건은 "어디에 짓고 싶은가"가 아니라 **"어디서 전력을 끌어올 수 있는가"** 로 바뀌었다.

두 세계의 격차(2026년 10월 기준). 시설 전체로 보면 밀도는 아직 낮다. Uptime Institute의 2025년 조사(2025년 7월 발행)는 평균 랙 전력밀도가 천천히 오르고 있으며 그 동력은 10~30 kW 랙의 확산이고, 30 kW를 넘는 시설은 드물다고 적는다. 반면 AI 랙을 만드는 쪽은 그 열 배를 본다. NVIDIA는 2025년 5월 기술 블로그에서 랙이 200 kW를 넘으면 지금의 54 V 랙 배전이 물리적 한계에 닿는다고 쓰고, 2027년부터

1 MW 랙을 겨냥한 800 V 직류 배전으로 가겠다고 밝혔다(2절). 위 표의 「100 kW 이상」은 이 두 세계 가운데 AI 쪽을 가리킨다.

왜 밀도가 올라갔나 — 6회차에서 이어짐

이것은 임의의 변화가 아니라 앞 회차들의 직접적 귀결이다.

1. **연산·메모리를 가깝게 붙여야 성능이 난다** (2회차·3회차·6회차) → 하나의 패키지에 GPU와 HBM을 집적 → 패키지당 전력이 킬로와트급
2. **노드 간 통신이 병목이다** → 노드를 물리적으로 가깝게 배치해야 고대역 연결(NVLink 등)이 가능 → 랙 안에 밀집
3. **결과:** 고대역 연결로 한 묶음이 되어야 하는 GPU들(5절의 스케일업 도메인)은 흩어 놓으면 성능이 안 나오므로, 그 묶음 안에서는 밀도를 낮출 수 없다. **그 범위에서 밀도는 선택이 아니라 아키텍처의 요구다.** 묶음을 얼마나 크게 잡아 한 랙에 넣을지는 성능 목표에 따른 설계 분기로 남는다(실습 5 (d)의 랙당 8장 대 72장).

---

2. 전력 체인 — 계통에서 트랜지스터까지

```
[송전 계통 (수만~수십만 V)]
        ↓ 수전 변전소
[중압 배전 (수천~수만 V)]
        ↓ 변압기
[저압 배전 (수백 V)]
        ↓ UPS (무정전 전원) + 발전기 백업
[PDU / 버스바 (랙 배전)]
        ↓
[랙 전원 (직류 배전으로 가는 추세)]
        ↓ PSU
[보드 전원]
        ↓ VRM (전압 조정 모듈)
[칩 (1V 이하, 수백~수천 A)]
        ↓ 패키지 전력망 → 온다이 전력망
[트랜지스터]
```

각 단계에서 손실이 곱해진다

효율 η_i 인 단계가 여러 개면 전체 효율은 $\prod_i \eta_i$ 다. 각 단계가 97%여도 다섯 단계면 86%다. **손실은 전부 열이 되어 다시 식혀야 한다.** 그래서 배전 효율 개선이 냉각 부담 감소로 이중의 이득을 준다.

이것이 **고전압 직류(HVDC) 배전**과 **랙 수준 전원 통합**이 추진되는 이유다. 변환 단계를 줄이면 손실과 발열이 함께 준다.

800 V 직류 — 랙 하나가 메가와트가 될 때. 지금 AI 랙 안에서는 전원 선반이 직류를 만들고 구리 버스바가 그것을 컴퓨트 트레이로 나른다. 흔히 「48 V 배전」이라 부르는 방식이다(업계 표준 OCP Open Rack V3의 공칭 전압이 48 V다). NVIDIA 문서는 랙 안 버스 전압을 54 V로 적는다. 이 회차는 둘을 한데 묶어 48 V급 체계라 부른다. NVIDIA는 2025년 5월 기술 블로그에서 이 방식이 랙 200 kW를 넘으면 한계에 닿는다고 보고, **2027년부터 1 MW 이상의 IT 랙을 지원하는 800 V 직류(800 VDC) 배전으로** 가겠다고 밝혔다. 같은 글에 따르면 54 V로 1 MW 랙 하나를 먹이려면 구리 버스바만 최대 200 kg이 든다. 전압을 올리면 같은 전력에서 전류가 줄고, 저항 손실은 전류의 제곱으로 준다(실습 4 풀이에서 숫자로 본다).

같은 글은 종단 간 전력 효율 최대 5% 개선, 총소유비용 최대 30% 절감을 내세우는데, 이것은 **NVIDIA의 주장이고** 독립 검증된 값이 아니다.

마지막 1cm가 가장 어렵다 — 5회차와 연결

칩에는 **1V 이하의 전압으로 수백~수천 암페어의 전류를** 공급해야 한다. 전력 $P = VI$ 에서 V 가 작으면 I 가 커진다.

문제: 배선 저항 R 에서 전압 강하 $\Delta V = IR$ 이 생긴다(IR drop). 전류가 크면 이 강하가 무시할 수 없어져, 칩 중앙부의 트랜지스터가 규정 전압을 못 받는다. 또한 저항 손실 $P_{\text{loss}} = I^2R$ 도 I 의 제곱으로 커진다.

대응

- 온다이 전력망(PDN)에 많은 금속층을 할애 → 신호 배선과 자리 다툼
- 디커플링 커패시터로 순간 부하 변동 흡수
- **후면전력공급(BSPDN)** — 5회차에서 본 그것. 전력망을 웨이퍼 뒷면으로 옮겨 앞면 혼잡을 풀고 IR drop을 줄인다

> **★ 이 책의 순환이 여기서 닫힌다.** 데이터센터의 전력 문제가 트랜지스터 구조(BSPDN)를 바꾸고, 트랜지스터 구조 변화가 다시 공정 난도(웨이퍼 박막화·정렬)를 올린다. **이 회차와 5회차는 서로의 원인이자 결과다.**

---

3. 냉각의 물리 — 왜 공기로는 안 되는가

기본 식

냉각 유체가 단위 시간에 가져갈 수 있는 열, 곧 열전달률 Q [W]는

$$Q = \dot{m}\, c_p\, \Delta T$$

- $\dot{m}$: 질량유량 [kg/s]
- c_p : 비열 [J/(kg·K)]
- ΔT : 유체의 입출구 온도차

단위를 곱하면 kg/s × J/(kg·K) × K = J/s = W다. 이 회차의 Q 는 열에너지[J]가 아니라 열이 흐르는 속도다.

그리고 표면에서 유체로 넘어가는 열전달률은

$$Q = hA\Delta T_{\text{surface}}$$

- h : 대류 열전달계수 [W/(m²·K)]
- A : 열전달 면적

공기 vs 물

물성	공기	물	비
밀도 (kg/m³)	~1.2	~1000	**약 830배**
비열 c_p (J/kg·K)	~1005	~4180	약 4배
체적 열용량 ρc_p (kJ/m³·K)	~1.2	~4180	**약 3,500배**
열전도율 (W/m·K)	~0.026	~0.6	약 23배
대류 열전달계수 h (전형)	10~100	1,000~10,000	**10~100배**

같은 부피 흐름으로 물은 공기의 약 3,500배의 열을 나른다. 공기로 100 kW를 식히려면 비현실적인 풍량이 필요하고, 그 풍량을 만드는 팬 전력 자체가 막대해진다(팬 전력은 대략 풍량의 세제곱에 비례한다).

> **간단 실습:** 랙 하나 30 kW를 공랭으로 식히는데 입출구 온도차를 15K로 잡으면 필요한 공기 유량은? $\dot{V} = Q / (\rho c_p \Delta T) = 30000 / (1.2 \times 1005 \times 15) \approx 1.66\ \text{m}^3/\text{s} \approx 3{,}500$ CFM. 100 kW면 그 3배 이상이 한 랙에 필요하다. **랙 앞면 면적을 생각하면 풍속이 비현실적이 된다.**

냉각 방식의 스펙트럼

방식	적용 밀도(대략)	원리	장단점
공랭 (CRAC/CRAH)	대략 30 kW/랙 안팎 (조건에 따라 다르다)	냉기·열기 통로 분리, 팬 순환	단순·저렴. 밀도 한계 명확
후면도어 열교환기 (RDHx)	~40 kW/랙	랙 뒷문에 수냉 코일	기존 공랭 인프라와 병용 가능
직접칩냉각 (D2C, 콜드플레이트. **직접 액랭이라고도 한다**)	수십~175 kW/랙	칩 위에 냉각판을 붙여 냉각수 순환	**AI 랙의 사실상 표준.** CDU·누수 감지·배관 필요. 칩 외 부품은 여전히 공랭 필요
단상 침지냉각 (**액침냉각이라고도 한다**)	높음	유전성 액체에 서버를 담금	팬 불필요, 균일 냉각. 유지보수·유체 비용·호환성 문제
이상(비등) 침지냉각	매우 높음	액체가 끓으며 잠열로 흡열	열유속 대응 최고. 유체 규제(PFAS 등)·시스템 복잡

2026년 현황: 고밀도 AI 랙에서는 직접칩냉각이 사실상 표준이 되었고, 침지냉각은 특정 용도에 제한적으로 쓰인다.

공랭 한계를 한 숫자로 못 박지 않는 이유. 공랭의 상한은 급기 방식(이중마루 타일인가, 열 통로 봉쇄인가, 열 사이 냉각기인가)과 허용 온도차에 따라 달라서 자료마다 다르다. ASHRAE TC 9.9의 2021년 백서는 40~50 kW 랙이 분당 최대 5,000 ft³(CFM)의 공기를 먹을 수 있는데 최고 성능의 바닥 타일 하나가 1,900 CFM이라고 적는다. 위 간단 실습의 식에 공기 밀도 1.18 kg/m³를 넣어 다시 재면 40 kW를 $\Delta T = 15$ K로 빼는 데 약 4,760 CFM이 들어 이 값과 맞는다. 간단 실습의 30 kW(약 3,500 CFM)만 해도 타일 하나로는 모자란다. 백서는 이런 랙이면 통로 공간이 허락하는 타일 수로 풍량이 실제로 충분한지 따져 보라고 하고, 머리 위·열 사이 냉각이나 후면도어 열교환기 같은 보조 냉각을 대안으로 든다. 그래서 이 회차는 공랭 한계를 **바닥 타일 급기만으로는 대략 랙당 30 kW 안팎으로** 읽는다. 40~50 kW 랙은 타일 여러 장으로 필요한 풍량을 댈 수 없는 배치라면 보조 냉각이 있어야 한다. 위 1절의 Uptime 조사에서 현장 배치가 30

kW 아래에 몰려 있는 것은 물리적 한계가 아니라 배치 현황이지만, 방향은 같다.

찬물을 열원 대신 「열 버리는 곳」으로 — 수열 냉각. 냉각탑 대신 댐 심층수나 깊은 호수·바다의 찬물을 끌어와 열교환기로 냉각수를 식히는 방식도 있다. 원리는 같은 식 $Q = \dot{m} c_p \Delta T$ 다. 가정값으로 찬물 1 m^3/s(약 1,000 kg/s)를 끌어와 5 K만 데워 돌려보낸다고 하면 $1{,}000 \times 4{,}186 \times 5 \approx 20.9$ MW를 실어 나른다. 실습 5의 1만 장 클러스터 시설 전력 15.6 MW를 이 조건으로 빼려면 계산하면 약 745 kg/s가 필요하다. 수온이 연중 낮고 일정하면 냉동기 부하가 줄어든다. 대신 취수·방류 허가와 방류수 온도 상승이 생태에 주는 영향이 따라붙는다. 4절 함정 4(물과 전기의 교환)가 여기서도 그대로다.

6회차와의 연결 — 열은 칩 안에서부터 문제였다

3D 적층에서 아래 칩의 열이 위 칩을 통과해야 하는 문제(6회차 7절)가, 여기서는 칩의 열이 콜드플레이트를 거쳐 냉각수로, 다시 CDU와 냉각탑으로 가는 문제가 된다. **열저항이 직렬로 연결된 하나의 사슬**이며, 어느 한 곳의 열저항이 크면 전체가 막힌다.

$$\Delta T_{\text{총}} = Q \times (R_{\text{junction-case}} + R_{\text{TIM}} + R_{\text{coldplate}} + R_{\text{fluid}} + \cdots)$$

TIM(열계면물질) 이 의외로 중요한 이유가 여기 있다. 칩과 냉각판 사이의 미세한 틈이 사슬 전체의 병목이 될 수 있다.

---

4. 효율 지표를 비판적으로 읽기

정의

$$\text{PUE} = \frac{\text{데이터센터 총 전력}}{\text{IT 장비 전력}}$$

이상값은 1.0이고, 값이 클수록 냉각·배전 등 부대설비가 많이 쓴다는 뜻이다.

$$\text{WUE} = \frac{\text{연간 물 사용량 [L]}}{\text{IT 장비 에너지 [kWh]}}, \qquad \text{CUE} = \frac{\text{총 } CO_2 \text{ 배출량}}{\text{IT 에너지}}$$

PUE의 구조적 함정 — 이 절의 핵심

함정 1: 분모가 IT 전력이다. 부대설비에 IT 전력과 무관한 고정 부하가 있으면, 비효율적인 서버를 쓸수록 분모가 커져서 **PUE가 오히려 좋아진다**(실습 3). 극단적으로, 아무 일도 안 하면서 전력만 쓰는 서버를 넣으면 PUE가 개선된다. **PUE는 유용한 일의 양을 전혀 반영하지 않는다.**

함정 2: 측정 경계(boundary)가 자의적이다.

- 발전 손실을 포함하는가? 송전 손실은?
- 자가발전·현장 신재생을 어떻게 계산하는가?
- 사무 공간, 보안 설비는 포함되는가?
- IT 전력을 어느 지점에서 재는가(UPS 출력? PDU? 서버 입력?)

측정 지점이 다르면 같은 시설이 다른 PUE를 갖는다.

실제 숫자 둘을 나란히 놓으면 경계 문제가 보인다. Uptime Institute의 2025년 조사에서 응답 시설(681곳)의 연평균 PUE는 가중평균 1.54였고, 이 대표값은 6년째 거의 움직이지 않았다. 같은 조사에서 2020년 무렵 이후 준공한 시설만 보면 평균 1.48이다. Google은 2025년 자사 함대 평균 PUE를 1.09로 공개한다. 직전 12개월 이동평균(TTM)이고, 모든 오버헤드를 포함하며, 대상은 안정 가동에 들어간 대형 시설이라고 경계를 밝혀 둔다. 둘의 차이에는 실제 효율 차이도 들어 있지만, 한쪽은 설문 응답의 가중평균이고 다른 쪽은 한 회사가 스스로 정한 경계 안의 실측이다. 표본도 경계도 다른 두 값을 빼서 「Google이 얼마나 더 효율적인가」로 읽기 전에 그 차이부터 확인해야 한다. 참고로 실습 5의 IT 13 MW 클러스터에 두 값을 넣으면 시설 전력은 1.54일 때 20.02 MW, 1.09일 때 14.17 MW다(계산하면).

함정 3: 시점과 조건. 설계값(design PUE)인가 실측 연평균인가? 어느 계절인가? 부하율이 얼마일 때인가? **부분 부하에서는 PUE가 크게 나빠진다.**

함정 4: 물과 전기의 교환. 증발식 냉각을 쓰면 전력은 적게 쓰지만 물을 많이 쓴다. PUE만 보면 좋아 보이지만 WUE는 나빠진다. **하나의 지표를 최적화하면 다른 자원으로 비용이 이전된다.**

> **통계 전공자를 위한 프레임:** PUE는 전형적인 **굿하트의 법칙(Goodhart's Law)** 사례다 — 지표가 목표가 되면 지표로서의 기능을 잃는다. 그리고 측정 경계의 자의성은 **비교가능성(comparability)의 문제**다. 서로 다른 경계로 측정된 값을 비교하는 것은 통계적으로 무의미하다.

더 나은 지표는?

- **작업당 에너지** (예: 학습 잡 1회당 kWh, 추론 토큰 100만 개당 kWh) — 유용한 일을 분모에 둔다
- **탄소 집약도 시간대별 매칭** — 연간 총량 상쇄가 아니라 시간대별로 청정 전력과 매칭했는가
- **한계 배출량** — 이 데이터센터가 추가로 유발한 발전이 무엇이었는가 (평균이 아니라 한계)

---

5. 네트워크 — 스케일업과 스케일아웃

두 축

	스케일업 (SCALE-UP)	스케일아웃 (SCALE-OUT)
범위	노드/랙 내부	노드 간, 랙 간
기술	NVLink, 독자 인터커넥트	InfiniBand, 이더넷(RoCE)
대역폭	매우 높음	상대적으로 낮음
지연	매우 낮음	높음
확장 한계	물리적 근접성 (수십 개 GPU)	수만 개

병렬화 전략과의 대응 (AI 전공자에게 익숙한 부분)

- **텐서 병렬(TP):** 레이어 하나를 여러 GPU에 쪼갬. 매 레이어마다 all-reduce가 필요 → **통신량이 엄청나므로 반드시 스케일업 도메인 안에서**
- **파이프라인 병렬(PP):** 레이어를 그룹으로 나눠 다른 GPU에 배치. 통신량이 적음 → 스케일아웃 가능. 대신 버블(유휴) 발생
- **데이터 병렬(DP):** 배치를 나눔. 스텝 끝에 그래디언트 all-reduce → 통신을 계산과 겹칠 수 있음
- **전문가 병렬(EP, MoE):** all-to-all 통신 패턴 → 통신 패턴이 불규칙해 네트워크 설계가 까다로움

> **왜 이것이 데이터센터 설계를 결정하는가:** 텐서 병렬을 쓰려면 GPU들이 물리적으로 가까워야 한다. 그래서 랙 안에 최대한 밀집시키고, 그러면 랙 전력밀도가 올라가고, 그러면 액체 냉각이 필요해진다. **모델 병렬화 전략이 건물 설계를 결정하는 인과 사슬이다.**

집합통신이 병목이 되는 지점

N 개 GPU의 ring all-reduce에서 각 GPU가 주고받는 데이터량은 대략 $2\frac{N-1}{N}\times$ (파라미터 크기)로, N 이 커져도 수렴한다. 문제는 **지연**이 홉 수에 비례해 늘어난다는 것이다. 그래서 대규모에서는 계층적 all-reduce(랙 내부 먼저, 그다음 랙 간)를 쓴다.

광연결 (CPO — Co-Packaged Optics)

전기 신호는 거리가 멀어질수록 손실과 전력이 급증한다(6회차 9절). 광은 거리에 따른 손실이 훨씬 적다.

CPO의 발상: 광 트랜시버를 스위치 칩과 **같은 패키지에** 넣어, 전기 신호가 가는 거리를 최소화한다. 전력을 크게 줄일 수 있다.

과제: 광소자의 온도 민감성, 레이저 신뢰성, 수리 가능성(패키지에 들어가면 개별 교체 불가), 비용. **6회차의 패키징 기술이 네트워크 영역으로 확장되는 지점이다.**

상용화는 어디까지 왔나(2026년 10월 기준, 제조사 발표로만). Broadcom은 2025년 10월 8일 102.4 Tb/s 이더넷 스위치 칩에 광엔진을 함께 패키징한 3세대 CPO 스위치(Tomahawk 6 – Davisson)를 출하하기 시작했다고 발표했다. 광 연결 전력을 기존 착탈식(플러거블) 모듈보다 70% 줄인다는 것은 회사의 주장이다. NVIDIA는 2025년 3월 InfiniBand용 Quantum-X Photonics와 이더넷용 Spectrum-X Photonics CPO 스위치를 발표하며 전자는 2025년 안, 후자는 2026년 출시를 예고했다. 실제 출하가 시작됐는지는 2026년 10월 현재 NVIDIA 공식 발표로 확인하지 못했다. 두 회사의 발표는 모두 스위치에 광엔진을 붙이는 제품이다.

---

6. 거시 수치와 그 불확실성

왜 숫자가 이렇게 갈리는가

데이터센터 전력 소비 추정치는 기관·연도·정의에 따라 큰 편차를 보인다. 기관 전망 하나를 기준점으로 잡으면, 국제에너지기구(IEA)가 2026년 4월 낸 *Key Questions on Energy and AI* 는 전 세계 데이터센터 전력 소비가 2025년 485 TWh에서 2030년 약 950 TWh로 거의 두 배가 되어 2030년 세계 전력 수요의 약 3%를 차지한다고 본다(기준 시나리오인 Base Case). 2025년 한 해 증가율은 17%였다. 미국만 떼어 보면 로렌스버클리국립연구소(LBNL)가 미국 에너지부 지원으로 2024년 12월 낸 보고서가 2023년 176 TWh(미국 전력의 4.4%), 2028년 325~580 TWh(6.7~12%)로 추정한다. 2028년 값이 두 배 가까운 폭의 범위로 나온다는 것 자체가 불확실성의 크기다. 1회차 6절 오해 6이 같은 수치로 효율과 총량을 구분한다. 규모 감각을 위해 하나만 덧붙이면, 1 GW 발전소가 이용률 90%로 한 해 내는 전력량은 계산하면 약 7.9 TWh다. 그런데 민간 조사기관 수치는 이와 상당히 다르다.

차이의 원인

1. **암호화폐 채굴 포함 여부** — 이것만으로 수십~수백 TWh가 오간다
2. **범위** — 상업용 코로케이션만인가, 기업 사내 서버룸까지인가, 통신사 시설은?
3. **AI 전용 vs 전체** — "AI 데이터센터"를 어떻게 구분하는가. 혼합 워크로드 시설은?
4. **추정 방법** — 출하된 가속기 수 × 정격전력 × 가동률로 추정하는 방식은 가정에 매우 민감하다
5. **발표 시점** — 전망치가 반기마다 크게 수정된다. 수정이 없어도 인용되는 숫자는 바뀐다. IEA의 2025년판 *Energy and AI* 는 2024년 약 415 TWh에서 2030년 약 945 TWh로 가는 전망을 냈다. 1년 뒤 판도 본문에서는 기준 시나리오를 「2024년 약 415 TWh에서 2030년 약 950 TWh」로 다시 적는데, 요약에서는 2025년 실적 485 TWh를 출발점으로 삼아 「485 TWh에서 950 TWh로 거의 두 배」라고 쓴다(부록 표는 2024년 416, 2025년 485, 2030년 945 TWh). IEA 스스로 중심 전망이 2025년판 궤적과 가깝다고 밝힌다. 바뀐 것은 2030년 전망이 아니라 어느 해를 출발점으로 인용하느냐와 반올림이다. 숫자 둘을 나란히 놓고 「전망이 올랐다」고 읽기 전에 기준 연도부터 맞춰야 한다
6. **발표 주체의 이해관계** — 규제 논의, 투자 유치, 환경 캠페인 각각의 유인이 다르다

> **1회차의 교훈이 반복된다.** 이 책에서 확정할 것은 **추세와 구조**이지 특정 수치가 아니다. 확정 가능한 것: ① 데이터센터 전력 소비가 빠르게 증가하고 있다, ② 그 증가의 주 동인이 AI 워크로드다, ③ 랙 전력밀도가 급격히 상승하고 있다, ④ 그 결과 액체 냉각이 표준이 되었다.

전력 계통이라는 새 병목

칩과 패키징 캐파가 확보되어도, **전력 계통 접속 대기열**이 병목이 된다. 신규 송전선 건설과 계통 연계 승인에는 수년이 걸린다. IEA는 2026년 4월 보고서에서 계통 접속 대기가 많은 관할에서 5~10년에 이를 수 있다고 적고, 미국에서는 접속이 늦어 현장 가스 발전을 붙인 데이터센터 사업이 늘고 있다고 쓴다. 그래서:

- 발전소 인근 부지 선호
- 자체 발전(가스터빈, 연료전지) 검토
- 원자력(기존 원전 전력 구매계약, SMR 장기 계획)에 대한 관심

- 유연부하(demand response) — 학습 워크로드는 지연에 덜 민감하므로 계통 여유에 맞춰 조절할 수 있다는 아이디어

이 지점의 경제·정책·법적 쟁점은 이 책에서 다루지 않는다.

---

7. 다시 메모리 시장으로 — 마지막 고리

이 책이 마지막 회차에서 메모리 시장으로 돌아오는 것은 순환이 여기서 닫히기 때문이다.

인과 사슬

```
AI 데이터센터 증설
      ↓ 가속기 수요 증가
가속기 하나당 HBM 스택 여러 개 필요
      ↓
HBM 수요 폭증
      ↓ HBM은 일반 DRAM보다 다이 면적을 훨씬 많이 소모
      │  (같은 비트 수를 만드는 데 필요한 웨이퍼가 더 많다)
일반 DRAM 공급 구축(crowding-out)
      ↓
메모리 전반의 가격 상승
      ↓
다시 데이터센터 구축 원가 상승
```

핵심 메커니즘: HBM 생산은 같은 웨이퍼 캐파에서 일반 DRAM보다 훨씬 적은 비트를 만들어낸다(TSV 면적, 적층 수율 손실, 더 큰 다이). 즉 **HBM으로 캐파를 돌리면 전체 DRAM 비트 공급이 줄어든다.** 이것이 AI 수요가 메모리 시장 전반에 파급되는 경로다.

마지막 토론 — 사이클은 반복되는가

"이번엔 다르다" 쪽 논거

- 추론 수요는 학습과 달리 **지속적·운영비적**이다. 한 번 배포된 서비스는 계속 돌아간다
- HBM은 고객과 장기 계약으로 물량을 확정하는 구조라 전통적 현물 시장과 다르다
- GPU 임대 쪽도 장기 약정으로 굳어 있다. GPU 클라우드 업체 CoreWeave의 2025 회계연도 연차보고서(10-K)를 보면, 약정 계약은 고객이 쓰든 안 쓰든 정해진 용량 값을 치르는 take-or-pay 방식이고, 약정 계약이 매출에서 차지한 비중은 2023년 88%, 2024년 96%, 2025년 98% 이상이었다. 가중평균 계약기간은 약 5년이고, 회사는 이 약정을 담보로 자산 단위 부채를 일으켜 설비를 산다
- 진입 장벽이 훨씬 높아졌다(첨단 패키징, 적층 기술)

"결국 같다" 쪽 논거

- 증설 리드타임(2~3년)과 수요 신호의 시차라는 구조는 그대로다
- 3사가 모두 HBM 캐파를 크게 늘리고 있다
- AI 자본지출 자체가 순환적일 수 있다. 그리고 그 계획은 몇 달 사이에도 움직인다. Meta는 2026년 1월 28일 실적 발표에서 2026년 자본지출(금융리스 원금 상환 포함)을 1,150억~1,350억 달러로 제시했고, 7월 29일 발표에서는 1,300억~1,450억 달러로 바꿨다(7월 발표는 이를 직전 전망치 1,250억~1,450억 달러를 좁힌 것이라고 적는다). 범위 중간값으로 계산하면 반년 사이 10% 올랐다. 올라갈 때 이만큼 빨리 움직이는 숫자는 내려갈 때도 빨리 움직일 수 있다
- 과거에도 매번 "이번엔 구조적 수요"라는 주장이 있었다

모임에서 읽는다면: 양쪽으로 팀을 나눠 논쟁시키면 일곱 회차의 내용이 자연스럽게 총동원된다. 수율(4회차), HBM 원가 구조(3회차), 패키징 캐파(6회차), 전력 제약(이 회차)이 모두 논거가 된다.

---

8. 자주 하는 오해

오해 1: "PUE가 낮으면 친환경적이다" → PUE는 부대설비 효율만 본다. 유용한 일의 양도, 전력의 출처(석탄인지 재생인지)도, 물 사용도 반영하지 않는다.

오해 2: "액체 냉각은 위험하고 특수한 기술이다" → 고밀도 AI 랙에서는 이미 표준이다. 메인프레임 시대에도 수냉이 쓰였다.

오해 3: "데이터센터를 추운 지역에 지으면 냉각이 공짜다" → 외기 냉방(free cooling)은 도움이 되지만, 고밀도 랙에서는 칩에서 냉각수까지의 열저항이 지배적이라 외기 온도만으로 해결되지 않는다. 또한 전력 계통·네트워크 지연·인력 문제가 있다.

오해 4: "전력만 확보하면 데이터센터를 지을 수 있다" → 계통 접속 승인, 부지 인허가, 물 공급, 지역 수용성이 모두 제약이다.

오해 5: "AI 데이터센터가 전력망을 붕괴시킨다 / 미미한 영향이다" → 둘 다 과장이다. 총량 기준으로는 전체 전력의 일부지만, **지역적으로 집중되면 그 지역 계통에는 큰 충격**이다. 총량과 국지적 영향을 구분해야 한다.

오해 6: "HBM이 잘 팔리면 메모리 회사는 다 좋다" → HBM 캐파 전환은 일반 DRAM 공급을 줄인다. 시장 전체 구조가 바뀌며, 승자와 패자가 갈린다.

---

9. 정량 실습

실습 1 — 랙 사양에서 냉각 요구 도출 랙 하나가 120 kW를 소비할 때: (a) 공랭으로 $\Delta T = 15$K를 유지하려면 필요한 풍량은? 랙 정면 면적을 0.6×2m로 가정할 때 풍속은? 현실적인가? (b) 수냉으로 $\Delta T = 10$K를 유지하려면 필요한 냉각수 유량은? (L/min)

실습 2 — 전력 체인 손실 변압(99%) → UPS(96%) → PDU(99%) → PSU(94%) → VRM(90%)일 때 계통에서 칩까지의 종합 효율은? IT 전력 1 MW를 공급하려면 계통에서 얼마를 받아야 하는가? 손실은 모두 열이 되므로 냉각 부하에 어떻게 더해지는가?

실습 3 — PUE 계산과 조작 IT 전력 10 MW, 냉각 3 MW, 배전 손실 1 MW인 시설의 PUE는? 이제 IT 장비를 효율화해서 같은 일을 8 MW로 한다면 PUE는 어떻게 되는가? 두 경우를 각각 풀어라. (가) 냉각과 배전 손실이 IT 전력에 완전히 비례해 준다. (나) 냉각 3 MW 가운데 1 MW, 배전 손실 1 MW 가운데 0.2 MW는 IT 전력과 무관한 고정 부하이고 나머지만 비례해 준다. **일은 같은데 PUE가 그대로이거나 오히려 나빠지는 것을 확인하라.**

실습 4 — IR drop 칩에 1V, 500A를 공급할 때, 전력망의 등가 저항이 0.2 mΩ이면 전압 강하는? 코어에 실제로 걸리는 전압은? 이 강하를 절반으로 줄이려면 무엇을 해야 하는가?

실습 5 — 데이터센터 규모 추정 (종합) 가속기 10,000장 규모의 학습 클러스터를 가정하고 다음을 추정하라. (a) 총 IT 전력 (가속기당 1 kW + 부대 30% 가정) (b) PUE 1.2일 때 시설 총 전력과 연간 전력량 (c) 필요한 HBM 총 용량과 스택 개수 (d) 랙당 8장씩 넣을 때 랙 수와 랙 전력밀도, 필요한 냉각 방식 **이 문제 하나에 일곱 회차의 내용이 전부 들어온다.**

실습 6 — 추론 서비스 원가 2회차의 디코딩 상한식과 위 전력 수치를 결합해, 토큰 100만 개 생성의 전력 원가를 개략 추정하라. 배치 크기를 바꾸면 어떻게 달라지는가?

---

10. 토론 질문

1. AI 데이터센터의 진짜 병목은 칩인가, 패키징인가, 메모리인가, 전력인가? 시간이 지나며 어떻게 이동해 왔고, 다음은 어디인가?
2. PUE를 대체할 지표를 설계한다면 무엇을 분자·분모에 두겠는가? 조작 가능성을 어떻게 막겠는가?
3. 학습 워크로드의 유연부하 특성을 계통 안정화에 활용할 수 있는가? 어떤 조건이 필요한가?
4. 밀도가 계속 올라가면 냉각의 물리적 한계는 어디인가? 그 한계에 도달하면 아키텍처는 어떻게 바뀌어야 하는가?
5. 메모리 사이클은 이번에 다른가? (7절 참조 — 양쪽 논거를 모두 세운 뒤 판단할 것)
6. **총정리:** 1회차에서 그린 밸류체인 지도에 일곱 회차에서 배운 병목들을 표시해 보라. 어느 지점에 자원을 투입하는 것이 가장 효과적인가?

---

11. 후속 공부

- **데이터센터 인프라:** ASHRAE 데이터센터 자료(Datacom 시리즈) — 온습도 기준과 냉각 설계의 산업 표준. 액체 냉각 전환의 배경은 TC 9.9 백서 「Emergence and Expansion of Liquid Cooling in Mainstream Data Centers」(2021)
- **효율 지표:** The Green Grid의 PUE·WUE 백서 (지표를 만든 주체의 정의 문서. 예: 「PUE: A Comprehensive Examination of the Metric」). Uptime Institute의 연례 데이터센터 조사 (2025년판)
- **에너지 전망:** IEA, *Energy and AI* (2025)와 그 후속판 *Key Questions on Energy and AI* (2026-04) — 가장 널리 인용되는 기관 전망. **방법론 장을 반드시 함께 읽을 것.** LBNL(로렌스버클리국립연구소)의 미국 데이터센터 에너지 보고서(2024-12)는 방법론이 투명해 참고 가치가 높다
- **네트워크·시스템:** Barroso, Hölzle & Ranganathan, *The Datacenter as a Computer* (3rd ed.) — **웨어하우스 규모 컴퓨팅의 표준 교재. 전산 배경 독자에게 특히 권한다**
- **분산학습 통신:** Megatron-LM(arXiv:1909.08053), DeepSpeed ZeRO(arXiv:1910.02054) 논문의 병렬화 전략 절
- **광연결:** OFC(Optical Fiber Communication Conference)의 CPO 관련 발표

자습 부록 A — 정량 실습 풀이

실습 1 풀이 — 랙 사양에서 냉각 요구 도출

▾ 풀이

(a) 공랭, 120 kW, $\Delta T = 15$ K

$$\dot{m} = \frac{Q}{c_p \Delta T} = \frac{120{,}000}{1{,}005 \times 15} = 7.96 \text{ kg/s}$$

공기 밀도 1.18 kg/m³ (약 25 °C)로 환산하면

$$\dot{V} = \frac{7.96}{1.18} = 6.75 \text{ m}^3/\text{s} = 405 \text{ m}^3/\text{min} \approx 14{,}300 \text{ CFM}$$

랙 정면 $0.6 \times 2 = 1.2$ m²에서의 풍속:

$$v = \frac{6.75}{1.2} = 5.6 \text{ m/s}$$

현실적인가? 아니다. 세 가지 이유가 동시에 걸린다.

문제	내용
소음	팬 소음은 대략 풍속의 5~6제곱에 비례한다. 같은 팬으로 풍속을 두 배로 올리면 소음은 수십 배(2^5~2^6 = 32~64배)가 된다
팬 전력	팬 소비전력은 유량의 세제곱에 비례한다. 냉각하려고 쓴 전력이 다시 열이 되는 악순환
압력손실	서버 내부를 그 속도로 통과시키려면 정압이 급증하고, 앞뒤 랙 간 재순환·바이패스가 심해진다

실무 한계는 자료와 조건에 따라 다르지만, 바닥 타일 급기만으로는 대략 랙당 30 kW 안팎이다(3절 「공랭 한계를 한 숫자로 못 박지 않는 이유」). ASHRAE TC 9.9 백서에 따르면 40~50 kW 랙의 풍량은 최대 5,000 CFM으로, 최고 성능 바닥 타일 하나(1,900 CFM)의 2.6배다. 120 kW 랙의 약 14,300 CFM은 타일 하나의 약 7.5배, 40~50 kW 랙 최대 풍량의 약 2.9배라서, 바닥 타일 급기로 식히는 것은 실용 범위를 벗어난다.

(b) 수냉, $\Delta T = 10$ K

$$\dot{m} = \frac{120{,}000}{4{,}186 \times 10} = 2.867 \text{ kg/s} \approx 2.87 \text{ L/s} = 172 \text{ L/min}$$

두 방식의 부피유량 비교

방식	부피유량
공랭	6.75 m^3/s
수냉	0.00287 m^3/s
비	**약 2,350배**

배관 지름을 40 mm로 잡으면 유속은 약 2.3 m/s로 권장 범위(1~3 m/s)에 들어온다. 25 mm로 줄이면 5.8 m/s가 되어 침식·소음 문제가 생긴다.

얻어 갈 것: 냉각 방식의 전환은 **물성이 강제한 결과**다. 그리고 이 전환은 데이터센터 설계 전체를 바꾼다 — CDU, 이차 배관 루프, 누수 감지, 유지보수 절차가 전부 새로 필요하다.

아래 그림은 120 kW 랙을 공랭 $\Delta T = 15$ K(공기 밀도 1.18 kg/m^3)와 수냉 $\Delta T = 10$ K로 식히는 경우다. 화면본에서는 랙 전력, 두 온도차, 공기 밀도, 랙 정면 면적, 배관 지름을 바꿔 풍량·풍속·냉각수 유량이 어떻게 변하는지 ASHRAE 백서의 두 기준(5,000·1,900 CFM)과 함께 볼 수 있다.



그림 7-1. 120 kW 랙을 공랭(ΔT 15 K)으로 식히려면 6.75 m³/s(약 14,300 CFM), 랙 정면 풍속 5.6 m/s가 필요하다. 수냉(ΔT 10 K)은 172 L/min이면 된다. 두 점선은 ASHRAE 백서의 5,000·1,900 CFM이다. — 화면본에서는 이 자리에서 값을 바꿔 볼 수 있다(화면본 7회차 실습 1 풀이).

실습 2 풀이 — 전력 체인 손실

▾ 풀이

종합 효율

$$0.99 \times 0.96 \times 0.99 \times 0.94 \times 0.90 = 0.796$$

단계별로: $0.99 \times 0.96 = 0.9504 \rightarrow \times 0.99 = 0.9409 \rightarrow \times 0.94 = 0.8844 \rightarrow \times 0.90 = 0.7960$.

칩에 1 MW를 전달하려면

$$\frac{1}{0.796} = 1.256\ \text{MW}$$

계통에서 1.256 MW를 받아야 한다. **손실 256 kW는 전부 열이 된다.**

이 열이 어디에 더해지는가 — 위치가 중요하다

손실 지점	효율	손실	열이 나오는 곳
변압	99%	12.6 kW	전기실
UPS	96%	49.7 kW	전기실
PDU	99%	11.9 kW	전산실 주변
PSU	94%	70.9 kW	**랙 안**
VRM	90%	111.1 kW	**보드 위, 칩 옆**

랙 안에서 빼야 할 열은 칩 1 MW가 아니라

$$\frac{1}{0.94 \times 0.90} = 1.182 \text{ MW}$$

즉 18%가 더 붙는다. PSU 손실 70.9 kW와 VRM 손실 111.1 kW는 랙 안에서 나오므로 랙의 냉각 부하에 들어가고, 실습 1의 랙 냉각 요구를 그만큼 키운다. 그 열을 어느 경로로 빼는지는 구조에 달려 있다. VRM을 콜드플레이트에 함께 물리면 냉각수가 가져가고, 그렇지 않으면 3절 표의 「칩 외 부품은 여전히 공랭 필요」 몫으로 남는다.

시설 전체의 열부하는 1.256 MW다 — 들어온 전력은 하나도 빠짐없이 열이 되어 나간다(연산은 에너지를 저장하지 않는다).

설계적 함의: 효율 개선은 두 번 이긴다. 전기요금을 줄이고, **식혀야 할 열(냉각 부하)까지 같은 크기로 줄인다.** VRM 효율을 90% → 94%로 올리면 VRM 손실이 111 kW → 약 64 kW로 약 47 kW 줄고(앞 단계들이 나르는 전력도 줄어 계통에서 받는 전력은 1.256 → 1.203 MW, 약 53 kW 준다), 시설이 빼야 할 열도 같은 53 kW만큼 준다. 줄어드는 것은 열부하이지 냉각 설비가 쓰는 전력 그 자체가 아니다. 냉각 설비의 소비전력이 얼마나 주는지는 냉동기의 성능계수(COP), 펌프·팬의 고정 부하에 달려 있어 이 실습의 가정만으로는 정해지지 않는다.

아래 그림은 다섯 단계 효율 99·96·99·94·90%로 칩에 1 MW를 보내는 경우다. 화면본에서는 단계별 효율과 칩 전력을 바꾸거나 다섯 단계를 모두 97%로 두어, 손실이 어느 단계에서 얼마나 나오고 그 열이 어디서 나오는지 볼 수 있다.



그림 7-2. 변압·UPS·PDU·PSU·VRM 효율이 99·96·99·94·90%면 종합 효율은 0.796이고, 칩에 1 MW를 보내려면 계통에서 1.256 MW를 받아야 한다. 손실 256 kW 가운데 VRM 몫 111 kW가 가장 크다. — 화면본에서는 이 자리에서 값을 바꿔 볼 수 있다(화면본 7회차 실습 2 풀이).

실습 3 풀이 — PUE 계산과 조작

▾ 풀이

원래 시설

$$\text{PUE} = \frac{10 + 3 + 1}{10} = \frac{14}{10} = 1.40$$

(가) IT를 8 MW로 효율화 — 부대설비가 완전 비례한다면

냉각 $3 \times 0.8 = 2.4$, 배전 $1 \times 0.8 = 0.8$

$$\text{PUE} = \frac{8 + 2.4 + 0.8}{8} = \frac{11.2}{8} = 1.40$$

PUE가 그대로다. 총 전력은 14 → 11.2 MW로 20% 줄었는데 지표는 아무 변화가 없다.

(나) 그런데 현실에서는 더 나쁘다 — 부대설비에는 고정 부하가 있다

냉각 3 MW를 고정 1 MW + 변동 2 MW로, 배전 1 MW를 고정 0.2 + 변동 0.8로 나눠 보자.

항목	IT 10 MW	IT 8 MW
IT	10	8
냉각	3.0	$1 + 2 \times 0.8 = 2.6$
배전	1.0	$0.2 + 0.8 \times 0.8 = 0.84$
총계	**14.0**	**11.44**
PUE	**1.40**	**1.43**

일은 똑같이 하면서 총 전력을 18% 줄였는데 PUE는 나빠졌다.

PUE의 구조적 함정 — 세 가지

1. **분모가 IT 전력이다.** 부대설비가 IT에 완전히 비례하면 IT를 효율화해도 PUE는 그대로이고((가)), 고정 부하가 남으면 분모만 더 줄어 지표가 나빠진다((나)). **PUE는 IT 효율을 측정하지 않을 뿐 아니라, 고정 부하가 있는 시설에서는 IT 효율 개선에 불이익을 준다.**
2. **일의 양이 들어 있지 않다.** 같은 PUE 1.2라도 토큰을 10배 만드는 시설과 아닌 시설이 구별되지 않는다.
3. **경계를 어디에 긋느냐로 조작 가능하다.** 변압 손실을 IT 쪽으로 넣느냐 부대 쪽으로 넣느냐, 발전 손실을 포함하느냐에 따라 값이 달라진다.

대안 지표

지표	내용
연산당 에너지	J/token, J/FLOP, 학습 1회당 MWh — **일의 양이 분모에 들어간다**
총 전력 절대값	계통·부지 제약을 논할 때는 비율이 아니라 절대량이 답이다
TUE (total-power usage effectiveness)	PUE를 IT 장비 안으로 연장한 지표. ITUE(IT 장비 총 전력 ÷ 그 안의 연산 부품에 닿은 전력, PSU·VRM·팬 손실을 담는다)와 PUE를 곱해 시설 총 전력 ÷ 연산 부품 전력을 본다. 예컨대 PUE 1.2, IT 장비 안 전달 효율 0.9면 ITUE ≈ 1/0.9이고 TUE ≈ 1.2/0.9 ≈ 1.33이다
WUE / CUE	물과 탄소. 지역 제약을 반영한다

판단에 쓸 때의 교훈: 비율 지표는 분모가 정책 변수일 때 언제나 조작 가능해진다. **"이 지표를 좋게 만드는 가장 쉬운 방법이 무엇인가"를 먼저 묻는 습관** — 1회차 4절에서 시작한 훈련이 여기서 완성된다.

아래 그림은 IT 10 MW(냉각 3 MW = 고정 1 + 변동 2, 배전 1 MW = 고정 0.2 + 변동 0.8)를 같은 일을 하며 8 MW로 줄이는 경우다. 화면본에서는 고정분·변동분과 줄인 IT 전력을 바꾸고, 「완전 비례」와 「고정 부하 있음」을 견줘 볼 수 있다.

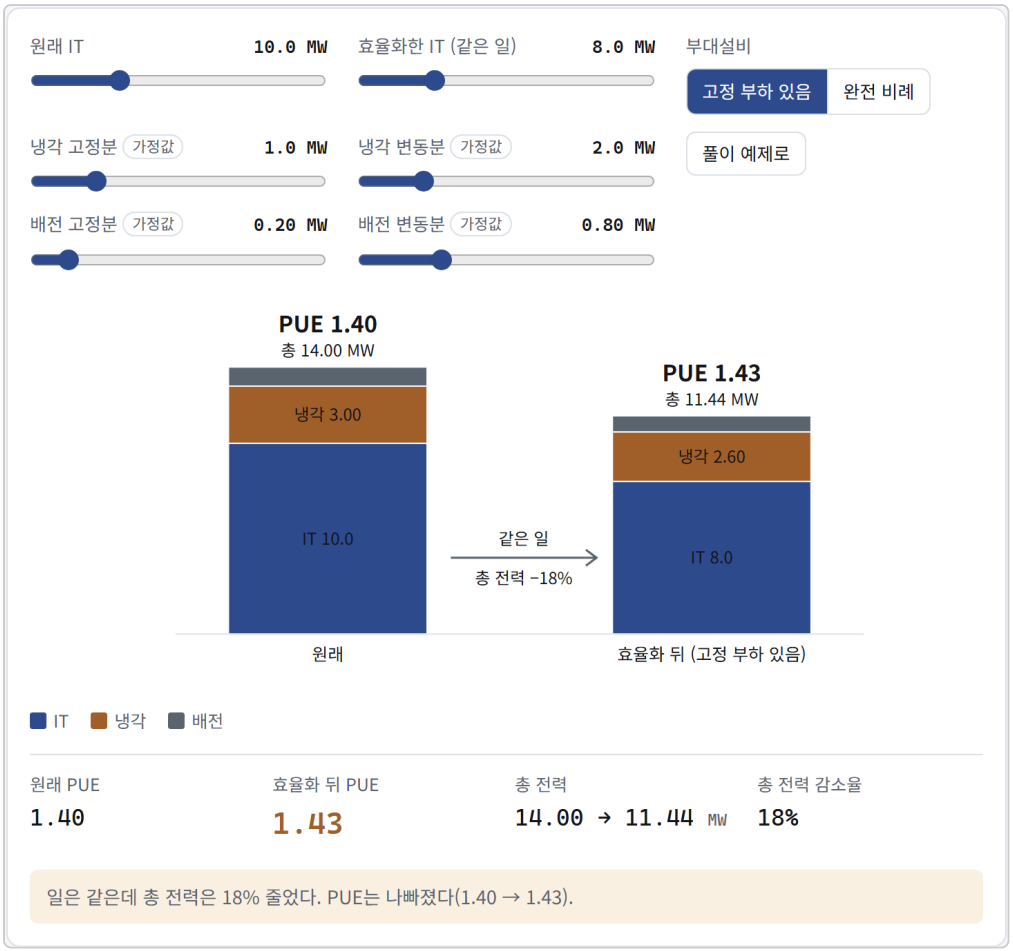


그림 7-3. IT 10 MW를 같은 일을 하며 8 MW로 줄이면 총 전력은 14 → 11.44 MW로 18% 줄지만, 부대설비의 고정 부하 때문에 PUE는 1.40에서 1.43으로 나빠진다. — 화면본에서는 이 자리에서 값을 바꿔 볼 수 있다(화면본 7회차 실습 3 풀이).

실습 4 풀이 — IR drop

▾ 풀이

전압 강하

$$\Delta V = IR = 500 \times 0.2 \times 10^{-3} = 0.1\ \text{V}$$

코어에 실제로 걸리는 전압

$$1.0 - 0.1 = 0.9\ \text{V} \quad (10\%\ \text{강하})$$

왜 심각한가

영향	내용
속도 저하	트랜지스터 구동 전류는 대략 $(V - V_{th})^{\alpha}$ 에 비례한다. V_{th} 가 0.3 V라면 유효 구동전압이 0.7 → 0.6 V로 **14% 감소**한다
타이밍 마진 붕괴	강하가 균일하지 않고 위치·시간에 따라 다르므로(스위칭 노이즈), 최악 조건을 가정해 클럭을 낮춰야 한다
전력 손실	$I^2R = 500^2 \times 0.2 \times 10^{-3} = 50$ W — 칩이 500 W를 쓴다면 그중 10%가 전력망에서 열로 사라진다

강하를 절반으로 줄이는 세 가지 방법

방법	원리	대가
저항 R 을 절반으로	더 두껍고 넓은 전력 배선, 더 많은 금속층	신호 배선 공간을 잡아먹는다 — 앞면 배선 자원의 경쟁
후면전력공급 (BSPDN)	전력 배선을 웨이퍼 **뒷면**으로 옮겨 앞면을 신호에 전부 내준다. 경로가 짧고 굵어져 R 이 크게 준다	새로운 공정(웨이퍼 박화, 뒷면 정렬) 필요
전압을 올리고 전류를 줄인다	같은 전력에서 V 를 2배로 하면 I 는 절반, $\Delta V = IR$ 도 절반, I^2R 손실은 **1/4**	마지막 강압을 칩 가까이서 해야 함 → 온패키지 전압조정기(IVR), 48 V 직류 배전

5회차와의 연결 — BSPDN이 왜 GAA만큼 중요한가

5회차 3절에서 후면전력공급을 "다음 분기점"이라고만 했다. 그 이유가 이 계산에 있다. 전류가 계속 늘어나는 상황($P = VI$ 에서 V 는 못 올리고 P 는 오르므로 I 만 는다)에서 $\Delta V = IR$ 을 잡으려면 R 을 줄이는 수밖에 없고, 앞면 배선만으로는 한계에 왔다. **BSPDN은 성능 향상 기술이 아니라 전력 전달의 물리적 한계를 푸는 기술이다.**

그리고 이것이 데이터센터까지 이어진다. 랙 배전이 48 V로 가는 것도, 전압조정기를 보드 위로 올리는 것도 전부 같은 식($\Delta V = IR$, 손실 $= I^2R$)의 서로 다른 층위에서의 적용이다.

랙 수준에서 같은 계산을 하면. 2절의 800 V 직류 이야기를 숫자로 옮긴다. IT 1 MW 랙을 48 V급 체계의 버스 전압 54 V로 먹이면 전류는

$$I_{54} = \frac{1{,}000{,}000}{54} \approx 18{,}519 \text{ A} \approx 18.5 \text{ kA}$$

800 V로 먹이면

$$I_{800} = \frac{1{,}000{,}000}{800} = 1{,}250 \text{ A}$$

전류는 약 14.8배 차이다. 같은 저항의 도체라면 저항 손실 I^2R 은

$$\left(\frac{800}{54}\right)^2 \approx 219$$

배 차이가 난다. 거꾸로 말하면 54 V로 손실을 800 V 수준에 맞추려면 도체 저항을 약 1/219로, 곧

같은 길이라면 단면을 약 219배로 키워야 한다. NVIDIA가 54 V로 1 MW 랙을 먹이면 구리 버스바가 최대 200 kg 든다고 한 것이 이 계산의 물리적 모습이다. 실제 설계에서는 전압마다 도체 굵기가 다르므로 219배는 「같은 도체를 쓴다면」이라는 조건이 붙은 비교다.

아래 그림은 이 풀이의 랙 배전 부분만 다룬다. 앞부분 칩 IR drop의 0.1 V·0.9 V·50 W는 그림에 없다. 그림은 IT 1 MW 랙을 54 V와 800 V로 먹이는 경우다. 화면본에서는 랙 전력과 두 전압(OCP 공칭 48 V 포함)을 바꿔 전류와, 같은 도체를 쓸 때의 저항 손실 비를 볼 수 있다.

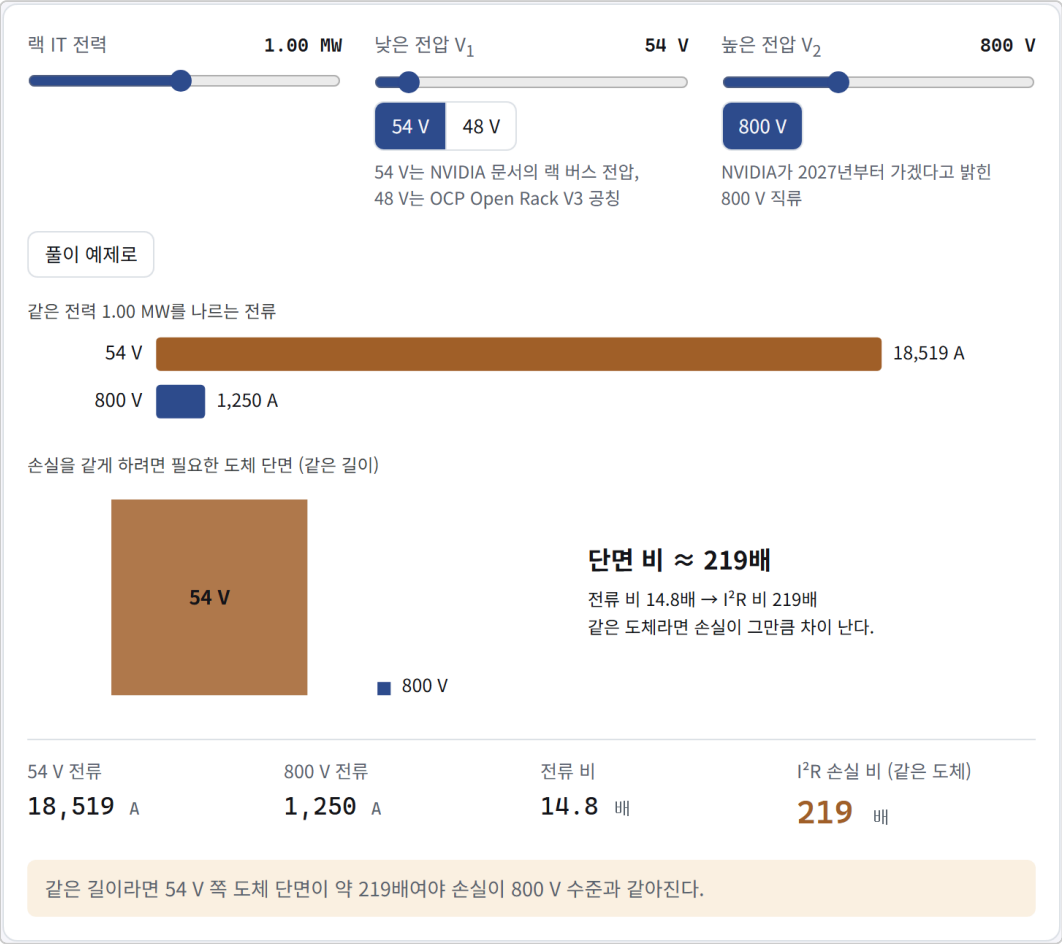


그림 7-4. 1 MW 랙을 54 V로 먹이면 약 18.5 kA, 800 V면 1,250 A다. 전류가 14.8배라 같은 도체의 저항 손실은 약 219배다. — 화면본에서는 이 자리에서 값을 바꿔 볼 수 있다(화면본 7회차 실습 4 풀이).

실습 5 풀이 — 데이터센터 규모 추정 (종합)

▾ 풀이

이 문제가 이 책의 최종 점검이다. 가속기 10,000장 규모 학습 클러스터.

(a) 총 IT 전력

항목	계산	값
가속기	10,000 × 1 kW	10 MW
부대(CPU·스토리지·네트워크) 30%	10 × 0.3	3 MW
IT 합계		**13 MW**

실제 시스템과 견주기. 「가속기당 1 kW + 부대 30%」는 계산 연습용 가정값이다. 제조사 시스템 사양으로 바꿔 보면, DGX B200은 GPU 8장 시스템의 최대 전력을 약 14.3 kW로 적는다. 1만 장이면 1,250대이고 계산하면 최대 17.875 MW다. GPU 8장에 2.1 TB 메모리를 단 DGX B300은 약 14 kW로 적으므로 같은 계산으로 17.5 MW다. 시스템 경계의 최대값이라 가정값 13 MW보다 크다. 비교할 실물 하나: 플로리다대(UF)의 HiPerGator 4세대는 DGX B200 63대(GPU 504장)로 2025년 가을 학기에 가동되었다. 63대에 14.3 kW를 곱하면 IT 최대 약 0.90 MW다(계산하면). 대학 연구용 슈퍼컴퓨터의 GPU 부분(DGX B200 63대)이 최대 1 MW 안쪽이고, 이 실습의 클러스터는 GPU 수로 그 약 20배(10,000 ÷ 504 ≈ 19.8)다.

(b) 시설 총 전력과 연간 전력량

$$P_{\text{시설}} = 13 \times 1.2 = 15.6\ \text{MW}$$

$$E_{\text{연간}} = 15.6\ \text{MW} \times 8{,}760\ \text{h} = 136{,}656\ \text{MWh} \approx 137\ \text{GWh}$$

가동률 90%를 가정하면 약 123 GWh.

규모 감 잡기: 한국 가구 한 곳의 연간 전력소비를 어림값 3,000 kWh로 잡으면

$$\frac{136{,}656{,}000\ \text{kWh}}{3{,}000} \approx 45{,}600\ \text{가구}$$

가속기 1만 장짜리 클러스터 하나가 중소도시 한 곳만큼 전기를 쓴다.

(c) HBM 총 용량과 스택 개수

가속기당 192 GB(B200급)로 두면

항목	계산	값
총 HBM 용량	10,000 × 192 GB	**1.92 PB**
스택 개수	가속기당 8스택 × 10,000	**80,000 스택**
DRAM 다이	스택당 8단 → 80,000 × 8	**640,000 다이**

192 GB는 칩 공표값이다. NVIDIA의 Blackwell 아키텍처 기술 개요(V2.1)는 HGX B200의 GPU 메모리를 「최대 192 GB」로 적지만, 완제품 서버 DGX B200의 사양은 8장 합계 1,440 GB, 계산하면 GPU당 180 GB다(2회차 6절 「용량도 칩과 시스템이 다르다」). 이 실습은 칩 공표값으로 1.92 PB를 쓴다. DGX 시스템 가용량으로 세면 10,000 × 180 GB = 1.80 PB(1,250대 × 1,440 GB와 같다)가 된다. 같은 방식으로 DGX B300(8장 합계 2.1 TB)이면 1,250대 × 2.1 TB = 2.625 PB다

(계산하면).

3회차 실습 1을 여기에 적용해 보라. 다이 수율 0.95로 8단을 KGD 없이 쌓으면 스택 수율이 66%다. 80,000 스택을 확보하려면 121,000 스택분을 시도해야 하고, 그 과정에서 정상 다이 수십만 개가 버려진다. 이 66%는 실제 수율이 아니라 「쌓기 전에 다이를 골라내지 않으면」이라는 가정의 값이다. 이 예제가 보여 주는 것은 다이 선별과 적층 수율이 공급량을 얼마나 깎을 수 있는가이고, "HBM이 없어서 가속기를 못 만든다"는 뉴스가 말하는 실제 공급 병목을 이 숫자 하나로 설명할 수는 없다.

(d) 랙 수와 밀도 — 여기서 설계 분기가 생긴다

	랙당 8장 (전통형)	랙당 72장 (고밀도형)
랙 수	**1,250 랙**	**139 랙**
랙 전력	약 10.4 kW	약 94 kW
냉각 방식	**공랭 가능**	**수냉 필수** (실습 1)
백색공간 (랙당 4 m² 환산)	약 5,000 m²	약 560 m²
노드 간 통신 거리	길다 → 광 트랜시버 비중↑	짧다 → 구리로 가능한 구간↑

밀도 선택 하나가 냉각 방식·부지 규모·네트워크 비용을 동시에 결정한다. 고밀도가 유리한 이유는 공간 절약이 아니라 **통신 거리 단축**이다 — 학습은 노드 간 통신이 성능을 좌우하므로, 물리적으로 붙여 놓는 것이 곧 성능이다. 6회차 실습 2의 "거리가 에너지를 정한다"는 논리가 랙 스케일에서 반복된다.

(e) 실제 병목 — 여기까지 계산하면 드러나는 것

15.6 MW는 중형 산업단지급 수전 용량이다. 그런데

항목	리드타임
가속기 조달	수개월~1년
건물 신축	1~2년
변전소·송전선 신설	**3~7년**

표의 리드타임은 자릿수를 보기 위한 어림값이다. IEA는 2026년 4월 보고서에서 변압기 납기만 평균 2~3년이고, 많은 관할에서 계통 접속 대기 자체가 5~10년에 이를 수 있다고 본다(6절). 위 표의 3~7년보다 길 수도 있다는 뜻이다.

가장 긴 것이 전력 계통이다. 그래서 신규 데이터센터 입지가 "인터넷이 빠른 곳"이 아니라 **"전기를 이미 끌어올 수 있는 곳"** 으로 결정되고, 자가발전·전력구매계약(PPA)·원자력 인접 부지 논의가 나온다.

1회차의 질문으로 돌아오기: 병목은 EUV → CoWoS → HBM으로 이동했다. 이 계산이 말하는 것은 **다음 병목이 반도체 바깥, 즉 전력 계통에 있다**는 것이다. 그리고 그것은 자본으로도 기술로도 빨리 풀리지 않는다.

아래 그림은 실습 5의 가정(가속기 1만 장, 장당 1 kW + 부대 30%, PUE 1.2, 장당 192 GB)일 때다. 화면본에서는 IT 전력의 경계를 DGX B200·B300 시스템 최대값으로, PUE를 4절의 1.54·1.48·1.09로, 장당 메모리를 DGX B200 시스템의 180 GB나 DGX B300으로 바꿔 볼 수 있다. 가구 수, 1 GW 발전소 한 해 전력량에 대한 몫, 3절 수열 냉각의 유량도 함께 움직인다.



그림 7-5. 가속기 1만 장(장당 1 kW + 부대 30%)은 IT 13 MW, PUE 1.2에서 시설 15.6 MW, 한 해 약 137 GWh다. DGX B200·B300 시스템 최대값으로 세면 IT는 17.875·17.5 MW이고, 장당 192 GB면 HBM은 1.92 PB다. — 화면본에서는 이 자리에서 값을 바꿔 볼 수 있다(화면본 7회차 실습 5 풀이).

실습 6 풀이 — 추론 서비스 원가

▾ 풀이

설정: 70B INT8 모델, H100 8장 노드(총 $B = 26.8$ TB/s), 노드 전력 6.4 kW, PUE 1.2 → 시설 기준 7.68 kW. 전기료 $0.08/kWh.

노드 전력 6.4 kW는 계산 연습용 가정값이다. 제조사 사양인 DGX H100의 최대 입력 전력 10.2 kW와 다르다. 10.2 kW는 시스템 경계의 최대값이다(2회차 실습 4가 칩과 서버의 경계를 같은 방식으로 구분한다). 10.2 kW를 넣으면 아래 전력 원가는 모두 약 1.6배가 되지만, 자본비가 지배한다는 결론은 그대로다.

2회차 디코딩 상한식 적용

$$\text{tokens/s} \lesssim \frac{26.8\times10^{12}}{7\times10^{10}}\times B_{sz} = 383\times B_{sz}$$

실제 대역폭 이용률을 60%로 두면 실효 $230\times B_{sz}$ tok/s.

배치별 100만 토큰 생성 원가

배치	실효 처리량	소요 시간	전력량	전력 원가
1	230 tok/s	4,353 s (1.21 h)	9.29 kWh	**$0.743**
8	1,840	544 s	1.16 kWh	**$0.093**
32	7,350	136 s (0.038 h)	0.290 kWh	**$0.023**

배치 1 대비 배치 32가 정확히 32배 싸다. 전력 원가는 배치에 반비례한다 — 같은 시간에 같은 전력을 쓰면서 32배의 일을 하기 때문이다.

그런데 전력은 원가의 일부일 뿐이다 — 자본비를 넣어 보자

H100 8장 노드를 25만 달러, 3년 상각, 가동률 80%로 두면

$$\frac{250{,}000}{3\times8{,}760\times0.8} = \$11.9\ /\text{시간}$$

배치 32에서 100만 토큰은 0.038시간이므로 **자본비 $0.45.**

원가 항목	100만 토큰당 (배치 32)	비중
전력	$0.023	4.9%
자본비(감가상각)	**$0.45**	**95.1%**

자본비가 전력비의 약 20배다.

이 결과가 말하는 것 — 이 책의 실질적 결론 중 하나

1. **"GPU를 놀리지 마라"가 최우선 원칙인 이유가 이것이다.** 지배적 원가가 시간당 고정비이므로, 이용률이 곧 원가다. 배치를 키우고 대기 시간을 줄이는 서빙 최적화의 경제적 가치가 전력 절감보다 훨씬 크다. 위 자본비는 달력 시간당 250,000 / (3 × 8,760) ≈ $9.51 을 가동률로 나눈 값이다. 가동률이 80%면 쓴 시간당 $11.9, 50%로 떨어지면 $19.0이 된다. **이 위험을 누가 지는가는 계약이 정한다.** GPU 클라우드를 take-or-pay 약정으로 빌린 고객은 쓰든 안 쓰든 약정 용량 값을 낸다(7절의 CoreWeave 사례에서 2025년 매출의 98% 이상이 이런 약정에서 나왔다). 그러면 놀린 GPU의 원가는 임대 업체가 아니라 고객의 몫이 되고, 업체는 그 약정을 담보로 돈을 빌려 다음 GPU를 산다. 위험은 사라지지 않고 고객에게, 그리고 약정이 몇몇 큰 고객에게 몰려 있으면 그 고객의 신용으로 옮겨 간다(같은 10-K에서 2025년 매출의 약 67%가 최대 고객 한 곳에서 나왔다).
2. **전력이 중요하지 않다는 뜻은 아니다.** 전력은 원가가 아니라 **제약**으로 작동한다 — 실습 5에서 본 대로, 전기를 못 끌어오면 가속기를 살 수 있어도 설치할 수 없다.
3. **양자화와 배치 최적화가 왜 그렇게 공격적으로 추진되는지**가 정량적으로 설명된다. 둘 다 같은 자본으로 더 많은 토큰을 뽑는 수단이다.

> **주의:** 위는 전력·자본만 본 것이다. 실제 서비스 원가에는 인건비, 네트워크, 스토리지, 미사용 캐파, 안전 여유가 붙고, API 가격에는 마진과 개발비 회수가 더해진다. **이 계산은 가격이 아니라 하한을 준다.**

아래 그림은 배치 32, 노드 전력 6.4 kW(가정값), 가동률 80%일 때다. 화면본에서는 배치, 노드 전력(DGX H100 최대 10.2 kW), 가동률(take-or-pay 문단의 50% 포함)과 단가를 바꿔 100만 토큰당 전력 원가와 자본비가 어떻게 갈리는지 볼 수 있다.



그림 7-6. 70B INT8 모델을 H100 8장 노드(26.8 TB/s, 대역폭 이용률 0.6)에서 배치 32로 돌리면 토큰 100만 개의 전력 원가는 약 $0.023, 자본비는 약 $0.45로 자본비가 95.1%다. — 화면본에서는 이 자리에서 값을 바꿔 볼 수 있다(화면본 7회차 실습 6 풀이).

자습 부록 B — 토론 질문 생각의 실마리

질문 유형	핵심 축	실마리
전력이 진짜 상한인가	**무엇이 대체 가능한가**	반도체 병목은 자본과 시간으로 풀린다. 전력 계통은 **입지·인허가·계통 안정도**라는 비시장적 제약이 걸려 있어 성질이 다르다. 실습 5(e)의 리드타임 표가 근거다
효율 개선이 총수요를 줄이는가	**제번스 역설**	연산당 에너지가 개선되면 단가가 내려가고 수요가 늘어난다. 역사적으로 IT에서 효율 개선이 총 전력 소비를 줄인 사례는 드물다. 다만 **성장률이 물리적 상한에 부딪히면** 이야기가 달라진다
PUE 대신 무엇을 볼까	**분모에 무엇을 넣나**	실습 3의 결론 — 일의 양이 들어간 지표(J/token)와 절대량(총 MW)을 병용해야 한다. 규제·공시 맥락에서는 **비율 지표만 요구하면 반드시 조작된다**는 점이 중요하다
국내 입지의 제약	**계통과 냉각수**	수도권 집중과 계통 여유는 반비례한다. 수냉 전환은 물 사용(WUE)까지 변수로 만든다. 지역 분산은 통신 지연 때문에 학습 클러스터에는 제약이 크고, 추론 서비스에는 오히려 유리하다 — **워크로드 성격이 입지를 정한다.** 요금도 변수다. 분산에너지 활성화 특별법 제45조(지역별 전기요금, 2026년 9월 18일 시행 현행본)는 전기판매사업자가 국가균형발전 등을 위하여 기본공급약관을 정할 때 송전·배전 비용 등을 고려해 전기요금을 달리 정할 수 있게 한다
메모리 시장과의 되먹임	**AI가 일반 시장을 잠식한다**	7절의 주제. HBM 생산이 일반 DRAM 캐파를 잠식하면 소비자용 메모리 가격이 오른다. 과거 사이클에는 없던 **수요 간 구축(crowding-out)** 구조이고, 1회차 토론 5의 가장 강한 근거다

자습 부록 C — 핵심 용어

용어	영문	한 줄 정의
PUE	power usage effectiveness	시설 총 전력 ÷ IT 전력. 1.0이 이상값
WUE / CUE	water / carbon usage effectiveness	IT 에너지당 물 사용량 / 탄소 배출량
백색공간	white space	랙이 실제로 놓이는 전산실 바닥 면적
랙 전력밀도	rack power density	랙 하나가 소비하는 전력 [kW/rack]
공랭 / 수냉	air / liquid cooling	공기로 / 액체로 열을 실어 내는 방식
직접칩냉각 (직접 액랭)	direct-to-chip liquid cooling	콜드플레이트를 칩에 직접 붙이는 수냉
침지냉각 (액침냉각)	immersion cooling	서버를 절연 유체에 담그는 방식
CDU	coolant distribution unit	일차·이차 냉각 루프를 분리·분배하는 장치
콜드플레이트	cold plate	칩에 밀착해 열을 냉각수로 넘기는 금속판
TIM	thermal interface material	칩과 콜드플레이트 사이의 열전달 물질. 사슬의 흔한 병목
열저항	thermal resistance	$R = t\,/\,(kA)$ [K/W]. 직렬로 더해진다
UPS / PDU / PSU / VRM	—	무정전전원 / 배전반 / 파워서플라이 / 전압조정모듈
IR drop	—	배선 저항 때문에 생기는 전압 강하 $\Delta V = IR$

용어	영문	한 줄 정의
BSPDN	backside power delivery network	전력 배선을 웨이퍼 뒷면으로 옮기는 방식
PPA (전력)	power purchase agreement	발전사업자와 직접 맺는 장기 전력구매계약
RDMA	remote direct memory access	다른 노드의 메모리를 상대 CPU를 거치지 않고 네트워크 카드가 직접 읽고 쓰는 전송 방식. InfiniBand와 RoCE가 쓴다. NVIDIA의 GPUDirect RDMA는 네트워크 카드 같은 PCIe 장치가 GPU 메모리와 직접 데이터를 주고받는 경로를 연다
take-or-pay	—	쓰든 안 쓰든 약정한 용량 값을 치르는 계약. 이용률 위험이 공급자에서 구매자로 넘어간다
제번스 역설	Jevons paradox	효율이 좋아지면 오히려 총 소비가 느는 현상

자습 부록 D — 자가진단

☐ $Q = \dot{m} c_p \Delta T$ 로 임의의 랙 전력에 대한 필요 유량을 계산할 수 있다 → 실습 1 ★

☐ **공랭이 왜 120 kW 랙에서 실용적이지 않은지** 세 가지 이유로 설명할 수 있다 → 실습 1 ★

☐ 물이 공기보다 부피 기준 3,500배 유리하다는 것을 물성으로 설명할 수 있다 → 자습 안내

☐ 전력 체인 효율을 곱해 계통 소요 전력을 계산하고, **손실이 어디서 열이 되는지** 구분할 수 있다 → 실습 2

☐ **PUE가 IT 효율 개선에 불이익을 준다**는 것을 숫자로 보일 수 있다 → 실습 3 ★

☐ PUE의 대안 지표를 두 개 이상 제시하고 이유를 말할 수 있다 → 실습 3

☐ IR drop을 계산하고, BSPDN이 왜 필요한지 5회차와 연결해 설명할 수 있다 → 실습 4 ★

☐ **가속기 1만 장 클러스터의 전력·HBM·랙 수를 스스로 추정할 수 있다** → 실습 5 ★★

☐ 데이터센터의 진짜 병목이 전력 계통 리드타임임을 설명할 수 있다 → 실습 5(e)

☐ 추론 원가에서 **자본비가 전력비의 약 20배**임을 계산으로 보이고, 그 함의를 말할 수 있다 → 실습 6 ★

★★ 표시(실습 5)를 혼자 힘으로 끝까지 풀 수 있으면 이 책의 목표는 달성된 것이다.

마치며 — 일곱 회차를 한 줄로 꿰기

혼자 읽었다면 마지막으로 이 사슬을 **보지 않고** 복원해 볼 것. 이것이 최종 시험이다.

1. 트랜스포머는 행렬곱 덩어리다(1회차).
2. 그런데 디코딩은 토큰 하나마다 가중치 전체를 읽으므로 연산이 아니라 **대역폭**이 성능을 정한다(2회차).
3. 대역폭을 올리려면 넓고 느린 버스가 필요하고, 그러려면 적층과 TSV가 필요하며, 그 적층 수율이 HBM 가격을 만든다(3회차).
4. 한편 칩 자체는 면적에 대해 **초선형으로** 비싸지므로 큰 칩을 만들 수 없다(4회차).
5. 노광의 물리(레티클 한계·확률적 결함)가 그 한계를 더 조인다(5회차).
6. 그래서 **쪼개서 붙이는** 수밖에 없고, 붙이는 거리가 다시 전력을 정한다(6회차).
7. 그 모든 전력은 열이 되고, 열을 뺄 수 있는 능력과 전기를 끌어올 수 있는 능력이 **최종 상한**이 된다(7회차).

한 단계에서 다음 단계로 넘어가는 화살표(1 → 2, 2 → 3 …)마다 숫자 하나씩을 붙일 수 있다면, 이 책의 공부는 끝난 것이다.

화살표	붙일 숫자
1 → 2	$I^{*} \approx 295$ FLOP/byte, 배치 32면 이용률 11%
2 → 3	4 TB/s를 고속 핀으로 만들면 스택당 96 W
3 → 4	12단 적층 시 정상 다이의 43%가 낭비
4 → 5	면적 8배에 원가 20배, $A_{50} = \ln 2 \,/\, D_0$
5 → 6	High-NA 필드 429 mm², 스텝당 132 ppm
6 → 7	배선 거리 200배 = 에너지 200배, 계면 열저항이 실리콘의 20배
7 → 끝	15.6 MW = 45,600 가구, 변전소 리드타임 3~7년

출처

스터디 커리큘럼이 고른 입문 자료 ("AI 데이터 센터" 항목)

커리큘럼 소주제	링크	이 책에서의 활용
AI 데이터 센터	IBM Think — "AI 데이터 센터" (국문)	AI 데이터센터 개념 입문. 1절의 기초. **국문 개관으로 발제 도입에 적합**
AI 데이터 센터	Bloom Energy — "Data Center Cooling for Hyperscale and AI Workloads"	냉각 방식 비교. 3절의 직접 배경. **연료전지 업체 자료이므로 자가발전 부분은 이해관계를 감안해 읽을 것**
AI 데이터 센터	Bloom Energy — 2026 Data Center Power Report: Mid-Year Pulse	2026년 데이터센터 동향. 6절 거시 수치. **위와 같은 유보 적용**
AI 데이터센터와 메모리 시장 분석	Abdullayev, 「The Hungry Intelligence: How AI Cannibalized the Global Memory Market and Priced Out the Consumer World (2022–2026)」, SSRN Working Paper (2026)	AI 데이터센터 수요가 범용 DRAM 공급과 가격에 미친 영향을 기업 사례 중심으로 다룬 경제학 워킹페이퍼. **7절의 직접 배경. 동료심사 전 단계이고 정성 분석이 중심이므로 수치는 원 출처로 다시 확인할 것**
AI 데이터센터와 메모리 시장 분석	Probity Market Insights — HBM 시장 보고서	HBM 시장 규모·전망. **시장조사기관 자료로, 1회차 4절의 "수치 비판적 읽기" 실습 대상**
AI 데이터센터와 메모리 시장 분석	Counterpoint Research — Global DRAM and HBM Market Share (분기별)	분기별 점유율 추적. **정기 갱신 자료이므로 조회 시점 명시 필요**
AI 데이터센터와 메모리 시장 분석	ScienceON 보고서 TRKO202500001876	한국과학기술기획평가원(KISTEP), 「AI반도체를 활용한 K-클라우드 기술개발사업」 예비타당성조사 최종보고서 (2024-08, 주관부처 과학기술정보통신부). 국산 AI 반도체 기반 클라우드 사업의 편익·경제성 분석을 다룬다. 메모리 시장 분석 자료는 아니다

2.0판에서 더한 1차 출처 (2026-10-08 확인)

- IEA, *Key Questions on Energy and AI* (2026-04) — 6절 세계 전망(485→약 950 TWh, 약 3%, 2025년 +17%), 원인 5의 두 판 비교(부록 표 A.4), 계통 접속 대기 5~10년
- LBNL, 「Berkeley Lab Report Evaluates Increase in Electricity Demand from Data Centers」 (2025-01-15) — 6절 미국 수치
- Uptime Institute Global Data Center Survey 2025 (2025-07) — 1절 랙 밀도, 4절 PUE 1.54
- Google 데이터센터 PUE 공개 페이지 — 4절 PUE 1.09와 측정 경계
- ASHRAE TC 9.9, 「Emergence and Expansion of Liquid Cooling in Mainstream Data Centers」 (2021) — 3절·실습 1 공랭 한계
- NVIDIA 기술 블로그 「NVIDIA 800 VDC Architecture Will Power the Next Generation of AI Factories」 (2025-05-20) — 1·2절, 실습 4 풀이
- NVIDIA DGX B200·DGX B300·HGX 사양 페이지 — 실습 5 (a)·(c)(HGX 페이지는 HGX B200 합계를 1.4 TB, 곧 시스템 가용량으로 적는다. 칩 공표값 192 GB의 출처는 Blackwell 아키텍처 기술 개요 V2.1). NVIDIA DGX H100/H200 사용자 안내서 — 실습 6 풀이 10.2 kW

- Broadcom 보도자료 「Tomahawk 6 – Davisson」(2025-10-08)·NVIDIA 보도자료 「Spectrum-X Photonics」(2025-03-18) — 5절 CPO
- CoreWeave FY2025 Form 10-K (회사 IR 게시본) — 7절, 실습 6 풀이
- Meta 2025년 4분기 실적 보도자료(2026-01-28)·2026년 2분기 실적 보도자료(2026-07-29) — 7절 자본지출 가이던스
- 국가법령정보센터 「분산에너지 활성화 특별법」 — 부록 B 제45조
- UF News, HiPerGator 4세대(2025-05)·HiPerGator 공개(2025-10) — 실습 5 비교 예
- NVIDIA GPUDirect RDMA 문서 — 부록 C
- Patterson 외, 「TUE, a New Energy-Efficiency Metric Applied at ORNL's Jaguar」, ISC 2013 (Springer, 372–382쪽) — 실습 3 풀이 TUE·ITUE. 초록으로 두 지표의 관계(ITUE는 IT 장비 안의 PUE, TUE는 둘을 합쳐 총 에너지와 연산에 쓰인 에너지의 비)를 확인했다. 본문(구독)의 세부 측정 경계는 확인하지 못했다

검증 완료 (2026-08-02 복수 출처 대조, 2026-10-08 1차 출처 재확인·갱신)

- 공기와 물의 물성치(밀도·비열·열전도율) 및 체적 열용량 비 — 표준 물성값
- 냉각 방식별 대략적 적용 밀도 구간(공랭 대략 30 kW 안팎, RDHx ~40 kW, D2C 수십~175 kW) — 복수 산업분석에서 유사한 구간 제시. **다만 구간 경계는 자료마다 다르므로 "대략"으로 서술.** 공랭 구간은 이전 판의 「~15 kW」를 2026-10-08에 고친 것이다. 근거는 ASHRAE TC 9.9 백서(40~50 kW 랙 최대 5,000 CFM 대 바닥 타일 1,900 CFM)와 Uptime 2025(30 kW 초과 시설 드묾)이고, 하나의 공식 한계값은 1차 출처에 없다
- AI 랙에서 직접칩냉각(D2C)이 사실상 표준이 되었다는 점 — 복수 출처 일치
- IEA 데이터센터 전력 전망 — 2026-10-08 *Key Questions on Energy and AI*(2026-04) 원문으로 갱신: 2025년 485 TWh → 2030년 약 950 TWh(Base Case, 부록 표 945 TWh), 2030년 세계 전력의 약 3%. 2026년판도 본문(24쪽)에서 Base Case를 2024년 약 415 TWh → 2030년 약 950 TWh로 적고, 부록 표 A.4는 2024년 416 TWh(2026-10-22 확인). 2025년판의 2024년 약 415 TWh → 2030년 약 945 TWh는 원인 5의 실례로 남겼다
- 미국 데이터센터 전력 176 TWh(2023, 4.4%), 325~580 TWh(2028 추정, 6.7~12%) — LBNL 보도자료 원문(2026-10-08)
- PUE 실측: Uptime 2025 가중평균 1.54(n=681, 6년째 거의 정체), Google 2025 함대 1.09(TTM, 모든 오버헤드 포함, 안정 가동 대형 시설) — 각 원문(2026-10-08). 같은 조사의 조사 시점 5년 이내(2020년 무렵 이후) 준공 시설 평균 PUE 1.48(Uptime 2025 보고서 PDF 8쪽)은 4절에서 쓴다
- DGX B200 GPU 메모리 1,440 GB·최대 약 14.3 kW, DGX B300 2.1 TB·약 14 kW, HGX B300 2.1 TB — NVIDIA 사양 페이지(2026-10-08)
- 800 VDC: 54 V 랙 배전, 200 kW 초과 시 한계, 1 MW 랙·2027년, 54 V 1 MW 랙 구리 버스바 최대 200 kg, 효율 최대 5%·TCO 최대 30%(회사 주장) — NVIDIA 기술 블로그 원문(2026-10-08)
- CoreWeave take-or-pay·약정 매출 비중(88%·96%·98% 이상)·가중평균 계약기간 약 5년·최대 고객 약 67% — FY2025 10-K 원문(2026-10-08)
- 랙 전력밀도가 급격히 상승 중이라는 추세 — 복수 출처 일치. 2026-10-08 Uptime 2025(일반 시설 10~30 kW 확산)와 NVIDIA(AI 랙 200 kW 초과, 1 MW 목표)로 두 세계를 나눠 날짜와 함께 서술

미확인 · 편차가 큼 (본문에서 단정하지 않음)

- **데이터센터 전력 소비 총량의 특정 수치** — 기관별 편차가 매우 크다. 세계는 IEA, 미국은 LBNL 수치만 출처를 명시해 제시하고(LBNL 2028년은 범위 그대로), 다른 수치는 인용하지 않았다
- **평균 랙 전력밀도의 구체적 값** — 조사 주체·표본·연도에 따라 크게 다르다. 본문은 Uptime 2025의 분포 서술(10~30 kW 확산, 30 kW 초과 드묾)만 옮기고 평균값은 쓰지 않았다
- **최신 AI 시스템의 최대 랙 전력** — 제품·구성에 따라 다르다. 본문은 「100 kW 이상」과 NVIDIA가 밝힌 설계 방향(200 kW 초과, 2027년 1 MW)만 서술
- **AI 데이터센터가 전체 데이터센터 전력에서 차지하는 비율** — 정의 문제로 편차가 크다
- **9절 실습의 전력·효율·가속기 사양** — 문제의 값은 전부 **계산 연습용 가정값**이며 특정 제품·시설의 실제 값이 아니다. 풀이에 비교용으로 덧붙인 DGX B200·B300·H100 사양과 HiPerGator 구성만 제조사·기관 공표값이고, 수열 냉각의 유량·온도차는 가정값이다
- **HBM 캐파 전환이 일반 DRAM 공급에 미치는 구축 배율** — 메커니즘은 확립되어 있으나 정량적 배율은 업체별 영업비밀. 본문은 방향성만 서술
- **실습 5(b)의 가구당 연간 전력소비 3,000 kWh** — 계산 연습용 어림값. 전국 공식 평균은 확인하지 못했다. 서울시 에너지정보(한전 통계 기반)의 서울 가구당 전력소비량은 2023년 월 285 kWh(연 약 3,400 kWh)라, 실제 가구 수 환산은 「45,600 가구」보다 작아질 수 있다(2026-10-22 확인)
- **실습 5(e) 리드타임 표(수개월~1년·1~2년·3~7년)** — 자릿수 어림값. 1차 출처는 IEA(2026-04)의 변압기 납기 2~3년·계통 접속 대기 5~10년뿐이다

부록

부록 I — 공식 요약과 실습 색인

이 책에서 실제로 쓰는 식을 회차별로 한곳에 모으고, 식마다 쓸 때의 조건을 붙였다. **이 표를 보지 않고 재구성할 수 있으면 이 책의 목적은 달성된 것이다.**

1. 연산과 대역폭 (2회차)

식	의미	쓸 때의 주의
$P_{\text{attainable}}(I) = \min(P_{\text{peak}}, B \times I)$	**루프라인.** 성능 상한은 연산 한계와 대역폭 한계 중 작은 쪽	정상상태 커널의 상한. 지연·통신·오버헤드는 담지 못한다
$I^* = P_{\text{peak}} / B$	**리지 포인트.** 연산 바운드와 대역폭 바운드가 갈리는 지점	반드시 dense 기준 스펙으로 계산할 것
$I \approx M$ (프리필)	한 번 읽은 가중치로 M 개 토큰을 처리. I 의 단위는 FLOP/byte	가중치 읽기가 메모리 이동을 지배하고 가중치가 2바이트(fp16)일 때의 근사로, 활성화 읽고 쓰기를 뺐다(활성화를 온칩에 두거나 퓨전할 때). 일반적으로 $I \approx 2M / b$. M 이 크므로 연산 바운드
$I \approx B_{\text{sz}}$ (디코딩)	한 번 읽은 가중치로 배치 크기만큼만 처리	위와 같은 조건(2바이트 가중치, 가중치 읽기 지배). 일반적으로 $I \approx 2B_{\text{sz}} / b$ 이고, 연산 바운드로 넘어가는 경계는 $B_{\text{sz}} = (b / 2)\, I^*$ (연산은 BF16 피크 기준). 배치가 작으면 대역폭 바운드
$\text{tokens/s} \lesssim \frac{B}{P \cdot b} \times B_{\text{sz}}$	**디코딩 상한.** P 파라미터 수, b 파라미터당 바이트	B 는 모델을 나눠 실은 칩 N 장의 대역폭 합. 실제로는 이용률 60~80%에 통신 오버헤드가 더 붙는다
$\text{KV} = 4 \cdot L \cdot S \cdot d \cdot B_{\text{sz}}$ 바이트	fp16 KV 캐시 크기	d 는 KV 헤드 차원 총합. GQA면 크게 줄어든다

기억할 숫자: 최신 가속기의 I^* 는 **300 FLOP/byte 근처.** 디코딩 배치 32면 이용률 약 11%.

2. 수율과 원가 (4회차)

식	의미
$Y = e^{-AD_0}$	**푸아송 수율**. A 다이 면적[cm²], D_0 결함 밀도[결함/cm²]
$Y = \left(1 + \frac{AD_0}{\alpha}\right)^{-\alpha}$	**음이항 수율**. 결함 클러스터링 반영. $\alpha \to \infty$ 면 푸아송
$\mathrm{GDPW} \approx \frac{\pi d^2 / 4}{A} - \frac{\pi d}{\sqrt{2A}}$	웨이퍼당 다이 수. 둘째 항이 가장자리 손실. d 는 웨이퍼 지름이고, A 는 d 와 같은 길이 단위의 제곱으로 넣는다($d = 30$ cm면 A 는 cm², $d = 300$ mm면 mm²)
다이당 원가 $= \frac{W}{\mathrm{GDPW} \times Y}$	W = 웨이퍼 한 장 값
$A_{50} = \frac{\ln 2}{D_0}$	수율이 50%가 되는 다이 면적
$N^* = \frac{C_{\mathrm{NRE}}}{(c_G + o_G) - (c_A + o_A)}$	ASIC 손익분기 물량. ASIC 한 대가 GPU 한 대의 일을 같은 처리량·품질로 대신한다(1:1 대체)는 가정
$Y_{\mathrm{stack}} = y^n \times \prod y_{\mathrm{bond}}$	**적층 수율** (3회차)

기억할 숫자: $D_0 = 0.1$ 에서 800 mm² 다이 수율 **45%**. 면적 8배에 원가 **20배**. D_0 가 두 배면 같은 목표 수율에서 허용되는 다이 면적은 절반.

왜 클러스터링이 수율을 올리는가: e^{-x} 가 볼록하므로 젠센 부등식에 의해 $E[e^{-\Lambda}] \geq e^{-E[\Lambda]}$. 결함 총량이 같아도 뭉치면 나머지 다이는 깨끗하다.

---

3. 노광 (5회차)

식	의미
$\mathrm{CD} = k_1 \frac{\lambda}{\mathrm{NA}}$	**레일리 판정식**. 해상 가능한 최소 선폭
$\mathrm{DOF} = k_2 \frac{\lambda}{\mathrm{NA}^2}$	초점심도. **NA의 제곱**에 반비례
$E = hc / \lambda$	광자 에너지. 같은 dose(단위 면적당 에너지)에서는 파장이 짧을수록 광자 수가 적어진다
상대 요동 $= 1 / \sqrt{N}$	확률적 결함의 근원
$y_{\mathrm{step}} = Y^{1/n}$	스텝 수율. n 은 공정 단계 수

기억할 숫자: $k_1 = 0.3$ 에서 ArF 액침 **42.9 nm**, EUV **12.3 nm**, High-NA **7.4 nm**. High-NA DOF는 EUV의 **0.36배**. EUV 광자 수는 ArF의 **7%**, 요동은 **3.78배**. 800단계에서 라인 수율 90%를 맞추려면 스텝당 불량 **131.7 ppm**.

---

4. 패키징과 열 (6회차)

식	의미
$E = \frac{1}{2}CV^2$	배선 전송 에너지. C 는 배선 길이에 대략 비례
$R = \frac{t}{kA}$	열저항. t 두께, k 열전도율, A 면적
$\Delta T = Q \times \sum R$	같은 열유량 Q[W]가 직렬 경로를 모두 지날 때 열저항은 **직렬로 더해진다.** 층마다 발열이 따로 있어 구간마다 지나는 열유량이 다르면 $\Delta T = \sum Q_i R_i$
필요 면적 $= N_{\text{배선}} \times (\text{피치})^2$	면적 배열 배선의 소요 면적

기억할 숫자: 기판 배선(10 mm)과 TSV(50 μm)의 에너지 비 **200배.** 4층 적층에서 실리콘 층에 걸리는 온도 차 ΔT 누적 **3.3 K** 대 접합 계면 **66.7 K** — 병목은 계면이다. 실리콘 열전도율 약 **150 W/(m·K)**, 언더필·범프층은 **1~5.**

---

5. 데이터센터 (7회차)

식	의미
$Q = \dot{m}\, c_p\, \Delta T$	유체가 실어 나르는 열전달률(W)
$\text{PUE} = \frac{\text{시설 총 전력}}{\text{IT 장비 전력}}$	1.0이 이상값
$\Delta V = IR$, 손실 $= I^2R$	IR drop과 전력망 손실
종합 효율 $= \prod \eta_i$	전력 체인은 곱셈으로 깎인다

기억할 숫자: 공기 비열 **1,005**, 물 **4,186** J/(kg·K). **부피당 열용량은 물이 공기의 약 3,500배.** 120 kW 랙: 공랭 6.75 m^3/s(랙 정면 풍속 약 5.6 m/s, 실용 범위 밖) 대 수냉 **172 L/min.** 전력 체인 종합 효율 약 **80%.** 가속기 1만 장 = **15.6 MW** = 연 137 GWh.

---

6. 정량 실습 색인 (전 32문항)

회차	실습	무엇을 계산하나	난도
1	1	점유율이 90%와 70%로 갈리는 정의 차이	쉬움
1	2	200% 성장과 25% 성장이 모순이 아닌 이유	쉬움
1	3	세 관찰로부터 병목 위치 추론	보통
2	**1**	**리지 포인트 세 세대 비교**	쉬움 ★
2	**2**	**디코딩 상한 — 모델×정밀도×장수**	보통 ★

회차	실습	무엇을 계산하나	난도
2	3	KV 캐시 예산과 최대 배치	보통
2	4	ASIC 손익분기 물량	보통
3	**1**	**적층 수율과 낭비되는 정상 다이**	보통 ★
3	2	대역폭 설계와 pJ/bit 전력	보통
3	3	메모리 예산 배분 최적화	어려움
3	4	HBF 적합성 판정 기준	정성
4	1	수율 곡선과 50% 붕괴 면적	쉬움
4	**2**	**다이당 원가 — 면적 8배에 원가 20배**	보통 ★
4	3	푸아송 vs 음이항 (젠센 부등식)	어려움
4	4	칩렛 손익과 역전 지점	보통
4	5	ASIC 판단과 현실적 조직 규모	쉬움
5	1	레일리 판정식 세 세대	쉬움
5	2	초점심도의 대가	보통
5	3	확률적 결함 — 광자 수와 dose	어려움
5	**4**	**스텝 수율 — 800단계의 잔인함**	쉬움 ★
5	5	노광 필드와 최대 다이 크기	보통
6	1	칩렛 최적 분할 수 탐색	어려움
6	2	배선 에너지 200배	쉬움
6	3	인터포저 면적과 레티클 한계	보통
6	4	3D 적층 열 — 실리콘 대 계면	보통
6	5	배선 수와 필요 피치	보통
7	1	냉각 유량 — 공랭 대 수냉	보통
7	2	전력 체인 손실과 열부하 위치	쉬움
7	3	PUE의 구조적 함정	보통
7	4	IR drop과 BSPDN	쉬움
7	**5**	**데이터센터 규모 종합 추정**	어려움 ★★
7	6	추론 원가 — 자본비 대 전력비	보통

★ 표시 다섯 문항이 이 책의 뼈대다. **★★(7-5)를 혼자 힘으로 끝까지 풀 수 있으면 목표 달성이다.**

7. 일곱 회차를 한 줄로 꿰기

> 트랜스포머는 행렬곱 덩어리다(1회차) → 그런데 디코딩은 토큰 하나마다 가중치 전체를 읽으므로 연산이 아니라 **대역폭**이 성능을 정한다(2회차) → 대역폭을 올리려면 넓고 느린 버스가 필요하고, 그러려면 **적층과 TSV**가 필요하며, 그 적층 수율이 HBM 가격을 만든다(3회차) → 한편 칩 자체는 면적에 대해 **초선형으로** 비싸지므로(4회차) 큰 칩을 만들 수 없고 → 노광의 물리가 그 한계를 더 조인다(5회차) → 그래서 **쪼개서 붙이는** 쪽으로 기울고, 붙이는 거리가 다시 전력을 정한다(6회차) → 그 모든 전력은 열이 되고, 열을 뺄 수 있는 능력과 전기를 끌어올 수 있는 능력이 **최종 상한**이 된다(7회차).

화살표	붙일 숫자
1 → 2	$I^* \approx 295$ FLOP/byte, 배치 32면 이용률 11%
2 → 3	4 TB/s를 고속 핀으로 만들면 스택당 96 W
3 → 4	12단 적층 시 정상 다이의 43%가 낭비
4 → 5	면적 8배에 원가 20배, $A_{50} = \ln 2 \,/\, D_0$
5 → 6	High-NA 필드 429 mm², 스텝당 132 ppm
6 → 7	배선 거리 200배 = 에너지 200배, 계면 열저항이 실리콘의 20배
7 → 끝	15.6 MW = 45,600 가구, 변전소 리드타임 3~7년

부록

부록 II — 용어 색인

일곱 회차에 나온 용어를 한자리에 모았다. **회차** 열은 그 용어가 처음 정의된 곳이다 — 설명이 부족하면 그 회차의 「자습 부록 C」로 가면 된다.

용어	영문	뜻	회차
가속기	accelerator	특정 연산에 특화된 칩(GPU·TPU·NPU 등)의 총칭	1
개구수	numerical aperture (NA)	렌즈가 빛을 모으는 각도의 척도	5
검사 / 계측	inspection / metrology	결함이 있는지 찾는 것 / 치수·두께를 재는 것	5
결함 밀도	defect density D_0	단위 면적당 치명 결함 수 [결함/cm²]	4
공랭 / 수냉	air / liquid cooling	공기로 / 액체로 열을 실어 내는 방식	7
구조화 ASIC	structured ASIC	일부 층만 맞춤 제작해 NRE를 낮춘 중간 형태	4
굿다이	good die	정상 동작하는 다이. GDPW × 수율	4
노드	(process) node	"3nm" 같은 공정 세대 명칭. 물리 치수가 아니라 마케팅 명칭	1
다이	die	웨이퍼에서 잘라 낸 칩 하나	4
다이싱	dicing	웨이퍼를 잘라 개별 다이로 분리하는 공정	6
단일 실패점	single point of failure	그곳 하나가 멈추면 전체가 멈추는 지점	1
대역폭 / 용량	bandwidth / capacity	초당 옮기는 양 / 담아 두는 양. **3회차 내내 구분해야 하는 두 축**	3
대역폭 바운드	memory-bound	메모리 전송이 한계인 상태 ($I < I^*$)	2
듀얼 다마신	dual damascene	절연막에 비아와 배선 홈을 파고 구리를 한 번에 채운 뒤 CMP로 덜어 내는 배선 공정	5
랙 전력밀도	rack power density	랙 하나가 소비하는 전력 [kW/rack]	7
레일리 판정식	Rayleigh criterion	$\mathrm{CD} = k_1\lambda$ / NA. 해상도의 근본 한계	5
레지스트	photoresist	빛에 반응해 성질이 바뀌는 감광막	5
레티클 한계	reticle limit	노광기가 한 번에 찍는 최대 면적(약 26×33 mm)	4
루프라인 모델	roofline model	성능 상한을 연산 한계와 대역폭 한계의 최솟값으로 나타낸 모형	2
리지 포인트	ridge point	연산 바운드와 대역폭 바운드가 갈리는 연산강도 $I^* = P_{\text{peak}} / B$	2
리프레시	refresh	DRAM 커패시터의 전하가 새기 전에 다시 써 주는 동작	3
마이크로범프	microbump	다이 사이를 접합하는 미세 솔더 돌기	3·6
멀티패터닝	multi-patterning	한 층을 여러 번 나눠 찍어 해상도 한계를 우회	5
메모리 계층화	memory tiering	접근 빈도에 따라 데이터를 서로 다른 매체에 배치하는 기법	3
배치	batch size	동시에 처리하는 시퀀스 수	2

용어	영문	뜻	회차
백색공간	white space	랙이 실제로 놓이는 전산실 바닥 면적	7
뱅크 / 채널	bank / channel	DRAM 내부의 병렬 접근 단위	3
비닝	binning	다이를 시험해 등급을 나누는 일. 일부 불량 코어를 끄고 하위 등급으로 파는 결함 비닝(수율 회수)과, 결함 없는 다이를 최대 동작 주파수·전력으로 나누는 속도·전력 비닝이 있다	4
수율	yield	만든 것 중 정상 동작하는 비율	1
수율 학습곡선	yield learning	시간이 지나며 D_0 가 내려가는 과정	4
시스톨릭 어레이	systolic array	데이터를 격자로 흘리며 곱셈-누산을 반복하는 연산 구조	2
식각	etching	필요 없는 부분을 제거하는 공정	5
신호 무결성	signal integrity	신호가 왜곡·잡음 없이 전달되는 정도	6
액침냉각	immersion cooling	서버를 절연 유체에 담그는 방식	7
양자화	quantization	가중치를 더 적은 비트로 표현해 용량·대역폭을 아끼는 기법	1
언더필	underfill	범프 사이를 채워 기계적으로 보강하는 수지	6
연산 바운드	compute-bound	연산 유닛이 한계인 상태 ($I > I^*$)	2
연산강도	arithmetic intensity	메모리에서 옮긴 1바이트당 수행하는 연산 수 [FLOP/byte]	1
열산화	thermal oxidation	실리콘을 고온에서 산화시켜 SiO_2 막을 기르는 공정. 두께는 Deal–Grove 식을 따른다	5
열저항	thermal resistance	$R = t \,/\, (kA)$ [K/W]. 직렬로 더해진다	7
워피지	warpage	열팽창 차이로 패키지가 휘는 현상	6
음이항 수율 모형	negative binomial	$Y = (1 + AD_0 \,/\, \alpha)^{-\alpha}$. 클러스터링을 반영	4
이방성 / 선택비	anisotropy / selectivity	수직으로만 깎이는 정도 / 원하는 물질만 깎이는 정도	5
이온주입	ion implantation	불순물 이온을 때려 넣어 전기적 성질을 바꾸는 공정	5
이종집적	heterogeneous integration	서로 다른 공정·기능의 다이를 한 패키지에 통합	6
인터포저	interposer	여러 다이를 얹어 서로 연결해 주는 중간 기판	3·6
적층 수율	stack yield	스택 전체가 정상일 확률. y^n 으로 급락한다	3
전공정	front-end process (FEOL + BEOL)	웨이퍼 위에 트랜지스터(FEOL)와 그 위 배선층(BEOL)을 만드는 웨이퍼 공정 전체. 웨이퍼를 자르고 붙이는 후공정과 구분된다	5
제번스 역설	Jevons paradox	효율이 좋아지면 오히려 총 소비가 느는 현상	7
증착	deposition	필요한 막을 쌓는 공정	5
직접 액랭	direct-to-chip liquid cooling	콜드플레이트를 칩에 직접 붙이는 수냉	7
채찍효과	bullwhip effect	수요 변동이 공급망 상류로 갈수록 증폭되는 현상	1
초점심도	depth of focus (DOF)	초점이 유지되는 깊이 범위. $\propto \lambda \,/\, \mathrm{NA}^2$	5
축소 투영	reduction projection	마스크 패턴을 줄여 웨이퍼에 찍는 방식. 4배가 표준, High-NA는 4배·8배	5

용어	영문	뜻	회차
칩렛	chiplet	큰 칩을 여러 작은 다이로 나눠 만든 뒤 붙이는 방식	4
커널 퓨전	kernel fusion	여러 연산을 합쳐 중간 결과를 메모리에 쓰지 않는 최적화	2
콜드플레이트	cold plate	칩에 밀착해 열을 냉각수로 넘기는 금속판	7
클러스터링 파라미터	α	작을수록 결함이 뭉친다. ∞ 이면 푸아송	4
타일링·블로킹	tiling / blocking	데이터를 온칩에 올릴 크기로 잘라 재사용하는 최적화	2
텐서 병렬	tensor parallelism	한 층의 행렬을 여러 장의 가속기에 쪼개 얹는 방식	2
파운드리	foundry	타사 설계를 위탁 생산하는 공장	1
팹리스	fabless	설계만 하고 생산은 위탁하는 반도체 회사	1
푸아송 수율 모형	Poisson yield model	$Y = e^{-AD_0}$. 결함이 완전 무작위라는 가정	4
프리필 / 디코딩	prefill / decode	입력 전체를 한꺼번에 처리하는 단계 / 토큰을 하나씩 생성하는 단계	2
플립칩	flip chip	다이를 뒤집어 범프로 기판에 직접 붙이는 방식	6
하이브리드 본딩	hybrid bonding	범프 없이 구리-구리를 직접 접합하는 방식. 피치와 열 저항 모두에서 유리	3·6
하이퍼스케일러	hyperscaler	초대형 클라우드 사업자(Google·AWS·Microsoft·Meta 등)	1
학습 / 추론	training / inference	모델을 만드는 과정 / 만들어진 모델을 쓰는 과정	1
확률적 결함	stochastic defect	광자 수 부족에서 오는 무작위 결함	5
후공정	back-end (assembly & test), packaging	웨이퍼를 잘라 칩을 패키지로 만들고 검사하는 단계. 웨이퍼 공정 안의 배선층 형성인 BEOL과 다르다	6
희소성	structured sparsity	가중치 일부를 0으로 만들어 연산을 줄이는 기법. 스펙시트 수치 부풀리기의 주범	2
2.5D	—	다이를 인터포저 위에 나란히 놓는 방식	6
3D 적층	3D stacking	다이를 수직으로 쌓는 방식	6
ALD	atomic layer deposition	원자층 단위로 쌓아 두께를 정밀 제어하는 증착	5
ASIC	application-specific integrated circuit	용도를 고정해 만든 맞춤형 칩	1
BSPDN	backside power delivery network	전력 배선을 웨이퍼 뒷면으로 옮기는 방식	5·7
CD	critical dimension	만들어야 하는 최소 선폭	5
CDU	coolant distribution unit	일차·이차 냉각 루프를 분리·분배하는 장치	7
CFET	complementary FET	n형과 p형을 수직으로 쌓는 차세대 구조	5
CMP	chemical mechanical polishing	화학·기계적으로 표면을 평탄화하는 공정	5
CoWoS	chip-on-wafer-on-substrate	TSMC의 대표적 2.5D 첨단 패키징 방식	1·6

용어	영문	뜻	회차
EDA	electronic design automation	칩 설계용 소프트웨어 도구	1
EMIB / 브릿지	embedded bridge	필요한 구간만 작은 실리콘으로 잇는 방식	6
EUV	extreme ultraviolet	파장 13.5 nm의 극자외선 노광	5
FlashAttention	—	어텐션을 퓨전해 HBM 왕복을 줄인 대표적 커널	2
GAA	gate-all-around	게이트가 채널을 사방에서 감싸는 트랜지스터 구조	5
GDPW	gross die per wafer	웨이퍼 한 장에서 나오는 다이의 총 개수(수율 적용 전)	4
GEMM	general matrix multiply	일반 행렬-행렬 곱셈	2
GQA / MQA	grouped / multi-query attention	KV 헤드 수를 줄여 KV 캐시를 줄이는 어텐션 변형	2
HAR 식각	high aspect ratio etch	깊고 좁은 구멍을 뚫는 식각. 3D NAND·TSV의 난제	5
HBF	high bandwidth flash	NAND를 같은 방식으로 적층해 대용량을 노리는 제안	3
HBM	high bandwidth memory	DRAM을 수직 적층해 대역폭을 극대화한 메모리	1·3
High-NA EUV	—	NA를 0.33에서 0.55로 올린 차세대 EUV. 필드가 절반	5
High-k	—	유전율이 높은 게이트 절연막. 누설전류를 막는다	5
IDM	integrated device manufacturer	설계·제조·판매를 모두 하는 종합 반도체 회사	1·4
IR drop	—	배선 저항 때문에 생기는 전압 강하 $\Delta V = IR$	7
KGD	known good die	적층 전에 검증을 마친 정상 다이	3·6
KV 캐시	key-value cache	이미 만든 토큰의 중간 결과를 저장해 재계산을 피하는 메모리	1
MoE	mixture of experts	토큰마다 일부 전문가만 활성화하는 희소 모델 구조	2
NRE	non-recurring engineering	칩 하나를 개발하는 데 드는 일회성 비용	2·4
OPC	optical proximity correction	회절 왜곡을 미리 상쇄하도록 마스크를 변형	5
OSAT	outsourced semiconductor assembly and test	후공정(패키징·테스트)을 전문으로 하는 외주 업체	1
PDK	process design kit	파운드리가 제공하는 공정별 설계 규칙·모델 묶음	4
PIM / PNM	processing in / near memory	연산기를 메모리 안에 / 메모리 옆에 두는 구조	3
PPA (전력)	power purchase agreement	발전사업자와 직접 맺는 장기 전력구매계약	7
PUE	power usage effectiveness	시설 총 전력 ÷ IT 전력. 1.0이 이상값	7
RDL	redistribution layer	패드 위치를 재배치하는 재배선층	6
TGV	through-glass via	유리 기판을 관통하는 배선. 실리콘 대체 후보	6
TIM	thermal interface material	칩과 방열판 사이의 열전달 매개 물질	6·7
TSV	through-silicon via	실리콘을 관통하는 수직 배선	3·6
UCIe	universal chiplet interconnect express	칩렛 간 인터페이스 표준	6

용어	영문	뜻	회차
UPS / PDU / PSU / VRM	—	무정전전원 / 배전반 / 파워서플라이 / 전압조정모듈	7
WUE / CUE	water / carbon usage effectiveness	IT 에너지당 물 사용량 / 탄소 배출량	7
pJ/bit	picojoule per bit	1비트 전송 에너지. 대역폭 설계의 실질적 제약	3

부록

부록 III — 인터랙티브 그림 목록

화면본(HTML)의 다음 자리에서는 그림의 값을 직접 바꿔 볼 수 있다. 인쇄본에는 같은 자리에 처음 상태의 정지 그림을 실었다. 단독 페이지는 설명과, 그림의 계산을 본문·풀이의 값과 맞춰 본 대조표를 함께 보여 준다.

그림	자리	내용	단독 페이지
그림 2-1	2회차 4절	4×4 가중치 고정 시스톨릭 어레이에 토큰 하나(M = 1, 배치 1 디코딩)를 흘린 넷째 박자. 활성화가 대각선으로 지나가는 동안 16개 PE 가운데 대각선의 4개만 일한다.	widgets/systolic.html
그림 2-2	2회차 6절	루프라인 위의 디코딩 워크로드. H100에서 배치 32, fp16 가중치면 연산강도가 약 32로 리지 포인트(약 295)에 못 미쳐 대역폭 바운드에 놓이고, 연산 유닛 이용률 상한은 약 11%다.	widgets/roofline.html
그림 2-3	2회차 실습 3 풀이	H100 80GB 한 장에 70B급 모델을 INT8 가중치(70 GB)로 올리고 GQA·KV fp16으로 S = 4,096 시퀀스를 받으면 KV 캐시 자리가 시퀀스 7개분뿐이다. 배치가 7이면 연산도 INT8로 한다고 볼 때 연산 유닛 이용률은 약 2.4%다.	widgets/kv-cache.html
그림 2-4	2회차 실습 4 풀이	칩 경계(GPU 700 W, ASIC 300 W)와 실습 4의 가정값(NRE 1억 달러, 단가 3,000·30,000달러, PUE 1.3, 가동률 80%, $0.08/kWh, 3년)에서 두 총비용 직선은 N* ≈ 3,590대에서 만난다. 분모 27,875달러 가운데 전력비 차이는 875달러(3%)다.	widgets/asic-breakeven.html
그림 3-1	3회차 6절	읽는 바이트의 10%를 HBF(0.4 TB/s, 가정값)에서 읽으면 HBM4 스택(3.3 TB/s)과 섞은 실효 대역폭은 약 1.91 TB/s로 HBM만 읽을 때의 58%이고, 용량은 548 GB다. 띠는 HBF 공개 사양의 범위 0.4~3.0 TB/s다.	widgets/hbm-hbf-tier.html
그림 3-2	3회차 실습 1 풀이	다이 수율 0.95로 12단을 선별 없이 쌓으면 스택 수율은 54.0%이고, 불량 스택 하나에 멀쩡한 다이가 평균 10.7개 묶여 버려진다. 막대는 무작위로 뽑은 스택 100개다.	widgets/hbm.html
그림 3-3	3회차 실습 2 풀이	스택당 4 TB/s를 1,024비트 버스로 내려면 핀당 31.25 Gbps, 2,048비트로 내면 15.6 Gbps가 필요하다. 비트당 에너지를 3과 0.75 pJ/bit로 두면 스택 8개분 메모리 인터페이스 전력은 (a)가 800 W에 육박하고 (b)는 약 190 W다(0.5~1 pJ/bit 범위면 약 130~260 W).	widgets/hbm.html
그림 4-1	4회차 실습 2 풀이	300 mm 웨이퍼에 800 mm² 다이를 깔면 직접 센 개수는 68개(격자 위치에 따라 66~71개)이고, √(2A) 근사식은 64.8개다. D_0 = 0.1이면 수율 44.9%, 굿다이 29.1개, 다이당 원가는 웨이퍼 값의 0.03435배로 100 mm² 다이의 약 19.9배다. 결함 점은 무작위 표본 하나다.	widgets/die-cost.html
그림 4-2	4회차 실습 3 풀이	λ = 0.8에서 다이를 칸 10개로 쪼개면 $(1-\lambda/n)^n$ = 0.4344로 극한 exp(−0.8) = 0.4493의 97%에 이른다. 같은 λ에서 음이항(α = 2) 수율은 51.0%로 푸아송 44.9%보다 높고, 결함 밀도의 분산은 D_0^2/α = 0.005다.	widgets/die-cost.html
그림 4-3	4회차 실습 5 풀이	실습 5의 가정값(NRE 1.5억 달러, ASIC 4,000·GPU 30,000달러, 3년 전력비 2,000·6,000달러)이면 N* = 5,000대로, 「대기업 사내 AI 인프라 — 수백~수천 대」 띠 안에 놓인다.	widgets/asic-breakeven.html

그림	자리	내용	단독 페이지
그림 5-1	5회차 실습 2 풀이	$k_1 = 0.3$에서 CD는 ArF 액침 42.9 nm, EUV 12.3 nm, High-NA EUV 7.4 nm다. $k_2 = 0.5$로 둔 초점심도는 53·62·22 nm로, High-NA에서 EUV의 0.36배가 된다.	widgets/litho.html
그림 5-2	5회차 실습 3 풀이	같은 dose에서 EUV 광자 수는 ArF의 7%, 상대 요동은 3.78배다. 요동을 맞추려고 dose를 14.3배 올리면 같은 광원 출력에서 노광 시간이 14.3배가 되고, 노광 시간을 지키려면 광원 출력이 250 W에서 약 3.6 kW로 올라가야 한다.	widgets/litho.html
그림 5-3	5회차 실습 4 풀이	800단계에서 라인 수율 90%를 내려면 스텝당 허용 불량률이 131.7 ppm, 1,200단계면 87.8 ppm이다.	widgets/litho.html
그림 5-4	5회차 실습 5 풀이	마스크 위 패턴 영역 104 × 132 mm를 4배씩 줄이면 웨이퍼 필드는 26 × 33 mm(858 mm^2), 한 방향만 8배로 줄이는 High-NA는 26 × 16.5 mm(429 mm^2)다. $D_0 = 0.1$이면 최대 다이 수율은 42.4%와 65.1%다.	widgets/litho.html
그림 6-1	6회차 6절	다이 수율 0.9, 본딩 수율 0.995(가정값)로 8층을 쌓으면 W2W 수율은 0.416, 양품만 골라 붙이는 D2W는 0.966으로 2.32배 차이가 난다.	widgets/chiplet.html
그림 6-2	6회차 7절	1,000 W 칩을 θJC 0.015, θTIM 0.025, θSA 0.020 K/W(가정값) 사슬로 35 °C 주변에 내보내면 단계별 온도 차가 15·25·20 K이고 접합 온도는 95 °C다.	widgets/thermal.html
그림 6-3	6회차 7절	활성화 에너지 0.7 eV, 시험 125 °C, 사용 55 °C에서 가속계수는 약 77.7이고 HTOL 1,000시간은 약 8.87년에 해당한다. 0.5 eV면 약 2.56년, 0.9 eV면 약 30.7년이다.	widgets/reliability.html
그림 6-4	6회차 실습 1 풀이	800 mm^2 실리콘을 n개로 나눌 때 시스템당 비용(웨이퍼 값의 배수). $D_0 = 0.1$, 분할마다 오버헤드 5%, 다이 하나 추가마다 조립 수율 0.99면 n = 4에서 0.01970으로 가장 싸고, 단일 다이보다 42.6% 싸다.	widgets/chiplet.html
그림 6-5	6회차 실습 2 풀이	배선 10 mm와 TSV 50 µm는 길이 비 200배이므로 비트당 에너지도 200배다. TSV 쪽 0.05 pJ/bit이면 초당 10^{12} 비트에서 0.05 W 대 10 W다.	widgets/thermal.html
그림 6-6	6회차 실습 4 풀이	1 W/mm^2 칩 4층을 쌓으면 실리콘(50 µm)에서 생기는 온도 차는 합 3.33 K, 범프·언더필 같은 계면(20 µm, 3 W/(m·K))에서는 66.7 K로 20배다.	widgets/thermal.html
그림 7-1	7회차 실습 1 풀이	120 kW 랙을 공랭(ΔT 15 K)으로 식히려면 6.75 m^3/s(약 14,300 CFM), 랙 정면 풍속 5.6 m/s가 필요하다. 수냉(ΔT 10 K)은 172 L/min이면 된다. 두 점선은 ASHRAE 백서의 5,000·1,900 CFM이다.	widgets/datacenter.html
그림 7-2	7회차 실습 2 풀이	변압·UPS·PDU·PSU·VRM 효율이 99·96·99·94·90%면 종합 효율은 0.796이고, 칩에 1 MW를 보내려면 계통에서 1.256 MW를 받아야 한다. 손실 256 kW 가운데 VRM 몫 111 kW가 가장 크다.	widgets/datacenter.html
그림 7-3	7회차 실습 3 풀이	IT 10 MW를 같은 일을 하며 8 MW로 줄이면 총 전력은 14 → 11.44 MW로 18% 줄지만, 부대설비의 고정 부하 때문에 PUE는 1.40에서 1.43으로 나빠진다.	widgets/datacenter.html
그림 7-4	7회차 실습 4 풀이	1 MW 랙을 54 V로 먹이면 약 18.5 kA, 800 V면 1,250 A다. 전류가 14.8배라 같은 도체의 저항 손실은 약 219배다.	widgets/datacenter.html

그림	자리	내용	단독 페이지
그림 7-5	7회차 실습 5 풀이	가속기 1만 장(장당 1 kW + 부대 30%)은 IT 13 MW, PUE 1.2에서 시설 15.6 MW, 한 해 약 137 GWh다. DGX B200·B300 시스템 최대값으로 세면 IT는 17.875·17.5 MW이고, 장당 192 GB면 HBM은 1.92 PB다.	widgets/datacenter.html
그림 7-6	7회차 실습 6 풀이	70B INT8 모델을 H100 8장 노드(26.8 TB/s, 대역폭 이용률 0.6)에서 배치 32로 돌리면 토큰 100만 개의 전력 원가는 약 $0.023, 자본비는 약 $0.45로 자본비가 95.1%다.	widgets/datacenter.html

간기(刊記)

만든 이

항목	내용
스터디	일요일은 AI · 반도체 스터디
책 이름	AI 인프라와 반도체
엮은이	권보원(Bowon Kwon) · 대구지방법원 부장판사
판	2.0.1판 · 2026. 10. 22.
이용 허락	이 책에서 엮은이에게 권리가 있는 부분은 CC BY-NC 4.0(저작자표시-비영리 4.0 국제). 요약은 「이 책의 사용법」 첫 상자에 있고, 조건은 라이선스 원문을 따른다. 화면본 인터랙티브 그림의 코드(HTML·CSS·자바스크립트)는 PolyForm Noncommercial 1.0.0이고, 상업적 이용은 따로 협의한다. 원문에서 옮기거나 다시 그린 표·인용·그림은 이 라이선스들의 대상이 아니며 각 원저작자의 권리를 따른다. 권리자가 알려 오면 고치거나 뺀다. © 2026 권보원 (Bowon Kwon)
구성	이 책의 사용법 · 1~7회차 · 부록 3종(공식 요약과 실습 색인, 용어 색인, 인터랙티브 그림 목록) · 간기
분량	정량 실습 32문항 · 인터랙티브 그림 12종 · 용어 114개

이 책에 대하여

이 자습서는 「**일요일은 AI**」 스터디의 반도체 회차를 위해 만들었다.

이 책은 스터디 교재로 출발했지만, **모임 밖에서 혼자 읽는 독자도 끝까지 갈 수 있도록** 다시 엮었다 — 회차마다 읽는 순서와 시간 배분을 적어 두었고, 정량 실습에는 풀이를 전부 붙였으며, 회차 끝에는 스스로 점검할 체크리스트를 두었다.

어떻게 만들었나

이 책은 엮은이가 Claude Code와 함께 만들었다. 생성형 AI로 만든 글이므로 1차 출처와 대조하는 검증을 거쳤어도 틀린 곳이 남아 있을 수 있다.

- 스터디 발표와 토론에서 나온 내용은 제조사 발표·표준 문서·기관 보고서·논문 같은 1차 출처로 다시 확인한 것만 새로 써서 넣었다.
- 회차마다 독립 검증을 했다. 원고를 고친 쪽과 다른 쪽이 고친 원고와 1차 출처만 받아 바뀐 문장을 대조하고, 틀린 것은 고쳤다. 계산 실습의 숫자는 다시 계산했다.
- 마지막으로 다른 회사의 AI 모델(OpenAI Codex)이 작업 경위를 모르는 상태에서 원고·인쇄본·그림 코드를 통독했다. 지적 125건은 원문 대조와 재계산으로 하나씩 확인해 고치거나, 고치지 않는 이유를 남겼다.
- 인터랙티브 그림은 본문에 있는 식과 표의 변수만 움직인다. 그림마다 계산 결과를 본문 풀이와 자동으로 대조하고, 하나라도 어긋나면 책이 만들어지지 않게 했다. 인쇄본의 정지 그림은 화면본 그림을 처음 상태로 띄워 찍은 것이다.

- 수식은 MathML로 변환해 넣었다. 외부 스크립트나 폰트를 부르지 않는다.
- 도해는 인라인 SVG로 그렸다. 화면·인쇄·라이트·다크에서 같은 모양으로 나온다.
- 인쇄본의 쪽번호·머리글·목차 쪽번호는 조판 단계에서 계산했다.

판 이력

판	날짜	개요
교안	2026. 8. 2.	「일요일은 AI · 반도체 스터디」 일곱 회차 교안을 처음 작성했다(회차별 학습목표·본문·정량 실습)
교안 보강	2026. 8. 23.	혼자 읽는 독자를 위해 회차마다 자습 안내, 실습 풀이(32문항), 토론 질문의 실마리, 용어, 자가진단을 더했다
1판	2026. 8. 23. (표지 날짜) ~ 8. 24.(최종 빌드)	**스터디 내부 배포본(공개하지 않았다).** 일곱 회차를 HTML·PDF 한 권(149쪽)으로 묶고, 사용법·공식 요약·용어 색인 111개를 새로 쓰고, 도해 7개를 다시 그렸다
2.0판	2026. 10. 8.	**첫 공개판.** 1판을 바탕으로 스터디 발표·토론에서 나온 내용을 1차 출처로 확인해 더하고, 인터랙티브 그림 12개(화면본. PDF는 같은 자리에 정지 그림)와 원자료 링크를 넣었다. 시점이 지난 수치를 다시 확인하고, 다른 AI 모델의 전권 통독으로 계산·서술 오류와 지나친 단정을 고쳤다. 용어 색인은 114개다
2.0판 정정	2026. 10. 13.	이용 허락 문구를 고쳤다(권리가 있는 범위, 법적 조건·기술적 조치에 의한 추가 제한 금지, 요약의 지위, 제3자 자료 제외). 인터랙티브 그림의 코드는 CC BY-NC 4.0에서 PolyForm Noncommercial 1.0.0으로 바꾸고 상업적 이용은 따로 협의하기로 했다. 그 전에 받은 사본에는 종전 조건이 그대로 적용된다. 본문 내용은 바뀌지 않았다
2.0.1판	2026. 10. 22.	TSMC A16 공정의 양산 시기를 바로잡았다. 2024년 발표 때의 「2026년 생산 계획」만 적었던 것을, 2026년 4월 기술 심포지엄에서 보도된 TSMC 임원 발언에 따른 「양산 2027년 시작 예상」으로 고쳤다(4회차 4절 표, 5회차 3절과 출처·검증 상태). 영문판을 번역하고 원자료와 대조하는 과정에서 확인된 사실·서지 오류도 함께 바로잡았고, 회사 이름은 각 회사의 현재 표기(Sandisk·Arm·ZEISS)로 맞췄다

감사의 말

엮은이는 2022년 1월부터 한국방송통신대학교 통계·데이터과학과 졸업생과 재학생으로 구성된 「일요일은 AI」 스터디에 함께하고 있다. 한 주제가 끝날 때마다 새로운 주제를 제안해 주시고, 바쁜 가운데에도 발표를 돌아가며 맡아 매주 가르침을 나누어 주시는 스터디원들께 큰 빚을 지고 있다. 늘 감사드린다. 발표와 토론에서 나온 내용은 1차 출처로 다시 확인해 반영했고, 남은 오류는 엮은이의 몫이다.

이용할 때

- **수치를 그대로 인용하지 않는다.** 계산 실습의 가격·전력비 등은 결론의 방향을 보이기 위한 가정값이고, 가속기 사양은 공개 스펙시트 기준이라 세대가 바뀌면 달라진다. 자세한 구분은 앞머리 「수치에 관한 일러두기」에 있다.
- **수치의 시점.** 양산·출하 현황, 전력 수요 전망, 표준과 제품 사양은 2026년 10월 8일 기준으로 다시 확인했다. 그 뒤의 발표는 반영되어 있지 않다.
- **링크.** 본문의 링크는 2026년 10월 8일에 확인했다. 기업 블로그와 보도자료는 사라지거나 주소가 바뀔 수 있다.
- **틀린 곳과 개정.** 특히 계산 실습의 산술과 단위, 인용한 사양 수치를 눈여겨봐 주기 바란다. 원문과 다른 서술을 찾으면 공개 저장소의 이슈(Issues)로 알려 주기 바란다. 고친 내용은 판과 날짜를 올려 반영한다.
- **개인의 학습 자료.** 이 책은 엮은이가 개인적으로 공부하며 엮은 것이다. 엮은이가 속한 기관의 견해나 업무와는 관계가 없다.

일요일은 AI · 반도체 스터디 · 권보원 엮음 · 2.0.1판 · 2026. 10. 22.