MODULE 07 · 심화

AI 반도체 생태계 — NVIDIA와 도전자들

모듈 1에서 예고한 심화편입니다. 왜 GPU 한 회사가 세계 AI 인프라를 쥐고 있는지, 누가 어떤 전략으로 도전하는지, 그리고 한국의 K-NPU는 지금 어디까지 왔는지. 장비 도입 논의에 참여하려면 이 지형도가 머리에 있어야 합니다. (시장 상황은 빠르게 변합니다 — 이 내용은 2026년 상반기 기준입니다.)

01 왜 NVIDIA 독주인가 — 성벽은 칩이 아니라 소프트웨어다

NVIDIA의 H100·H200 같은 데이터센터 GPU는 성능도 뛰어나지만, 성능만으로 독점이 만들어지지는 않습니다. 진짜 성벽은 **CUDA**입니다. 2007년부터 쌓아온 GPU 프로그래밍 생태계 — 개발 도구, 수천 개의 라이브러리, 그 위에 올라간 거의 모든 AI 프레임워크, 그리고 CUDA로 훈련된 전 세계 개발자들 — 이 20년 치 축적이 경쟁사가 칩을 잘 만들어도 넘기 힘든 벽이 됩니다.

> 💡 **비유**
>
> 아무리 좋은 키보드를 새로 발명해도, 전 세계가 이미 쿼티 자판에 익숙하다면 바꾸게 만들기 어렵습니다. CUDA는 AI 업계의 쿼티 자판입니다 — 더 나은 대안이 있어도 "모두가 이미 이것으로 일하고 있다"는 사실 자체가 해자입니다.

그런데 여기에 더해 공급 부족까지 겹치면서 가격은 치솟았습니다. 수억 원짜리 장비가 몇 달 사이 수천만 원씩 더 비싸지는 시장 — 공공기관이 최신 GPU를 제때 확보하기 어려운 구조적 이유입니다.

02 판이 갈라지는 지점 — 학습과 추론은 다른 시장이다

그런데 도전자들에게 틈이 열리고 있습니다. AI 연산의 무게중심이 **학습(training)에서 추론(inference)으로** 이동하고 있기 때문입니다. 모델을 만드는 것은 한 번이지만, 만들어진 모델로 서비스하는 것은 매일, 수백만 번입니다. 업계 추정으로는 전체 AI 연산에서 추론 비중이 이미 절반 안팎에 이르렀고, 뚜렷한 과반

전환은 2026년경으로 보는 시각이 많습니다(Deloitte 2026 전망 등 — 소스마다 추정이 갈리는, 빠르게 움직이는 수치입니다).

- **학습**은 CUDA 생태계 의존이 깊어 NVIDIA를 벗어나기 어렵습니다.
- **추론**은 다릅니다 — 정해진 모델을 빠르고 싸게 돌리면 되므로, 서빙 소프트웨어 하나만 잘 지원해도 진입할 수 있습니다.
- 그래서 도전자 대부분이 **추론 특화 칩**으로 승부합니다. 전력 대비 처리량과 가격이 무기입니다.

03 해외 도전자들 — 세 가지 다른 전략

도전자	전략	비고
AMD (MI 시리즈)	정면 대결 — GPU 대 GPU	하드웨어는 근접, 소프트웨어(ROCm, AMD판 CUDA)의 성숙도가 관건
Groq, Cerebras	추론 속도 극단 특화	초고속 추론 클라우드로 틈새 확보
Tenstorrent	개방으로 우회	아래 상세

이 중 **텐스토렌트(Tenstorrent)**는 눈여겨볼 가치가 있습니다. AMD의 Zen, 애플 A칩, 테슬라 자율주행 칩을 설계한 **짐 켈러(Jim Keller)**가 이끄는 회사인데, 전략이 독특합니다. NVIDIA의 폐쇄적 생태계와 정반대로 — 개방형 명령어 규격(칩에게 "이렇게 계산하라"고 지시하는 공통 문법으로, CUDA처럼 특정 회사 전유물이 아니라 누구나 가져다 쓸 수 있게 공개돼 있습니다)인 RISC-V를 기반으로 하고, 소프트웨어 스택을 오픈소스로 풀고, 칩 설계 자체를 다른 회사에 라이선스로 판매합니다. "성벽을 정면 돌파하는 대신, 성벽이 없는 개방 진영을 만들겠다"는 접근입니다. 한국과의 접점도 많습니다 — 팹리스(공장 없이 설계만 하는 반도체 회사)답게 삼성 파운드리(설계는 다른 회사가 하고, 생산만 대신해주는 반도체 공장)에서 칩을 생산하고, 현대차그룹·삼성·LG 계열이 투자자로 참여했습니다(Tenstorrent 뉴스룸).

04 K-NPU — 한국의 도전, 그리고 왜 국가가 미는가

한국 정부가 국산 AI 반도체(K-NPU)를 강하게 지원하는 이유는 산업 육성만이 아닙니다. **AI 인프라 주권** 문제입니다. 공공·국방처럼 폐쇄망에서 돌아가야 하는 AI가 전부 수입 칩과 외국 클라우드에 의존하면, 공급이 끊기는 순간 선택지가 없어지니까요. 우리 연구회가 폐쇄망 로컬 AI를 연구하는 것과 정확히 같은 문제의식의 국가 버전입니다.

기업	주력	현재 위치 (2026 상반기 기준)
리벨리온 (Rebellions)	추론 칩 ATOM, 차세대 REBEL	사피온(SK 계열의 AI 반도체 회사)과 합병해 국내 대표 주자로 통합(보도). 국내 클라우드(KT 등) 상용 배치 경험
퓨리오사AI (FuriosaAI)	1세대 워보이 → 2세대 RNGD	RNGD가 LG AI 연구원의 엑사원(EXAONE) 서빙용으로 채택(2025.7)되며 실전 검증 단계 진입

정부는 K-클라우드 프로젝트 등으로 **국산 NPU 기반 데이터센터 실증**을 확대하고 있어서, 공공기관에 국산 NPU 장비가 도입되는 흐름은 앞으로 더 잦아질 겁니다. 즉 "우리 기관에 NPU 장비가 들어온다면 무엇을 확인해야 하는가"는 머지않아 실무 질문이 됩니다.

> ↗ 심화 학습
>
> 국산 NPU의 현장 적용 사례 — 포스코DX가 엣지 NPU 기업 모빌린트와 협약, 폐쇄망 제조 현장에서 LLM 구동 추진: **해설 블로그** · 개인 블로그·홍보성 내용이 섞여 있으니 "이런 흐름이 있다" 수준으로 참고하세요. 아직 PoC(시범 적용) 단계라는 점도 본문에 명시돼 있습니다.

05 냉정한 현재 수준 — 병목은 언제나 소프트웨어

도전자들의 공통 약점은 하드웨어가 아니라 소프트웨어 생태계입니다. TOPS(칩의 이론상 최대 연산 속도 단위) 같은 칩 스펙이 아무리 좋아도 —

- ✔ 내가 쓰려는 **모델**이 그 칩용으로 변환·최적화되어 있는가?
- ✔ 내가 쓰려는 **서빙 엔진**(vLLM 등)이 그 칩을 백엔드로 지원하는가?
- ✔ 문제가 생겼을 때 참고할 **커뮤니티와 문서**가 존재하는가?

이 세 가지가 안 되면 그 칩은 아직 "내 것"이 아닙니다. 현재 K-NPU를 포함한 추론 특화 칩들은 주요 오픈 모델 지원을 빠르게 넓히는 중이지만, CUDA 생태계의 폭에는 아직 못 미칩니다. 그래서 도입 판단의 기준은

벤치마크 표가 아니라 **"내 모델·내 엔진이 오늘 그 장비에서 실제로 도는가"의 실측**이어야 합니다.

NPU 도입 검토 체크리스트 (어느 기관이든 통하는 4문항)
① 우리가 쓸 모델(예: 한국어 오픈 sLM)의 지원 여부와 변환 절차는? ② OpenAI 호환 API 서빙이 되는가, 어떤 엔진으로? ③ 우리 태스크 골든셋으로 실측한 처리량·지연은? ④ 장애 시 지원받을 경로(업체 SLA(장애 대응 시간 등을 약속한 계약서)·커뮤니티)는?

06 모델은 어디로 가나 — 스스로 조절하고, 스스로 고친다

칩의 지형과 나란히, 모델 자체의 진화 방향도 하나로 수렴하고 있습니다. **"고정된 계산"에서 "입력에 따라 스스로 조절하는 계산"으로.** 세 흐름이 그 증거입니다.

- **필요한 만큼만 켠다 — MoE(혼합전문가).** 모든 파라미터가 항상 일하는 대신, 모델 안의 라우터(입력을 보고 어떤 전문가에게 보낼지 정하는 내부 분류기입니다 — 인터넷 공유기와는 다른 개념입니다)가 입력을 스스로 해석해 이 질문에 필요한 전문가들만 활성화합니다. 총량은 커지되 토큰당 비용은 작게 — 이미 최신 오픈 모델의 주류 구조입니다(우리 서버의 주력 모델 후보도 MoE입니다).
- **어려운 만큼만 생각한다 — 테스트타임 컴퓨트.** 모듈 2에서 본 thinking 모드가 이것입니다. 성능을 올리는 축이 "학습을 더 시키자"에서 "추론 시점에 더 생각하게 하자"로 이동했습니다. 쉬운 질문엔 짧게, 어려운 문제엔 길게 — 생각의 양 자체를 모델이 조절하는 방향으로 가고 있습니다.
- **스스로 검증하고 고친다 — 자기 피드백 루프.** 자기 출력을 재검토하고, 틀린 부분을 찾아 다시 시도하는 루프가 모델과 서비스 안으로 흡수되는 중입니다. 우리가 6월 데모에서 모델 바깥에 만들었던 "검증 → 재실행" 구조를, 업계는 모델 안으로 넣으려 하고 있는 겁니다.

세 흐름의 공통점을 보세요 — 전부 **추론 시점**의 이야기입니다. 모델의 미래가 이 방향인 한, 추론 인프라와 서빙 설계의 중요성은 계속 커집니다. 그리고 한 가지 단서: 모델이 루프를 내장하더라도, **"무엇이 우리 업무의 정답인가"를 정의하는 검증 게이트는 도메인 지식의 영역**으로 남습니다. 범용 자기검증은 모델이 해도, 우리 업무의 합격 기준은 우리만 압니다 — 하네싱 설계 역량이 사라지지 않는 이유입니다.

07 우리에게 의미 — 칩이 바뀌어도 산정 능력은 남는다

이 지형도에서 우리가 가져갈 결론은 세 가지입니다.

- **추론 특화 시대는 우리에게 유리합니다.** 우리가 하는 일(작은 모델을 효율적으로 서빙해 업무에 쓰는 것)이 바로 시장 전체가 이동하는 방향입니다.
- **NVIDIA 독주는 영원하지 않지만, 소프트웨어 생태계 확인 없는 대안 도입은 위험합니다.** 위의 4문항 체크리스트가 그 방어선입니다.
- **어떤 칩이 오든 판단 프레임은 동일합니다.** 메모리 용량·대역폭 확인 → 모델·양자화 산정 → 서빙 엔진 확인 → 골든셋 실측. 이 연구회에서 훈련하는 산정 능력은 GPU가 NPU로 바뀌어도 그대로 통합니다.

우리 연구회에선

공공 부문에 국산 NPU 실증이 확대되는 흐름 속에서, "그 장비로 어떤 모델을 어떻게 서빙할 수 있는지 산정할 줄 아는 실무자"는 기관 입장에서 드물고 귀한 사람이 됩니다. 이 모듈까지 소화했다면 여러분이 바로 그 후보입니다.

CHECK

이해했는지 확인해 봅시다

틀려도 괜찮습니다 — 오답을 고르면 왜 아닌지 설명이 나오고, 정답을 찾을 때까지 다시 고를 수 있습니다. 점수는 첫 시도 기준입니다.

문제 1 / 5

Q1. NVIDIA 독주의 가장 근본적인 원인은?

① 칩 제조 공정이 경쟁사보다 항상 한 세대 앞서서

② 20년 가까이 쌓인 CUDA 소프트웨어 생태계(도구·라이브러리·개발자)라는 성벽 때문에

③ HBM 메모리를 독점 공급받고 있어서

④ 특허로 경쟁사의 AI 칩 제조 자체를 막고 있어서

← 이전
모듈 6 · 평가

다음 →
자가진단 · 6레벨 질문