



주요 동향(2) : ICT

1 Hot Chips 2026: AI 반도체, 칩에서 시스템으로

- ➔ AI 반도체 경쟁력이 개별 칩의 속도보다 시스템 전체의 처리 효율로 결정
- Hot Chips 2026, AI 반도체의 핵심 의제가 연산칩에서 메모리·네트워크로 확대
 - Hot Chips 2026에는 반도체 기업과 대규모 AI 서비스를 운영하는 글로벌 빅테크가 참여해 GPU·CPU부터 메모리·네트워크·자동차용 SoC까지 다양한 신제품과 시스템 설계를 폭넓게 발표
 - 올해는 튜토리얼 발표 10건 중 6건이 메모리를 다루고 네트워크·인터커넥트도 별도 세션으로 편성되면서, 주요 논의 범위가 개별 연산칩에서 데이터 공급과 장치 연결을 포함한 시스템 구현 영역으로 확대
 - 이러한 의제 변화는 AI 시스템의 실제 성능이 연산칩의 처리 능력뿐 아니라 데이터를 제때 공급하고 장치 간 결과를 빠르게 전달하는 능력에 따라서도 크게 달라지는 최근 기술 경쟁 환경을 반영
 - AI 활용이 다양해지며 하나의 성능 기준으로 판단하기 어려워짐
 - AI 활용의 범위는 대규모 모델 학습과 실시간 추론 서비스를 넘어, 정보 검색·도구 호출을 연속 수행하는 에이전틱 AI와 로봇·자동차 등 물리 환경에서 작동하는 피지컬 AI까지 빠르게 확대
 - 대규모 학습은 가속기 간 연결과 확장 효율, 실시간 추론은 응답 지연과 메모리 대역폭, 에이전틱 AI는 전체 작업 시간과 수행 결과, 피지컬 AI는 제한된 전력과 일정한 처리 시간·안전성이 중요
 - 이처럼 워크로드마다 성능을 제약하는 병목과 우선순위가 달라지면서, 개별 연산칩의 최대 성능만으로 시스템 전체의 효율과 도입·운영 비용을 판단하기 어려운 상황
 - 주요 기업도 칩 단품보다 시스템 구성과 운영 효율을 앞세워 경쟁
 - 이에 따라 응답 지연과 동시 처리량, 전력 소모와 운영 비용을 실제 서비스 환경에서 함께 측정한 시스템 성능이 개별 칩의 제원보다 하드웨어의 실질적인 도입 효과를 판단하는 주요 기준으로 부상

- 주요 기업도 연산칩 단품보다 CPU·메모리·네트워크·소프트웨어를 결합한 시스템 단위로 제품을 소개하고, 전력당 처리량과 토큰당 비용 등 실제 서비스의 성능과 운영 효율을 보여주는 지표를 강조
- 결국 AI 반도체 경쟁은 개별 칩의 제원 향상 자체보다, 워크로드별 병목을 식별하고 구성요소의 성능을 조율해 실제 서비스의 처리 효율과 비용 절감으로 연결하는 역량에 따라 결정

➔ AI 활용 변화가 바꾼 반도체 설계와 경쟁 구도

1. (환경 변화) AI 활용 방식이 달라지며 필요한 반도체 성능도 세분화

● 대규모 AI 학습, 가속기를 늘릴수록 연결 효율이 성능을 좌우

- 대규모 AI 모델은 단일 가속기로 학습하기 어려워 수천 개의 가속기에 작업을 분산하며, 가속기 수가 늘어날수록 장치 간 계산 결과를 공유하고 학습 상태를 동기화하기 위한 통신량도 증가
- 이러한 통신 부담은 MoE 모델에서 더욱 두드러지며, 입력별로 일부 전문가 신경망만 사용해 연산량을 줄이는 대신 여러 가속기에 분산된 신경망 사이의 데이터 이동은 확대
- 따라서 대규모 학습의 성능은 개별 칩의 최대 연산량만으로 설명하기 어려우며, 많은 가속기를 안정적으로 연결해 장치 확대에 따른 성능 손실을 줄이는 확장 효율이 핵심 기준으로 부각

● 실시간 AI 서비스, 입력 처리와 답변 생성의 자원 요구가 분화

- 대규모 학습에서는 가속기 간 통신이 핵심 병목이었다면, 실시간 추론에서는 하나의 요청 안에서도 처리 단계(프리필, 디코드)에 따라 성능을 좌우하는 자원이 달라짐
- 구체적으로 프리필은 입력 문맥을 한꺼번에 계산하므로 병렬 연산 성능이 중요한 반면, 디코드는 모델을 반복해서 읽으며 답변을 순차적으로 생성하므로 메모리 대역폭이 처리 속도를 좌우
- 서비스 운영에서는 요청을 묶을수록 장치 활용률과 전체 처리량은 높아지지만, 이용자 대기시간도 길어질 수 있어, 추론 성능은 응답속도·처리량·토큰당 비용을 함께 고려해 판단

● 에이전틱·피지컬 AI, 빠른 연산을 넘어 작업 완료와 안전성까지 요구

- 실시간 추론이 개별 요청의 응답속도를 중시한다면, 에이전틱 AI는 정보 검색·도구 호출·결과 확인을 반복하므로 단계별 지연이 누적된 전체 작업 시간과 최종 성공 여부가 중요



- 한편 퍼지컬 AI는 제한된 전력·공간에서 인지·판단·제어 등 다양한 작업을 동시에 수행하므로, 작업 간 간섭을 차단하고 작업별 처리 시간을 일정하게 유지하는 설계가 안전성 확보의 전제
- 이처럼 워크로드마다 우선 지표가 확장 효율, 응답성, 작업 완수, 안전성으로 갈리면서, 가속기 설계도 하나의 범용 구조를 고도화하는 방식에서 병목별로 구조를 달리하는 방향으로 이동

2. (연산 구조) 학습·추론의 서로 다른 요구에 맞춰 AI 가속기 설계도 다변화

● GPU 경쟁, 개별 칩의 성능에서 랙 전체의 통합 효율로 확대

- 워크로드별 성능 목표가 세분화되자, GPU 진영은 개별 칩을 특정 용도에 맞춰 분리하기보다 기존 소프트웨어 생태계를 유지하면서 CPU·메모리·네트워크까지 결합한 랙 플랫폼으로 대응
- NVIDIA의 Vera Rubin NVL72와 AMD의 Helios는 각각 GPU 72개와 CPU·네트워크 칩을 한 랙에 묶어 대형 가속기처럼 운용해, 시스템 전체의 범용성과 운영 효율을 강화
- 이러한 랙 단위 통합은 연산뿐 아니라 전력, 냉각, 장애 복구까지 하나의 관리 체계로 묶어, 개별 GPU가 아닌 시스템 전체를 성능과 운영의 기본 단위로 설정

● 학습용은 대규모 연산, 서비스용은 빠른 응답에 맞춰 AI 칩 분화

- 다만 GPU를 랙 단위로 통합해도 학습과 서비스가 요구하는 연산량·메모리·응답 속도는 서로 달라, 기업들은 용도별 병목에 맞춰 연산·메모리·연결 구조가 다른 칩을 설계
- Google은 8세대 TPU를 대규모 학습용 8t와 다수 요청을 동시에 처리하는 추론·서비스용 8i로 분리해, 같은 세대에서도 연산·메모리·연결 구조를 용도별로 최적화
- 이는 학습은 대규모 연산과 가속기 확장성을, 서비스는 다수 요청의 동시 처리와 빠른 응답을 우선하면서 동일 세대의 AI 칩도 용도에 따라 내부 구조가 달라지고 있음을 의미

● 추론 서비스의 속도와 비용을 개선하는 전용 가속기 확대

- AI 추론은 문맥 처리·토큰 생성·데이터 이동 단계마다 성능을 제한하는 지점이 달라, GPU를 전면 대체하기보다 특정 구간의 지연과 데이터 이동을 줄이는 전용 가속기가 확대

- NVIDIA의 Groq 3 LPX는 응답 지연에 민감한 토큰 생성과 긴 문맥의 실시간 추론에 특화돼, 범용 GPU가 처리하던 추론 단계의 속도와 전력효율을 높이는 전용 가속기로 제시
 - Intel, SambaNova, Cerebras도 각각 도입 비용과 전력 부담, 반복 연산의 데이터 이동, 칩 외부 통신을 줄이기 위해 연산 회로·메모리 배치·칩 규모를 달리한 전용 가속기를 제시
 - 에이전틱 AI 확산으로 CPU의 데이터 처리·도구 실행 역할 확대
 - 처리 구간마다 다른 가속기가 배치되면서 CPU는 GPU에 데이터를 공급하는 보조 장치를 넘어, 작업 배정과 처리 결과의 연결을 뒷받침하는 시스템 조율 기반으로 역할 확대
 - 이에 NVIDIA는 자체 설계 코어를 탑재한 Vera CPU를 랙 플랫폼의 한 축으로 배치해, CPU를 GPU의 보조 부품이 아니라 에이전트 작업 완료 속도를 좌우하는 핵심 장치로 제시
 - Intel과 Arm도 단일 코어 성능과 함께 여러 작업의 동시 처리를 위한 코어 수, 가속기 연결을 위한 입출력 대역폭, 데이터를 공급하는 메모리 성능을 주요 지표로 제시
 - 가속기 분업 확대로 장치 간 작업 인계가 새로운 병목으로 부상
 - GPU·전용 가속기·CPU가 처리 단계를 나눠 맡으면 하나의 AI 작업이 여러 장치를 거치면서, 데이터와 중간 결과를 전달하고 개별 장치의 처리 시점을 맞추는 과정도 증가
 - 다만 처리 단계별로 특화된 장치를 사용하면 각 구간의 효율은 높아지지만, 장치가 바뀔 때마다 데이터와 중간 결과를 전달하고 다음 작업을 기다리는 시스템 부담도 함께 증가
 - 따라서 전체 처리 효율은 개별 가속기의 성능뿐 아니라 워크로드에 적합한 장치를 조합하고, 작업 인계 과정에서 데이터와 중간 결과를 지연 없이 전달하는 역량에 따라 결정
- 3. (데이터 이동) AI 연산이 빨라질수록 데이터를 제때 공급하고 옮기는 기술이 중요**
- 연산속도를 따라잡기 위해 HBM의 성능과 적층 기술을 함께 고도화
 - 가속기의 연산 성능은 빠르게 높아지는 반면, 메모리 대역폭은 이를 따라가지 못해, 필요한 데이터가 도착할 때까지 가속기가 대기하면서 장치 활용률과 전력효율이 떨어지는 문제가 확대



- 가속기와 가장 가까운 HBM도 적층 단수가 늘어나면 수직 연결부와 입출력 회로가 저장 공간을 차지하고 발열과 힘이 커져, 메모리를 계속 쌓아 용량과 대역폭을 높이는 방식이 한계에 접근
 - 이에 삼성전자는 베이스 다이에 데이터 처리·제어 기능을 배치해 불필요한 이동을 줄이고, SK하이닉스는 발열과 힘을 완화하는 패키징으로 HBM 추가 적층의 물리적 부담을 줄이는 방향을 제시
 - 데이터를 가속기로 옮기지 않고 메모리에서 직접 처리하는 기술 확대
 - 다만 HBM의 대역폭과 적층 수를 확대하더라도 데이터가 메모리와 프로세서 사이를 왕복하는 구조는 유지되므로, 단순 반복 작업에서는 계산보다 데이터 이동에 더 많은 시간과 전력이 소모
 - 이에 삼성전자는 LPDDR5X에 연산 회로를 결합한 PIM(Processing-in-Memory)을 공개해 반복 연산을 데이터가 저장된 위치에서 처리함으로써, 프로세서로 전송되는 데이터의 양과 왕복 횟수를 함께 축소
 - XCEN와 삼성전자가 공동 발표한 MX1은 CXL 확장 메모리에서 데이터를 미리 정리·선별하고 필요한 결과만 가속기로 보내, 데이터 이동의 해법을 전송속도 향상에서 전송량 축소로 확대
 - 가속기 수가 늘어날수록 장치 간 통신 속도가 시스템 성능을 좌우
 - 패키지 내부의 데이터 이동을 줄여도 여러 가속기가 하나의 모델을 나눠 처리하려면 중간 결과를 주고받고 처리 시점을 맞춰야 하므로, 가속기·서버 간 통신 지연이 성능을 제한
 - 가속기가 데이터 송수신과 보안·저장장치 접근까지 직접 처리하면 AI 연산에 사용할 자원이 줄어들기 때문에, 기업들은 데이터 입출력과 관리 업무를 전용 네트워크 칩으로 분리
 - 통신·보안·저장장치 접근을 전용 네트워크 칩에 맡기면 가속기는 AI 연산에 집중할 수 있지만, 가속기 수가 늘수록 연결 속도와 장애 대응력이 전체 성능을 좌우하므로 메모리·네트워크의 통합 설계가 중요
4. (경쟁 구도) AI 반도체 경쟁이 개별 칩에서 데이터센터 플랫폼 전체로 확대
- 자체 규격 중심 통합형과 개방형 표준 기반 플랫폼이 경쟁
 - 데이터 이동을 줄이는 과제는 연산칩 한 종류만으로 해결하기 어려워, 경쟁 단위도 개별 칩에서 CPU·메모리·네트워크·시스템 소프트웨어를 함께 최적화한 데이터센터 플랫폼으로 확대

- NVIDIA는 자체 GPU·CPU·네트워크와 NVLink를 결합한 통합 성능을 앞세우고, AMD는 공동 규격과 외부 부품을 조합하는 개방형 랙으로 구매자 선택권과 시스템 확장 유연성을 강조
- 자체 규격 중심의 통합형은 빠른 구축과 예측 가능한 성능에, 개방형 표준 기반 플랫폼은 공급선 다변화와 확장 유연성에 강점이 있어 운영환경에 따라 두 전략이 병행될 전망

● AI 서비스 기업도 자체 칩을 설계하며 플랫폼 경쟁에 참여

- 반도체 기업이 플랫폼 공급 범위를 넓히는 가운데 대규모 AI 서비스를 운영하는 빅테크도 내부 워크로드와 운영 데이터를 바탕으로 전용 칩을 직접 설계하며 플랫폼 경쟁의 주요 주체로 참여
- OpenAI는 Broadcom과 공동 개발한 첫 자체 설계 추론 칩 Jalapeño를 공개하고, 초기 설계부터 제조용 테이프아웃까지 약 9개월이 걸렸다고 밝혀 모델 기업의 빠른 칩 개발 가능성을 제시
- Microsoft와 Meta는 자체 칩을 핵심 AI 서비스에 확대 적용하고, Google은 용도별 TPU를 클라우드 고객에게 제공하면서 빅테크의 역할이 반도체 수요자에서 설계·운영 주체로 확대

● 칩을 연결하는 규격뿐 아니라 소프트웨어 호환성과 운영 비용이 성패를 결정

- 자체 규격 중심 플랫폼과 개방형 표준, 빅테크의 자체 칩이 공존하면서 패키지 내부의 다이부터 랙의 가속기까지 서로 다른 칩을 연결하는 규격의 주도권이 경쟁의 첫 관문으로 부상
- 그러나 물리적 연결 규격을 맞춰도 기존 AI 모델이 새로운 하드웨어에서 바로 작동하지는 않으므로, 모델의 연산을 각 칩에 맞는 명령으로 변환하는 컴파일러·런타임의 호환성이 플랫폼 채택을 좌우
- 전용 가속기는 AI 모델이 바뀔 때마다 소프트웨어를 다시 최적화해야 하므로, 성능과 전력 절감 효과가 개발 환경 전환과 유지·관리 비용보다 큰 시스템이 시장에서 선택될 가능성이 높음

➔ 시스템 경쟁에 대응해 국내 AI 반도체의 통합·검증 역량을 강화할 시점

● 국내 AI 반도체의 참여 영역이 메모리부터 자동차·저전력 AI까지 확대

- 삼성전자·SK하이닉스의 차세대 메모리에 이어 BOS반도체의 자동차용 시스템 온칩 Eagle-N, KAIST의 실물 에이전트용 저전력 공간추론 프로세서가 발표되며 국내 참여 분야가 응용별 연산칩으로 확대



- 완성 칩뿐 아니라 국내 팹리스·IP 기업도 추론 가속기와 메모리 서브시스템, 칩 간 연결 설계 자산을 선보이며 AI 시스템의 연산·메모리·인터커넥트 영역으로 참여를 확대
- 다만 글로벌 기업이 가속기·CPU·메모리·네트워크와 소프트웨어를 실수요에 맞춰 통합하는 반면, 국내에는 개별 기술을 하나의 시스템으로 묶을 주체가 부족하고 장기 운용·검증 경험도 제한적
- 공동설계·현장검증·수요 연계로 개별 기술을 경쟁력 있는 시스템으로 전환
 - (공동설계·소프트웨어) 자동차·로봇처럼 운영 조건과 요구 성능이 명확한 분야에서 모델·NPU·메모리·인터커넥트를 공동설계하고, 컴파일러·런타임을 확보해 국산 칩의 도입과 전환 비용을 절감
 - (실증·평가) 공동 설계한 시스템은 실제 서비스 조건에서 응답속도, 처리량, 전력, 비용과 장시간 가동 안정성을 검증해, 최대 연산 성능만으로 판단하기 어려운 도입 효과를 입증
 - (수요 연계) 검증이 일회성 실증에 그치지 않도록 수요기업이 개발 초기부터 목표 성능과 운영 조건을 제시하고 개선에 참여한 뒤, 성능이 확인되면 후속 구매로 이어지는 구조를 마련

출처 : Semiconductor Engineering 외(2026.8.)

<https://semiengineering.com/issues-stack-up-with-more-hbm-layers/>

<https://openai.com/ko-KR/index/jalapeno-first-results/>

<https://www.businesskorea.co.kr/news/articleView.html?idxno=275580>

<https://news.skhynix.com/en/tech-note-series-ep2/>

<https://www.mk.co.kr/en/business/12136672>

<https://www.mk.co.kr/news/it/12139820>

<https://it.donga.com/109414/>

<https://hotchips.org/>

<https://locsic.com/thinking/hotchips-2026-overview/>

<https://www.supercomputing.news/ai/hot-chips-2026-feeding-compute>

<https://www.thediligencestack.com/p/ais-inference-era-of-ferment>

<https://semiengineering.com/chip-industry-week-in-review-153/>

<https://www.techradar.com/pro/diamonds-crescents-and-wildcats-intel-shows-off-its-hardware-for-the-next-generation-of-agentic-ai-workloads>

<https://newsroom.intel.com/client-computing/intel-outlines-architectures-for-agentic-ai-at-hot-chips-2026>

<https://blogs.nvidia.com/blog/vera-rubin-lpx-spectrum-x-nvlink-fusion/>

<https://developer.nvidia.com/blog/how-nvidia-groq-3-lpx-unlocks-ultrafast-interactivity-at-long-context-on-nvidia-vera-rubin/>

<https://www.servethehome.com/amd-helios-mi400-system-architecture-at-hot-chips-2026/>

<https://techcommunity.microsoft.com/blog/azureinfrastructureblog/maia-200-software-defined-dataflow-and-all-ethernet-networking-for-efficient-inf/4548198>

<https://engineering.fb.com/2026/08/24/networking-traffic/mtia-300-meta-training-chip-built-in-nics/>

<https://www.cerebras.ai/blog/ultrafast-frontier-inference-cerebras-deep-dive-at-hot-chips-2026>

<https://newsletter.semianalysis.com/p/openai-jalapeno-better-than-nvidia>