2027년은 불연속적 AI 도약의 해, "지금 우리는 결정적 순간을 지나고 있다"

골드만삭스 AI 보고서, 1,000 Homes of Power in a Filing Cabinet - Rising Power Density Disrupts AI Infrastructure[1]

윤준영 태재미래전략연구원 선임연구원 (julia.yoon@fcinst.org)

AI 데이터센터는 하나의 산업
"완전히 새로운 공급망을 탄생시킬 것"

인공지능(AI) 기술의 급속한 발전과 함께 데이터센터의 전력 수요가 폭발적으로 증가하면서 기존 IT 인프라의 한계가 드러나고 있다. 불과 몇 년 전까지 단일 서버 랙(rack)이 소비하는 전력은 가정용 전자제품 수준이었지만, 이제는 수백 가구가 사용하는 전력을 소모하는 수준에 이르렀다. AI 모델의 고도화가 계속되면서 전력 소비는 더욱 가파르게 상승하고 있으며, 전문가들은 데이터센터가 더 이상 단순한 IT 시설이 아니라 하나의 산업이 되어가고 있다고 말한다. 초기에는 금보다 비쌌던 알루미늄이 결국 알루미늄 제련소와 같은 중화학 공업 시설 수준의 인프라로 변모하는 것에 비견할 수 있다는 것이다.

투자은행 골드만삭스의 글로벌 연구소(Goldman Sachs Global Institute)가 지난 5월 발표한 아티클 "1,000 Homes of Power in a Filing Cabinet – Rising Power Density Disrupts AI Infrastructure"은 데이터센터의 질적 변화를 통해 지금 우리가 어느 지점을 지나고 있는가를 보여준다. AI 담당 부사장인 프랭크 롱(Frank Long)은 보고서에서 "AI가 지배적 컴퓨팅 패러다임이 되면서 시장구조와 지정학을 뒤집을 수 있는, 맞춤형 인프라 지원을 위한 완전히 새로운 공급망을 탄생시킬 것"이라고 했다.

불과 몇 년 전만 해도 서버 랙 한 대가 쓰는 전력은 가정용 에어컨 몇 대 수준(5~15kW)에 불과했지만, 이제 단일 장비가 도시 한 구역이 쓸만한 전기를 먹어치우는 수준에 이르렀다. 2022년 ChatGPT 등장 시점의 AI 서버 랙은 20~40kW였고, 2024년 엔비디아(NVIDIA)의 '오베론(Oberon)' 시스템은 400kW까지 치솟았다. 2027년 등장을 예고한 엔비디아의 차세대 시스템 '카이버(Kyber)'는 576개의 GPU를 하나의 랙에 집적해 600kW를 요구할 전망이다. 심지어 업계 로드맵 상으로는 랙 당 1MW(1,000kW)까지 목표로 하고 있다.

이와 같은 극단적인 전력 소모 증가로 인해 업계에서는 "데이터센터가 더 이상 서버 보관 창고가 아니라, 제련소와 같은 산업 인프라로 변모하고 있다"고 진단한다. 실제로 알루미늄 제련소가 도시에서 대규모 발전소 인근으로 이전했던 것처럼, AI 데이터센터도 전력 공급이 풍부한 지역으로의 이전이 불가피해지고 있다는 것이다.

전통적 IT 인프라의 종언

데이터센터의 변화는 단순히 서버 성능의 고도화로만 설명되지 않는다. 현대 AI의 근본적 요구사항 때문이다. 지난 20년간 컴퓨팅 성능은 9만 배 증가했지만 데이터 전송 속도는 30배 향상에 그쳐, 심각한 병목 현상이 발생하고 있다.

1) https://www.goldmansachs.com/insights/articles/rising-power-density-disrupts-ai-infrastructure

현대 AI 모델은 수백 개의 GPU가 실시간으로 데이터를 주고받아야 하는데, 통신이 늦어지면 값비싼 GPU들이 데이터를 기다리며 유휴 상태에 빠진다. 이를 해결하는 유일한 방법은, 매우 단순한 얘기이긴 하지만 GPU를 물리적으로 최대한 가까이 배치하는 것이다. 수백 개의 GPU가 촘촘히 연결된 초고밀도 집적 구조는 바로 이런 필연성에서 나온다.

하지만 이는 전력 공급과 냉각 부담을 전례 없는 수준으로 끌어올리고 있다. 공기 냉각만으로는 감당이 불가능해지면서 액체 냉각, 전용 배관, 정교한 누수 감지 시스템까지 필수가 됐다. 전력 공급망 또한 가정용 전자기기 부품 수준에서 벗어나 전기차와 중공업 분야 수준의 설비가 요구된다. 바야흐로 전통적인 IT 인프라의 종언이다.

2025년, 결정의 해

특히 올해는 전환점으로 꼽힌다. 차세대 데이터센터 건설은 최소 18~24개월이 걸리는 만큼 오늘의 투자와 의사결정이 2030년대 AI 패권 경쟁의 지형을 좌우하게 된다는 전망에서다. 이에 기업은 전례 없는 리스크를 안고 역대 최대 규모 투자를 감수해야 하는 이중 부담에 직면했다. 목적 맞춤형 설계를 통해 전용화된 설비의 다른 용도 전환이 불가능해, 과거처럼 "AI 수요가 꺾이면 클라우드 서버로 전환한다"는 식의 안전망이 더는 존재하지 않기 때문이다. 이러한 리스크에도 불구하고 글로벌 기업들이 앞다퉈 대규모 자금을 투입하는 이유는 간단하다. 연산 자원 확장이 곧 AI 모델 성능의 기하급수적 향상으로 직결되는 만큼, AI 패권 경쟁에서 뒤처지는 순간 돌이킬 수 없는 격차가 벌어지기 때문이다. 일단 경쟁 스타트라인에 서기 위해 대규모 투자가 필수적이며, 기술 발전 단계상 바로 지금이 투자를 결정해야 할 결정적 시기라는 것이다. 골드만삭스는 "오늘의 자금 조달 결정이 향후 10년과 그 이후의 경쟁력을 결정할 것"이라고 표현했다.

초고밀도 아키텍처로 갈수록 연결 기술이 승부처

이 변화는 기업 전략에만 그치지 않고 지정학에도 큰 영향을 미칠 전망이다. 미국은 2022년 10월 이후 첨단 GPU와 함께 '인터커넥트'라는 칩 간 연결 기술까지 중국 수출 통제 대상으로 삼았다.[2] 단순히 연산 칩을 차단하는 수준을 넘어, 수백 개 GPU를 하나로 묶는 네트워크 자체를 차단하겠다는 의도다.

이에 엔비디아가 중국 전용으로 출시한 H800 칩은 기존 모델인 H100의 연산 성능은 유지했지만 칩 간 전송 속도인 '인터커넥트 대역폭'을 대폭 제한한 제품이었다. 이때 대역폭이란 단위 시간당 전송할 수 있는 데이터 용량으로, 도로 폭에 비유할 수 있다. 도로가 넓을수록 더 많은 차량이 동시에 오갈 수 있듯, 대역폭이 클수록 여러 GPU가 더 많은 데이터를 주고받을 수 있다.

현대 AI는 수백 개의 GPU를 하나의 시스템처럼 연결해 작동시키기 때문에 이 통신 능력이 제한되면 전체 성능이 크게 떨어진다. 아무리 개별 칩 성능이 뛰어나도 칩들 사이의 연결이 느리면 병목 현상이 발생해 시스템 전체가 비효율적으로 돌아가게 된다. 초기에는 중국이 이 제약을 우회하며 AI 모델 개발 속도를 유지했지만, 앞으로 초고밀도 서버 아키텍처가 본격화되면 격차는 눈에 띄게 벌어질 전망이다.

골드만삭스는 "AI 칩 자체보다 GPU를 묶는 연결 기술이 차세대 패권 경쟁의 관건"이라며 "수백 개 GPU를 유기적으로 연결하는 (차세대) 구조와 연결 속도가 제한된 소규모 GPU 클러스터의 성능 차이는 메울 수 없는 것"이라고 분석했다.

2) 2022년 10월 7일, 미 상무부 산하 산업보안국(BIS)은 중국을 겨냥한 고성능 AI 칩과 슈퍼컴퓨터 수출 통제를 시작했다. 중국이 첨단 AI 시스템을 구축하는 데 필요한 핵심 부품들을 차단하기 위한 조치였다. 이후 2023년 10월에는 이 조치를 더욱 정교화하여 총 처리 성능, 성능 밀도, 데이터센터용 설계 여부 등 다양한 기준을 추가 도입했다.

지리적 분포도 바꾸고 있는 데이터센터

게다가 전력과 냉각의 부담은 데이터센터의 지리적 분포도 바꾸고 있다. 현재까지 데이터센터 허브는 미국의 경우 버지니아 북부나 실리콘밸리에 집중돼 있었지만, 전력 수요가 폭증하면서 더 이상 감당할 수 없는 상황에 이르렀다. 향후에는 대규모 발전소와 가까운 지역, 에너지 공급 여력이 풍부한 신흥 지역이 새로운 거점으로 떠오를 가능성이 크다. 이는 디지털 인프라의 지도를 다시 그리며 새로운 지역에 경제적 기회를 창출할 수 있다.

골드만삭스는 "AI 경쟁은 이제 소프트웨어나 알고리즘이 아니라 전력과 냉각 인프라 확보가 승부처"라며 "2025년이 향후 10년간의 AI 리더십을 가르는 해가 될 것"이라고 강조했다. AI 인프라 전쟁은 이미 시작됐고, 이 전쟁은 기술 경쟁을 넘어 국가 안보와 산업 구조, 지정학적 균형까지 흔들고 있다.

TAEJAE's Insight

- AI 생태계의 3대 요소로 모델·데이터·인력을 꼽지만, 이 모든 것을 현실로 구현하는 기반이 데이터센터다. 데이터센터가 모델 학습, 데이터 저장·처리, 서비스 추론에 이르기까지 AI 구현의 전 과정을 물리적으로 뒷받침하기 때문이다. 그 과정에서 서버 랙의 고밀도 집적이 불가피해 산업 인프라에 버금가는 전력과 냉각을 요구하게 됐다. 이제 데이터센터는 AI 경쟁의 성패를 가르는 핵심 변수이자 가장 먼저 확보해야 할 전략 자산이 되었다.
- 이 같은 인식은 이미 세계 곳곳에서 정책으로 구체화되고 있다. 미국은 최근 대기오염 허가 절차인 청정대기법(Clean Air Act) 절차를 완화해 허가 전에도 데이터센터 건설을 시작할 수 있는 길을 열었다.[3] 중국은 2023년부터 세계 최초로 상업용 수중 데이터센터를 가동했으며, 올해에는 이를 해상풍력 발전과도 결합해 에너지 효율을 높일 계획이다.[4] 일본은 무리한 속도전보다 안정성을 중시해 데이터센터를 홋카이도 등 지방으로 분산해 재난 리스크와 부지 부족 문제를 동시에 풀어가려 한다.[5] 모두가 공통적으로 "데이터센터가 곧 에너지 인프라"라는 관점을 공유하며, 속도·효율·안정이라는 각기 다른 전략으로 대응 중이다.
- 그러나 지상에서의 한계가 명확하다 보니, 경쟁의 무대가 전력과 냉각 문제에서 자유로운 우주로 확장되고 있다. 미국 스타트업 스타클라우드는 구글과 손잡고 위성에 서버를 탑재해 생성형 AI를 궤도에서 가동하는 실험을 준비 중이다. 스타클라우드에 따르면 지상에서 40MW급 데이터센터를 10년간 운영하면 전력비만 1억 4천만 달러가 소요되지만, 궤도에서는 태양광 전력 덕분에 이를 200만 달러, 즉 현재의 1.4% 수준으로 줄일 수 있다. 냉각도 마찬가지다. 지상 데이터센터는 에너지의 40% 이상을 냉각에 쓰지만, 우주에서는 자연 복사 냉각을 통해 이 비용을 40~60% 절감할 수 있다. 발사·유지비 등 아직 변수는 남았지만, 전력·냉각이라는 지상 최대의 난제를 해결할 수 있다는 점에서 우주 데이터센터는 차세대 대안으로 부상하고 있다. 이를 입증하듯 론스타 데이터 홀딩스는 지난 3월 달 궤도에서 소형 데이터센터 운용 시험을 마쳤고,[6] 중국 ADA 스페이스는 2,800기 위성을 연결한 초대형 데이터센터 구상을 내놓으며 지난 5월 12기를 궤도에 올려 시범 운용에 들어갔다.[7]

3) Volcovici, V. (2025, September 10). Trump EPA seeks to speed up permitting for AI infrastructure. *Reuters.*
4) 한애란. (2025, July 19). 열나는 데이터센터, 바다에 빠뜨리자…美·中 이어 한국도 나선다 [딥다이브]. *동아일보.*
5) 김현재. (2024, December 26). 데이터센터의 지방 분산에 박차를 가하는 일본. *KOTRA 해외시장뉴스.*
6) Woollacott, E. (2025, April 10). The plans to put data centres in orbit and on the Moon. *BBC News.*
Lonestar Data Holdings, Inc. (2025, March 5). *Lunar data center achieves success en route to the Moon.*
7) 박찬. (2025, 5월 19일). 중국, '우주 AI 슈퍼컴퓨터' 위성 12기 발사…2800기 목표. *AI타임스.*

TAEJAE's Insight

- 반도체 기술 투자에 집중해온 한국은 차세대 전력·냉각 인프라 확보 없이는 한계에 직면할 것이다. 데이터센터가 국가 경쟁력의 새로운 전쟁터로 떠오른 만큼, 에너지 정책과 산업 정책을 연계한 종합적 접근이 필요하다.
- 기술은 빠르게 진화한다. 이를 따라잡을 금융 모델은 충분치 않다. 불확실성은 어느 때보다도 크다. 동시에 지정학적 공급망 경쟁을 따라잡아야 한다. 이런 모든 것들을 헤쳐 나가면서 투자하고 살아남아야 하는 시대다. 세계 모든 국가가 기업과 연계하게 된 이유다. 그렇게 한다고 해서 성공한다는 보장도 없다. 기술, 정책, 외교, 국가책략 모든 것을 총합한 원팀이 필수적이다.

초고밀도 AI 인프라 시대, 한국은 어떤 길을 선택해야 할지, 여러분의 의견을 들려주세요.