생성형 AI 생태계 기업전략과 반도체 역할

이선재 한국전자통신연구원 기술정책연구본부 선임연구원

1. 머리말

2022년 11월 말, 오픈에이아이(OpenAI)는 GPT-3.5를 기반으로 하는 대화형 인공지능 서비스 ChatGPT를 공개, 단 5일만에 100만 사용자를 확보하였다. 100만 사용자 확보 기준으로 아이폰이 74일, 페이스북이 10개월, 넷플릭스가 3.5년이 걸린 점을 고려하면 그야말로 열풍이라고 표현할 수 있다. 일반 사용자의 단순한 질문에 대한 대답부터 기술 및 산업에 대한 전문성 높은 주제에 이르기까지 활용 범위가 넓고, 프롬프트를 활용하는 사용 방법의 편리성으로 인하여 사용자는 계속 증가할 것으로 예상한다.

ChatGPT 이후의 AI 기술은 텍스트 기반의 이미지 생성, 비디오 및 3D 이미지 생성까지 가능한 멀티모달 AI 기술 개발 경쟁을 더욱 가속화시킬 것이다. 초거대 AI 모델 발전과 더불어, AI는 장기적으로는 사람처럼 인지·활동하고 새로운 환경에 적응·성장하는 인간 수준의 일반인공지능(GAI)으로 발전할 전망이다.

ChatGPT와 생성형 AI 열풍을 바라보면서 빅테크 및 반도체 기업들도 생존 전략을 수정하기 시작하였다. 반도체가 없이는 초거대 모델의 학습 및 클라우드를 통한 서비스를 구동할 수 없으므로 대규모 데이터 연산을 빠르고 효율적으로 실행하는 새로운 차원의 컴퓨팅 하드웨어인 AI 반도체가 필수적이다.

ChatGPT로 촉발된 빅테크 기업들의 초거대 AI 모델 경쟁은 파라미터의 지속적인 증가를 가져왔으며, 필연적으로 AI 반도체나 클라우드 등 컴퓨팅 자원의 중요성이 함께 강조되고 있다. 이에 본고에서는 생성형 AI 생태계 내 반도체의 역할과 관련 기업 동향에 대해 알아보고자 한다.

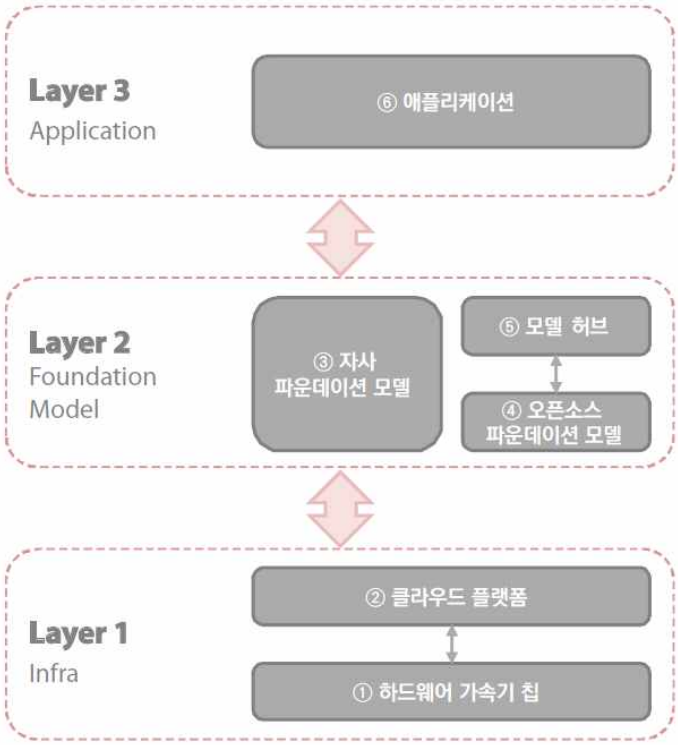
2. 생성형 AI 생태계 구조와 전개 방향

2.1 생성형 AI 생태계 구조

생성형 AI Tech Stack은 [그림 1]과 같이 크게 Infrastructure layer(이하 인프라), Foundation Model layer(이하 파운데이션 모델), Application layer(이하 애플리케이션) 등 3개로 나눌 수 있다. 인프라는 고성능 반도체 및 데이터센터 통신을 기반으로 클라우드에서 하이퍼스케일 컴퓨팅 파워를 제공한다. AI 데이터 처리용 반도체 등과 같은 컴퓨팅 하드웨어 제조사와 사용자가 컴퓨팅파워를 가상으로 빌려서 사용할 수 있는 클라우드 서비스 제공 기업을 포함한다.

파운데이션 모델은 애플리케이션에 사용되는 트랜스포머 기반의 초거대 언어 모델(LLM)로서 다양한 태스크에 적용 가능한 범용적 플랫폼이다. 파운데이션 모델을 생성, 호스팅, 운영하는 기

업은 LLM을 폐쇄형으로 만들어 독점 API를 통해 공급하는 Proprietary 모델과 LLM을 오픈소스로 만들어 모델 허브 호스팅을 통해 배포하는 Open-source 모델로 구분할 수 있다.
애플리케이션은 인공지능 모델(Layer 2)을 활용하여 사용자와 가장 밀접하게 상호작용하는 앱 형태의 인공지능 서비스이다. 자체 모델 파이프라인(End-to-End Apps)을 구축하거나, 타사 API를 자체 AI 제품에 통합하여 제공하는 서비스 방식이 있다.



출처 : https://a16z.com

[그림 1] 생성형 AI Tech Stack

2.2 생성형 AI Tech Stack별 전개 방향

2.2.1 Layer 1 - 컴퓨팅 인프라

대규모 설비투자를 통해 하이퍼스케일 데이터센터를 확충하고, 초거대 AI 특화형 반도체 개발을 통해 학습용 컴퓨팅 기반을 최적화한다. 1차적으로 자체 파운데이션 모델 확보를 위해 학습을 위한 AI 반도체 및 클라우드 역량에 집중하고, 2차적으로 '잠금(Lock-in) 전략'을 통해 진영을 구축해 생태계로 확장할 전망이다. 컴퓨팅 인프라 분야는 대규모 투자를 요하는 분야로서, 마이크로소프트, 구글(알파벳), 아마존, 엔비디아 같은 빅테크 기업이 주도하고 있다.

2.2.2 Layer 2 - 파운데이션 모델

초기에는 범용성을 지향하는 초거대 매개변수 확보 경쟁이 전개되고, 이를 바탕으로 특수 목적성에 부합하는 경량 파운데이션 모델의 등장이 전망된다. 기업의 기술 수준 및 수익 구조에 따라 Proprietary 또는 Open-Source 기반 파운데이션을 확보하고, 웹 또는 검색엔진 등과 결합·연계되면서 실시간 데이터 정보의 인프라화가 전개될 전망이다.

2.2.3 Layer 3 - 애플리케이션

초기에는 빅테크 전략 중심의 폐쇄형 통합 모델로 전개되다가, 점진적으로 스타트업이 진입하면서 비즈 모형 중심의 개방형 모델로 전개될 전망이다. 다양한 생성형 AI 앱이 난립하는 가운데 언어(텍스트) 중심 AI에서 멀티모달 중심 AI로 진화하고, 적용 분야가 전 산업으로 확대될 것이다.

3. 생성형 AI 기업의 비즈 모형과 반도체 역할

3.1 생성형 AI 생태계 기업의 방향성

생성형 AI 서비스는 결국 컴퓨팅 인프라+파운데이션 모델+애플리케이션의 최적화가 필요하다. 생성형 AI 붐으로 관련 기업은 <표 1>과 같이 자사의 강점을 기반으로 생성형 AI를 접목하는 수직 계열화 전략의 방향성을 가지고, 기존 사업 영역의 확장 또는 강화를 도모하고 있다. 기업들의 전략 방향은 자체 개발 또는 다른 기업과의 제휴를 통해 하드웨어부터 애플리케이션까지 모든 단계의 제품 및 서비스를 제공하는 풀스택(Fullstack) 확보이다.

<표 1> 생성형 AI 기업 풀스택 구성도

Layer		MicroSoft	OpenAI	Google	Amazon	Nvidia
3	Apps		ChatGPT	BARD AI		
2	Foundation Model	MT-NLG	OpenAI ChatGPT 4.0	LaMDA	Cohere Stability AI	NVIDIA NEMO
1	Cloud Platform	Microsoft Azure		Google Cloud	aws	엔비디아 DGX
	Compute Hardware	GPU	GPU	GPU, TPU	GPU, NPU	GPU

특히, 기업들은 대규모 연산을 최적화하기 위한 하드웨어의 필요성을 인지하고 개발에 집중하고 있다. 구글은 자사의 기계학습 맞춤형 가속기인 TPU를 생산하여 AI챗봇 바드(BARD)의 언어 모델 람다(LaMDA) 학습 및 서비스에 활용 중이다. 아마존은 클라우드 1위를 위한 전성비 특화 반도체에 집중하면서 '그래비톤 칩'으로 구성된 데이터 센터를 구축·운영하고 있다. 그래비톤 칩은 모바일 기기용 반도체에 특화된 ARM의 설계를 기반으로 만들어져 전성비가 높은 것이 특징이다. 마이크로소프트(이하 MS) 역시 'Singularity' 프로젝트(2022년)를 통해 AI 전용 데이터센터 구축을 목표로 하고 있으며, 이는 AI 가속기 직접 설계 및 운용 등의 내용을 포함하고 있다. 엔비디아는 강점인 대용량 컴퓨팅 파워를 클라우드 서비스로 제공하면서 시장 대응에 나서고 있다.

다음에서는 생성형 AI 시장에서 생존하기 위한 빅테크 기업의 전략을 자세히 알아보고 비즈 모형 내에서 반도체의 역할을 확인하고자 한다.

3.2 구글

생성형 AI의 풀스택 Tech(TPU-구글클라우드 플랫폼-LaMDA-Bard AI)를 기반으로 인프라를 고도화하고, 파운데이션 모델 및 앱까지 수직 계열화 전략을 구사한다. 맞춤화를 통해 주 수익원인 검색엔진 시장의 지위를 유지하고, AI로 웹 기반 생산성 도구인 Google Workspace를 개선하여 시장을 공략하고자 한다. 또한, 웹기반 실시간 검색 정보의 인프라화를 전개하고 있다.

구글은 2023년 2월 크롬에 자사 언어모델인 LaMDA 기반의 챗봇 바드를 탑재하여 대화형 검색 서비스를 개발하였으며, 바드를 검색엔진에 연동하여 실시간 검색정보의 데이터 인프라화를 추진하고 있다. 바드가 검색 엔진과 통합되면 구글의 검색 광고 수익에 악영향을 미칠 것을 고려하여 분리형 구조로 서비스를 제공하면서 검색시장의 자기잠식을 최소화하는 전략을 구사한다. 텍스트와 이미지 생성이 가능한 TPU 기반의 생성형 AI 특화 모델을 제공하면서 구글 클라우드의 차별화를 시도하고 있다. 구글 클라우드의 머신러닝 플랫폼 '버텍스 AI'가 생성형 AI를 지원하며, 생성형 AI 앱 빌더를 통해 수시간 내에 생성형 AI 애플리케이션을 구축할 수 있다.

3.3 아마존

생성형 AI 적용의 후발주자이지만, 기존 클라우드 시장의 강자로서 LLM 개발 기업과 협력해 사용자에게 낮은 비용으로 높은 성능의 생성형 AI 개발 서비스를 지원하는 전략을 구사한다.

BLOOM 기반의 개방형 파운데이션 모델 확보로 생성형 AI 생태계를 활성화하고, AI 기반 서비스 확대를 통해 클라우드 시장 1위를 수성하는 것이 목표이다. 또한 생성형 AI 친화형 서비스를 위하여 인공지능 반도체도 적용하고자 한다.

AWS는 AI 스타트업 허깅페이스와 장기 전략적 파트너십을 맺고 차세대 AI 개발에 협력('23.2.)한다. 허깅페이스의 소프트웨어 개발자는 아마존의 클라우드 컴퓨팅 파워와 트라니움, 인퍼런시아, 그래비톤 등 AI 작업용으로 설계된 아마존의 칩을 사용하는 등 주로 기업 간 연합 형태의 전략을 추진하고 있다. 이는 MS가 GPT를 활용하여 비즈니스 모델을 개선할 때 수행한 전략(자사 클라우드 + 타사 GPT 서비스)과 유사하다.

3.4 마이크로소프트

생산성 도구의 고도화 및 생성형 AI에 최적화된 서비스 제공으로 클라우드 시장 1위에 올라서고자 MS의 강점인 업무 생산성 도구 전반에 초거대 생성형 AI를 접목하고, 이를 클라우드 플랫폼에 얹어 서비스하는 전략을 구사한다. 특히 ChatGPT 기술을 접목하여 가장 먼저 서비스함으로써 선점 효과를 가져올 수 있었다. ChatGPT 기능을 Microsoft 365에 탑재('23.3.)한 Microsoft 365 Copilot은 Microsoft 365 앱에 내장되어 편의성과 활용성을 제고하면서, 생산성 도구를 고도화함으로써 시장을 공략한다. 또한, 오픈AI가 개발한 멀티모달 기능을 갖춘 'GPT-4'를 새로운 '빙' 검색에 도입('23.3.)하였으며, 검색 창에 질문을 입력하거나 별도로 마련된 채팅 화면에서 AI와 바로 대화하는 방식을 채택하며 사용 편이성을 확보하였다.

3.5 엔비디아

슈퍼컴퓨터 기반의 클라우드 서비스를 구독 형태로 제공함으로써 막대한 컴퓨팅 자원을 기반으

로 생성형 AI 생태계를 공략하고 있다. 아직 대부분의 AI는 GPU를 기반으로 돌아가고 있으므로, CUDA 기반 생태계 잠금효과를 통해 확고한 시장 지배력을 유지하고자 하는 전략이다. 자체 행사인 2023 GTC에서 발표한 DGX 클라우드는 컴퓨팅 파워를 클라우드 기반으로 제공하는 서비스로, 수억 달러가 투자되는 구축형(on premise) 인프라를 마련하는 대신 필요한 만큼 임대하는 새로운 방식을 통해 클라우드 시장에서 강점을 가질 것으로 예상된다. 또한, GDX 클라우드를 기반으로 생성형 AI 모델 학습이 가능한 '엔비디아 AI 파운데이션'을 통해 생성형 AI생태계를 공략하고 있다. AI 언어 모델 구축 서비스인 니모 모델도 제공한다. 기업은 자사 언어 데이터를 니모에 입력해 언어 모델을 개발할 수 있다. 피카소 모델은 이미지를 비롯한 비디오, 3D생성 모델을 구축하는 클라우드 서비스로, 텍스트를 이미지나 비디오, 3D로 변환해 준다.

4. 맺음말

컴퓨팅 인프라를 기반으로 혁신을 이루고 있는 AI는 시각·언어·음성과 같은 단일지능에서 트랜스포머를 기반으로 한 초거대 AI로 발전하고 있다. ChatGPT의 기반인 GPT 모델에서와 같이 수천억 개의 파라미터로 구성된 대규모 연산이 요구되고 있으며, 이에 따라 인공지능 연산을 위한 하드웨어인 AI 반도체 필요성이 증가하고 있다. ChatGPT로 촉발된 생성형 AI 열풍은 AI 반도체 수요 증가를 가속화할 것이다. 빅테크 기업들은 기술 수준 및 역량에 따라 자체 칩을 개발하거나 기업 간 연합을 통해 풀스택을 확보하고 생성형 AI 생태계를 구성하고자 노력 중이다.

그러나 아직도 대부분의 AI는 AI 알고리즘 처리를 목적으로 개발되지 않은 그래픽 처리 장치인 GPU 기반으로 작동하고 있는 실정이다. 효율적인 AI 알고리즘 처리를 위해 세계 각국의 기업은 AI 가속기 개발에 뛰어들었으나 AI 반도체 시장은 지배적인 사업자가 나타나지 않은 초기 시장에 머물러 있다.

대기업 및 스타트업은 생성형 AI 시장에서 각자의 한계를 보완하면서 장점을 최대한 활용하는 전략이 필요하다. 하드웨어, 소프트웨어, 서비스기업 간 장벽을 허물고 서로 협력해야 하며, 스타트업의 인력 확보 노력도 필요하다. 정부는 기업들이 공동으로 활용 가능한 컴퓨팅 자원을 지원하고, 수요처 제공에도 노력해야 한다. 국내 AI 반도체 산업의 경쟁력 확보를 위해서는 개별 기업의 NPU 개발 노력도 중요하지만, 수요처에서 확보된 다양한 레퍼런스가 필요하기 때문이다.

생성형 AI 붐은 국내 AI 반도체 산업의 경쟁력 확보에 분명히 좋은 기회가 될 것이다. 현재의 대기업 중심 캐치업 전략을 벗어나, 산학연 연구 저변 확대 및 공동연구 환경 지원을 통해 원천기술을 확보하고 기술 선도형으로 전환해야 할 것이다. 이를 통해 시장을 선도하는 AI 반도체 기업으로 성장한다면 AI 생태계와 반도체 생태계의 선순환을 통해 국내 AI 반도체 생태계의 지속가능성을 찾을 수 있을 것이다.

※ 본 연구는 한국전자통신연구원 연구운영지원사업의 일환으로 수행됨. [23ZR1400, 국가지능화 R&D 경쟁력 제고를 위한 기술정책 연구]

[주요 용어 풀이]

- LLM : Large Language Models
- GCP : Google Cloud Platform
- GPT : Generative Pre-trained Transformer
- TPU : Tensor Processing Unit

[참고문헌]

[1] 2023 인공지능 반도체 (KISTEP, 2023)
[2] AI와 AI 반도체 기술동향 (TTA, 2022)
[3] 기업 홈페이지(구글, MS, 엔비디아, 아마존)
[4] Generative AI의 tech stack, Daewoo Kim, https://moon-walker.medium.com
[5] Who Owns the Generative AI Platform?, Matt Bornstein et.al, https://a16z.com/

※ 출처: TTA 저널 제207호