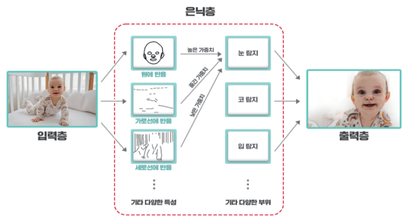
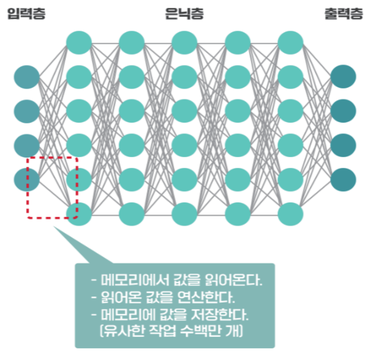
# 최신 인공지능 기술에서의 메모리 수요




**인공지능** 여러 계층의 신경 세포망으로 구성.

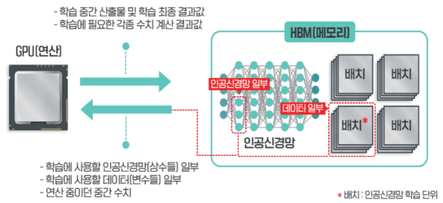
**학습** 세포망의 이전 단계의 값을 동시에 가져와 연산하고, 이를 다음 단계에 해당하는 각각의 신경 세포망에 저장하는 과정

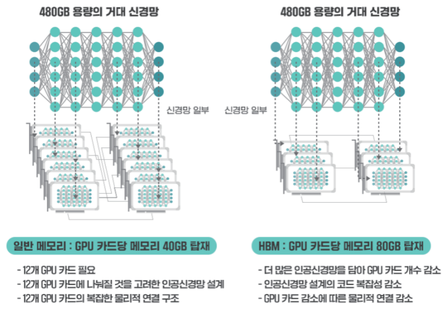
**추론** 문제해결 행위, 입력으로부터 최종 단계의 출력을 내놓는 과정으로, 이러한 추론 과정에서 각 세포마다 연결되는 강도를 의미하는 가중치 정보가 요구된다.

- 신경 세포망의 계층이 깊을수록, 즉 신경망의 크기가 클수록. 그리고 각 세포에 더 많은 데이터가 학습(저장) 되있을 수록 정확한 추론 가능
- 따라서, 학습 및 추론 과정에서 각 신경세포에 데이터가 저장되야 하므로 **많은 메모리 용량**이 요구됨.
- 또한, 각각의 신경세포에서 동시에 여러 개의 데이터를 불러들여 동시 연산이 이뤄져야 하므로, **높은 대역폭**과 병렬 연산처리 요구됨.

**참고 문헌** : https://news.skhynix.co.kr/post/jeonginseong-column-ai-1

# 인공지능 기술 구현 시 HBM3 제품의 역할



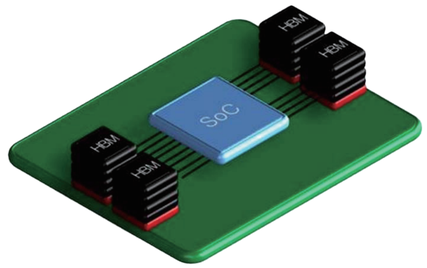

학습 과정에선 여러 입력에 대한 출력을 비교하고, 오답이 나타나면 신경망의 가중치를 조정하여 정답이 나오게 된다. 이러한 과정에서 세포 사이의 연결 강도는 수십만번 이상 변화

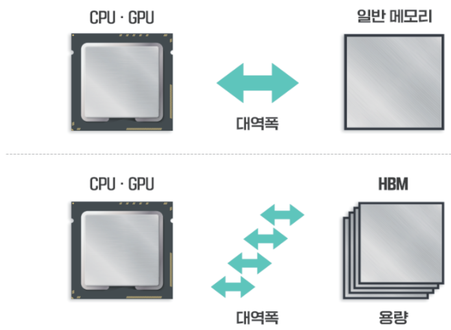
**HBM3의 역할 1** 메모리에는 인공 신경망 각 루트에서의 가중치 값, 각 신경세포에 저장되는 값이 모두 저장된다. 이 값을 가지고 GPU가 연산을 하여 결과값을 내놓으면 이 값은 다음 신경 세포에 저장되는데 이 또한 메모리에 저장 되는 값이다. 따라서 HBM3은 인공지능 구현 시 메모리 용량이 커서 이러한 인공신경망을 구현하기에 적합하다.

**HBM3의 역할 2** 인공지능 연산 과정에서 동시 다발적으로 메모리로부터 많은 데이터를 GPU와 주고 받아야 하므로, 높은 대역폭이 요구된다. HBM3는 높은 대역폭을 가지므로 이와 같은 인공지능 연산이 가능하게 한다.

**참고 문헌** : https://news.skhynix.co.kr/post/jeonginseong-column-ai-2

# SK Hynix HBM3 제품의 특징




**메모리 용량** 적층 구조를 통해 면적당 높은 용량의 확보

**높은 대역폭** 메모리가 오고 가는 통로의 수를 늘려 정해진 시간 동안 많은 데이터의 처리가 가능. 즉, 높은 대역폭의 확보

**GPU를 이용** GPU는 CPU와 달리 여러 작업을, 여러 연산을 동시에 진행하는데 최적화 된 Processor이다. 실제로 인공지능의 구현 과정에서 GPU를 쓰면 CPU대비 5배의 성능 향상이 나타남.

**참고 문헌1** : https://news.skhynix.co.kr/post/jeonginseong-column-ai-2
**참고 문헌2** : https://www.skcareersjournal.com/1282

# SK Hynix HBM3의 응용

사진 1

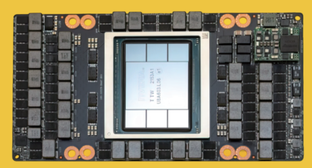


사진 2


사진 3


**사진 1** Nvidia의 GH100 보드에 사용된 모습

**사진 2** 이러한 보드를 이용하여 인공지능 및 데이터센터의 구현을 위한 HW 셋업 되있는 모습, 실제로 자사 HBM3 제품은 생성형 AI인 GPT-4의 기능 구현에 이용되고 있음.

**사진 3** 인공지능 및 데이터센터의 수요 증가로 인해 HBM3은 점점 시장 수요가 늘고, 이에 따라 더욱 주목 받고있는 제품입니다.

**notebook 이용하여 시연 예정** : MSFT bing image creator(GPT4 응용한 생성형 AI의 대표적 예시)

**참고 문헌1** : https://wccftech.com/nvidia-hopper-h100-gpu-pictured-worlds-first-4nm-hbm3-chip-for-datacenters/
**참고 문헌2** : https://www.servethehome.com/chatgpt-hardware-a-look-at-8x-nvidia-a100-systems-powering-the-tool-openai-microsoft-azure-supermicro-inspur-asus-dell-gigabyte/
**참고 문헌3** : https://www.sedaily.com/NewsView/29ODKRIOSV