자료분석학회 동계학술대회
2024년 12월 6일, 충남대학교

데이타, 인공지능, 시스템

- AI 시대 살아남기 -

원유집
카이스트

제 0 부.

소개

- 원유집
- KAIST 전기전자공학부 ICT 석좌교수
- 카이스트 스토리지연구센터장

- 운영체제, 시스템 소프트웨어, 분산시스템

- 서울대학교 계산통계학과 학사 (1990), 석사 (1992)
- 미네소타 주립대학교 박사 (1997)

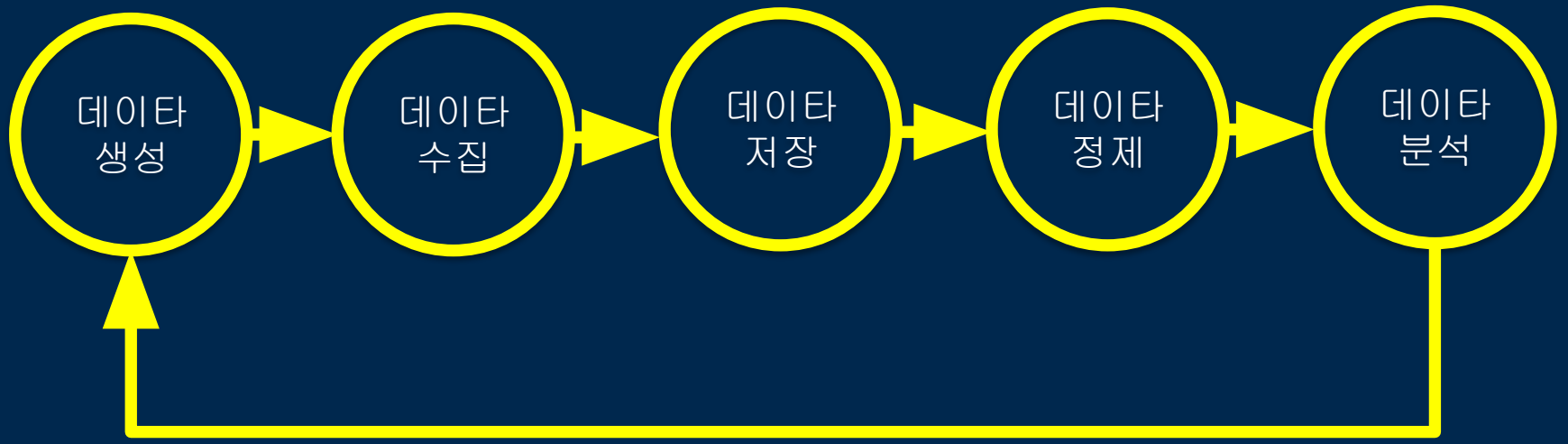
- 한국정보과학 회장 (2024.1 – 2024.12)
- 공학한림원 일반회원, 삼성전자 자문교수, LG전자 자문교수,
- 외교부 과기외교자문위원 (2023-2024)
- Program Chair(USENIX FAST 2024)
- General Chair (ACM SOSP 2025),
- 연구재단 전문위원(RB): 기초연구본부(2010-2012), 국책연구본부(2018-2022)

- 한양대학교 컴퓨터 공학부 교수 (1999- 2019), 인텔(Intel)사 연구원 (1997-1999)
- Senior Associate Editor, ACM Trans. On Storage

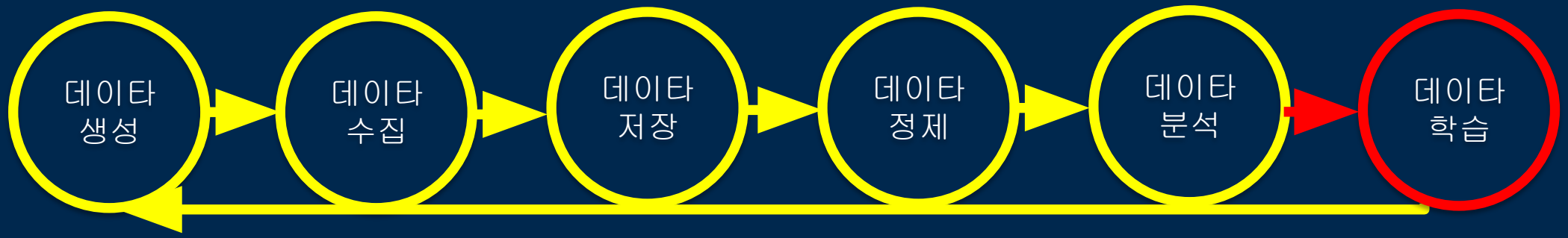
제 1 부. 데이타

데이터를 갖는자 세계를 얻는다.

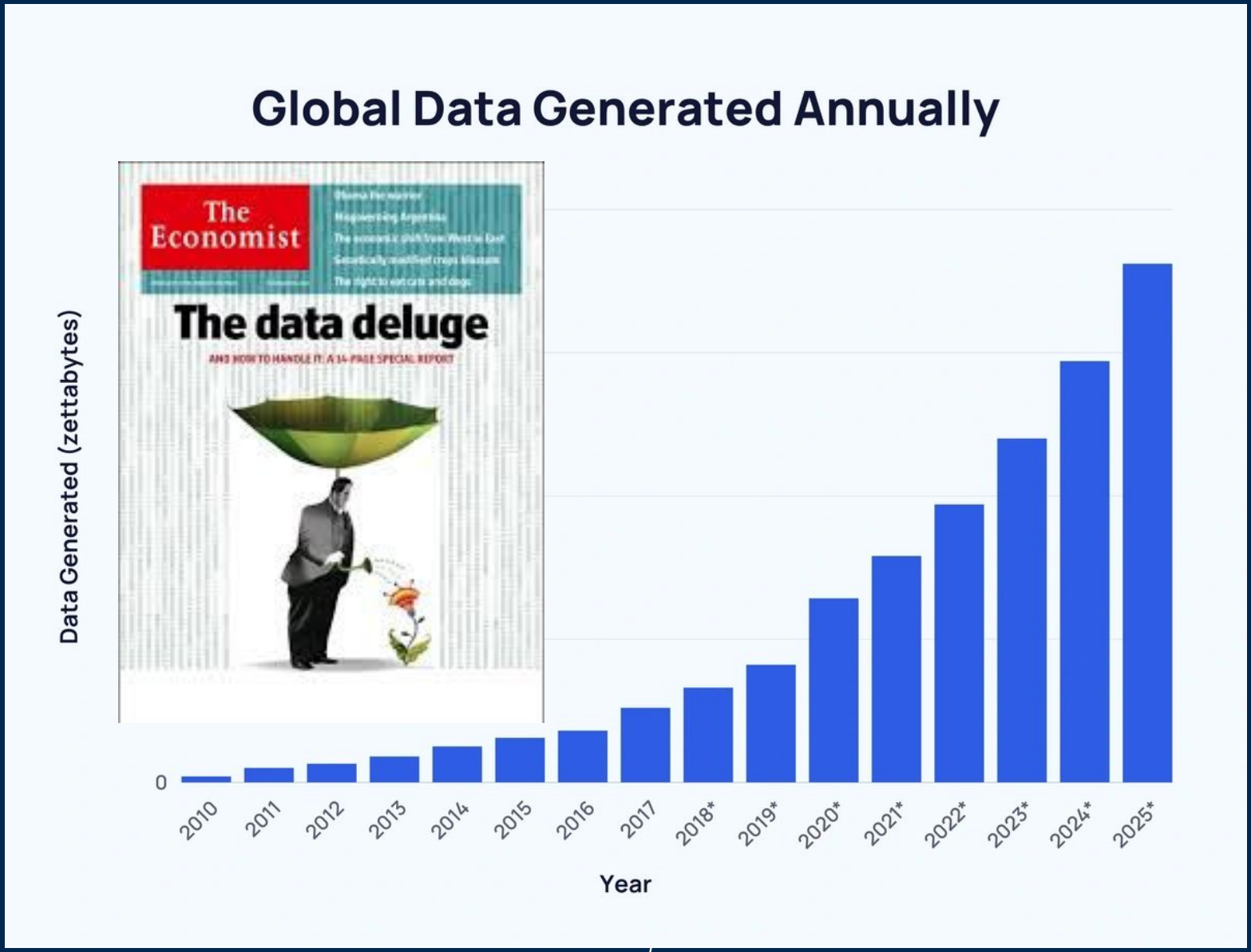
데이타 생명주기 (AI 이전)



데이타 생명주기 (AI이후)



데이터 홍수


Global Data Generated Annually
The Economist
The data deluge
Data Generated (zettabytes)
0
2010
2011
2012
2013
2014
2015
2016
2017
2018*
2019*
2020*
2021*
2022*
2023*
2024*
2025*
Year

일일 데이타 생성량 (2024)

- 402.74 million TB 가 생성됨. (노트북 4억대)
- 인터넷 트래픽중 이 절반 이상이 비디오
- 3330 억개의 이메일이 전송됨.
  - (세계인구: 82억명, 30억명이 일인당 매일 이메일 100통)
- The US has over 2,700 data centers

- 2024년에 149 zettabytes 가 생성 됨.
- 2025년에 181 zettabytes 가 생성될 것으로 예측

Connected Car

스마트 공장

데이터에...

데이타를...

데이타 생명주기

Trifacta, Alteryx, Datarobot,
Hyperconnect, Talend 등…

$$$

15%만 저장

5%만 분석

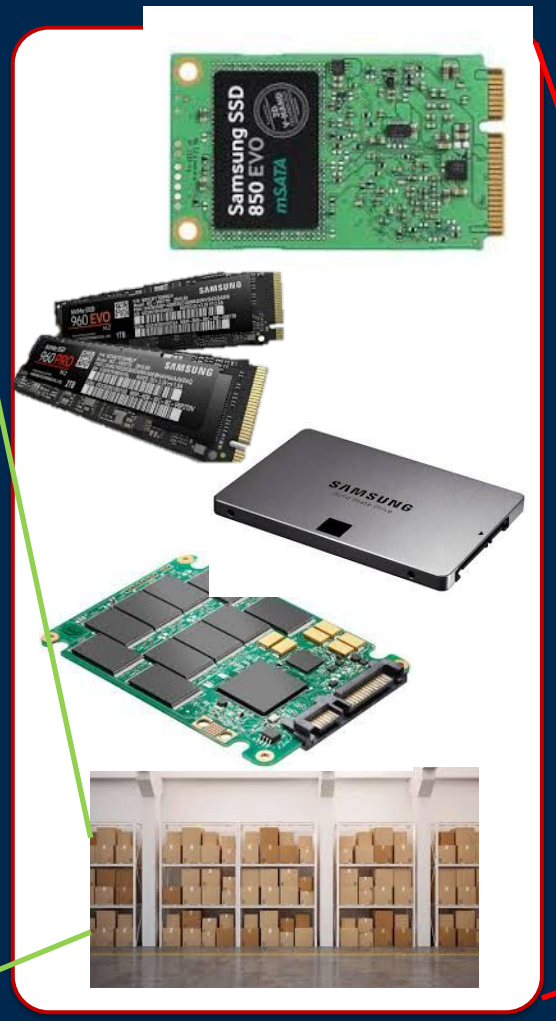

Market Summary > Palantir Technologies Inc
71.87 USD
+ Follow
+53.57 (292.73%) ↑ past year
Closed: Dec 5, 5:11 PM EST • Disclaimer
After hours 71.99 +0.12 (0.17%)
1D
5D
1M
6M
YTD
1Y
5Y
Max
21.62 USD 20 May 2024
80
60
40
20
0

제 2 부

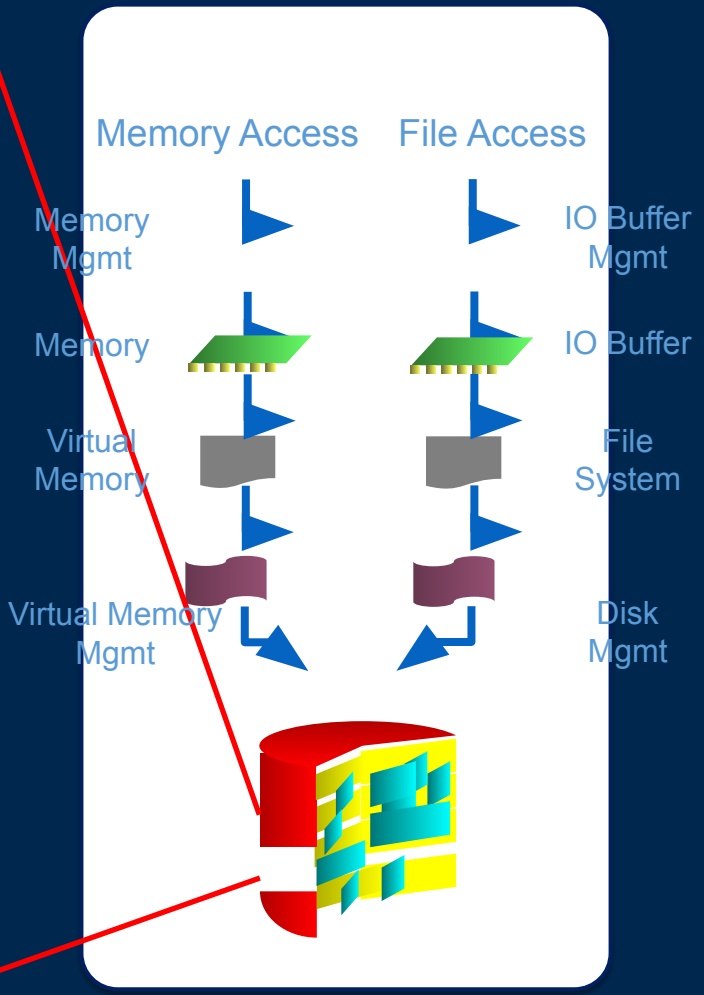
데이타의 발전사: 역사를 알면 미래가 보인다.

데이타, 저장장치, 컴퓨터 시스템

데이타



저장장치



시스템

기술 생태계

데이타

학습

사물 인터넷

AR/VR

BIO

자율주행

딥러닝

분석

key	value
firstName	Bugs
lastName	Bunny
location	Earth

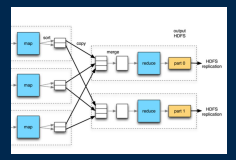
riak

Cassandra

CouchDB relax

mongoDB

축적



amazon web services™

Google bigtable

hadoop

UFS Universal Flash Storage

nvm EXPRESS

시스템

기술 메가트랜드

@1992 **PC** 시대

PERSONAL COMPUTERS

In last 9 years, PCs have undergone major changes and changed our lives

December 27, 1993 | By PETER H LEWIS

Napster, BMG in music pact

October 31, 2000: 4:49 p.m. ET

Controversial online music-sharing site mends fences with Bertelsmann

@1997 **네트웍** 시대
(미국 부통령 알고어, 텔레커뮤니케이션 빌, Information Super Freeway)

@2003 **모바일 기기**의 등장
데이터 축적 가속화, **빅데이터**의 시작

PRESS RELEASE
OCTOBER 23, 2001

Apple Presents iPod

iRiver offers 'Toblerone' 512MB MP3 player

Plus NEC's Centrino-style S800, VIA's latest mini-ITX board and more

By Tony Smith 16 Apr 2003 at 12:10 SHARE ▼

@2007 **스마트폰** 등장 (아이폰, 안드로이드)
데이타베이스

First iPhone ads published on Apple.com (Coming June 29)

By AppleInsider Staff
Sunday, June 03, 2007, 04:15 pm PT (07:15 pm ET)

…

First Android Phone: "T-Mobile G1 With Google"

Greg Sterling on September 23, 2008 at 10:48 am

…

@2016 알파고가 이세돌에 승리
인공지능

Artificial intelligence (AI)

AlphaGo seals 4-1 victory over Go grandmaster Lee Sedol

@2020 Google GPT3 **생성형 인공지능**

@2024 **인공지능용 데이터 센터, 에너지 장벽**

좀더 자세히: 빅데이타와 인공지능

Douglas Laney
VP Distinguished Analyst

6 years at Gartner , 30 years industry experience
Chicago, IL USA

@2001: 빅데이타의 개념등장 (Volume, Velocity, Variability)

@2004: 초대용량 데이타를 작은 컴퓨터를 연결하여 처리

Map Reduce, Jeff Dean et.al. OSDI 2004

@2006: 빅데이타 시스템 상용화: **Hadoop**

Doug Cutting and Mike Caferella (20조시장)

@2006: 구글 빅데이타 시스템

(google finanace, google earth, Orculus, Youtube)

Fay Chang et.al. OSDI 2006

@2008: 일일 처리량 폭발적 신장 (일일 **20 PB** 처리)

@2011: 구글 고성능 데이타 분석 **SW** 개발 **(levelDB)**

@2016: 알파고가 이세돌에 압승

Jeff Dean, Tensorflow, OSDI 2016

요소기술 측면에서 조명한 기술 메가 트렌드

.....

→ 인공지능 (80년대 말)

→시스템 (90년대 초)

→네트워크 (90년대 말)

→데이타베이스 (2010년)

→인공지능 (2015년)

→ (???)

인공지능의 새로운 이슈: 데이타센터

NEWS

Facebook parent Meta plans $5 billion AI data center in Louisiana: What we know

Greg Hilburn
Shreveport Times

Published 5:38 p.m. CT Nov. 19, 2024 | **Updated 6:58 a.m. C**

Microsoft launches two data center infrastructure chips to speed AI applications

By **Max A. Cherney**

November 19, 2024 10:38 PM GMT+9 · Updated 11 days ago

Aa

Elon Musk raises $6 billion for xAI's Memphis data center; will purchase 100,000 Nvidia chips to boost Tesla's full self-driving capabilities

Story by Wayne Williams • 1d • 2 min read

IT/과학

최신뉴스 | 과학 | 디지털 | 컴퓨터/인터넷 | 뉴미디어/통

Advertisement

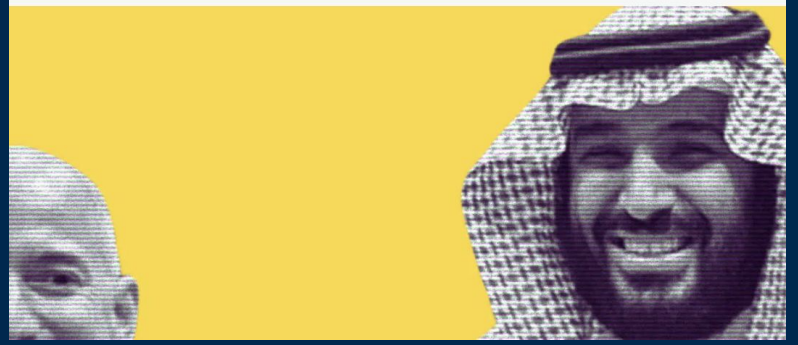
韓 AI의 심장 '4조 AI컴퓨팅센터', SPC로 추진…

이데일리 원문 | 기사전송 2024-11-26 17:07 최종수정 2024-11-27 08:43

LAST UPDATED NOVEMBER 29, 2024 • IN AI NEWS

Saudi Arabia Eyes $12.8 Bn to Build Data Centre

Big-tech companies like Oracle, AWS, IBM and ServiceNow are also taking part.



다시돌아본 기술의 메가트렌드 @ 2024

.....

→ 인공지능 (80년대 말): 추론, 생성

→시스템 (90년대 초) : 하드웨어의 운영

→네트워크 (90년대 말) : 데이타 전송

→데이타베이스 (2010년) : 데이타 저장,정리,정돈

→인공지능 (2015년) : 추론, 생성

→ 시스템 (2024): 하드웨어 운영 → AI 데이타 센터, 에너지

기술 생태계

인공지능 @2024

학습

사물 인터넷

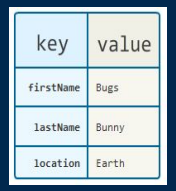
AR/VR

BIO

자율주행

딥러닝

추출



riak

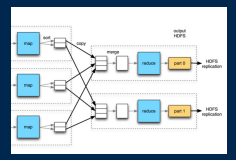
Cassandra

CouchDB relax

mongoDB

@2010

축적



amazon web services™

Google bigtable

hadoop

@2003

UFS Universal Flash Storage

nvm EXPRESS

시스템

제 2 부: 인공지능

A.I. 를 갖는 자 세계를 얻는다.

AI군비경쟁

The one key difference between the U.S. and China in the AI arms race

BY CLAY CHANDLER AND NICHOLAS GORDON
May 17, 2024 at 5:33 PM GMT+9

The AI Arms Race Is Changing Everything

14 MINUTE READ

How Apple Fell Behind in the AI Arms Race

After years of playing it safe with generative artificial intelligence, Apple is set to unveil new features next week

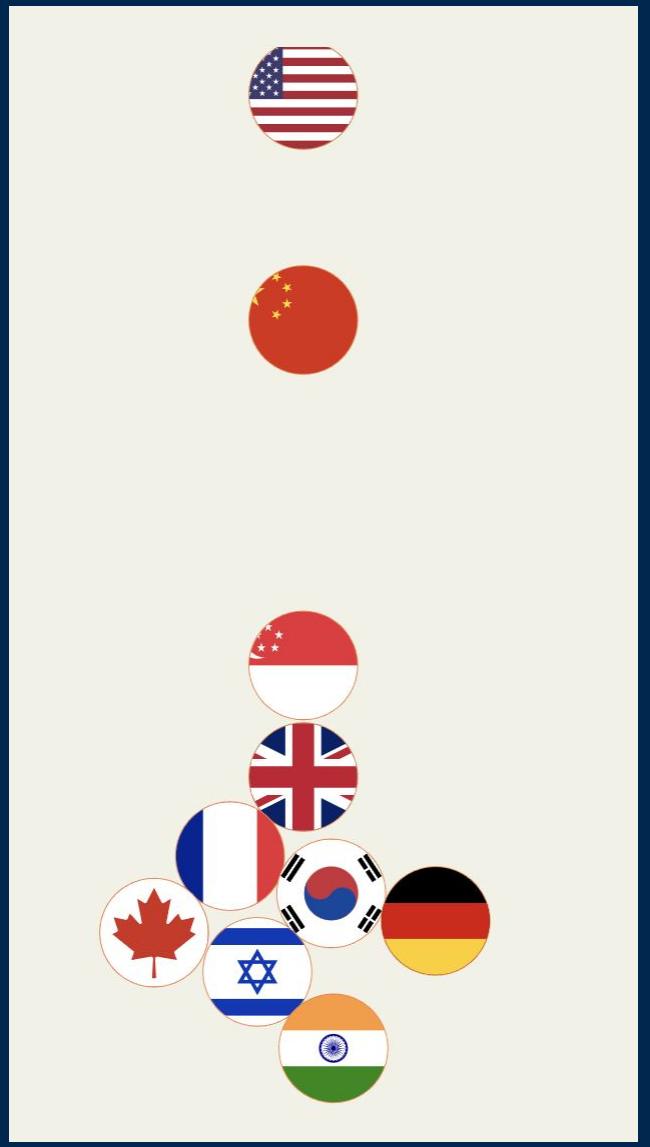
s itself in the unusual position of having to take risks. MICHAEL NAGLE/BLOOMBERG NEWS

By *Aaron Tilley* Follow
June 5, 2024 9:00 pm ET

Share Resize 339 Listen (2 min)

AI기반 신질서


31 OCTOBER 2024 • 7 MINUTE READ
White House issues "first-ever" National Security Memorandum on AI
New directives would have implications across a range of federal agencies



AI기반 신질서

1. 맨해튼 프로젝트 (1942~1946): 1945년 일본 히로시마와 나가사키에 원자폭탄 투하로 전쟁을 종결.

2. 아폴로 프로젝트 (1961~1972): 1969년 아폴로 11호를 통해 닐 암스트롱이 인류 최초로 달에 착륙.

3. 인터넷의 개발 (1970 – 1980): 인터넷 개발 및 상업화 성공, 전세계 네트워킹 프로토콜의 표준 주도

   PC 개발: 1990년대

   스마트폰 개발

4. 워프 스피드 프로젝트 (2020): mRNA 백신 개발

5. AI 맨해튼 프로젝트 (2020년대 이후): AI 패권 경쟁

결과 (진행 중): 미국 빅테크 기업(Google, Microsoft, Nvidia 등)이 AI 기술 주도. 글로벌 AI 정책 표준화를 주도하며 경제적, 군사적 영향력 확대.

[발췌: 정리 gpt4o, 프롬프트 강민구]

- AI 경쟁 == 아이디어? 기술 경쟁?

- # AI 경쟁 == 쩐의 ~~경쟁~~ 전쟁

메타의 2024년도 GPU 투자 계획: **35만대, 26조원!**

(네이버 시가총액 28조원)

Zuckerberg wants to build artificial general intelligence with 350K Nvidia H100 GPUs

Maybe the AGI can finish that Metaverse, haha – oh wait, they're serious

Katyanna Quach Sat 20 Jan 2024 // 01:58 UTC

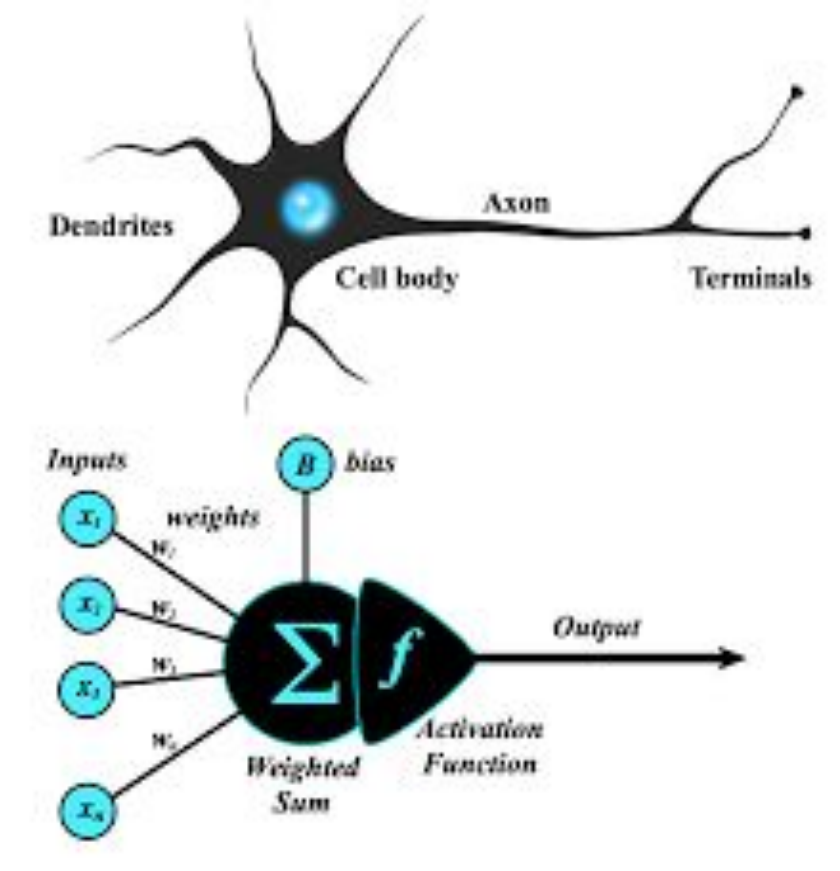
H100 가격: 5천4백만원, DGX 서버 (H100 8장): 한화 6억원

2023년도 2분기 H100 구매량

- Meta 150,000 대
- Microsoft 150,000 대
- Google 50,000 대
- Amazon 50,000 대
- Oracle 50,000 대
- Tencent 50,000 대
- CoreWeave 40,000 대
- Baidu 30,000 대
- Alibaba Cloud 25,000 대
- Bytedance 20,000 대
- Lambda 20,000 대
- Tesla 15,000 대

- 네이버클라우드: 1500대
- 삼성 SDS: 수 1000

어쩌다 쩐의 전쟁?



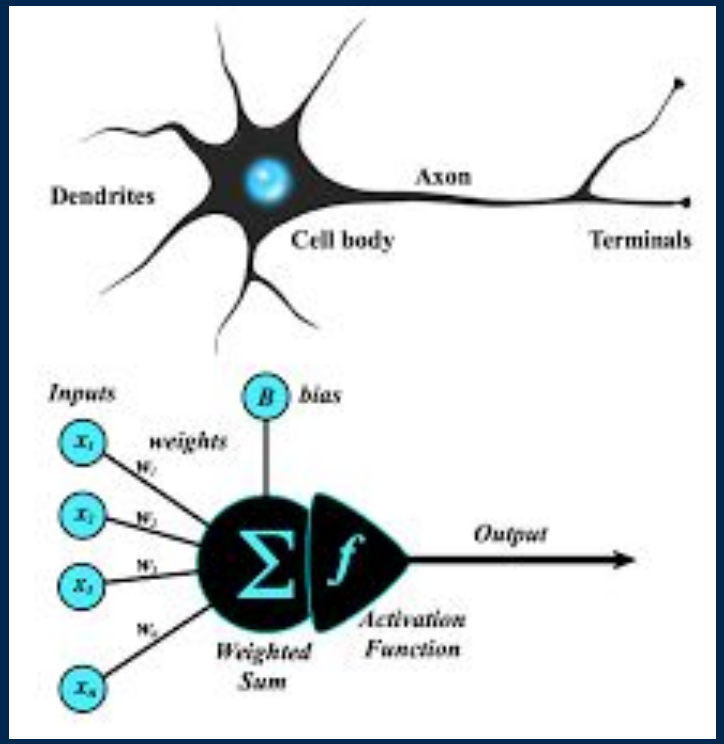
1943년: 워렌 맥쿨로치(Warren McCulloch)와 월터 피츠(Walter Pitts), 전기적 신호를 통해 뇌의 작동 모사.

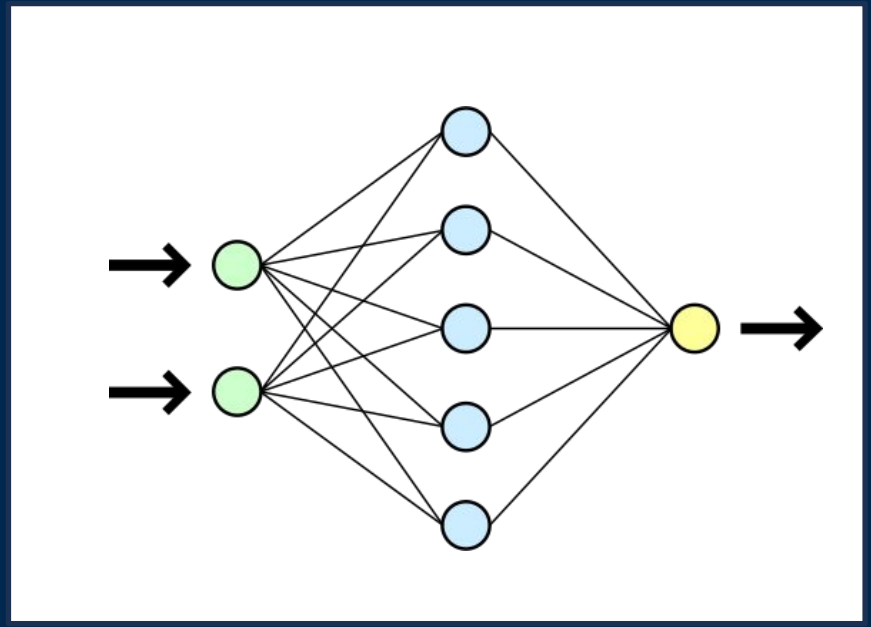
신호의 세기를 숫자로 표현

신호의 세기(가중치): weight, 왜곡도: bias

뇌의 동작을 가중치와 왜곡도로 표현

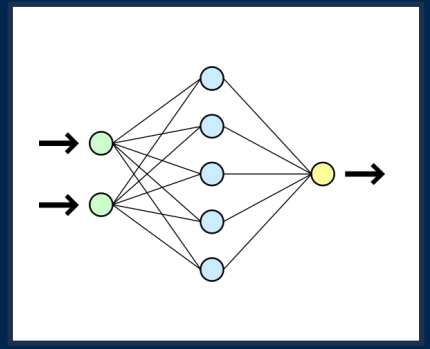
- 학습 (training): 데이타를 통해 모델의 가중치와 왜곡도를 구하는 것
- 추론(inference): 답을 구하는 것

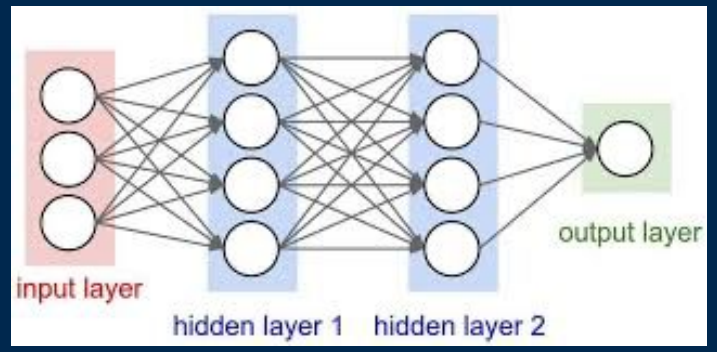


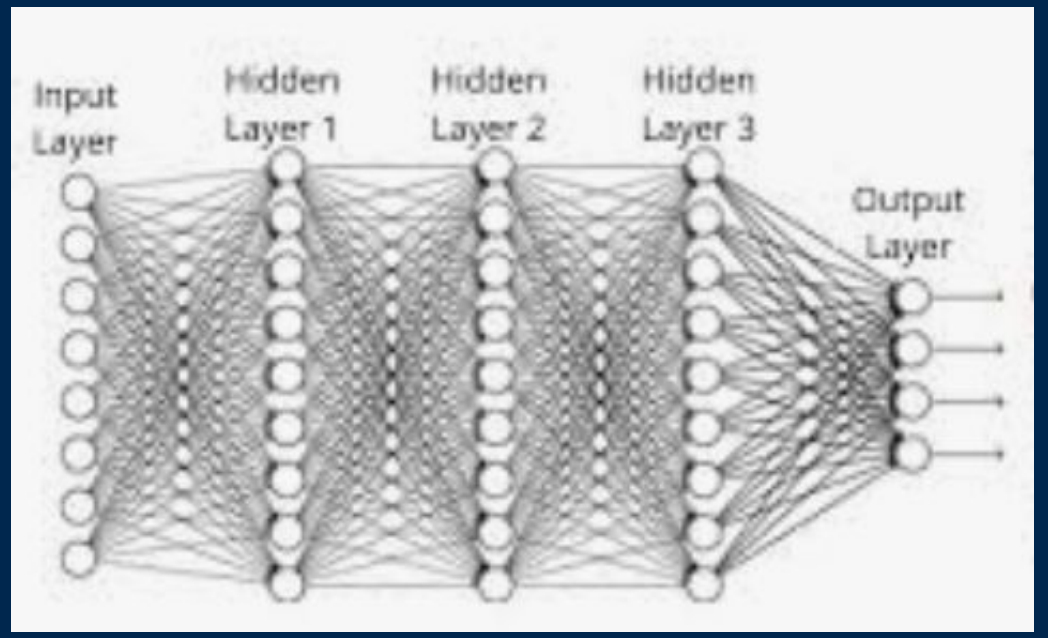


어쩌다 자본경쟁?

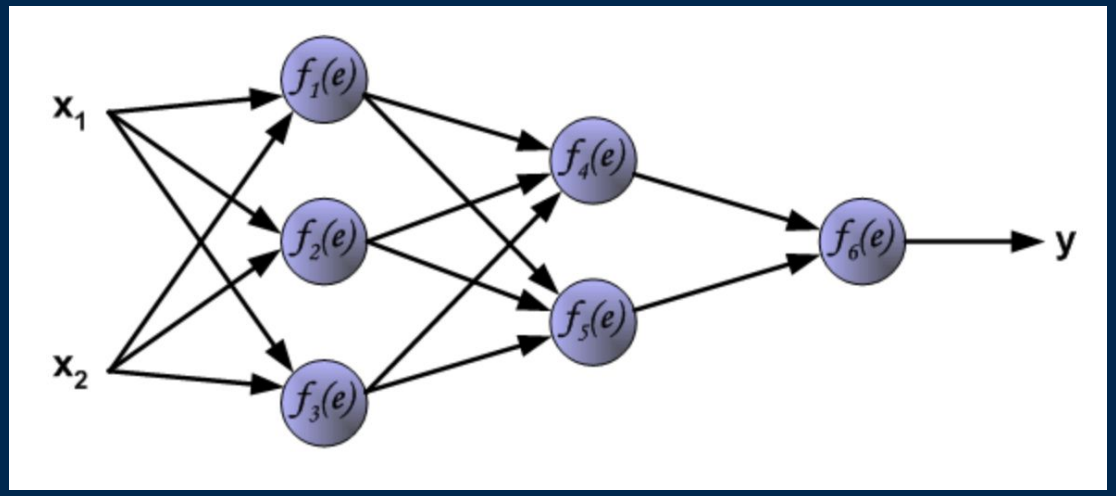
가중치와 왜곡도 갯수 == IQ

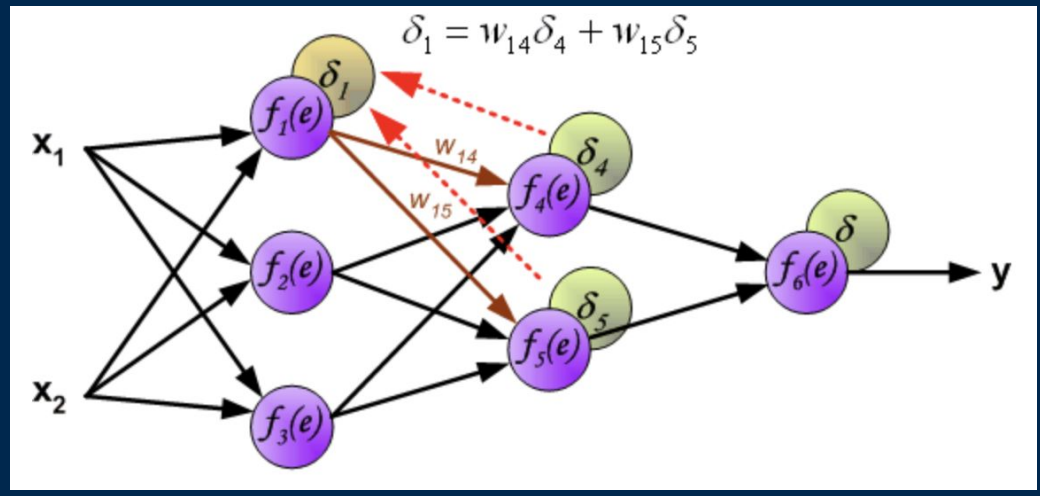




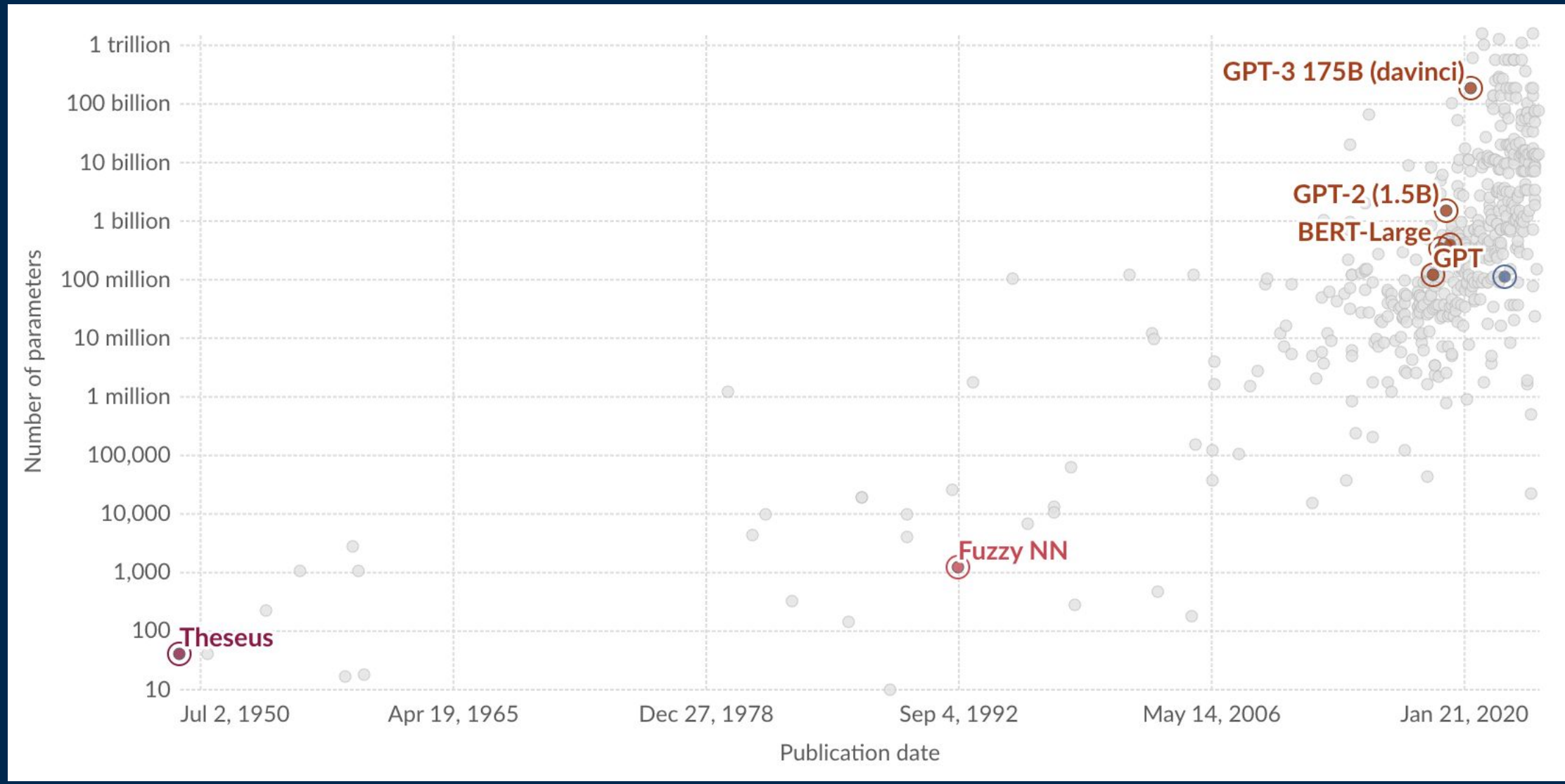


Backpropagation



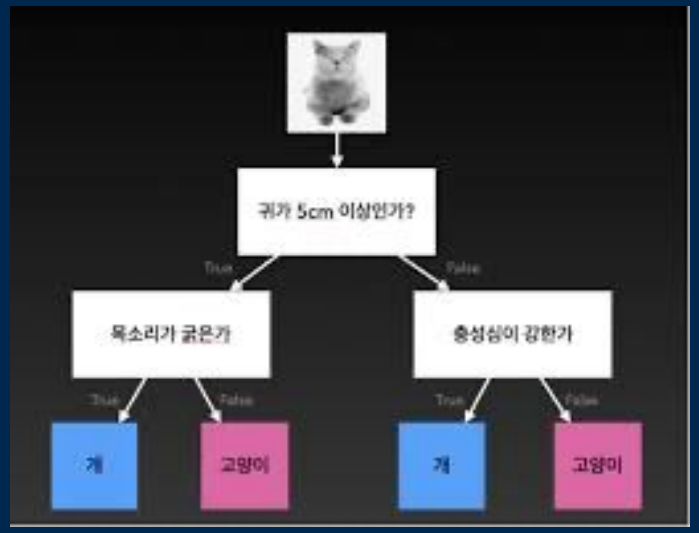


IQ 개선 □ 가중치와 왜곡도 증가 □ H100 추가!



인공지능 기술의 패러다임 전환

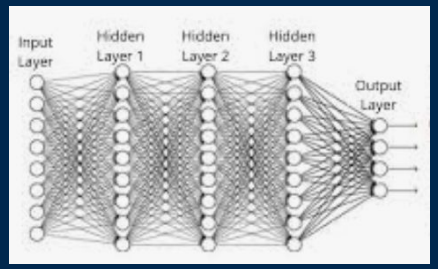
계산

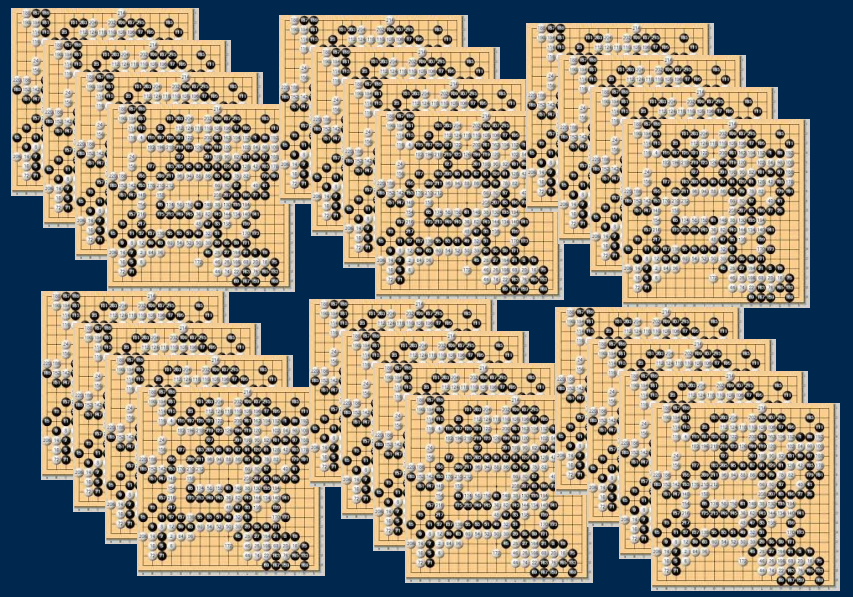


컴퓨터가 빨라짐

대량의 데이타가 축적됨

기억

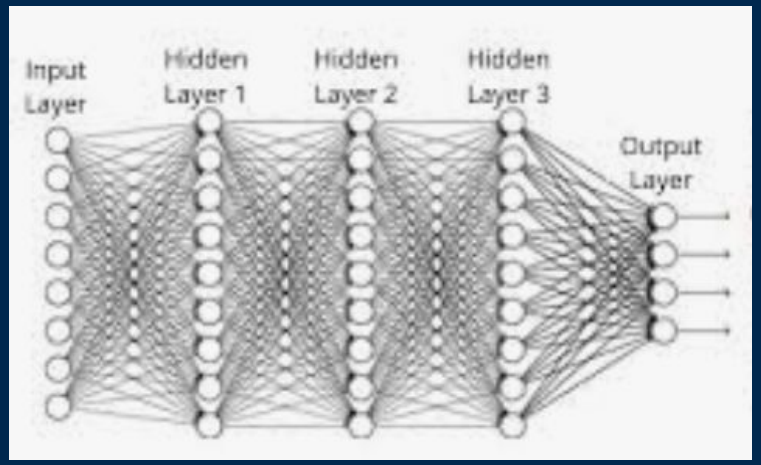




지능과 인공지능

GPT-4

- 인자의 갯수: 1.8 Tr (조)
- 학습데이타: 2021.9 까지
- 학습기간: 2022.5 – 2022.8

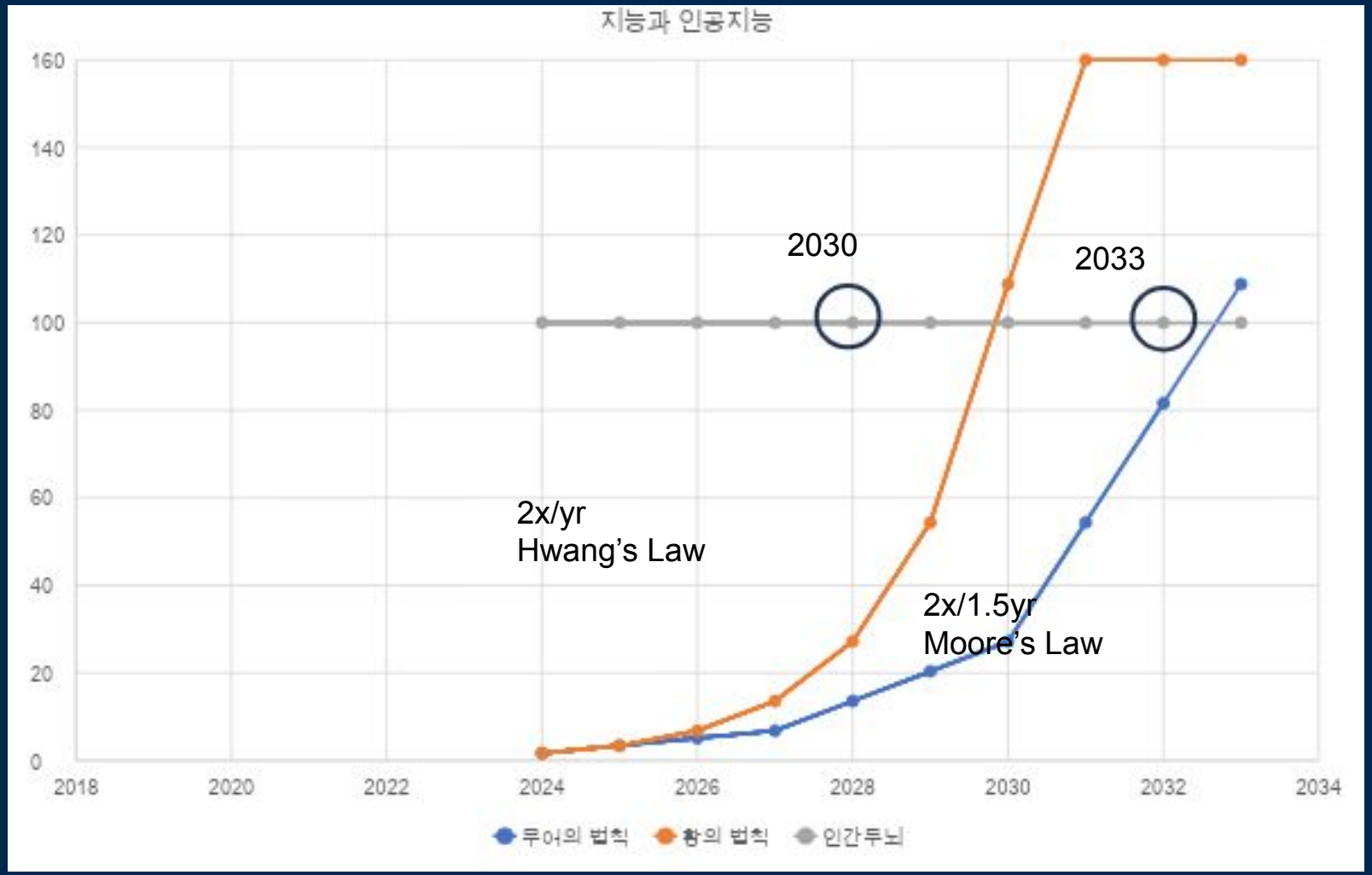


인간

- 시냅시스 갯수: 150 Tr (조)
- 학습데이타: 지금까지
- 학습기간: 항상

현재 GPT-4 용량은 인간두뇌의 1%

언제 인간두뇌와 같은 수의 모델인자를 같게 되는지? 늦어도 2033.



제 3 부: 인공지능 기술

NVIDIA 를 갖는 자 세계를 얻는다.

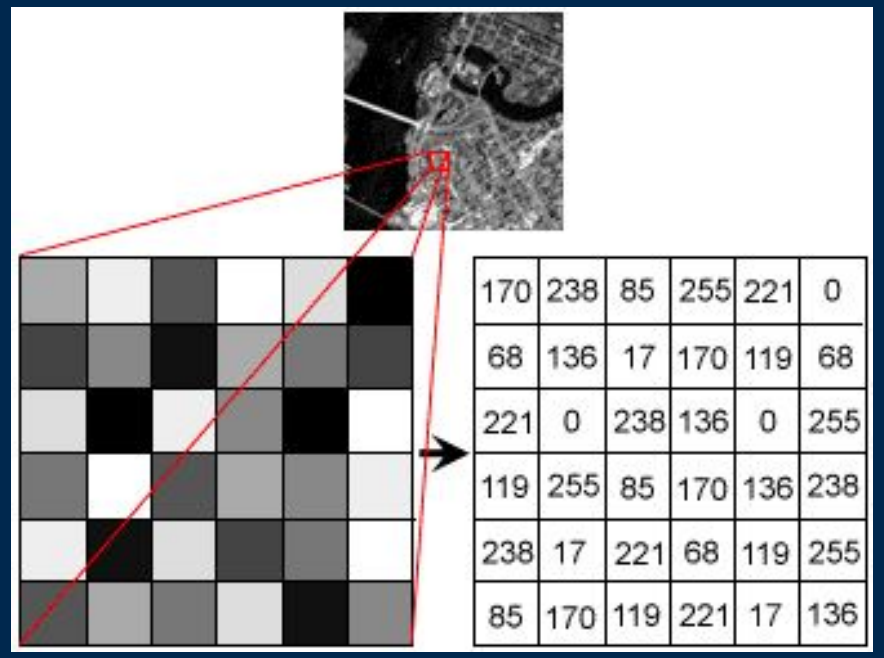
지난 5년간 32배


NVIDIA Corporation (NVDA)
NasdaqGS - Nasdaq Real Time Price • USD
138.25 +2.91 (+2.15%)
At close: 1:00 PM EST
138.12 -0.13 (-0.09%)
After hours: 4:59 PM EST
Comparisons
Indicators
Corporate Events
O 136.85 H 139.35 L 136.05 C 138.25 Vol 142m
yahoo!finance
NVIDIA
150.00
138.25
125.00
100.00
75.00
50.00
25.00
2020
Apr
Jul
Oct
2021
Apr
Jul
Oct
2022
Apr
Jul
Oct
2023
Apr
Jul
Oct
2024
Apr
Jul
Oct

GPU (Graphics Processig Unit)

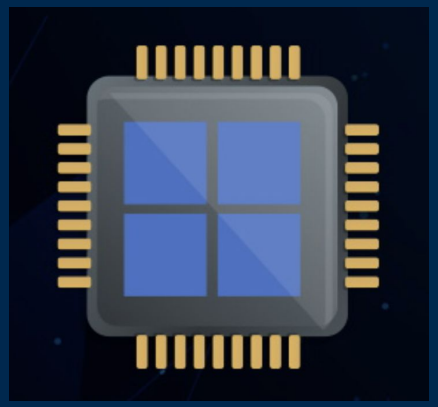
NVIDIA, AMD, ASUS, HP, Intel

Graphics Processing

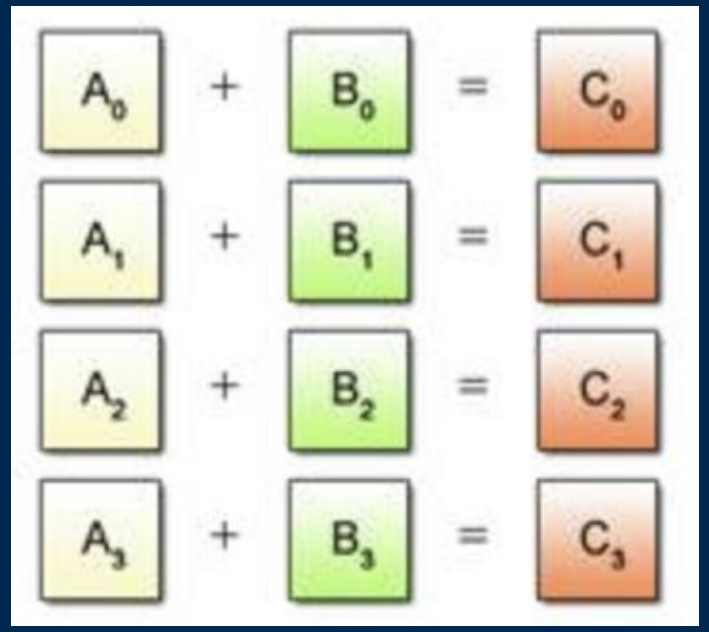


- 그래픽스 처리는 행렬 연산.

CPU



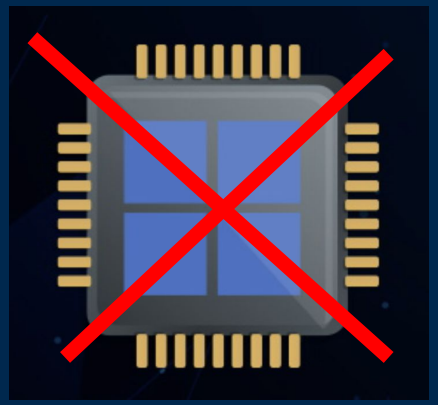
- 연산: 덧셈, 뺄셈, 곱하기, 나누기, 비교하기, <, >, = 등등
- 하나의 연산을 한번에 하나씩 매우 빠르게 하도록 설계되어 있음.

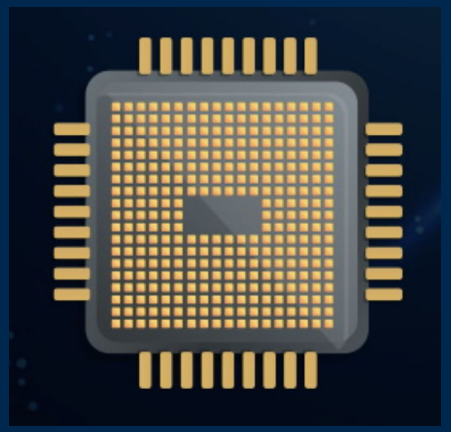


GPU

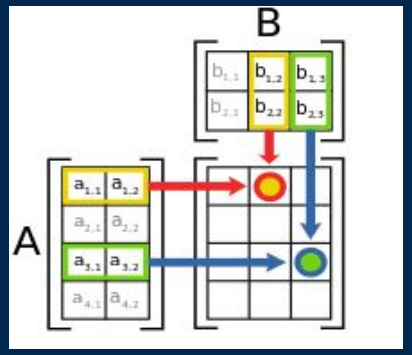
같은 연산을 동시에 여러개를 수행함.

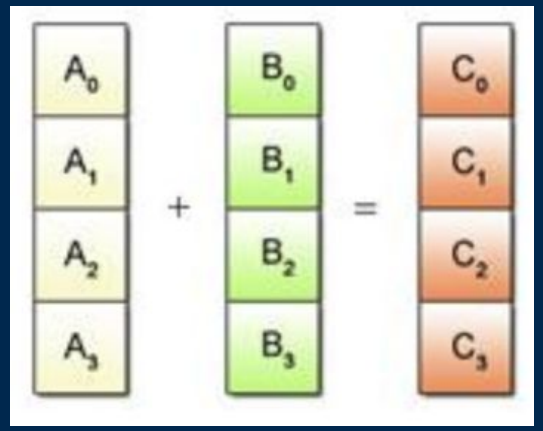
고성능 CPU 필요없음.





- 연산: 덧셈, 뺄셈, 곱하기,
  다수의 연산을 한번에
  여러개씩하도록.

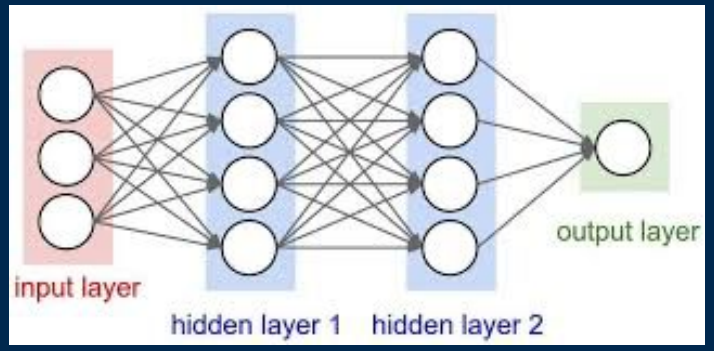


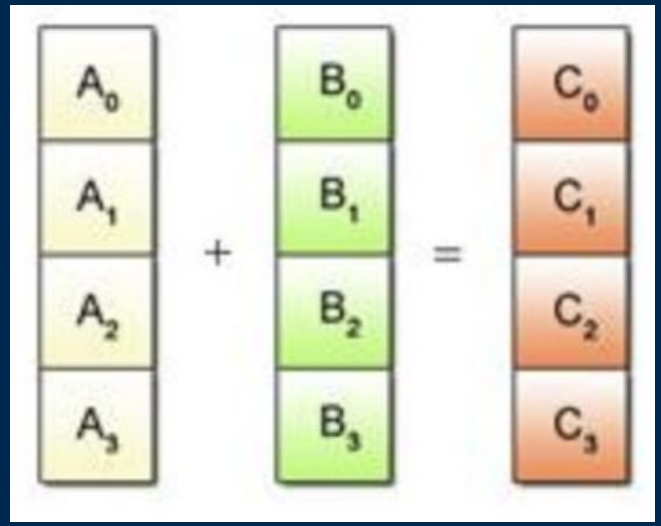


AI == GPU, 왜?

신경망 모델의 학습과 추론이 근본적으로 행렬연산임.

GPU가 신경망 모델의 학습과 추론에 아주 적합함.





왜 NVIDIA 의 GPU인가?

- GPU는 전용 개발 프로그램을 사용해야함.:
  - OpenCL vs. CUDA

OpenCL

복잡: 베우기 어려움
강력함
범용

CUDA

간단: 배우기 쉬움
개발하기 쉬움
성능이 좋음
엔비디아 전용

NVIDIA는 CUDA라는 소프트웨어 개발에 올인

- Ian Buck, Stanford PhD
- Author of CUDA, 2006

CUDA

- 쓰기 쉬움
- 다양한 도구

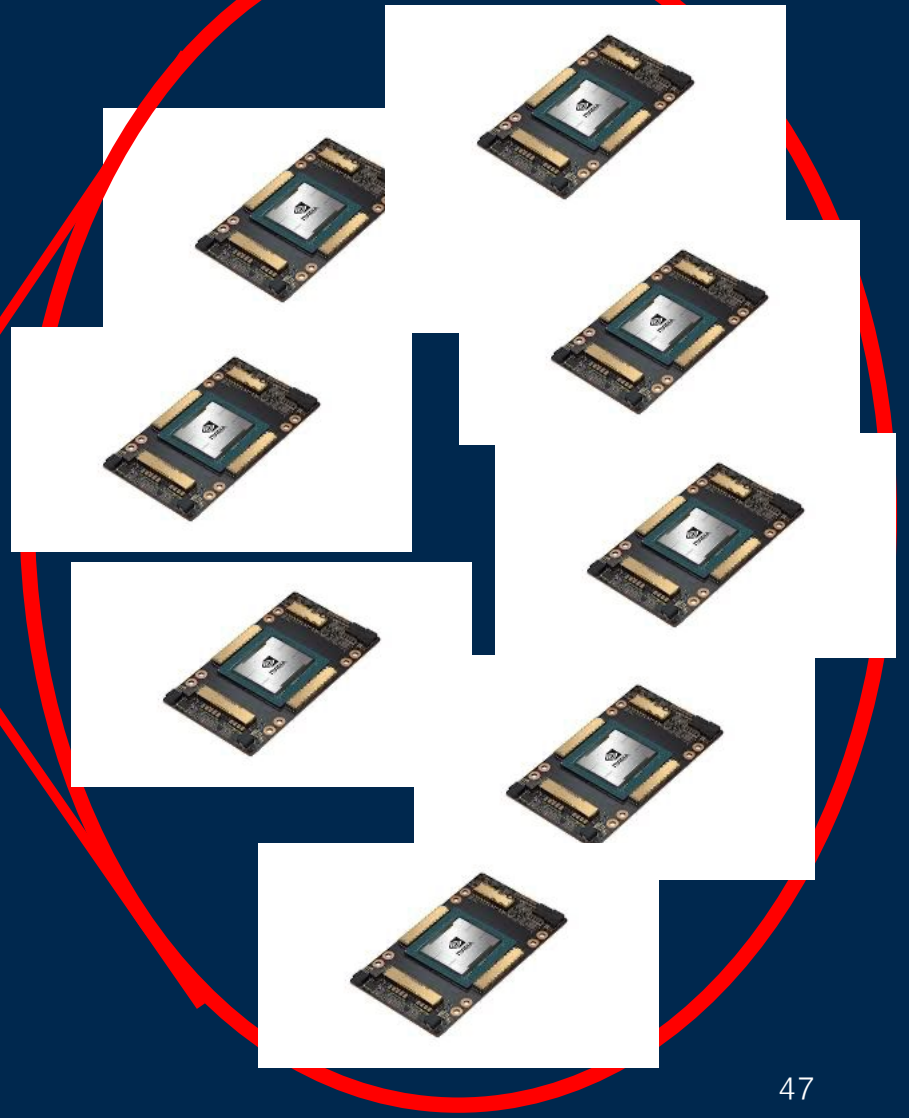
개발자

- HW는 무관심
- SW 도구가 중요

NVIDIA 생태계가 구축

CUDA

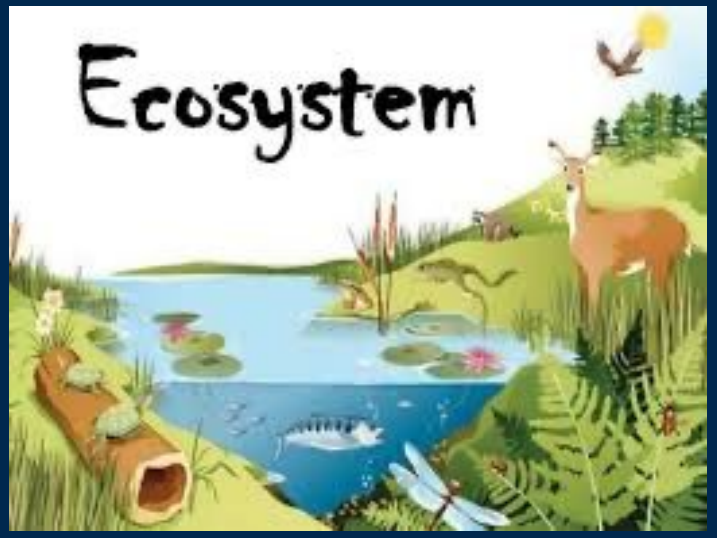
독점



생태계 구축이 관건

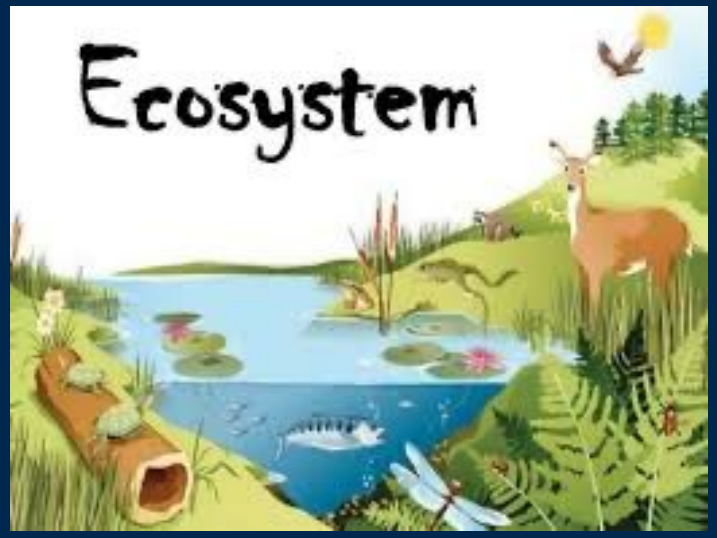
- Nvidia 생태계

  CUDA



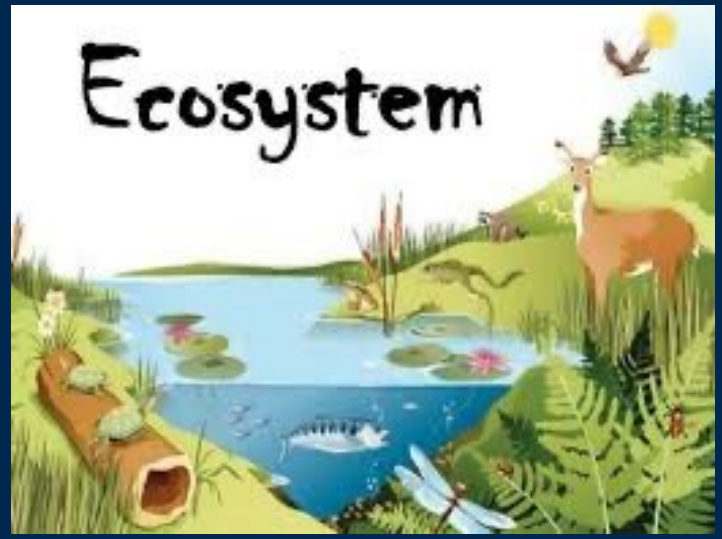
- TSMC 생태계

  디자인 툴킷



- ARM 생태계

  SW 툴체인



판매와 매출

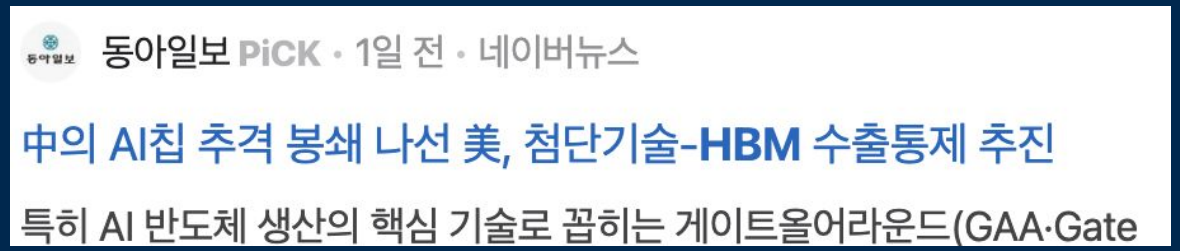
- 하드웨어로 매출, 소프트웨어로 판매

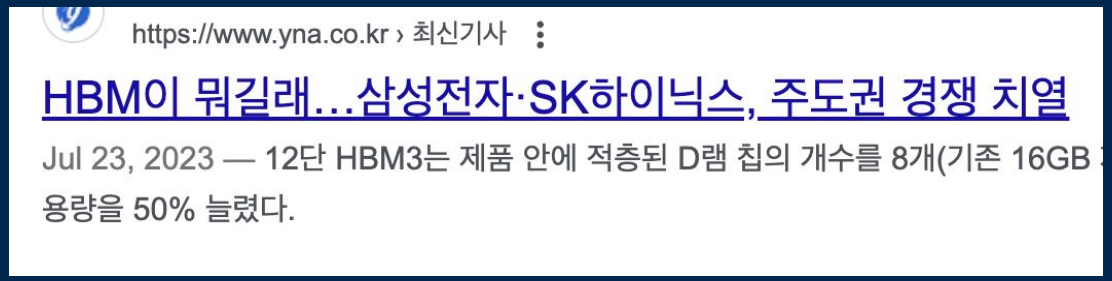
Apple WWDC

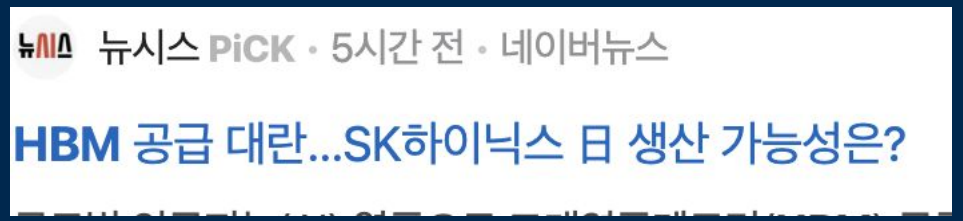
Nvidia GTC

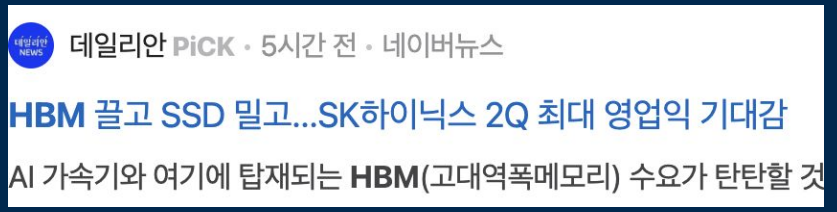
Google IO

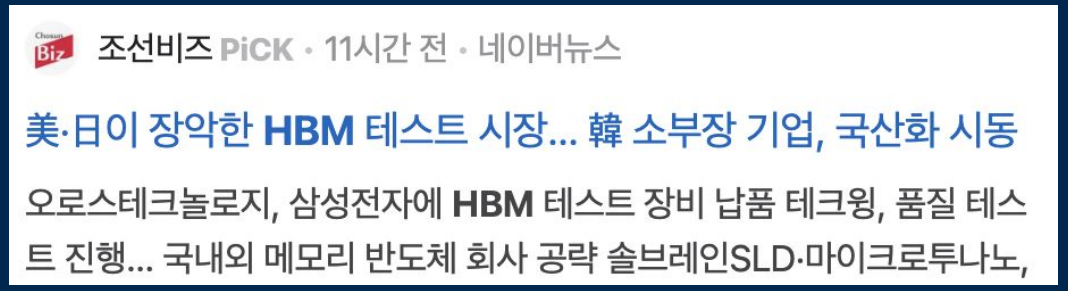
HBM (High Bandwidth Memory)



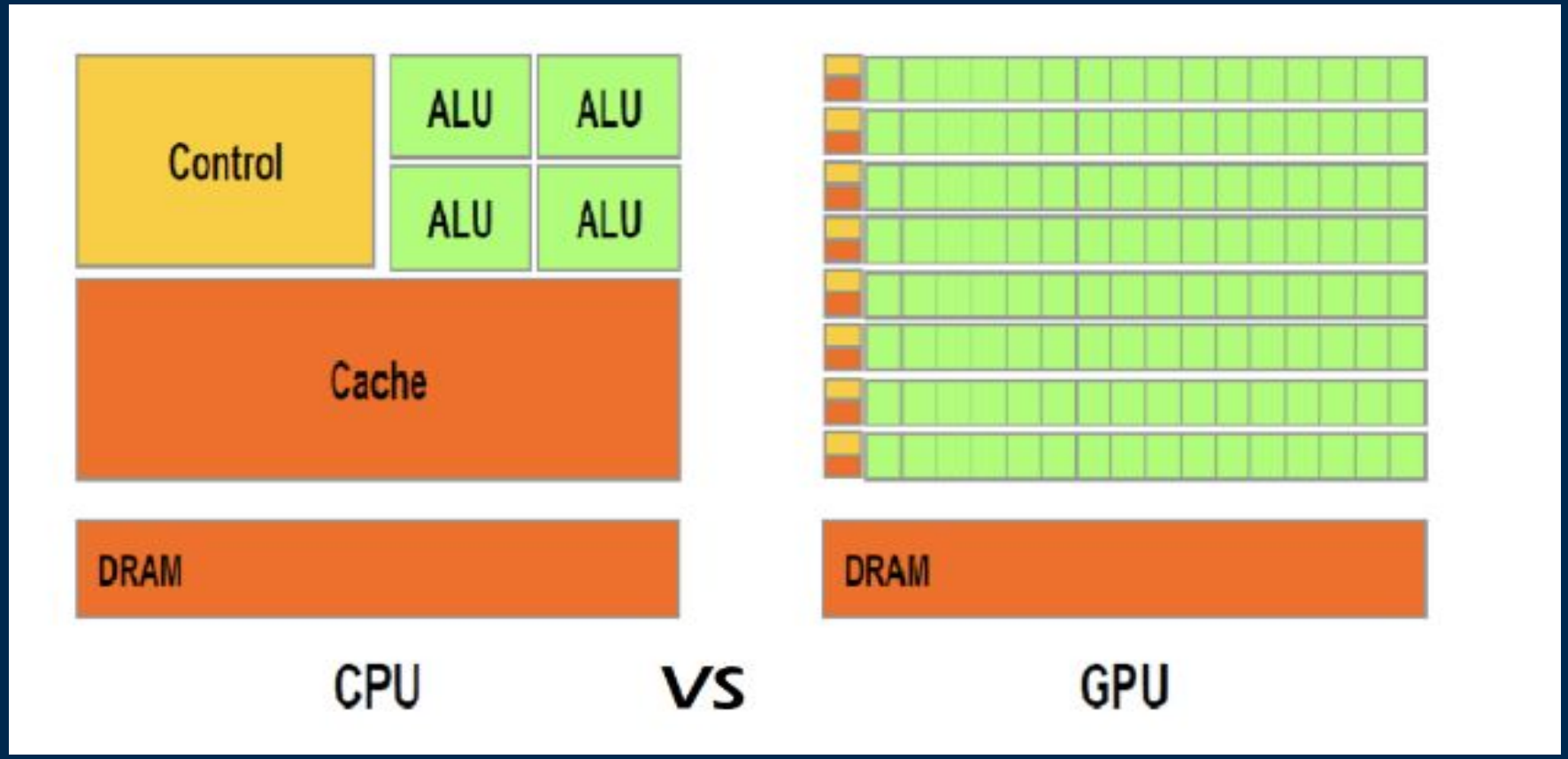




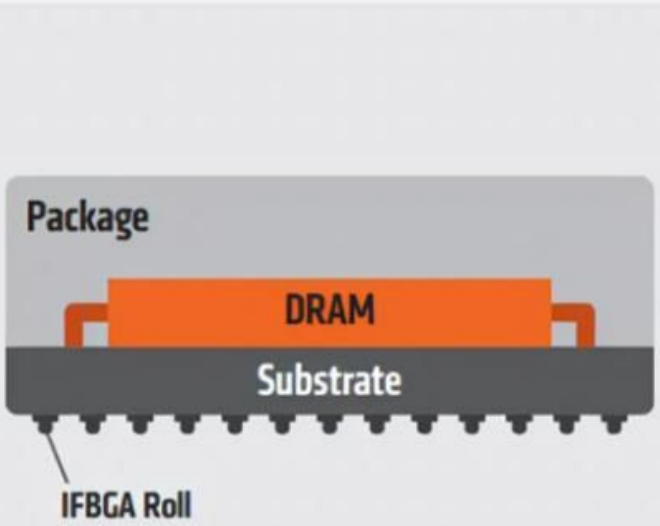


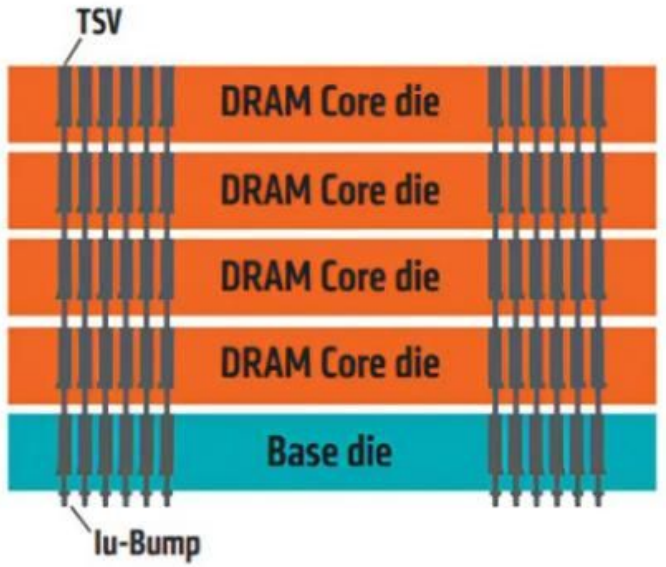


HBM (High Bandwidth Memory)



HBM (High Bandwidth Memory)





GDDR5	Per Package	HBM
32-bit	Bus Width	1024-bit
Up to 1750MHz (7GBps)	Clock Speed	Up to 500MHz (1GBps)
Up to 28GB/s per chip	Bandwidth	>100GB/s per stack
1.5V	Voltage	1.3V

CXL (Compute Express Link)

https://www.sedaily.com › NewsView

반도체 업계가 CXL에 주목하는 이유 [강해령의 하이엔드 .

Dec 23, 2023 — 기존 CPU와 D램이 위치하던 마더보드 내에서 확장된 **CXL** D램
하이닉스가 마치 SSD같은 저장 장치 모양으로 출시한 ' ...

https://biz.chosun.com › stock_general › 2024/02/05

AI 반도체와 함께 뜨는 'CXL'이 뭐길래... 창업 1년만에 1000 ...

Feb 5, 2024 — AI 반도체와 함께 뜨는 '**CXL**'이 뭐길래... 창업 1년만에 1000억 기업가치 받은 ...
지능(AI) 반도체와 함께 '컴퓨트익스프레스링크(**CXL**)'이 주목 ...

samsungsemiconductor.com
https://news.samsungsemiconductor.com › ai-시대를-이...

AI 시대를 이끌 차세대 D램, 'CXL 메모리'의 모든 것

CXL은 컴퓨팅 시스템에서 중앙처리장치(CPU)와 메모리, 그래픽 처리장치(GPU
적으로 활용하기 위한 새로운 인터페이스입니다. 기존에는 CPU를 ...

Samsung Semiconductor
https://semiconductor.samsung.com › 홈 › 뉴스

삼성전자, 업계 최초 'CXL 2.0 D램' 개발

'메모리 풀링(Pooling)'은 서버 플랫폼에서 여러 개의 **CXL** 메모
풀(Pool)에서 메모리를 필요한 만큼 나누어 사용할 수 있는 ...

지디넷코리아
https://zdnet.co.kr › view

삼성·SK하이닉스, 이번엔 CXL 기술 경쟁...제 2의 HBM 노린다

Mar 25, 2024 — SK하이닉스는 2022년 8월 **CXL** 2.0을 지원하는 96GB D램 샘플을 선보
월에는 업계 최초 **CXL** 기반 연산 기능을 통합한 메모리 솔루션 CMS을 ...

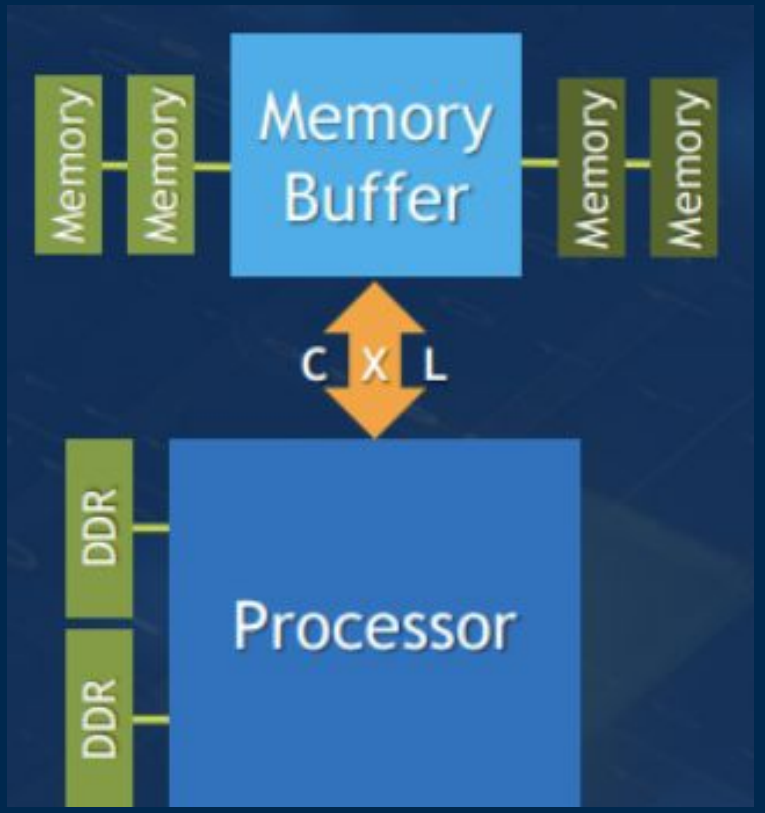
https://byline.network › 2024/01

[그게 뭔가요] 삼성과 하이닉스가 미는 그 기술, 'CXL'

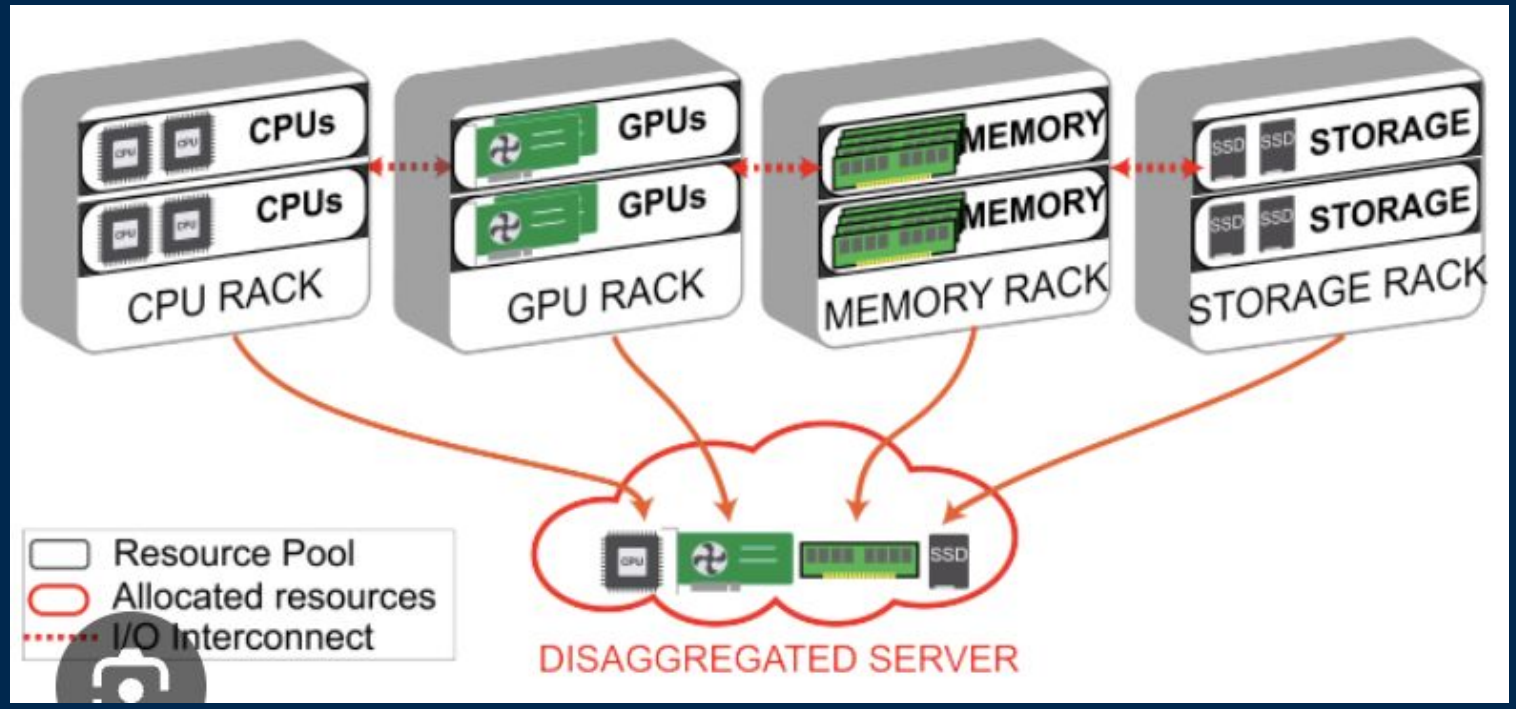
Jan 30, 2024 — 반도체 관련 뉴스를 살펴보다가, 요즘 매우 빈번하게 나오는
라고요. 컴퓨트 익스프레스 링크(Compute Express Link)의 ...

CXL (Compute Express Link)

단일 서버에 장착할 수 있는 디램 용량이 증가함.



CXL (Compute Express Link)



CXL (Compute Express Link)

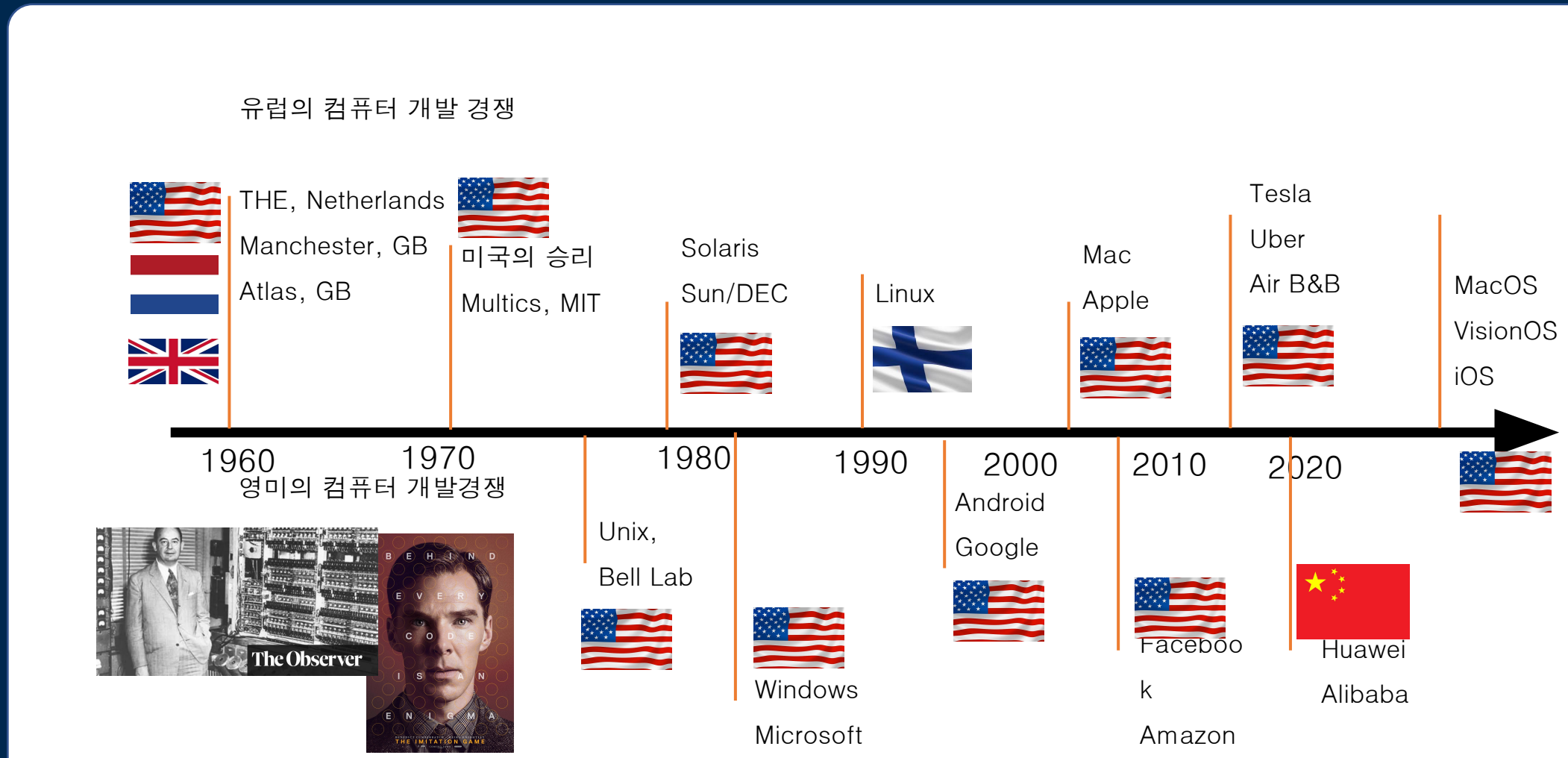
CXL을 미는 이유

- 디램 시장이 늘어난다.

CXL의 이면

- 디램 시장이 안늘어날 수 있다.
- 생각보다 느리다
- 분리형 시스템의 경우 CXL말고 다른 대안이 있다.

IT 패권과 원천기술: 운영체제에 대하여



제 3 부: 인공지능

다음은?

창발

100Tr 개의 LLM을 만들었을 때 어떤 일을 할 수 있을 것인가?

오답에 대한 원인 파악 불가 □ 오류수정이 매우 어려움

에너지와 물

Google Datacenter, Dalles, OR, Dec. 2022

ICT 시스템

막대한 에너지 사용 및 탄소 배출

2023년 세계 에너지 4-5%, 탄소 1-2%
2030년 세계 에너지 20% (전망)

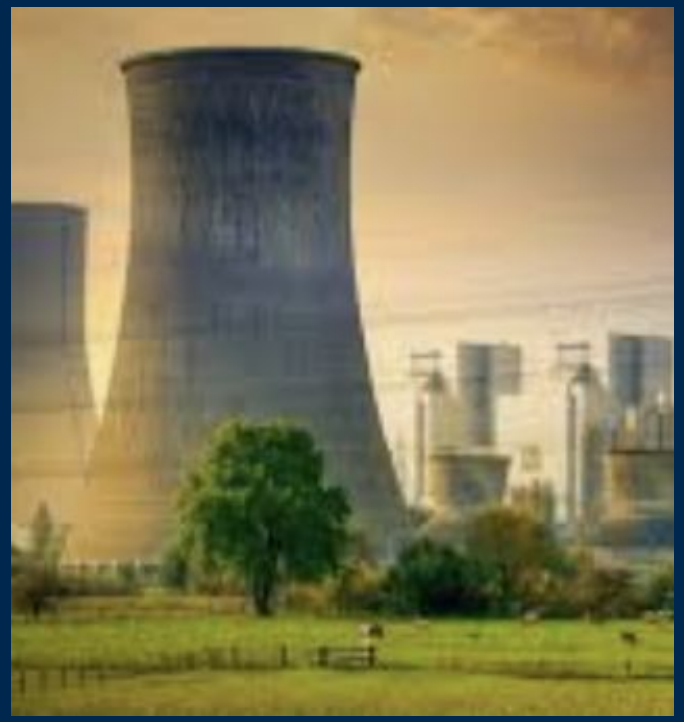
트렌드 1 | 인공지능

ChatGPT: 2023년 1월 한달간
17만 5천 명분의 전기 사용
[Ludvigsen, March, 2023]

트렌드 2 | 하드웨어 개선

서버의 경우 탄소 배출의 40%가
하드웨어 생산시 발생하는 “내재 탄소”

에너지와 물



에너지 장벽

전력생산

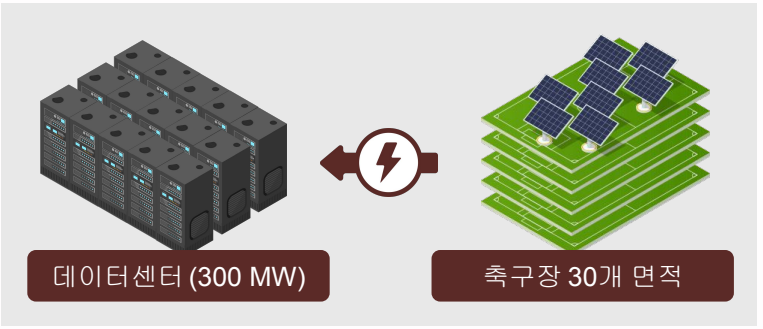
친환경 에너지 사용

Lancium, Verne Global 신재생 에너지 솔루션

Facebook Los Louna
친환경 에너지 데이터센터

새만금 재생에너지 단지
데이터센터 건축

한계 제한적 적용



낮은 에너지 밀도

배전/송전 시스템 포화

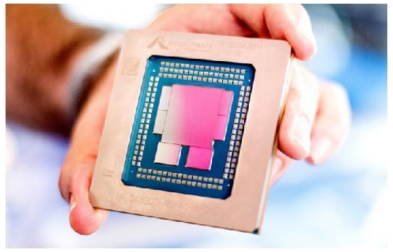
에너지 장벽

하드웨어

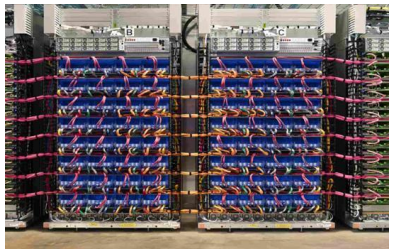
생산,폐기
데이터센터
건축/구축
공조

저전력 반도체

분리형 데이터 센터 구조



Amazon의 AWS Graviton 3

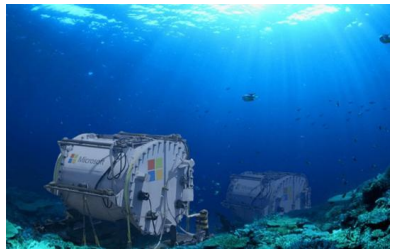


Google의 AI 가속기

수열 냉각

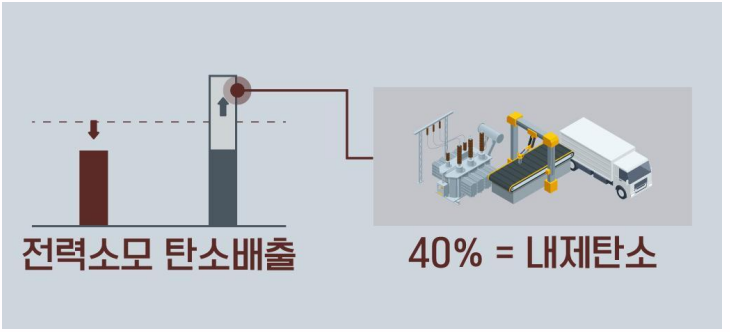
Nautilus의 Stockton 데이터센터

PRASA의 수열 냉각



Microsoft의 수중 데이터센터

한계 제한적 적용



내재 탄소 미고려

탄소 집약도 미고려

에너지

조선경제 > 테크

전기 먹는 하마 AI… 마크롱 "원전 늘려 미래 에너지 확보"

세계 경제 다보스 포럼, 최우석 기자 현장 취재

최우석 기자 임경업 기자

업데이트 2024.01.19. 06:45

AI 개발에 에너지 수요 폭증...바다에 빠뜨리고 태양 전력 사용하고[미래on]

AI 통한 응답 추

태양광·풍력 등

(서울=뉴스1) 서장원

홈 > 시사·교양 > 경제

[커버스토리] 인공지능의 미래 전기에 달렸다

장규호 기자 | 입력 2024.03.03 17:47 | 수정 2024.03.03 17:47 | 생글생글 838호

"전기 없인 챗GPT도 없다"...전력·에너지株로 번진 AI 열풍

박한신 기자

입력 2024.05.12 18:17 수정 2024.05.13 09:14 지면 A17

오늘의 주

Market Summary > Nuscale Power Corp

29.65 USD

+26.61 (875.33%) past year

Follow

Closed: Nov

After hours

1D Max

Open	28.64	Mkt cap	7.17B	52-wk high	32.30
High	32.30	P/E ratio	-	52-wk low	1.88
Low	28.46	Div yield	-		

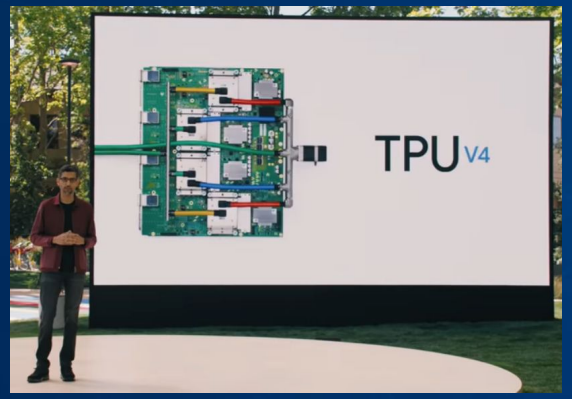
HW? vs. SW?

애플사가 자사의 iOS및 SW에 최적화된 CPU @ 2020
P.A. Semi 인수 @2008

What Is the Apple M1 Chip?

A processor that Apple developed in-house powers the newest Macs and iPad Pro. It's a mighty powerful chip, but it's not without a few quirks. Here's everything you need to know if you're in the market for an M1-powered Apple

Google TPU: Tensor Flow 전용 칩 @ 2020



Articles / News

Amazon's Custom Graviton2 Chips Arrive in AWS EC2

Tobias Mann | Editor
June 13, 2020 1:57 AM

Share this article:

Amazon: 자사 데이타센터용 칩 @ 2020
Annapurlab(Israel) 인수 @ 2016

■ March 17, 2021, 7:00 PM GMT+9

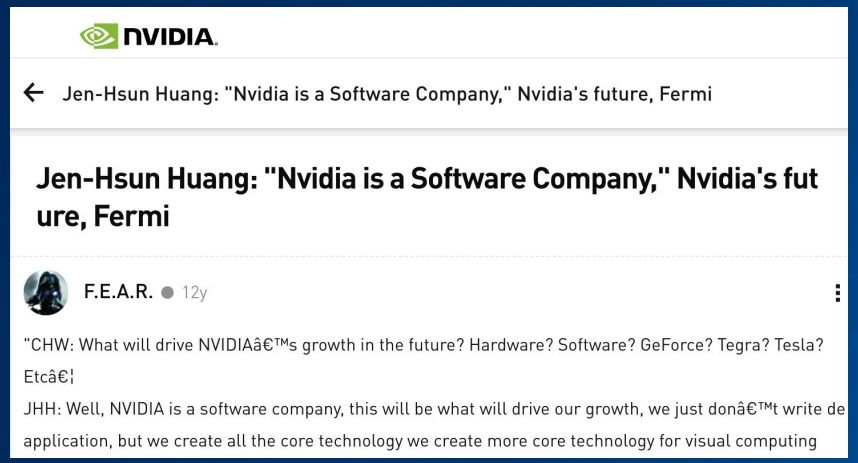
Why Amazon, Google, and Microsoft Are Designing Their Own Chips

● All the biggest tech companies are prioritizing custom designs, which adds to the growing problems facing the incumbents.

By Ian King and Dina Bass

ILLUSTRATION: SIMON LANDREIN FOR BLOOMBERG BUSINESSWEEK

SW 역량 확보 중요



Nvidia is a software company.

Apple is a software company.

Tesla is a software company.

마치며

- 현대 AI는 자본 전쟁이다.
- SW가 하드웨어를 판다.
- 생태계 구축은 최소 10년이상 투자가 필요하다.
- 다음의 AI 기술은 에너지

감사합니다.