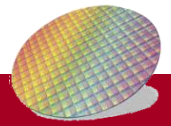
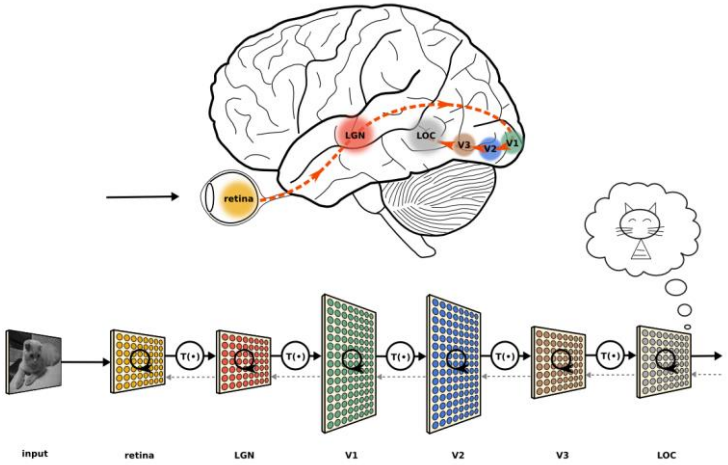
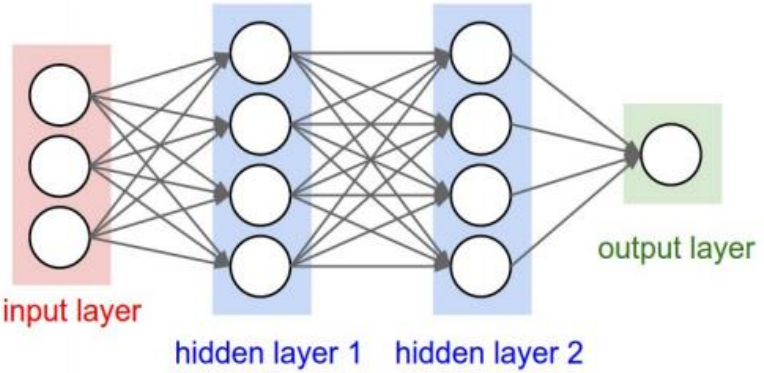
Deep learning IC Design
2025



What is deep learning?

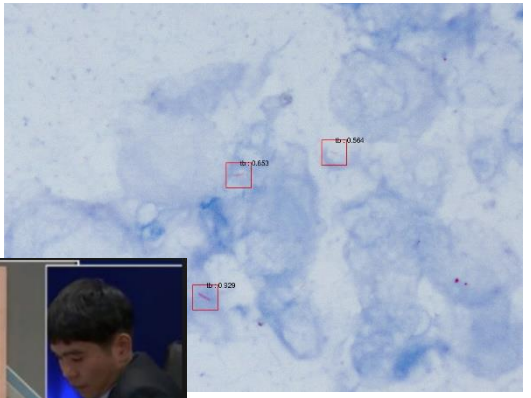
- The cutting edge of artificial intelligence
  - Approximates a complex function with a large number of simple trainable units
- Different Neural Network for Different Applications
  - Convolutional Neural Network(CNN) /Recursive Neural Network (RNN)/ Long Short Term Memory (LSTM)/Generative Adversarial Networks

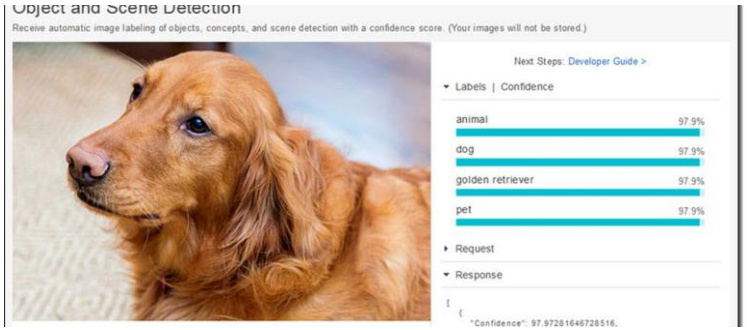


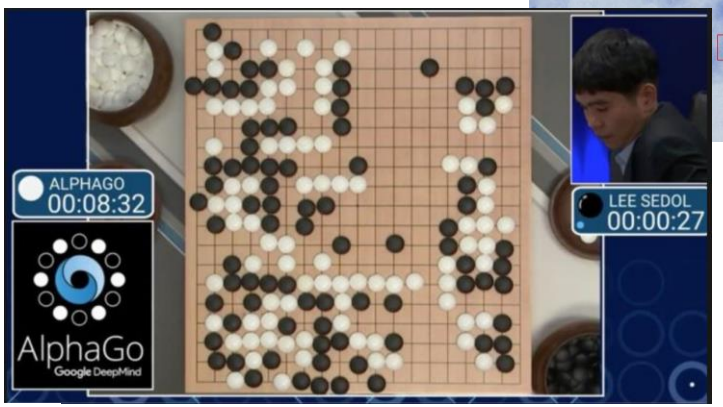


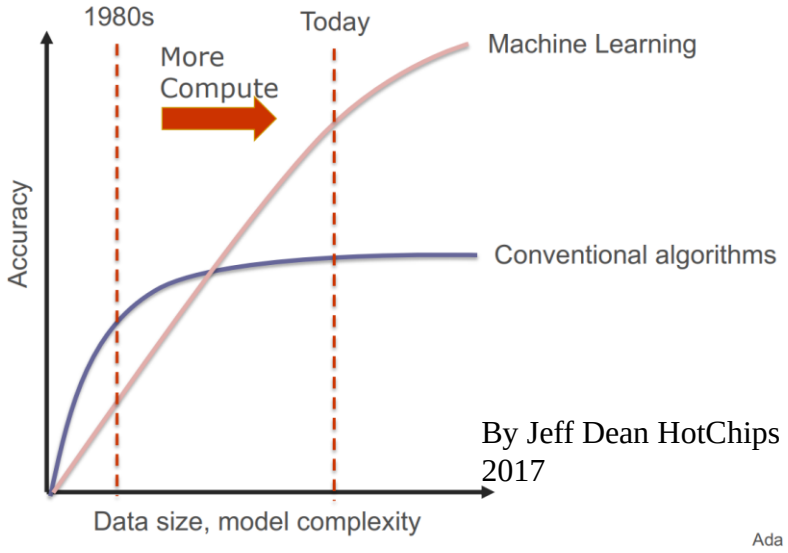
Why Deep Learning?

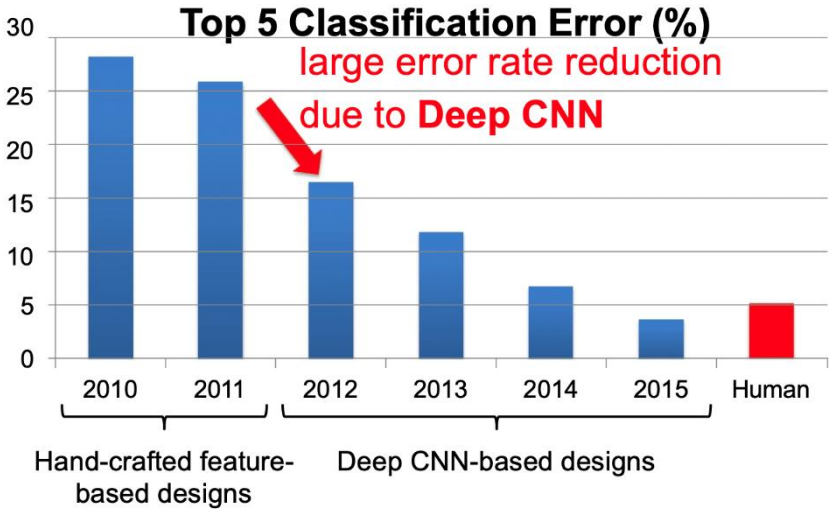
- AI, especially DNN, excels in many area:
  - Image classification, Object detection
  - Speech Recognition, and many others



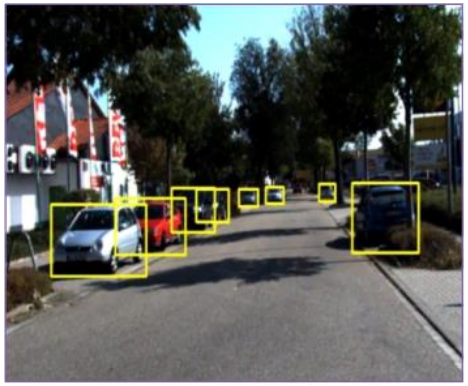




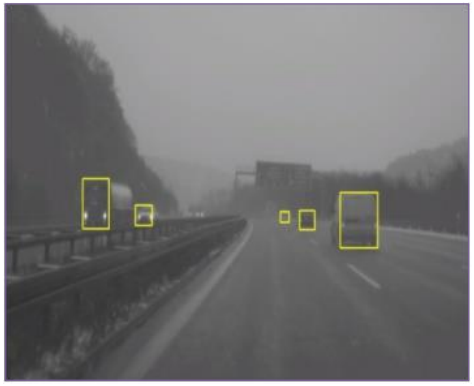




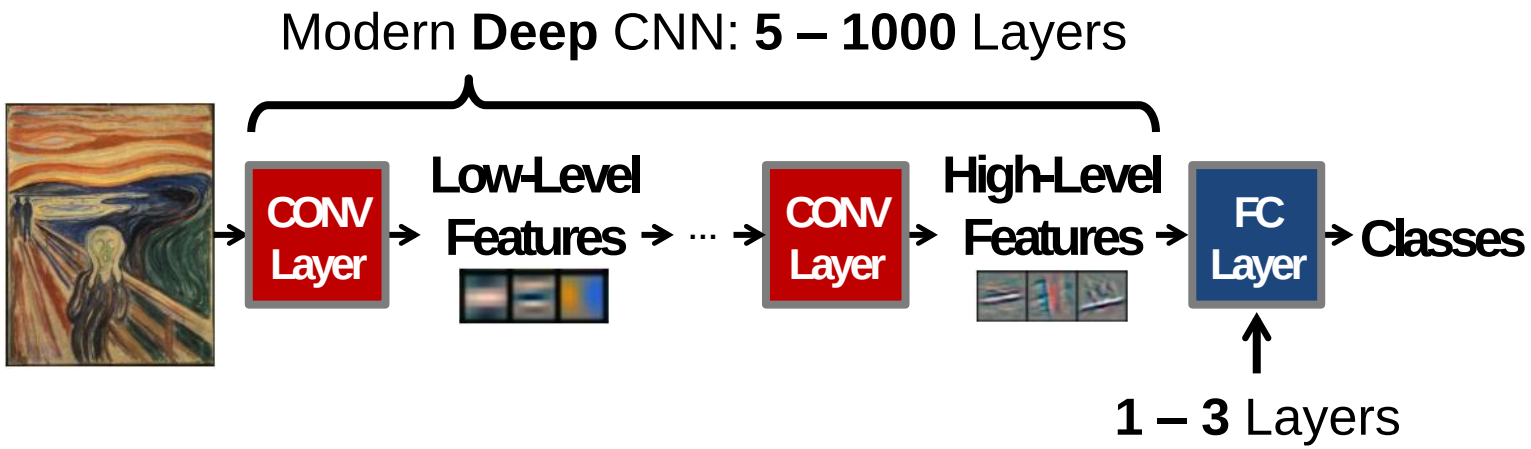
Deep Learning for Self-driving Cars



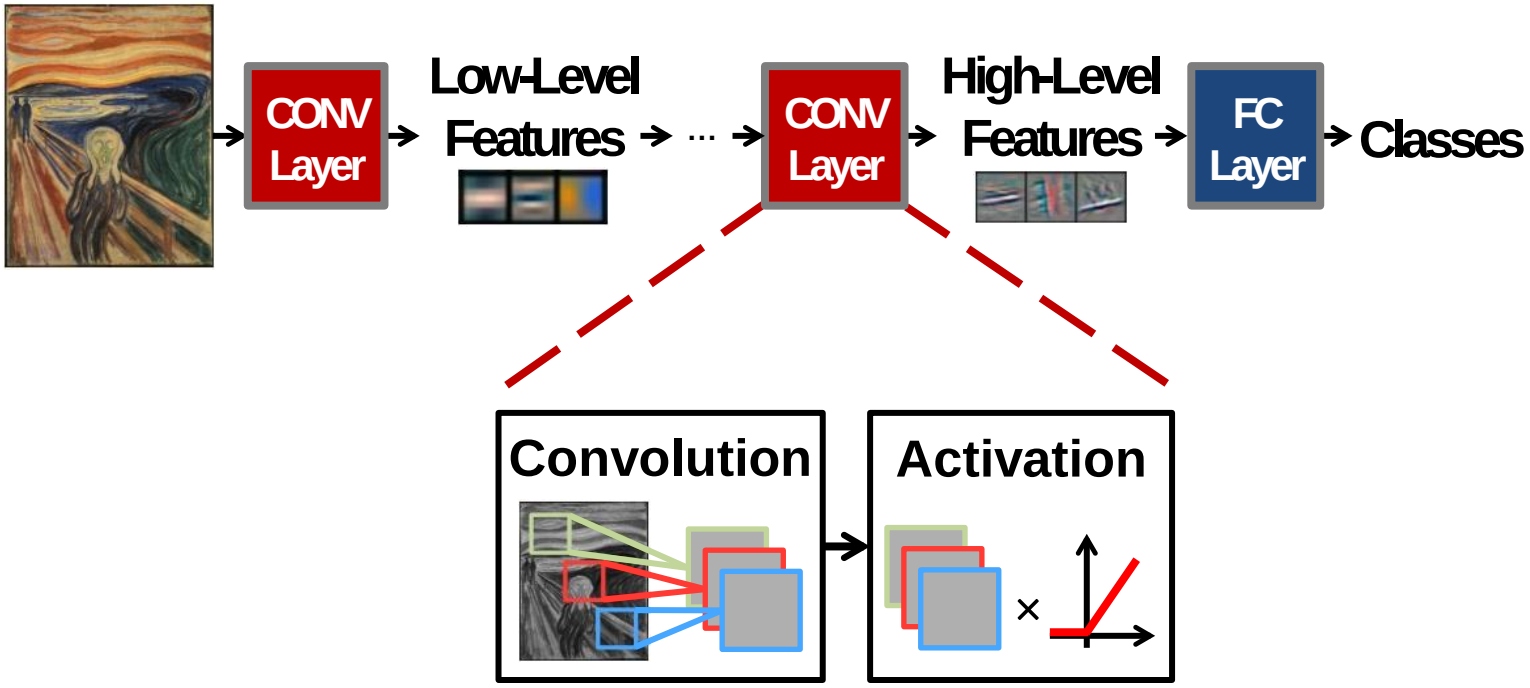




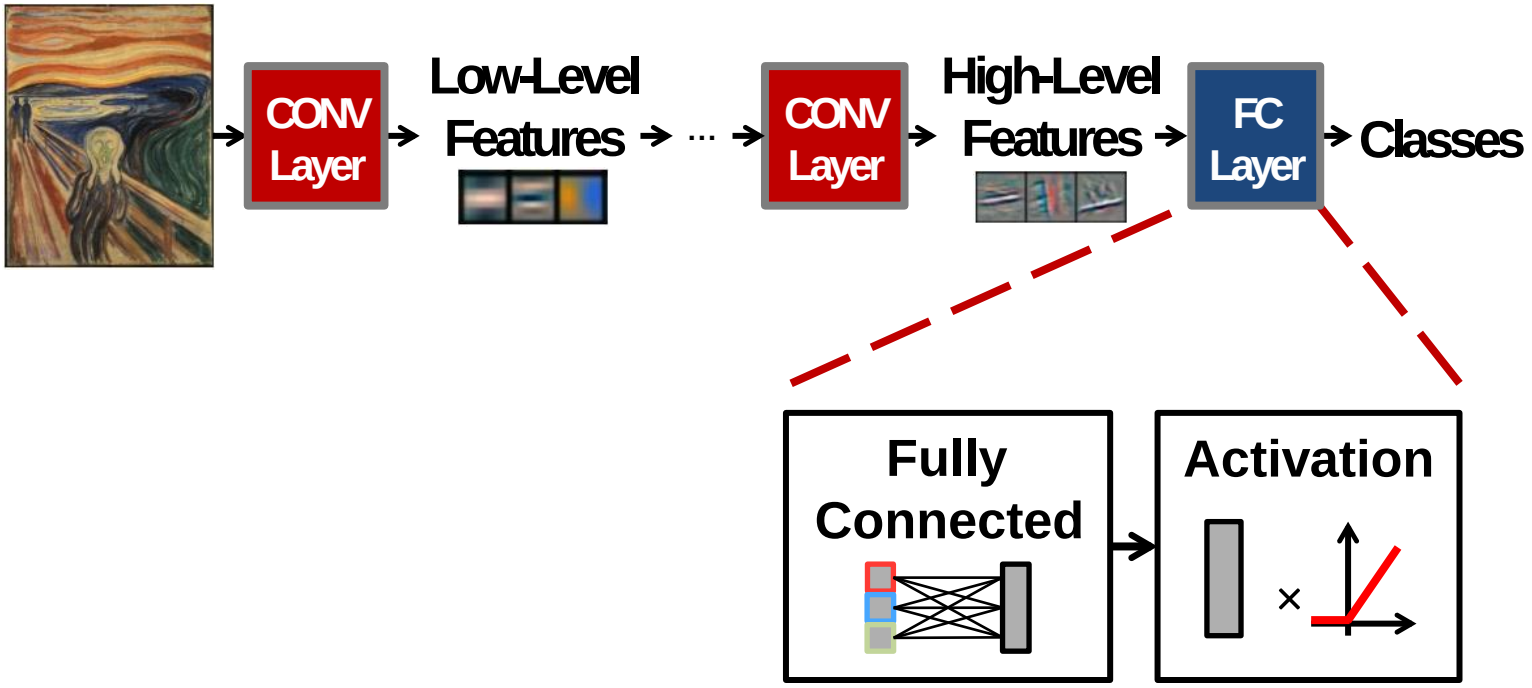
Deep Convolutional Neural Networks



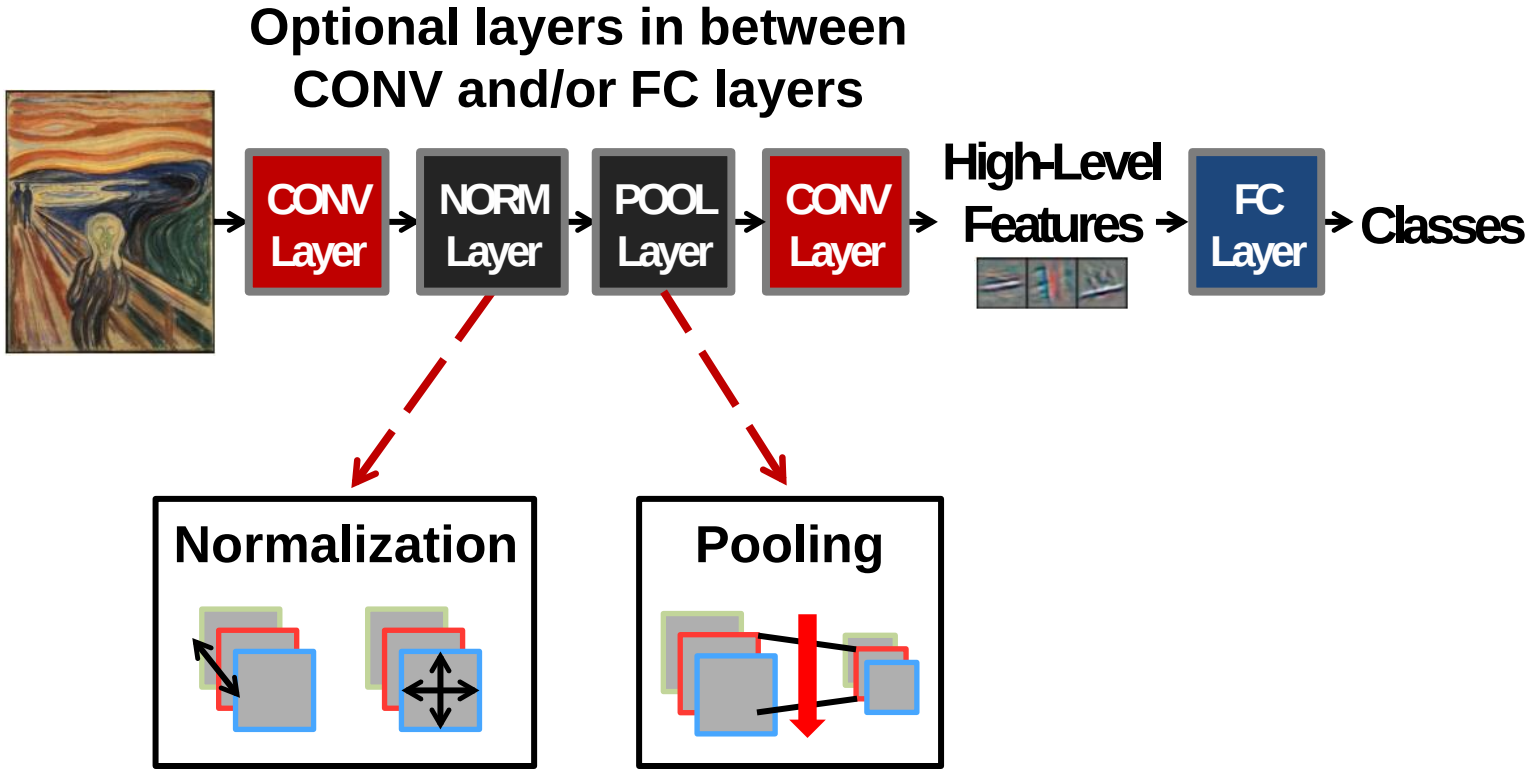
Deep Convolutional Neural Networks



Deep Convolutional Neural Networks



Deep Convolutional Neural Networks

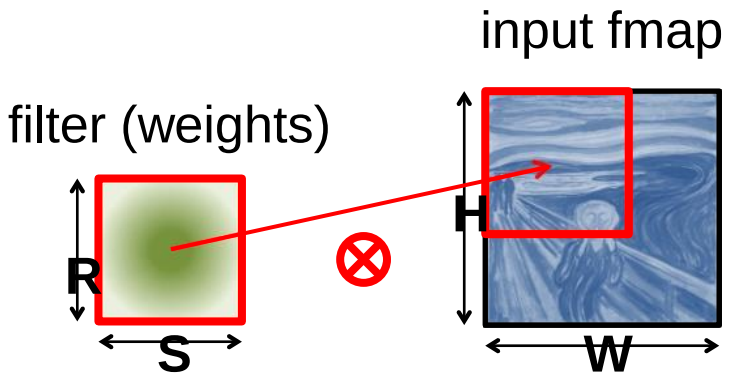


Convolution (CONV) Layer

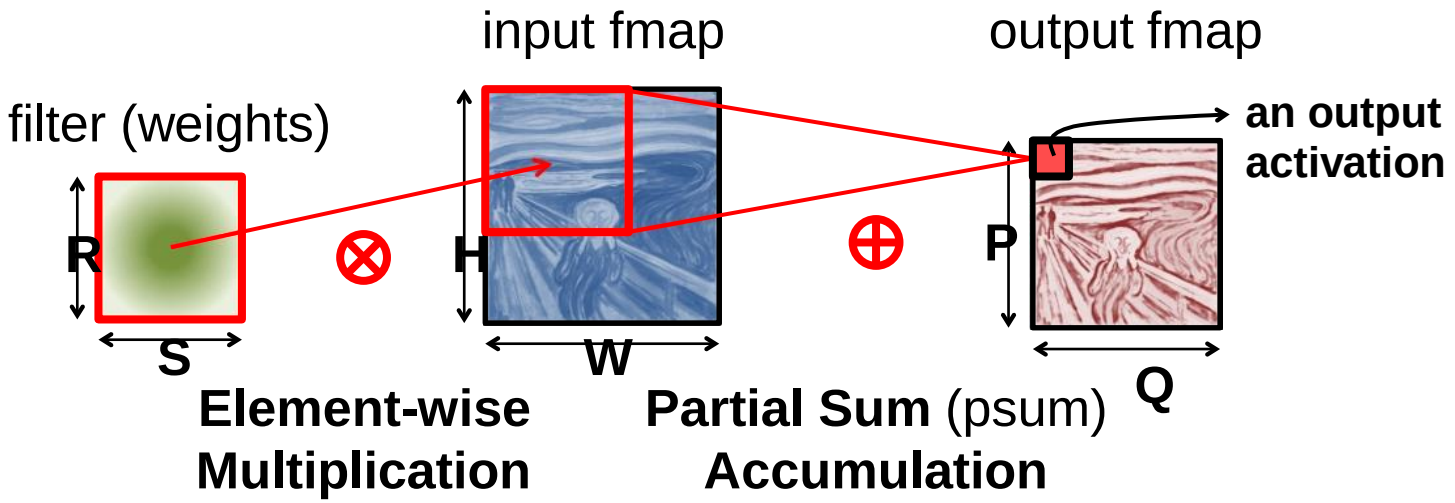
a plane of input activations
a.k.a. **input feature map (fmap)**

filter (weights)

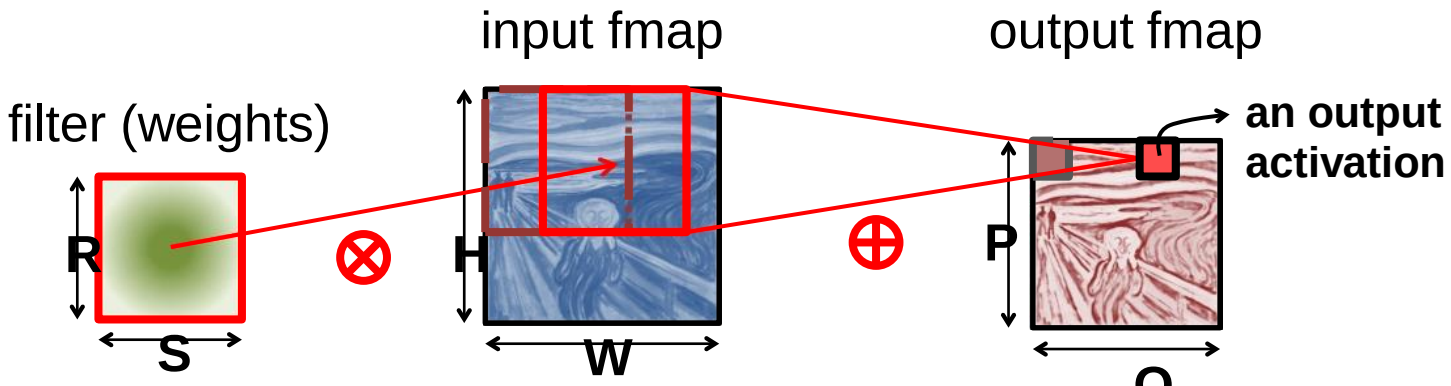
Convolution (CONV) Layer



Convolution (CONV) Layer

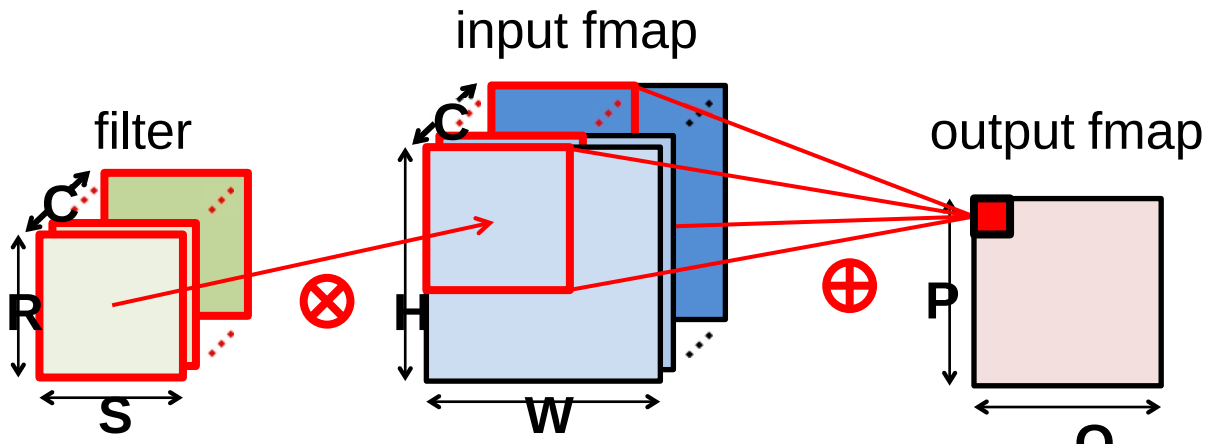


Convolution (CONV) Layer



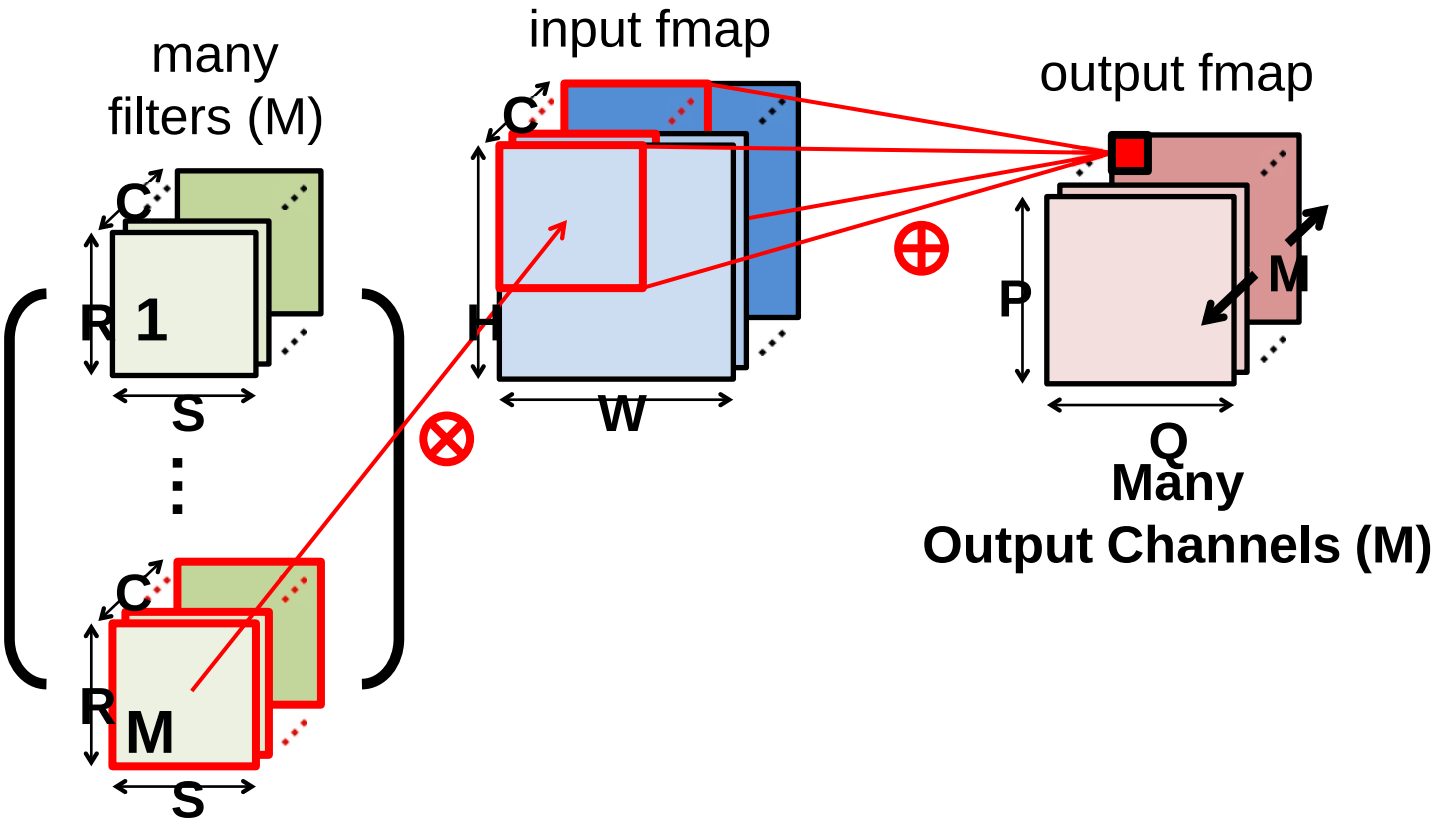
Sliding Window Processing

Convolution (CONV) Layer

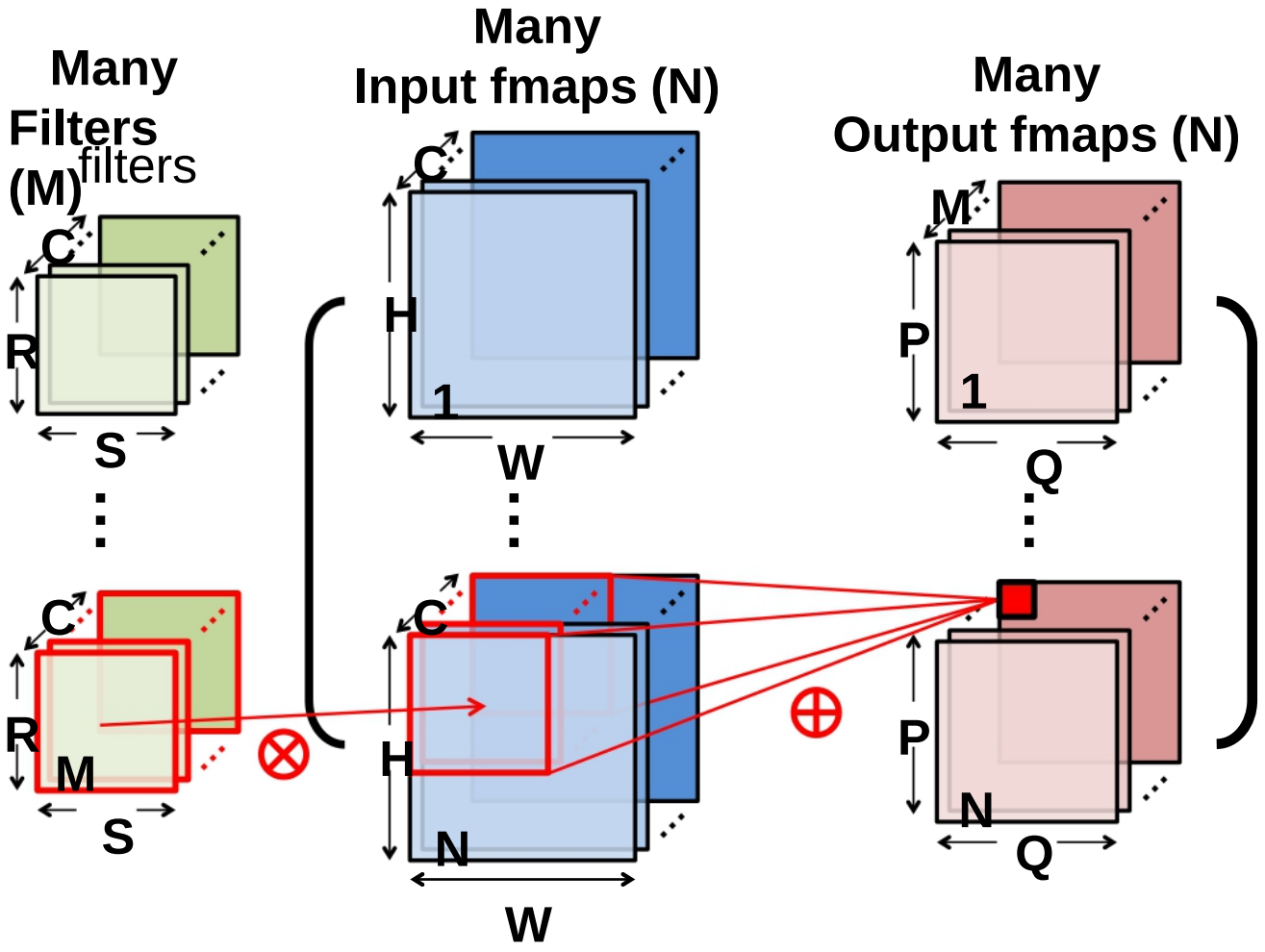


Many Input Channels (C)

Convolution (CONV) Layer

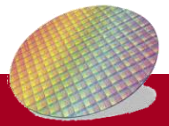


Convolution (CONV) Layer



CNN Decoder Ring

Shape Parameter	Description
N	Number of input fmaps Or Number of output fmaps (batch size)
M	Number of 2-D output fmaps Or Number of 3-D filter Or output channels
C	Number of 2-D input fmaps Or Number of 2-D filters Or Input channels
H	Height of input fmap
W	Width of input fmap
R	Height of 2-D filter (=H in FC)
S	Width of 2-D filter (=W in FC)
P	Hight of Ofmap
Q	Width of Ofmap



CONV Layer Implementation

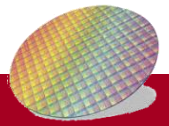
Naïve 7-layer for-loop implementation:

```
for (n=0; n<N; n++) {
    for (m=0; m<M; m++) {
        for (x=0; x<P; x++) {
            for (y=0; y<Q; y++) {
                O[n][m][x][y] = B[m];
                for (i=0; i<R; i++) {
                    for (j=0; j<S; j++) {
                        for (k=0; k<C; k++) {
                            o[n][m][x][y] += i[n][k][Ux+i][Uy+j] × f[m][k][i][j];
                        }
                    }
                }

                O[n][m][x][y] = Activation(o[n][m][x][y]);
            }
        }
    }
}
```

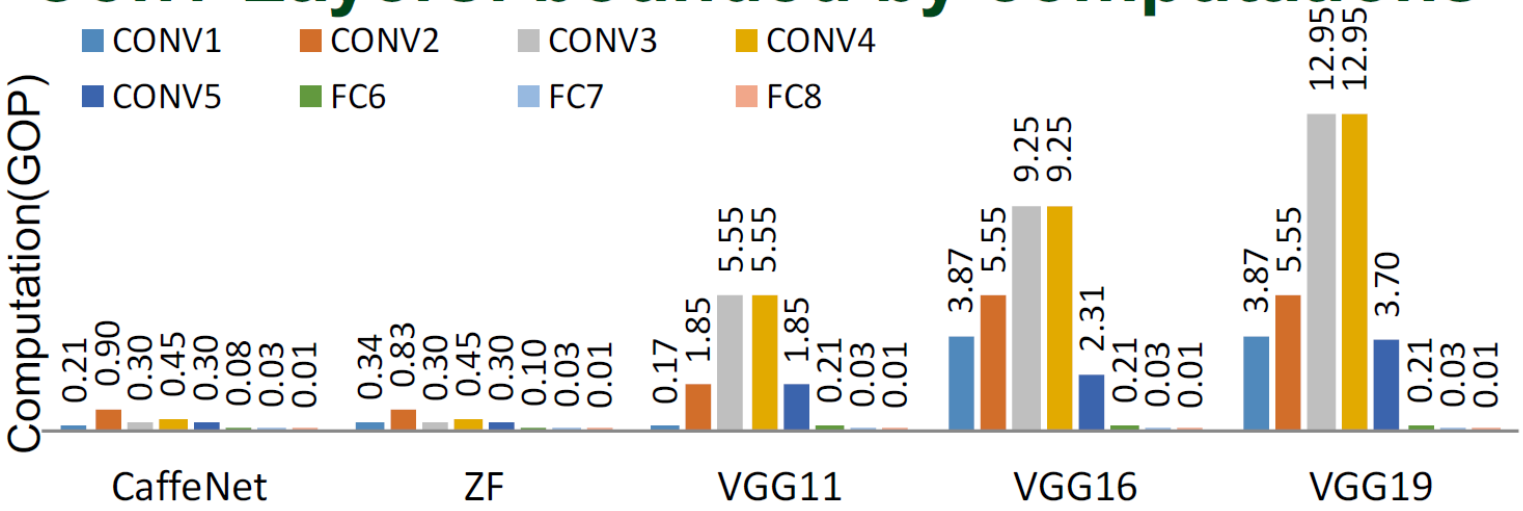
for each output fmap value

convolve a window and apply activation

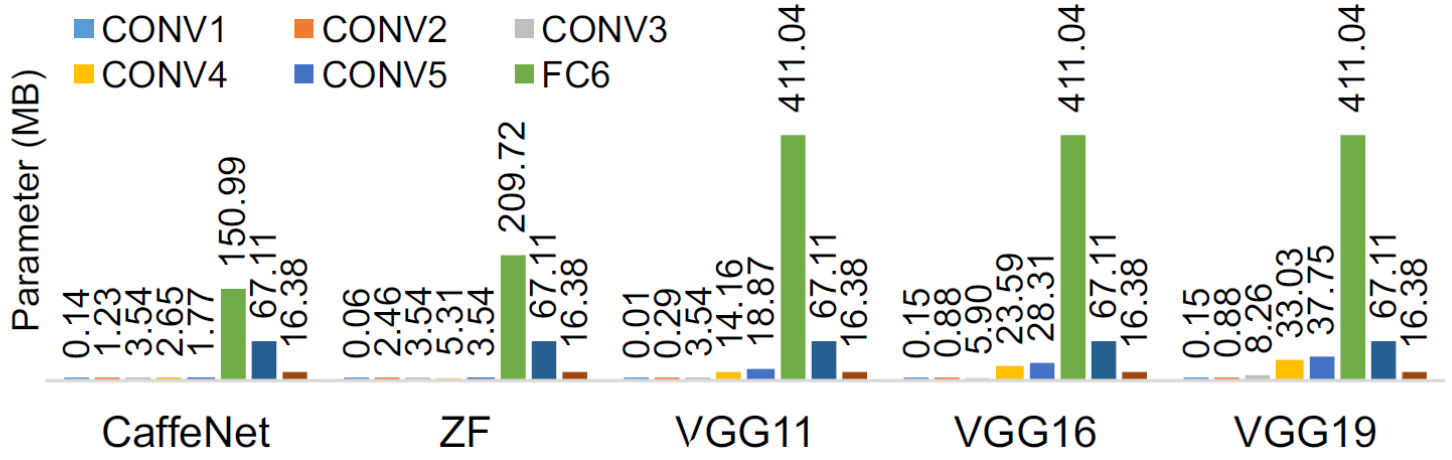


Machine Learning Has High Complexity: CNN as an example [FPGA 2016]

- **Conv Layers: bounded by computations**



- **FC Layers: bounded by memory access**



Machine Learning is power-hungry

Version	System	Match	Power	Time
AlphaGo Fan	1202*CPU+176*GPU	Vs. Fan 5:0	~200kW	2015.10
AlphaGo Lee	50*TPU	Vs. Li 4:1	~2kW	2016.3
AlphaGo Master	TPU in Single Box (4*TPU)	Vs. Ke 3:0	~200 W	2017.5

Alpha Go: V1 to V3, 1000X lower power

The computer in the autonomous car need large power and big fan to remove heat

AI Chip 是目前台灣半導體發展的主要方向

首頁 / 財經

台灣人工智慧晶片聯盟串聯百家產業鏈 打造AI晶片生態系

18:14 2020/09/21 | 工商 | 涂志豪

中華民國經濟部
Ministry of Economic Affairs, R.O.C.

認識經濟部 新聞與公告 政策計畫 法規及訴願 便民服務 經濟統計 資

本部新聞

2020-09-21 14:30 技術處

「台灣人工智慧晶片聯盟」愛台大進擊 串聯百家產業鏈 打造AI晶片全球應用服務大腦

點閱數：994

AI人工智慧預計是下一個十年最重要的技術。為了串連臺灣半導體供應鏈以全速搶攻AI市場，在行政院科技會報辦公室與經濟部技術處全力推動下，「台灣人工智慧晶片聯盟」(AI on Chip Taiwan Alliance，AITA，諧音愛台聯盟)於今(21)日舉行聯盟會員大會，現場匯集產、官、學、研及公協會代表出席，透過聯盟平台能量，集結國內外業者投入裝置端AI晶片技術的開發，並發表多項成果。

相關檔案

新聞附件：「台灣人工智慧晶片聯盟」107家會員廠商 PDF

相關圖片

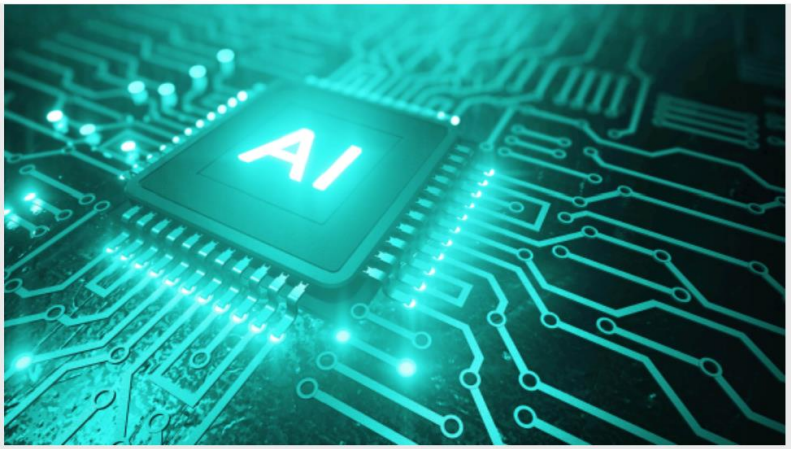
AI 決勝關鍵在於晶片

《科技與創新》AI 決勝關鍵在於晶片

讚

LINE

2019-05-20 08:00



在AI加持下，各產業創新產品與服務全面興起，背後靠的正是AI晶片的突破，有如汽車引擎一般，成為加速AI發展的關鍵！（圖片由工研院提供）

人工智慧（AI）能為人類做的事情愈來愈多，透過演算法，人工智慧不僅會跟人對話，能認出人臉，幫人看病，甚至還幫人開車。在AI加持下，各產業創新產品與服務全面興起。這些新興應用背後的運算能力，靠的是AI晶片的突破，有如汽車引擎一般，是加速AI發展的關鍵！

性能爆表、續航翻倍，蘋果自研晶片為 Mac 帶來哪些改變？

作者 愛范兒 | 發布日期 2020 年 11 月 12 日 8:30 | 分類 Apple , 晶片 , 筆記型電腦 分享 分享 Follow

[性能爆表、續航翻倍，蘋果自研晶片為 Mac 帶來哪些改變？ | TechNews 科技新報](http://technews.tw/)

http://technews.tw/2018/04/25/war-of-ai-chip/

反客為主！ 科技大咖相繼自產晶片 威脅半導體巨擘

全球科技巨頭爭相製造晶片，AI 晶片大戰即將開打？

作者 雷鋒網 | 發布日期 2018 年 04 月 25 日 8:15 | 分類 AI 人工智慧 , 晶片 , 零組件 Follow G+ 讚 407 分享

反客為主！ 科技大咖相繼自產晶片 威脅半導體巨擘 | 全球財經 | 全球 | 聯合新聞網 (udn.com)

https://www.ntdtv.com.tw Edge | edge://split-windov

新唐人亞太台 站內搜尋

•)) 鐵獻 對國家無幫助 柯文哲公布「戰友」郭台銘！與侯、馬拔河 2023 年 11 月 23 日 星期四

首頁 › 新聞 › 財經

AI晶片自研趨勢 六大科技巨頭下單台積電

更新時間：2023-11-21 12:57:35

udn / 產經 / 科技產業 聽新聞 0:00 / 0:00

微軟自研AI晶片…台積電操刀 首款產品「Maia」亮相

2023-11-17 06:20 經濟日報／編譯黃淑玲、記者尹慧中／綜合報導

外電報導，微軟規劃，Maia是專門設計用來運算大型語言模型，「訓練」AI模型與AI服務的「推論」，作為其月收30美元的AI助手「Copilot」基礎，並讓微軟Azure雲端客戶能製作客製化的AI服務，作為輝達晶片的替代選擇。

[微軟自研AI晶片…台積電操刀 首款產品「Maia」亮相 | 科技產業 | 產經 | 聯合新聞網 (udn.com)]

科技大廠自研晶片

科技大廠	名　稱	製程節點	初期估年投產晶圓
微　軟	Maia	5nm	1.2
谷　歌	TPUv5	5nm	4.0
	TPUv6	3nm	NA
亞馬遜	Trn1	7nm	5.5
	Trn2	5nm	1.2
Meta	MTIA	7nm	1.5
特斯拉	Dojo	7nm	1.0

資料來源：採訪整理、研究機構報告　　單位：萬片　　製表：科技新聞中心

Google 公開自研晶片「Tensor」，同步揭露 Pixel 6 系列重點特色 | TechNews 科技新報 (2021/8/3)
https://www.itho me.com.tw/news/ 127947

https://udn.com/news/story/6871/3067694

全球科技巨頭爭相製造晶片，AI 晶片大戰即將開打？

作者 雷鋒網 | 發布日期 2018 年 04 月 25 日 8:15 | 分類 AI 人工智慧 , 晶片 , 零組件 Follow G+ 讚 407 分享

Google TPU

- TPU Card to replace a disk
- Up to 4 cards / server

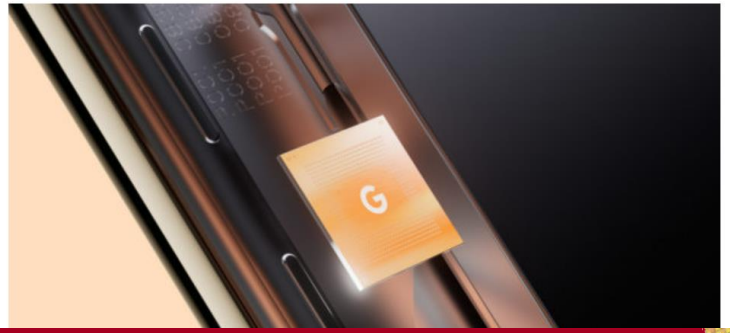
TPU Card & Package

[Google 公開自研晶片「Tensor」，同步揭露 Pixel 6 系列重點特色 | TechNews 科技新報](#) (2021/8/3)

零組件 行動裝置 網路 AI 人工智慧 尖端科技 生物科技 能源科技 系列專題 財經 財報快訊 拓墣觀點 市場動態

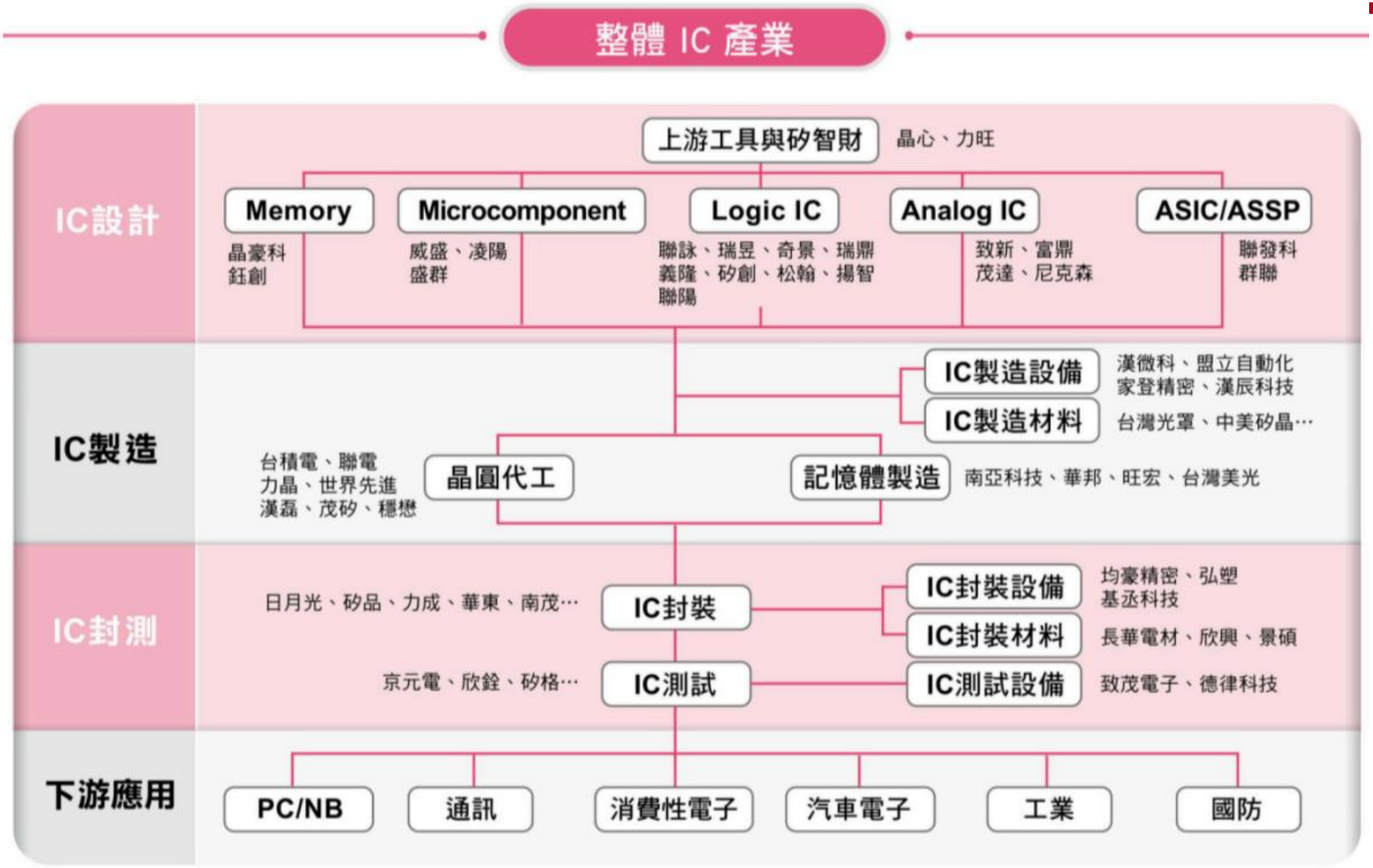
Google 公開自研晶片「Tensor」，同步揭露 Pixel 6 系列重點特色

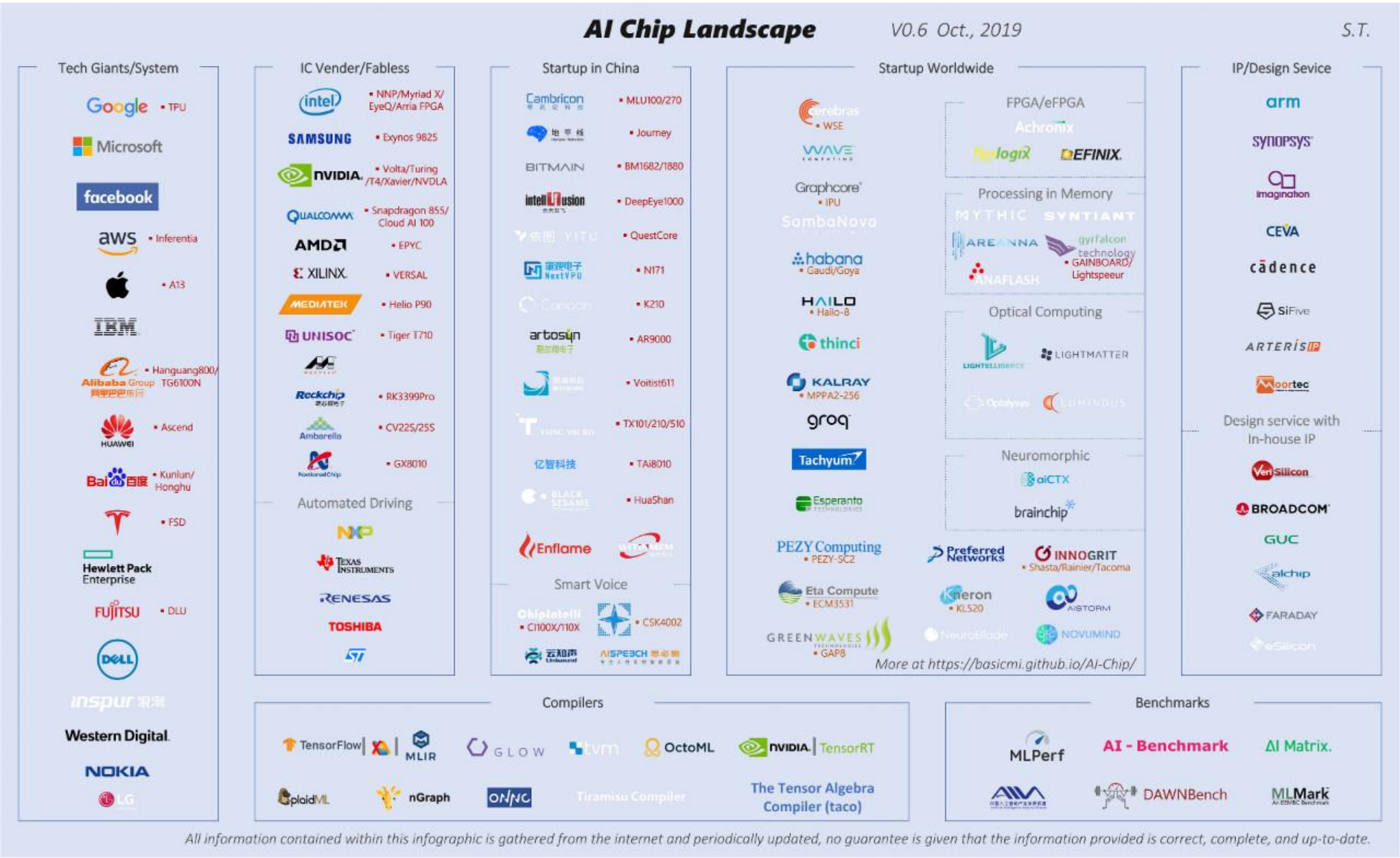
作者 陳 冠榮 | 發布日期 2021 年 08 月 03 日 3:00 | 分類 Android 手機 , Google , 晶片 分享 分享 Follow



[淺談 Google 的 TPU - UNWIRE.PRO](#)

台灣IC產業地圖


整體 IC 產業
IC設計
上游工具與矽智財
晶心、力旺
Memory
晶豪科
鈺創
Microcomponent
威盛、凌陽
盛群
Logic IC
聯詠、瑞昱、奇景、瑞鼎
義隆、矽創、松翰、揚智
聯陽
Analog IC
致新、富鼎
茂達、尼克森
ASIC/ASSP
聯發科
群聯
IC製造
IC製造設備
漢微科、盟立自動化
家登精密、漢辰科技
IC製造材料
台灣光罩、中美矽晶…
台積電、聯電
力晶、世界先進
漢磊、茂矽、穩懋
晶圓代工
記憶體製造
南亞科技、華邦、旺宏、台灣美光
IC封測
日月光、矽品、力成、華東、南茂…
IC封裝
IC封裝設備
均豪精密、弘塑
基丞科技
IC封裝材料
長華電材、欣興、景碩
京元電、欣銓、矽格…
IC測試
IC測試設備
致茂電子、德律科技
下游應用
PC/NB
通訊
消費性電子
汽車電子
工業
國防



AI-Chip | A list of ICs and IPs for AI, Deep Learning. (basicmi.github.io)

Course Administration

- Instructor: Ing-Chao Lin (林英超)
  - email: iclin@mail.ncku.edu.tw
  - Tel: +8866-2757575 ext. 62553

Course Website:

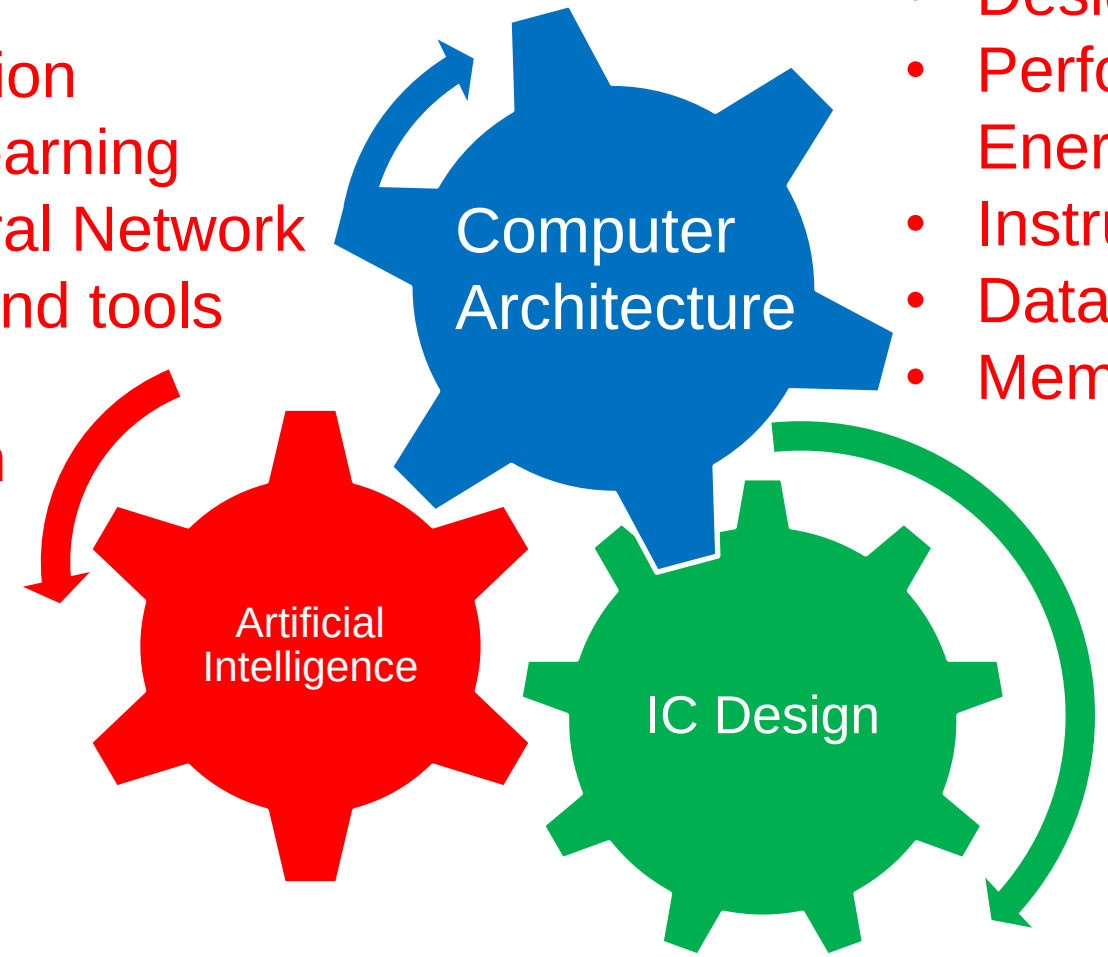
http://moodle.ncku.edu.tw

announces and slides will be posted here

submit your homework there

This course is interdisciplinary

- AI application
- Machine learning
- Deep Neural Network
- Software and tools
  - Python
  - Pytorch
  - ….

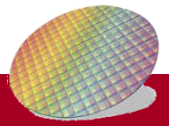


- Design Metrics
- Performance, Power, Energy estimation
- Instruction Set Archi.
- Data path
- Memory Hierarchy

- FPGA& VLSI Design & Design flow
- HDL
- Comb. & Seq. & FSM

Topics Covered-2025

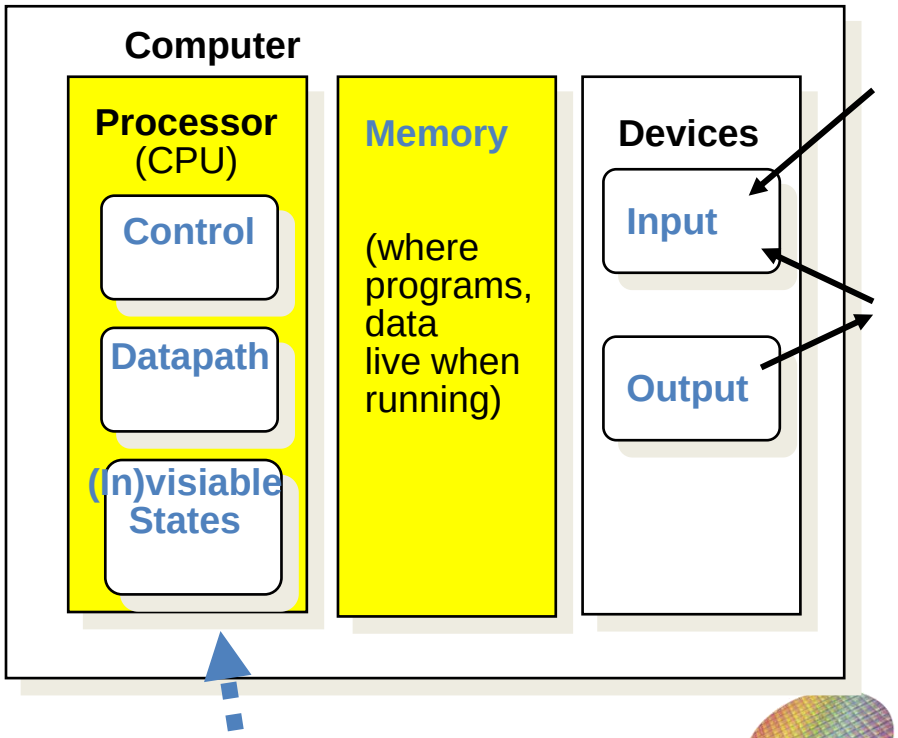
- Digital integrated circuits and Computer Architecture (7~8 Weeks)
  - Memory Hierarchy
  - Instruction Level Parallelism
  - Thread Level Parallelism
  - Data level Parallelism
- Deep Learning IC Design (7~8 Weeks)
  - DNN Background
  - DNN Development Resource
  - DNN Inference Background
  - DNN Hardware /DNN Dataflow
  - DNN Hardware Software Codesign
  - Case Study



Prerequisite

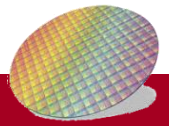
- Digital Design & Computer Organization
  - Combinational Logic and Synchronous Sequential Logic
  - Instruction Set
  - DataPath &Pipeline

- Verilog
  - You need to know how to write combinational, sequential, and FSM
  - We will only cover FPGA on labs
- Machine Learning
  - Basic knowledge of Neural Network, pytorch or other ML development language and tools



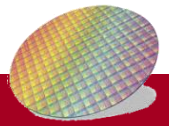
Online and Make-up class

- Online courses
  - 2/25 9:10
  - Will announce in advance
- Make-up class
  - 7:00 PM @ March 4 (Tuesday)



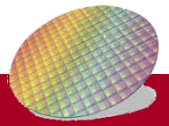
Textbook & Materials

- Textbook
  - Digital integrated circuits
    - Integrated Circuit Design, fourth edition, by N. Weste
  - Computer Architecture, 5th or 6th ed. by D. Patterson
  - Efficient Processing of Dee Neural Network
    - Efficient Processing of Deep Neural Networks | SpringerLink
- Selected Paper
- Zynq & Pynq:
  - http://www.pynq.io
  - Embedded Designs — Embedded Design Tutorials 2023.1 documentation (xilinx.github.io)



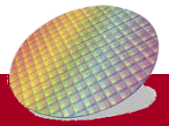
Tentative Grading (Spring 2025)

- Homework and Lab Assignment (25 %)
- Midterm Exam (20 %)
  - 4/22
- Final Exam (25 %)
  - 6/10
- Project (15%)
- Class Participation (15%)
  - Attendance, In-class quiz
  - Q&A, other activities (參加演講或是其他活動, 15%)
- Percentage of each part may vary slightly



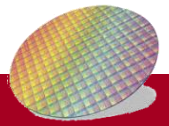
Note for Lab and Homework Assignment

- Grading for each homework
  - Report: You need to carefully explain your codes and results in you results. If your explanation is unclear, you will lose points
- Honesty is the best policy. If you copy your homework from someone, both will get zero points for your homework. In addition, both will lose points of your final grade for each violation.



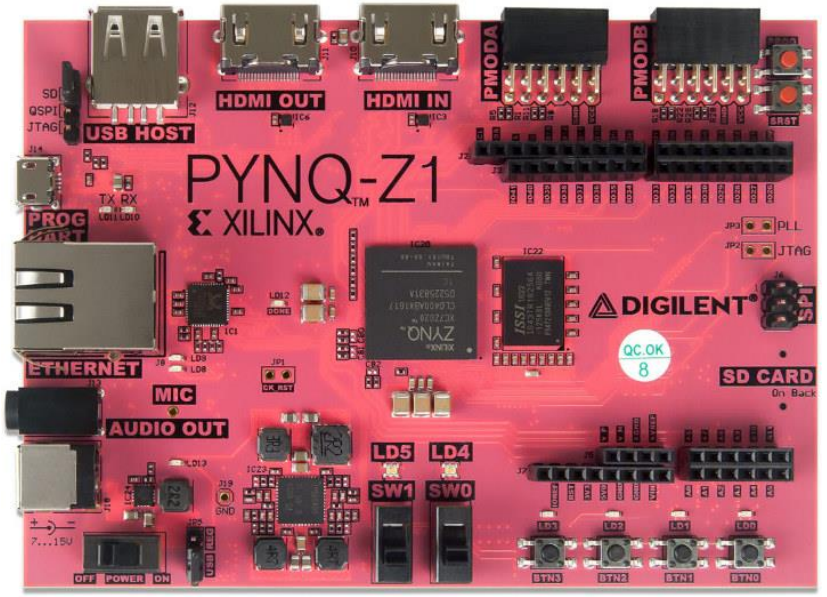
Note for Exam

- Finish it individually
- In-class exam
- Normally, it takes 3 hours
  - Part 1: closed book: Most questions are from the course material
  - Part 2: open book
- I will provide some exercises for you to prepare exams.



Notes for Project

- Two students form a group
- I will list some previous projects for your reference or you can find by yourselves
- Finish the project and present the project
- Higher score
  - If you use FPGA parts of the Pynq
  - if you do more modifications or create your won design



In-class quiz

- In-class quick test is a very simple test.
- Will be announced in a few days before. Test what you have learn.
- Normally take less than 20 minutes.
  - If you can't do anything, you will get 40 points
- Cover what I taught in the previous classes

Backup Slides