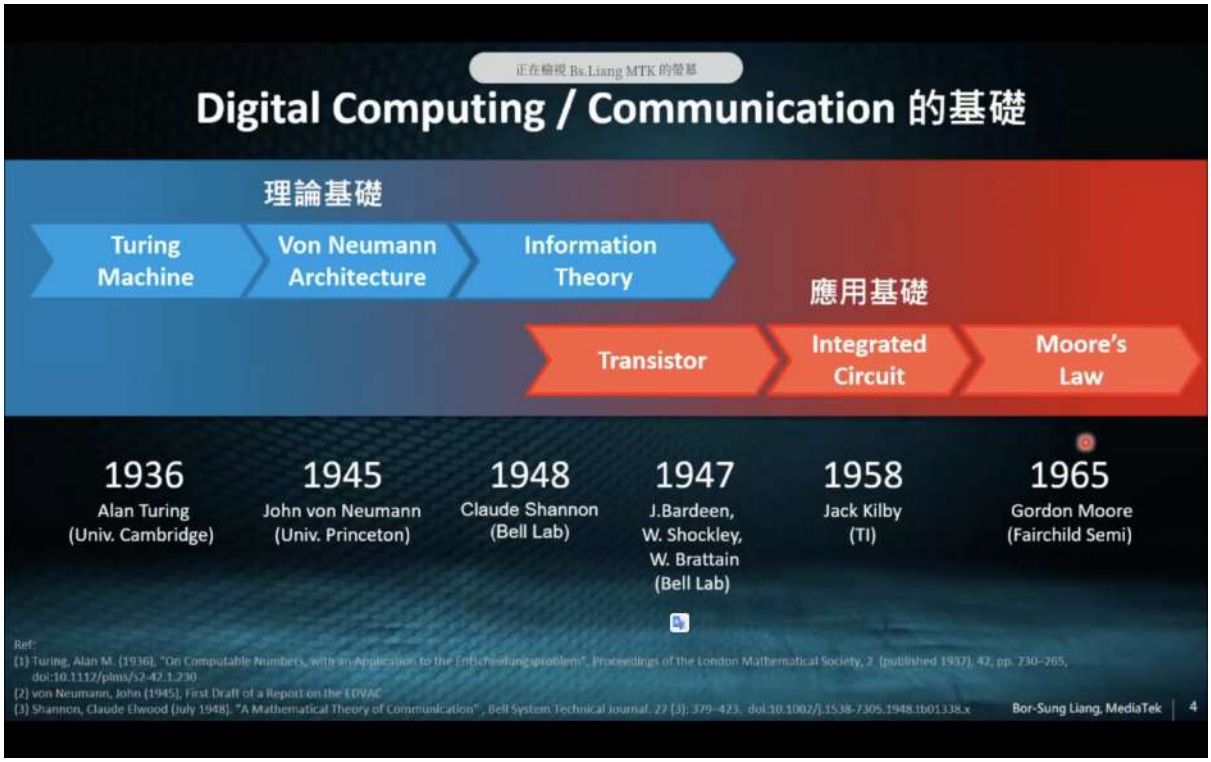
Course 1: IC運算平台趨勢：數位運算、人工智慧與量子運算

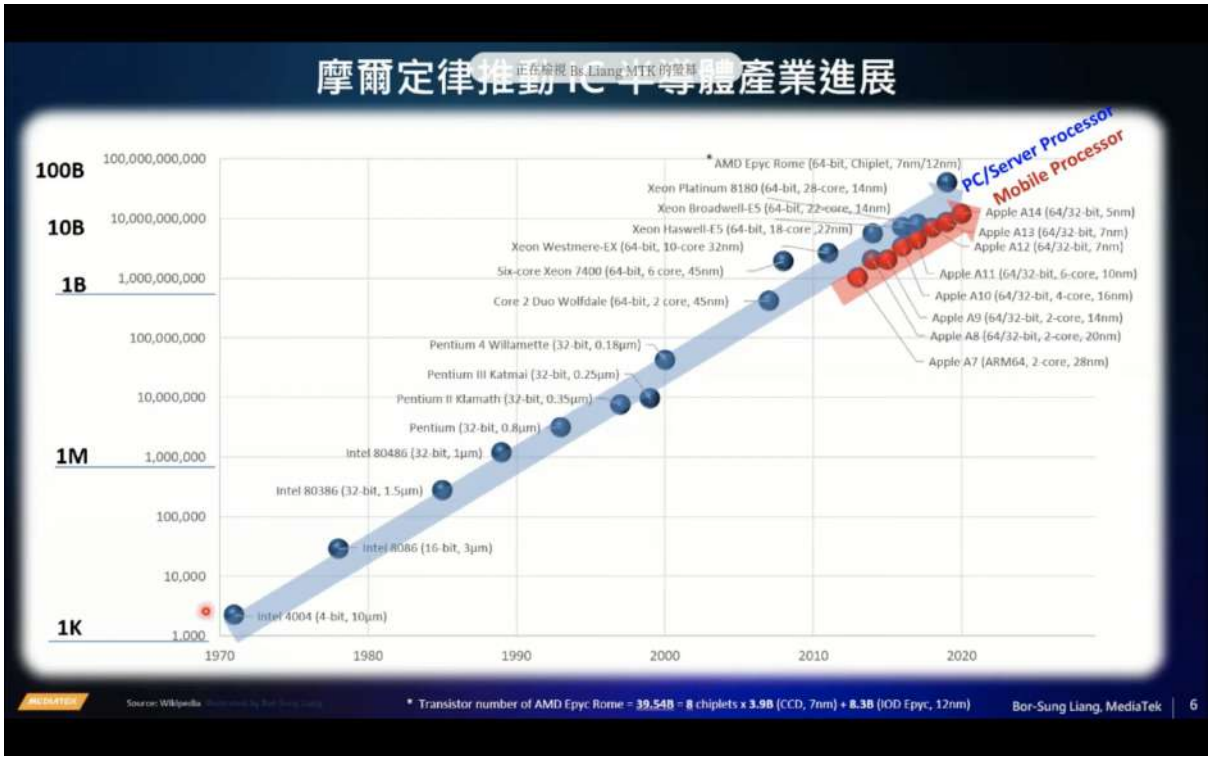


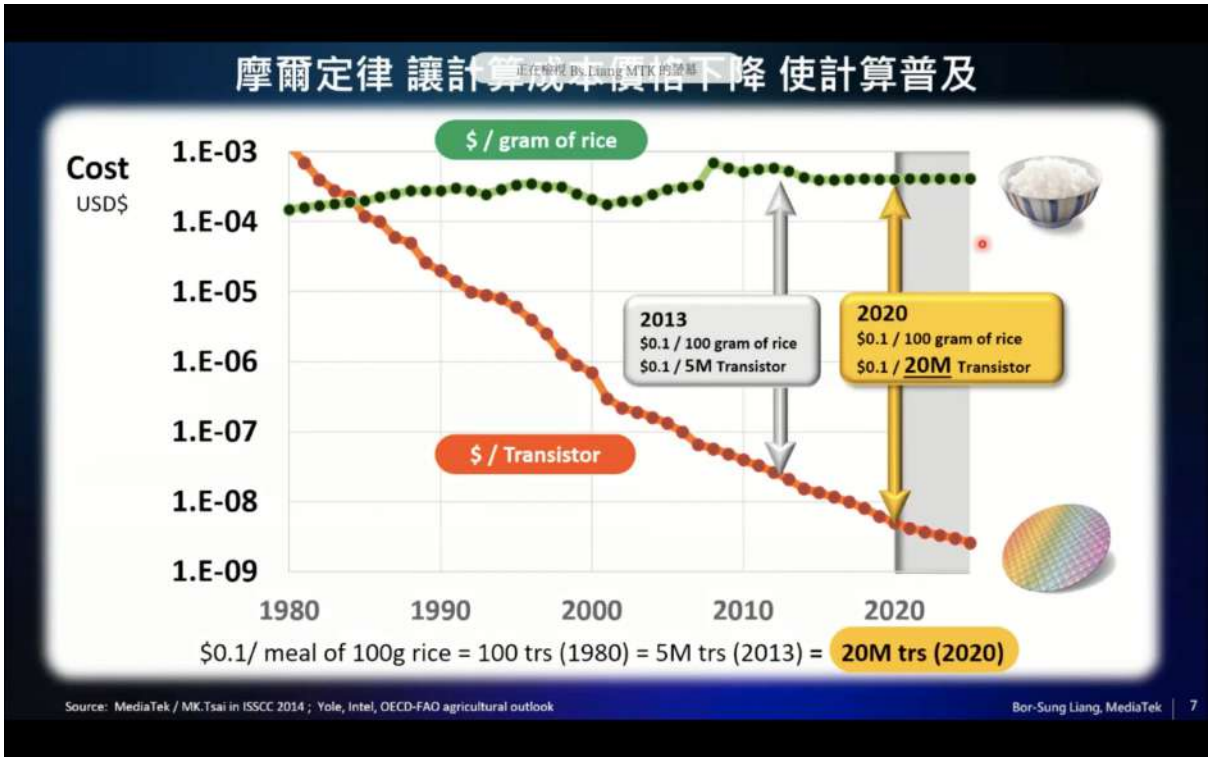
0与1的运算逻辑 → 计算功能

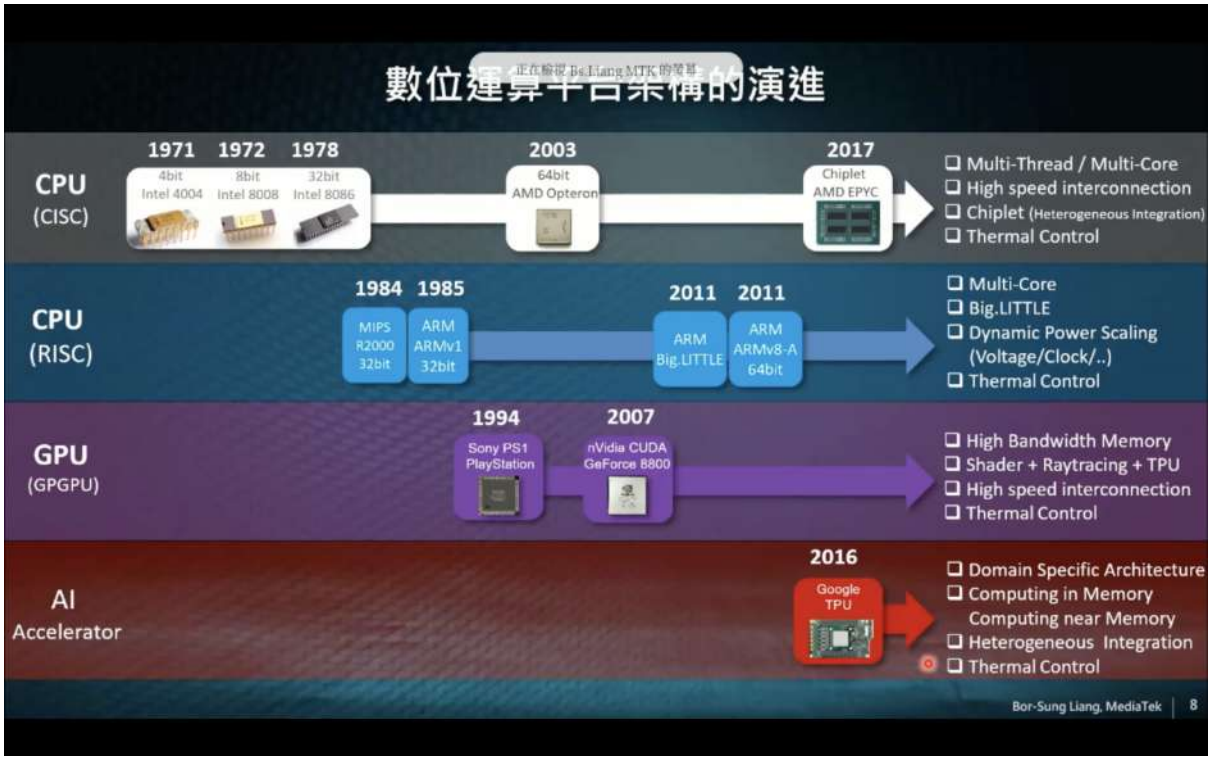
注意不是算盘（记录功能），无control功能

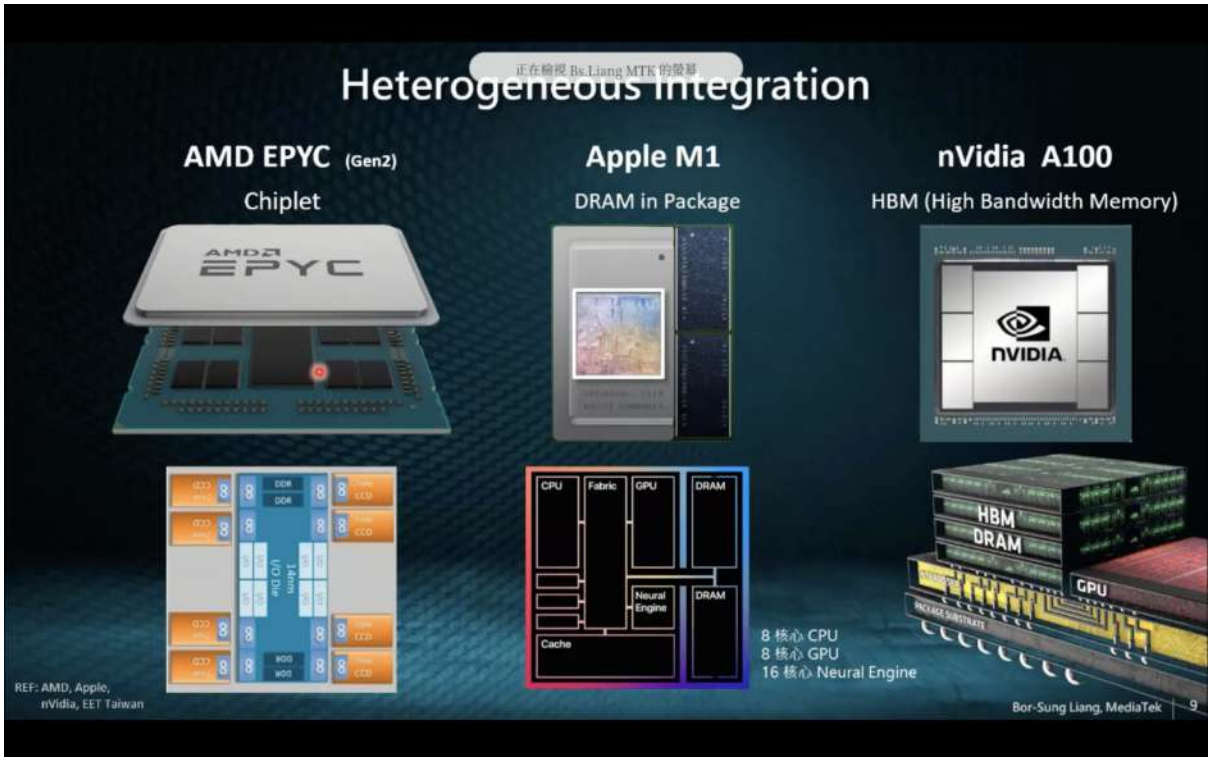
Charles Babbage Difference Engine是运算平台的起源

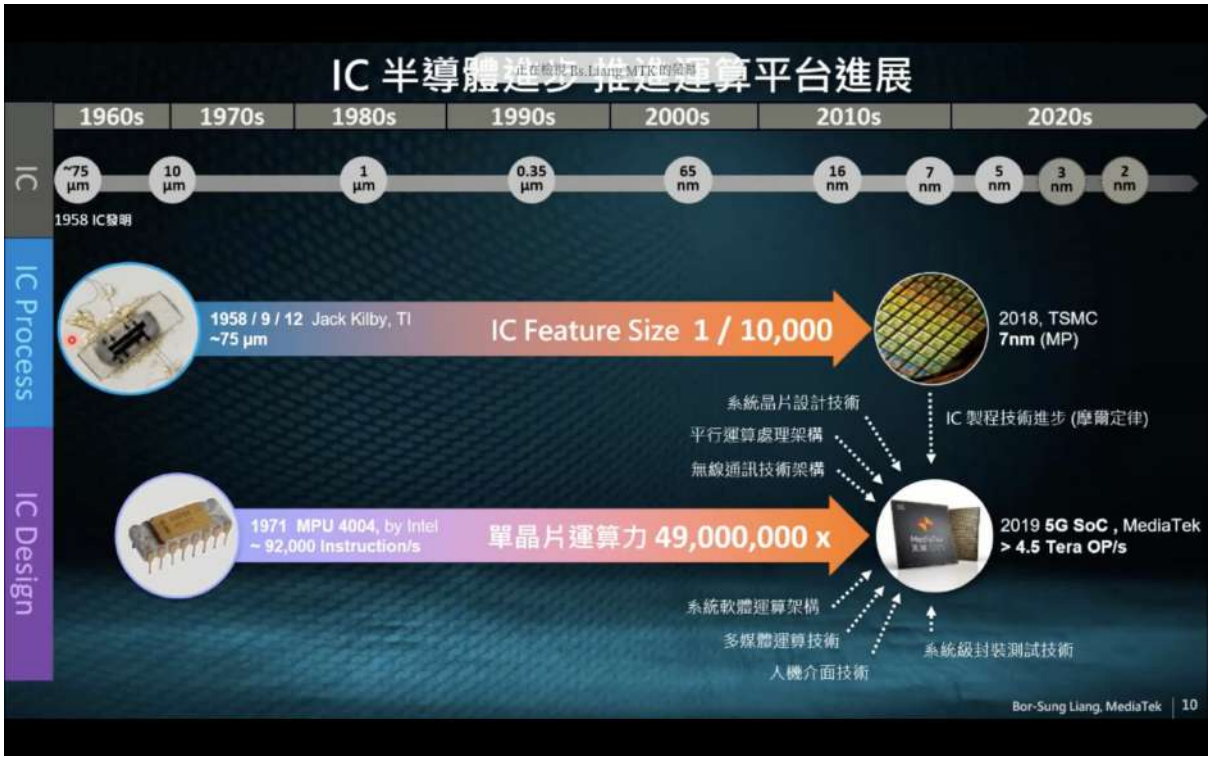

運算平台的起源
正在檢視 Bs Liang MTK 的螢幕
電子式電腦
半導體 Semiconductor
機械式開關
1822
Charles Babbage
Difference Engine
1904
(UK) John Ambrose Fleming
真空管 Vacuum tube
1946
US Army Ballistic Research Lab
ENIAC
1947 /12/23
(Bell Labs) Bardeen, Brattain, Shockley
電晶體 Transistor
1953
Univ of Manchester
Transistor Computer
1958 /9/12
(TI) Jack Kilby
積體電路 IC
1963
MIT Instrumentation Lab
Apollo Guidance Computer
AND
XOR
Bor-Sung Liang, MediaTek
5


摩爾定律推動 IC 半導體產業進展
正在檢視 Bs Liang MTK 的螢幕
100B
10B
1B
1M
1K
100,000,000,000
10,000,000,000
1,000,000,000
100,000,000
10,000,000
1,000,000
100,000
10,000
1,000
1970
1980
1990
2000
2010
2020
PC/Server Processor
Mobile Processor
AMD Epyc Rome (64-bit, Chiplet, 7nm/12nm)
Xeon Platinum 8180 (64-bit, 28-core, 14nm)
Xeon Broadwell-E5 (64-bit, 22-core, 14nm)
Xeon Haswell-E5 (64-bit, 18-core, 22nm)
Xeon Westmere-EX (64-bit, 10-core 32nm)
Six-core Xeon 7400 (64-bit, 6 core, 45nm)
Core 2 Duo Wolfdale (64-bit, 2 core, 45nm)
Pentium 4 Willamette (32-bit, 0.18µm)
Pentium III Katmai (32-bit, 0.25µm)
Pentium II Klamath (32-bit, 0.35µm)
Pentium (32-bit, 0.8µm)
Intel 80486 (32-bit, 1µm)
Intel 80386 (32-bit, 1.5µm)
Intel 8086 (16-bit, 3µm)
Intel 4004 (4-bit, 10µm)
Apple A14 (64/32-bit, 5nm)
Apple A13 (64/32-bit, 7nm)
Apple A12 (64/32-bit, 7nm)
Apple A11 (64/32-bit, 6-core, 10nm)
Apple A10 (64/32-bit, 4-core, 16nm)
Apple A9 (64/32-bit, 2-core, 14nm)
Apple A8 (64/32-bit, 2-core, 20nm)
Apple A7 (ARM64, 2-core, 28nm)
Source: Wikipedia
* Transistor number of AMD Epyc Rome = 39.54B = 8 chiplets x 3.9B (CCD, 7nm) + 8.3B (IOD Epyc, 12nm)
Bor-Sung Liang, MediaTek
6

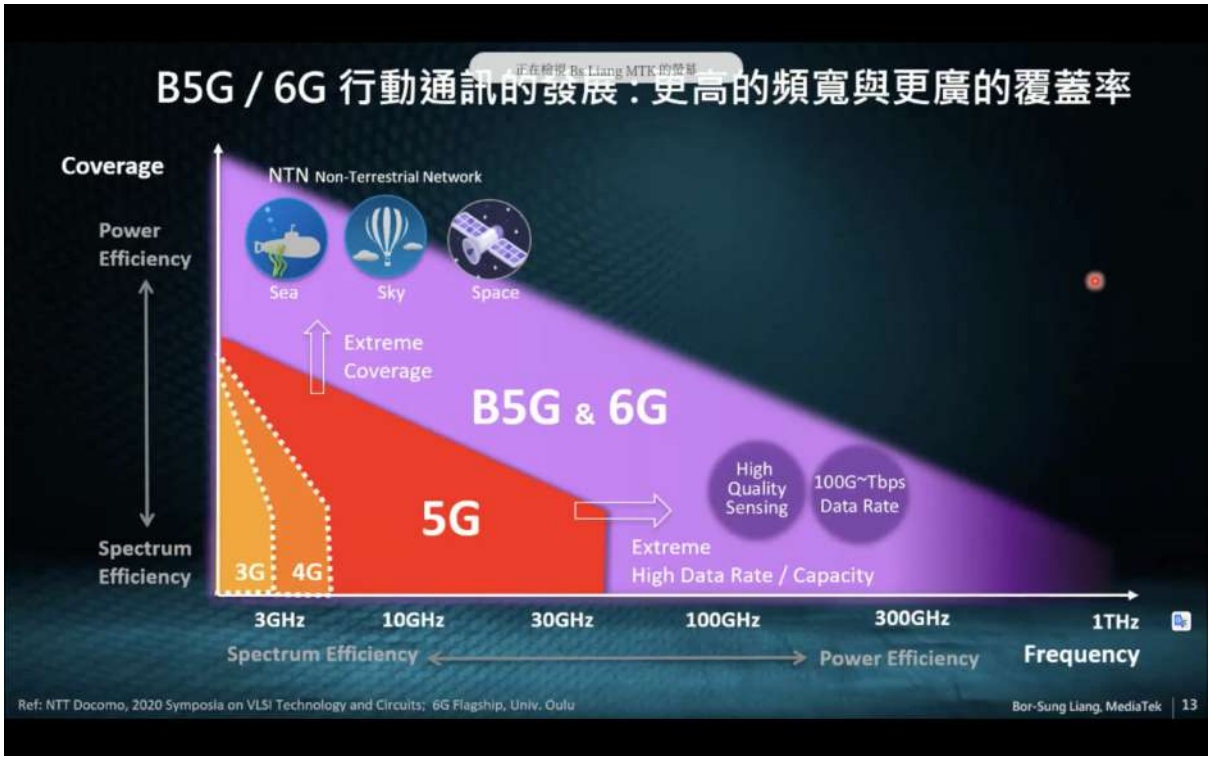

摩爾定律 讓計算成本價格下降 使計算普及
Cost
USD$
1.E-03
1.E-04
1.E-05
1.E-06
1.E-07
1.E-08
1.E-09
$ / gram of rice
$ / Transistor
2013
$0.1 / 100 gram of rice
$0.1 / 5M Transistor
2020
$0.1 / 100 gram of rice
$0.1 / 20M Transistor
1980
1990
2000
2010
2020
$0.1/ meal of 100g rice = 100 trs (1980) = 5M trs (2013) = 20M trs (2020)
Source: MediaTek / MK.Tsai in ISSCC 2014 ; Yole, Intel, OECD-FAO agricultural outlook
Bor-Sung Liang, MediaTek | 7


數位運算平台架構的演進
CPU
(CISC)
1971
4bit
Intel 4004
1972
8bit
Intel 8008
1978
32bit
Intel 8086
2003
64bit
AMD Opteron
2017
Chiplet
AMD EPYC
Multi-Thread / Multi-Core
High speed interconnection
Chiplet (Heterogeneous Integration)
Thermal Control
CPU
(RISC)
1984
MIPS
R2000
32bit
1985
ARM
ARMv1
32bit
2011
ARM
Big.LITTLE
2011
ARM
ARMv8-A
64bit
Multi-Core
Big.LITTLE
Dynamic Power Scaling
(Voltage/Clock/..)
Thermal Control
GPU
(GPGPU)
1994
Sony PS1
PlayStation
2007
nVidia CUDA
GeForce 8800
High Bandwidth Memory
Shader + Raytracing + TPU
High speed interconnection
Thermal Control
AI
Accelerator
2016
Google
TPU
Domain Specific Architecture
Computing in Memory
Computing near Memory
Heterogeneous Integration
Thermal Control
Bor-Sung Liang, MediaTek | 8


正在檢視 Bs.Liang MTK 的螢幕
Heterogeneous Integration
AMD EPYC (Gen2)
Chiplet
Apple M1
DRAM in Package
nVidia A100
HBM (High Bandwidth Memory)
AMD EPYC
NVIDIA
CPU
Fabric
GPU
DRAM
Neural Engine
DRAM
Cache
8 核心 CPU
8 核心 GPU
16 核心 Neural Engine
HBM
DRAM
GPU
PACKAGE SUBSTRATE
REF: AMD, Apple, nVidia, EET Taiwan
Bor-Sung Liang, MediaTek 9

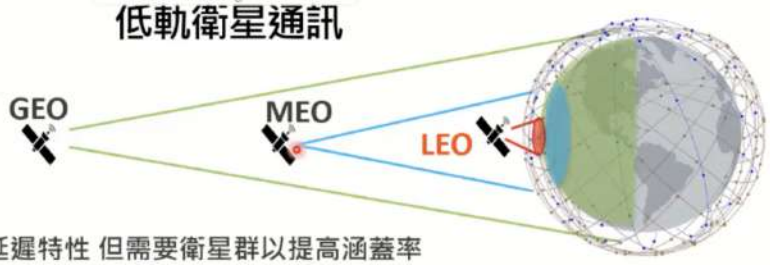

正在檢視 Bs.Liang MTK 的螢幕
IC 半導體進步推進運算平台進展
IC
IC Process
IC Design
1960s
1970s
1980s
1990s
2000s
2010s
2020s
~75 µm
10 µm
1 µm
0.35 µm
65 nm
16 nm
7 nm
5 nm
3 nm
2 nm
1958 IC發明
1958 / 9 / 12 Jack Kilby, TI
~75 µm
IC Feature Size 1 / 10,000
2018, TSMC
7nm (MP)
系統晶片設計技術
平行運算處理架構
無線通訊技術架構
IC 製程技術進步 (摩爾定律)
1971 MPU 4004, by Intel
~ 92,000 Instruction/s
單晶片運算力 49,000,000 x
2019 5G SoC , MediaTek
> 4.5 Tera OP/s
系統軟體運算架構
多媒體運算技術
人機介面技術
系統級封裝測試技術
Bor-Sung Liang, MediaTek 10


IC 推動 Computing 與 Communication 的發展
1960s
1970s
1980s
1990s
2000s
2010s
2020s
IC
~75 µm
10 µm
1 µm
0.35 µm
65 nm
16 nm
7 nm
5 nm
3 nm
2 nm
1958 IC發明
1971 Microprocessor (Intel)
1981 PC (IBM)
1995 Internet Commercial Use
2007 Smartphone (iOS, Android)
2016 AI Processor (Google TPU)
Computing
Consumer Electronics
PC
PC + Internet
Smartphone Mobile Internet
AI + IoT
Comm.
0G
1G ~2.4K bps
2G ~64K bps
3G ~384K bps
4G ~100M bps
5G >G bps
Analog Voice
Digital Voice SMS
Multimedia Social Media
Video Streaming Gaming
eMBB
mMTC
uRLLC
Taiwan
交大 電子所
RCA 技轉
tsmc
MEDIATEK
Feature Phone SoC
Smartphone SoC
5G SoC
1958
1974
1980
1987
1997
2004
2011
2019
Bor-Sung Liang, MediaTek 11


B5G / 6G 行動通訊的發展：更高的頻寬與更廣的覆蓋率
Coverage
Power Efficiency
Spectrum Efficiency
NTN Non-Terrestrial Network
Sea
Sky
Space
Extreme Coverage
B5G & 6G
5G
3G
4G
High Quality Sensing
100G~Tbps Data Rate
Extreme High Data Rate / Capacity
3GHz
10GHz
30GHz
100GHz
300GHz
1THz
Spectrum Efficiency
Power Efficiency
Frequency
Ref: NTT Docomo, 2020 Symposia on VLSI Technology and Circuits; 6G Flagship, Univ. Oulu
Bor-Sung Liang, MediaTek 13

正在檢視 Bs.Liang MTK 的螢幕

低軌衛星通訊



LEO 有低延遲特性 但需要衛星群以提高涵蓋率

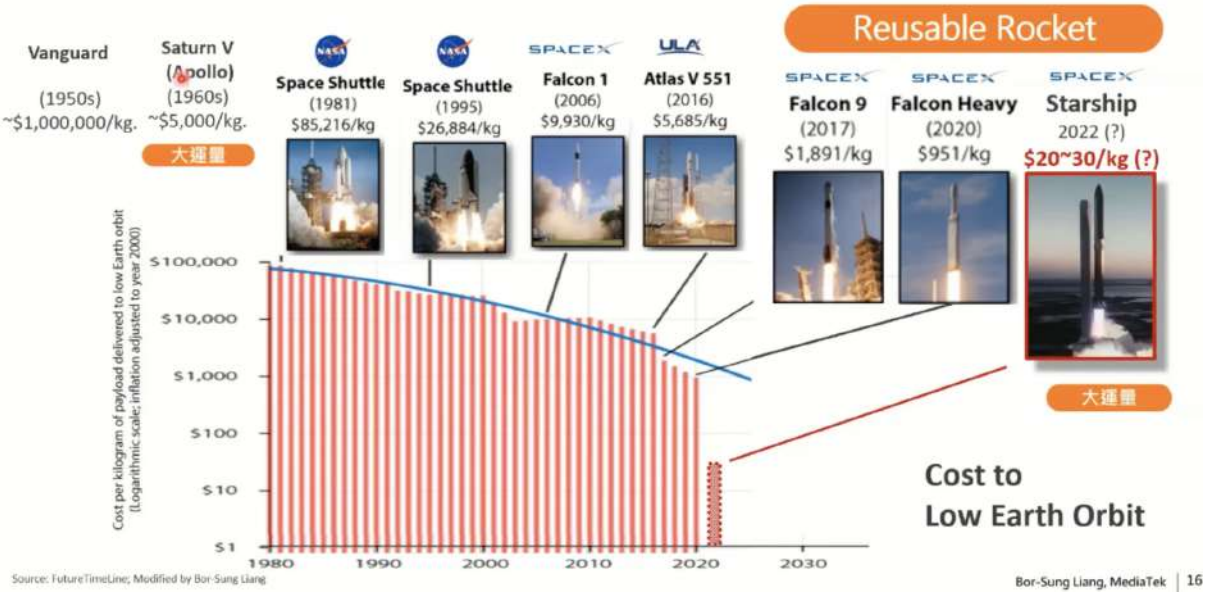
衛星軌道	LEO Low Earth Orbit	MEO Medium Earth Orbit	GEO Geostationary Earth Orbit
距地球高度(公里)	160~2,000	2,000~35,786	≧35,786
覆蓋地球 衛星數與覆蓋度	全球涵蓋需要上千顆衛星 覆蓋度可達100%	6個衛星， 覆蓋度達96%	3個衛星，可覆蓋近99%， 無法覆蓋極圈
軌道週期	約1.5~2小時	約12小時	24小時(地球自轉同步)
RTT(m sec)	30~50	125~250	600~800
應用	國際太空站、低軌衛星通訊 Iridium, Starlink, Kupier	衛星定位系統，美 GPS、中國北斗、 俄國GLONASS、歐盟 Galileo	衛星電視、 廣播、氣象觀測

資料來源：Avnet Silica、Telesat、DIGITIMES Research整理，2021/1

* RTT = Round Trip Time

正在檢視 Bs.Liang MTK 的螢幕

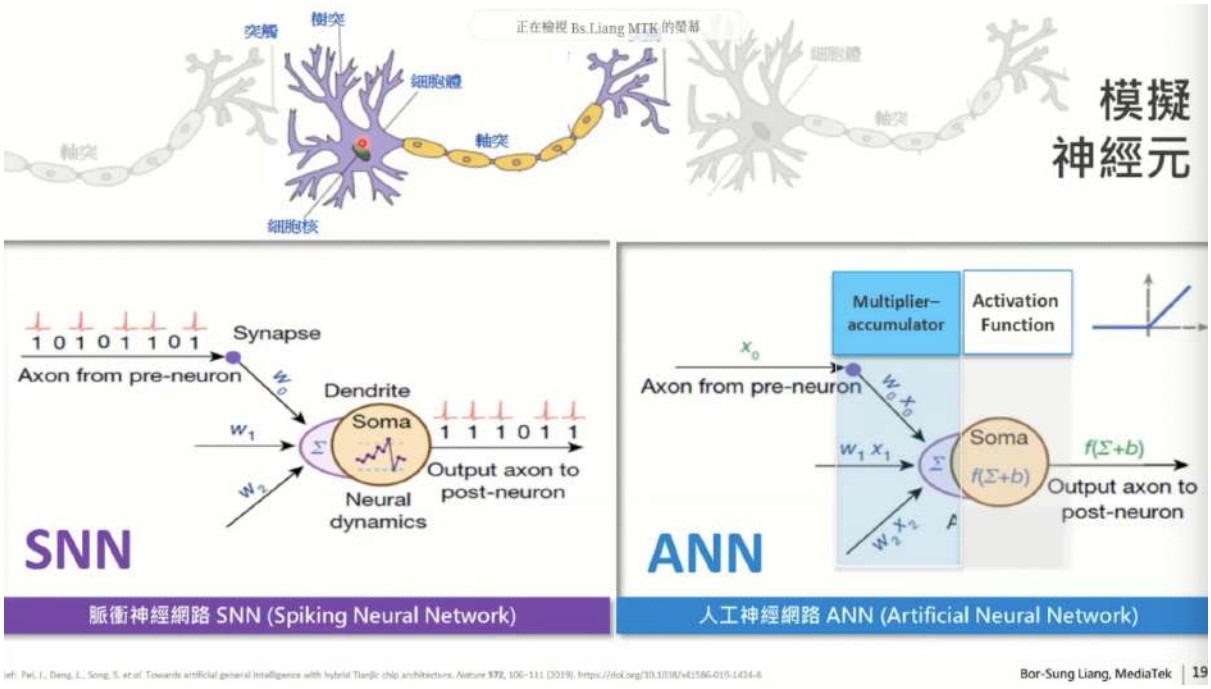
參考資料: 低軌衛星通訊的啟動關鍵 - 可重複使用火箭



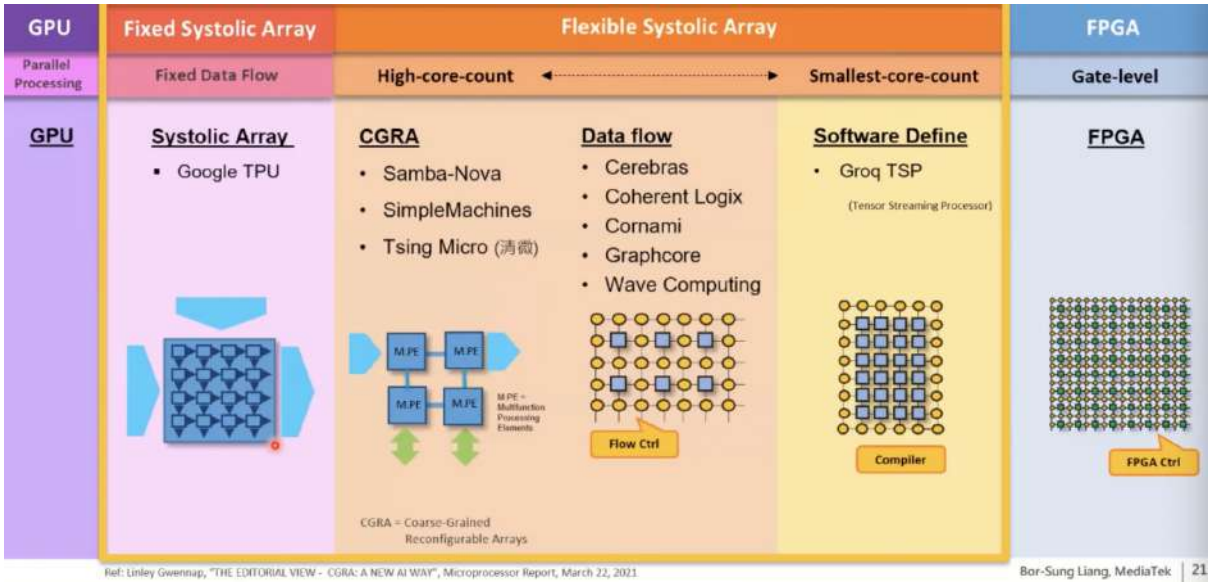
Source: FutureTimeLine; Modified by Bor-Sung Liang

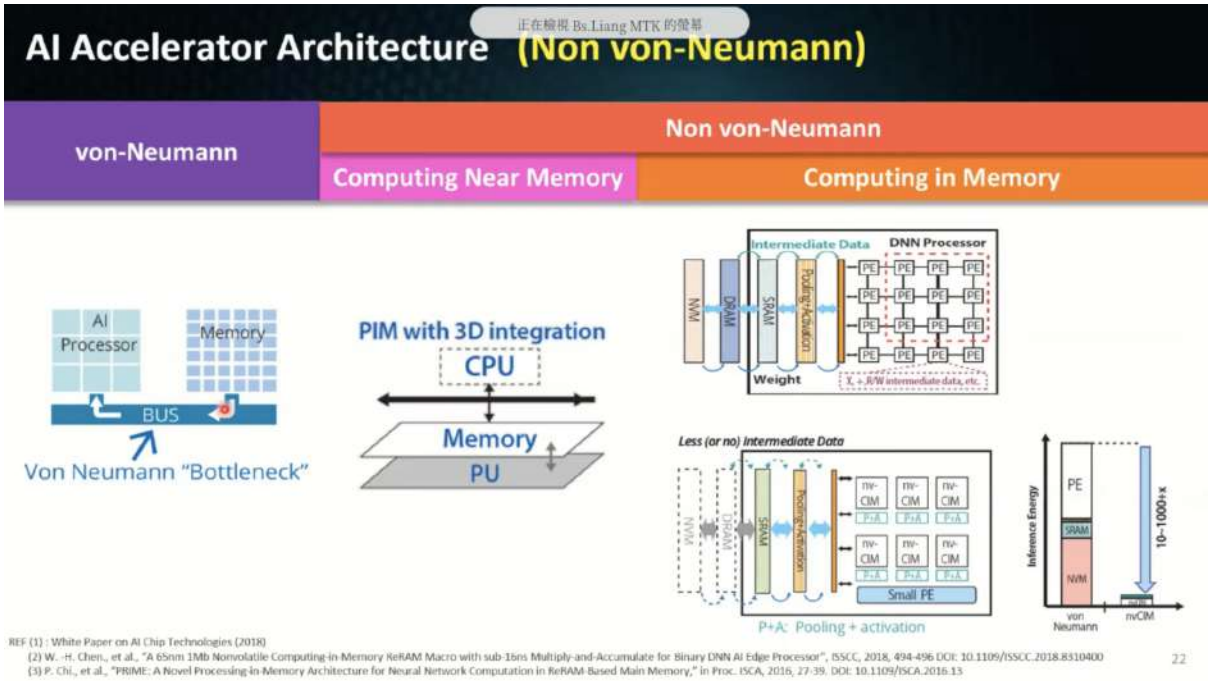

正在檢視 Bs Liang MTK 的螢幕
低軌衛星通訊
全球通信衛星星座計劃超過25個，計劃發射衛星數量超過10萬顆
美國 5~6 萬顆：Space X 佔 4 萬顆
中國大陸：低軌衛星系統未來10年規模達 3000-6000 顆
Satellite Constellation
核准衛星數
衛星高度
實際發射 (2021/4)
Freq.
Inter-satellite Comm.
Starlink SpaceX US 41943 (11943+30000) 550 km 1193 Ku (12–18 GHz) Ka (26.5–40 GHz) Laser
Iridium Next Iridium US 66 780 km 66 L band (1-2GHz) Ka band
OneWeb OneWeb UK/India 7088 (716+6372) 1,200 km 110 Ku (12–18 GHz) Ka (26.5–40 GHz)
Project Kuiper Amazon US 3236 590–630 km 0 (計劃發射中) Ka (26.5–40 GHz)
Telesat LEO Telesat Canada 1671 1,000–1,248 km 0 (計劃發射中) Ka (26.5–40 GHz)
鴻雁星座 CASC China 324 1100 km 1 (2018, 重慶號) 寬帶通信
虹雲工程 CASIC China 156+4+1 1000 km 1 (2018, 武漢號) 寬帶通信
行雲工程 CASIC China 80 800~1400 km 2 (2020) 12 (2021/f) 窄帶通信 (物聯網) Optical
九天微星星座 九天微星 China 72 700km 7 (2018,甄嬛系列) 物聯網
銀河航天 銀河航天 China 1 (星火03星) Q/V/Ka 衛星鏈路傳輸 1G bps
Source: Data from Companies; Wikipedia ; Table by Bor-Sung Liang
Bor-Sung Liang, MediaTek 15

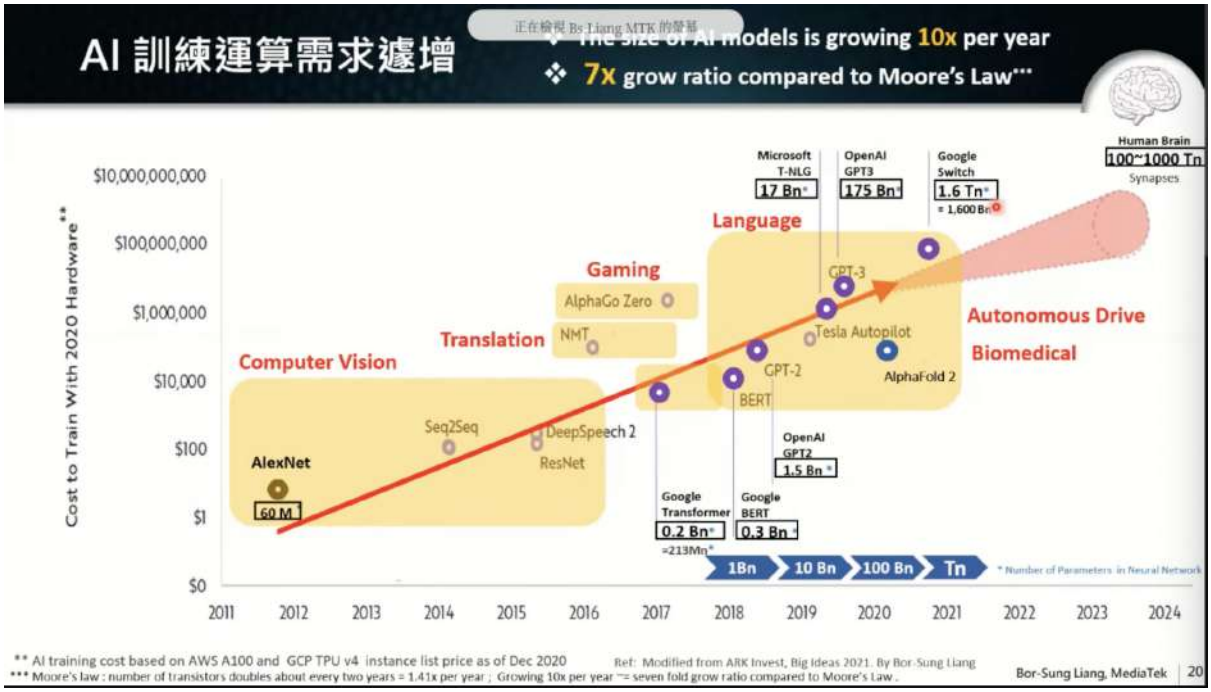

正在檢視 Bs Liang MTK 的螢幕
IC 運算功能成 AI 實用化
Prof. Hans Moravec, CMU (@Stanford, 1976) :
Computers were still millions of times too weak to exhibit intelligence
Apple-II (1977) : 0.5 MIPS
Super Computer Cray-1 : 130 MIPS
Alexnet Training @ 2012: 1 Exa-flops ~= 1,000,000,000,000 M FLOPS
'50~'60 AI
1969~1985 AI Winter
'80~'90 AI
1996~2006 AI Winter
2012 Alexnet
算力
IC / Semiconductor
算法
AI / Deep Learning Algorithm
大數據
Mobile / IoT Sensors
Global Economic Impact
Alexnet Training
Apple-II 63440 years
Cary-1 244 years
GPU x 2 6 Days
AI Science
AI Technologies
Ref (1) History of artificial intelligence, Wikipedia
(2) Yann LeCun, Facebook AI Research, "Deep Learning Hardware: Past, Present, & Future", ISSCC, 2019/02/18
Bor-Sung Liang, MediaTek 18

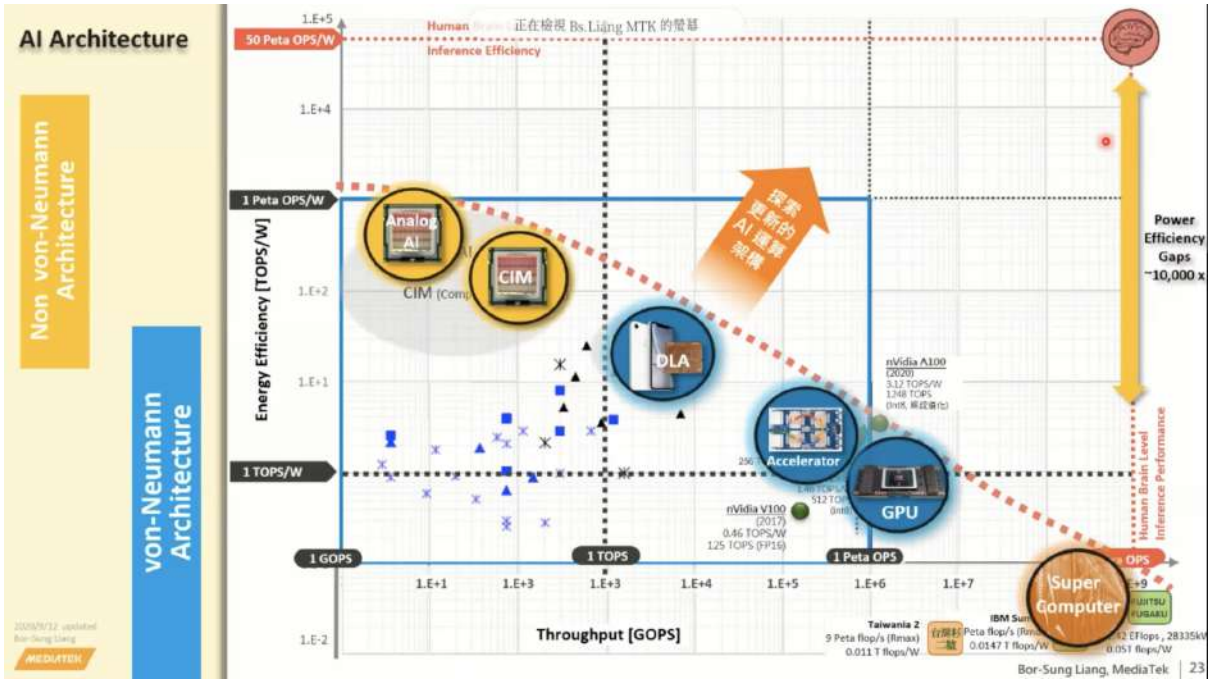


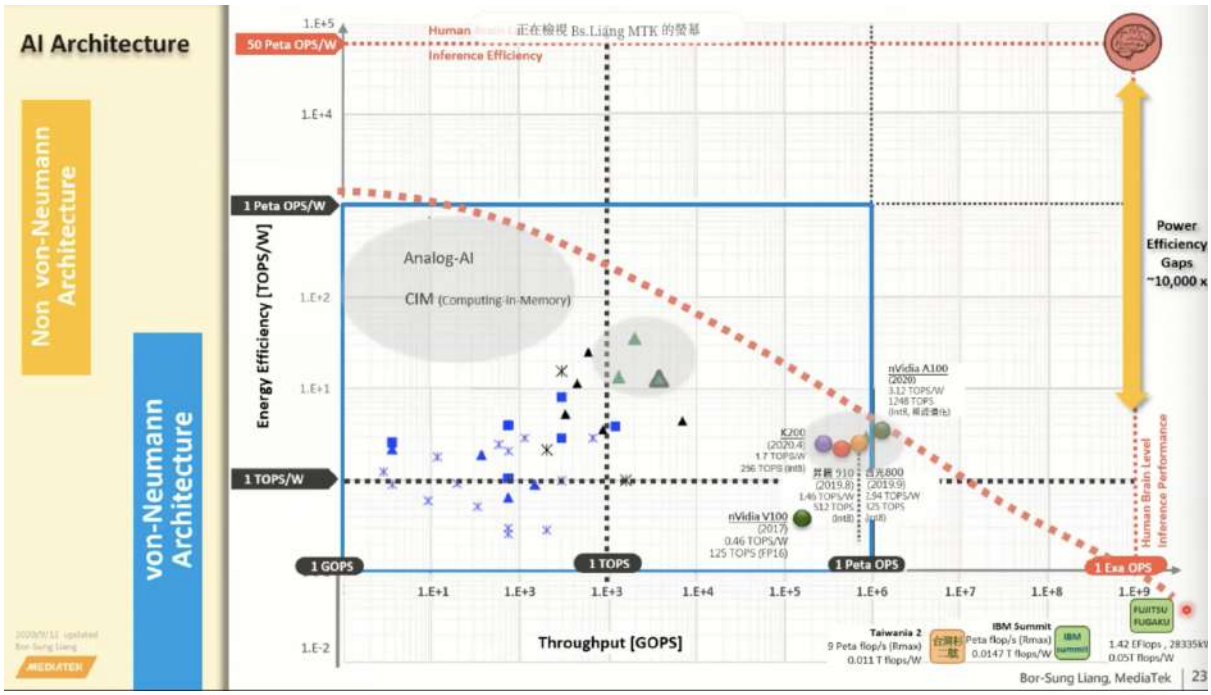
乘加 + Activation Function

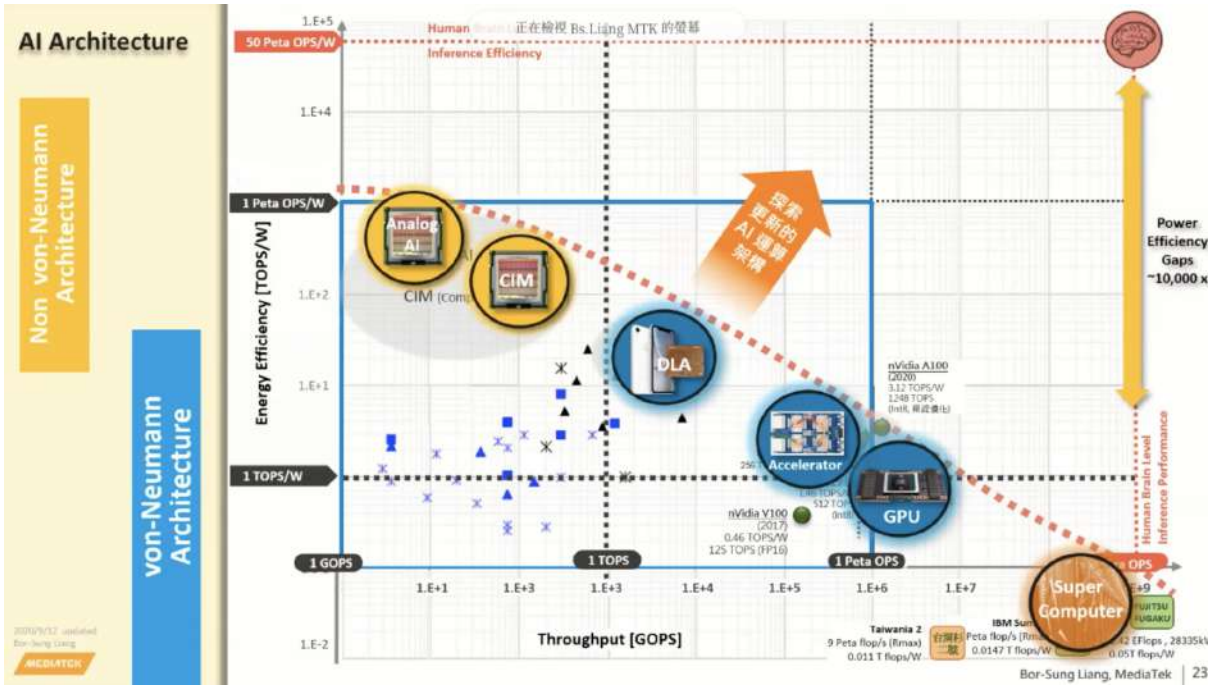


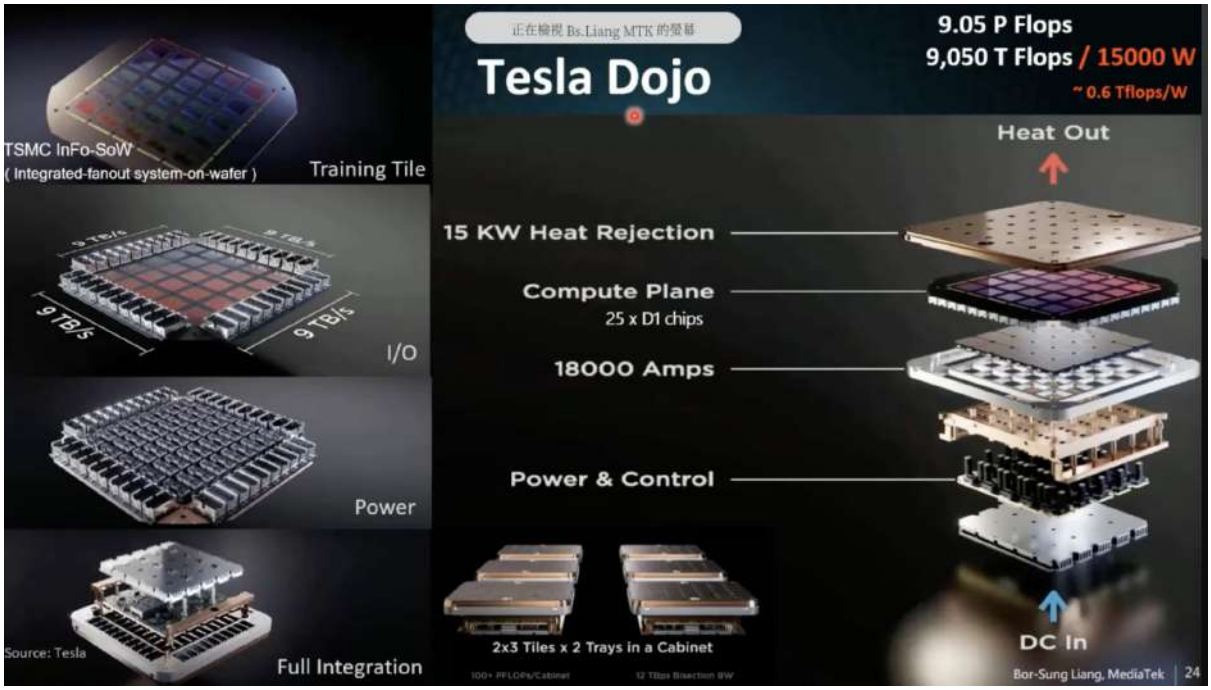

正在檢視 Bs.Liang MTK 的螢幕
AI Accelerator Architecture (Non von-Neumann)
von-Neumann
Non von-Neumann
Computing Near Memory
Computing in Memory
AI Processor
Memory
BUS
Von Neumann "Bottleneck"
PIM with 3D integration
CPU
Memory
PU
Intermediate Data
DNN Processor
NVM
DRAM
SRAM
Pooling+Activation
Weight
Less (or no) Intermediate Data
nv-CIM
Small PE
P+A: Pooling + activation
Inference Energy
PE
SRAM
NVM
10~1000x
von Neumann
nvCIM
REF (1) : White Paper on AI Chip Technologies (2018)
(2) W. -H. Chen., et al., "A 65nm 1Mb Nonvolatile Computing-in-Memory ReRAM Macro with sub-16ns Multiply-and-Accumulate for Binary DNN AI Edge Processor", ISSCC, 2018, 494-496 DOI: 10.1109/ISSCC.2018.8310400
(3) P. Chi., et al., "PRIME: A Novel Processing-in-Memory Architecture for Neural Network Computation in ReRAM-Based Main Memory," in Proc. ISCA, 2016, 27-39. DOI: 10.1109/ISCA.2016.13
22


正在檢視 Bs.Liang MTK 的螢幕
AI 訓練運算需求遽增
The size of AI models is growing 10x per year
7x grow ratio compared to Moore's Law***
Cost to Train With 2020 Hardware**
$10,000,000,000
$100,000,000
$1,000,000
$10,000
$100
$1
$0
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
Human Brain
100~1000 Tn
Synapses
Microsoft T-NLG
17 Bn*
OpenAI GPT3
175 Bn*
Google Switch
1.6 Tn*
= 1,600 Bn
Language
Gaming
AlphaGo Zero
Translation
NMT
Computer Vision
AlexNet
60 M
Seq2Seq
DeepSpeech 2
ResNet
GPT-3
Tesla Autopilot
GPT-2
BERT
AlphaFold 2
Autonomous Drive
Biomedical
OpenAI GPT2
1.5 Bn *
Google Transformer
0.2 Bn*
=213Mn*
Google BERT
0.3 Bn *
1Bn
10 Bn
100 Bn
Tn
* Number of Parameters in Neural Network
** AI training cost based on AWS A100 and GCP TPU v4 instance list price as of Dec 2020
*** Moore's law : number of transistors doubles about every two years = 1.41x per year ; Growing 10x per year = seven fold grow ratio compared to Moore's Law .
Ref: Modified from ARK Invest, Big Ideas 2021, By Bor-Sung Liang
Bor-Sung Liang, MediaTek
20


AI Architecture
Non von-Neumann Architecture
von-Neumann Architecture
50 Peta OPS/W
Human Inference Efficiency
1 Peta OPS/W
1 TOPS/W
Energy Efficiency [TOPS/W]
Analog AI
CIM
DLA
Accelerator
GPU
Super Computer
探索更新的AI運算架構
Power Efficiency Gaps ~10,000 x
Human Brain Level Inference Performance
nVidia A100 (2020) 312 TOPS/W 1248 TOPS
nVidia V100 (2017) 0.46 TOPS/W 125 TOPS (FP16)
1 GOPS
1 TOPS
1 Peta OPS
1 Exa OPS
Throughput [GOPS]
Taiwania 2 9 Peta flop/s (Rmax) 0.011 T flops/W
IBM Summit 0.0147 T flops/W
FUJITSU FUGAKU
Bor-Sung Liang, MediaTek | 23


AI Architecture
Non von-Neumann Architecture
von-Neumann Architecture
50 Peta OPS/W
Human Inference Efficiency
1 Peta OPS/W
1 TOPS/W
Energy Efficiency [TOPS/W]
Analog-AI
CIM (Computing-in-Memory)
Power Efficiency Gaps ~10,000 x
Human Brain Level Inference Performance
nVidia A100 (2020) 312 TOPS/W 1248 TOPS
nVidia V100 (2017) 0.46 TOPS/W 125 TOPS (FP16)
1 GOPS
1 TOPS
1 Peta OPS
1 Exa OPS
Throughput [GOPS]
Taiwania 2 9 Peta flop/s (Rmax) 0.011 T flops/W
IBM Summit 0.0147 T flops/W
FUJITSU FUGAKU
Bor-Sung Liang, MediaTek | 23

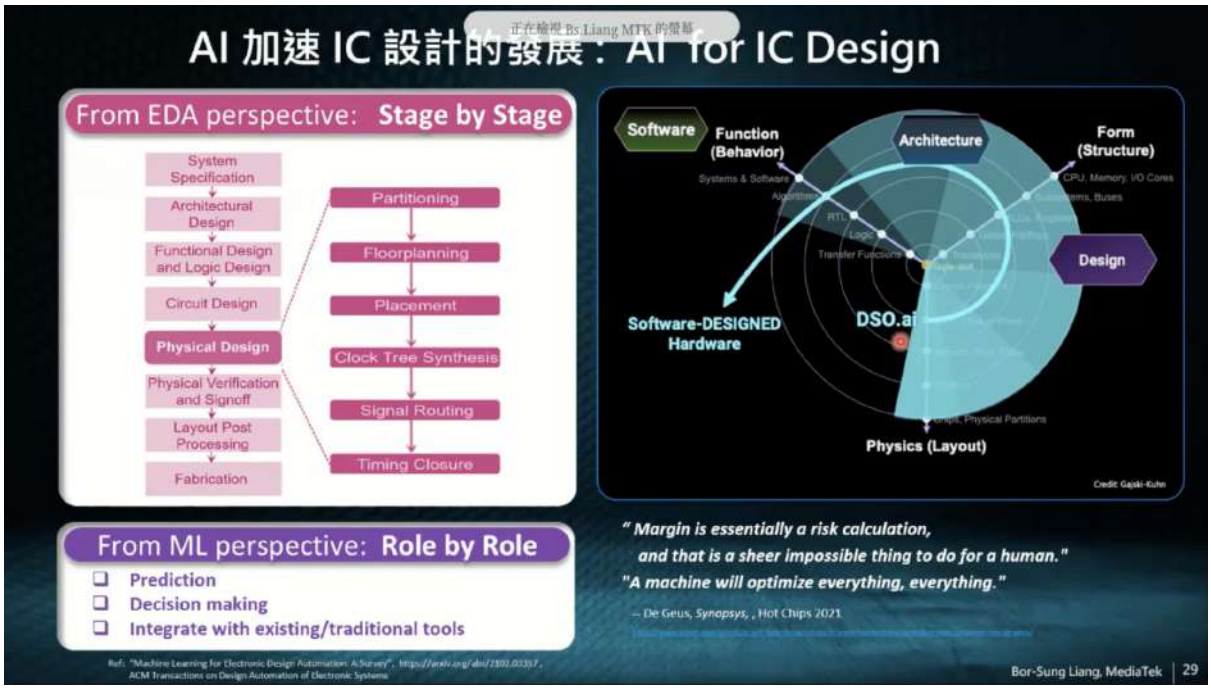

AI Architecture
Non von-Neumann Architecture
von-Neumann Architecture
50 Peta OPS/W
1 Peta OPS/W
1 TOPS/W
Energy Efficiency [TOPS/W]
Throughput [GOPS]
1 GOPS
1 TOPS
1 Peta OPS
Human Inference Efficiency
Analog AI
CIM
DLA
Accelerator
GPU
Super Computer
nVidia A100 (2020)
nVidia V100 (2017)
0.46 TOPS/W
125 TOPS (FP16)
Taiwania 2
9 Peta flop/s (Rmax)
0.011 T flops/W
Power Efficiency Gaps ~10,000 x
Human Brain Level Inference Performance
Bor-Sung Liang, MediaTek | 23

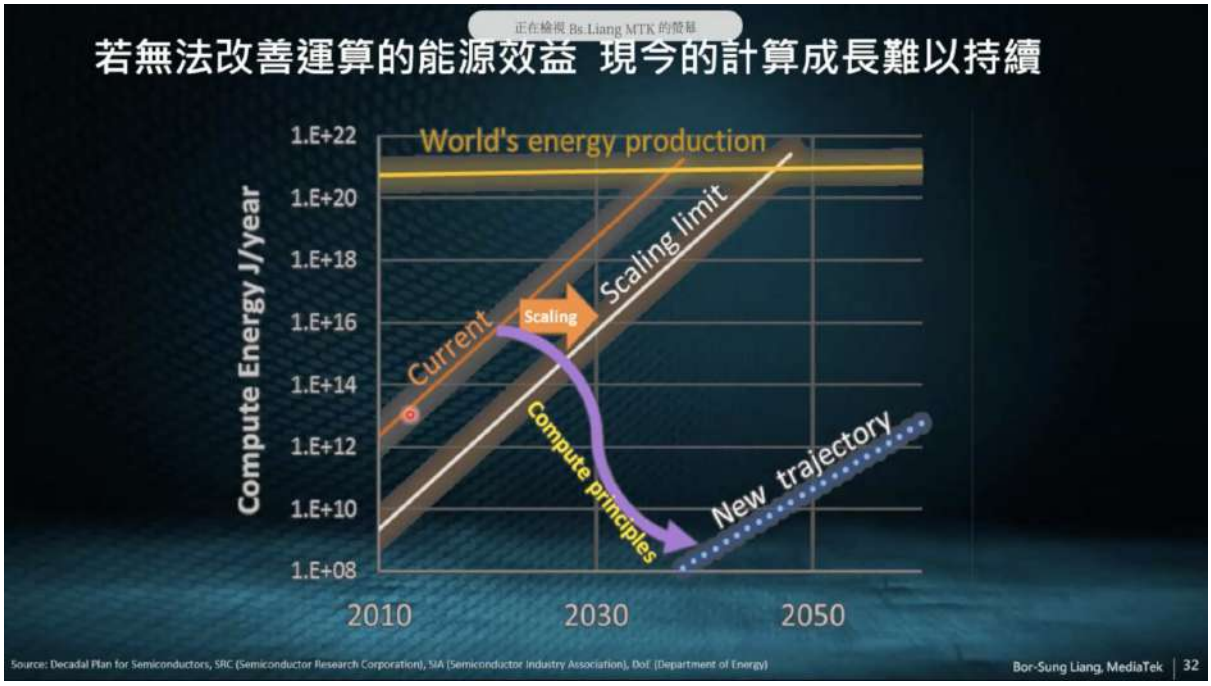

Tesla Dojo
9.05 P Flops
9,050 T Flops / 15000 W
~ 0.6 Tflops/W
TSMC InFo-SoW
(Integrated-fanout system-on-wafer)
Training Tile
9 TB/s
I/O
Power
Full Integration
Source: Tesla
Heat Out
15 KW Heat Rejection
Compute Plane
25 x D1 chips
18000 Amps
Power & Control
2x3 Tiles x 2 Trays in a Cabinet
DC In
Bor-Sung Liang, MediaTek | 24


Tesla Dojo 能源消耗
Dojo Interface Processors
1 Cabin = 2 x 3 x 2 =12 Tiles
9.05 P Flops x 12 = 108.6 P FLOPS
15000 W x 12 = 180,000 W
X 10 =
1086 P FLOPS = 1.086 Exa Flops
1,800,000 W = 1,800 KW = 1.8MW
每小時耗電 1800 度 (KWh)
每天耗電 43,200 度 / 電費= NT$ 101,952 元
每年耗電 15,768,000 度 / 電費= NT$ 37,212,480 元
Bor-Sung Liang, MediaTek 26


大腦半球
小腦
2% body mass
20-25% energy budget
大腦 516 Kcal / 全身 2400Kcal = 21.5%
人腦神經元數目
每分鐘葡萄糖消耗 (μmol/min)
每日葡萄糖總消耗 (g/day)
每日卡路里總消耗 (Kcal/day)
能量消耗功率 Watt (J/s)
860億
497.9
129
516
25 W
(24.98W)
1 Kcal = 4184 J 1 Day = 86400 s
Ref: Herculano-Houzel S (2011) Scaling of brain metabolism with a fixed energy budget per neuron: Implications for neuronal activity, plasticity and evolution. PLoS ONE 6: e17514. / In US National Library of Medicine, National Institutes of Health (NIH)
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0017514
Bor-Sung Liang, MediaTek 27


正在檢視 Bs.Liang MTK 的螢幕
世界超級電腦算力未來三年增長六倍
主導架構能力
X86, PowerPC
GPU (nVidia, AMD, Intel)
AI Accelerator (Google, ..)
發展自有架構
Alpha-like (神威)
DSP (天河2A)
Accelerator (阿里,百度,華為)
發展自有架構
ARM (富岳)
SPARC (京)
發展自有架構
RISC-V
2023
美國, 5,500
中國, 4,000
日本, 1,500
歐盟, 1,500
臺灣, 20
2020
美國, 639
中國, 566
日本, 528
歐盟, 340
臺灣, 10
三年內約成長六倍
0
2,000
4,000
6,000
8,000
10,000
12,000
Rmax Sum （ Peta Flops ）
12,662 Peta Flops
= 12.662 Exa Flops
2,226 Peta Flops
= 2.226 Exa Flops
Peta = 10 15 = 千兆
Exa = 10 18 = 百萬兆
= 1000 Peta
台灣預計將會有氣象局新購超級電腦, 因此暫估算算力增加 10 petaflop
估算 美國、中國、泛歐盟、日本與台灣的超級電腦算力增長。其他國家暫先假設算力持平未增長。
Bor-Sung Liang, MediaTek 28


正在檢視 Bs.Liang MTK 的螢幕
AI 加速 IC 設計的發展：AI for IC Design
From EDA perspective: Stage by Stage
System Specification
Architectural Design
Functional Design and Logic Design
Circuit Design
Physical Design
Physical Verification and Signoff
Layout Post Processing
Fabrication
Partitioning
Floorplanning
Placement
Clock Tree Synthesis
Signal Routing
Timing Closure
Software
Function (Behavior)
Architecture
Form (Structure)
Design
Software-DESIGNED Hardware
DSO.ai
Physics (Layout)
Credit Gajski-Kuhn
From ML perspective: Role by Role
Prediction
Decision making
Integrate with existing/traditional tools
" Margin is essentially a risk calculation, and that is a sheer impossible thing to do for a human."
"A machine will optimize everything, everything."
– De Geus, Synopsys, , Hot Chips 2021
Ref: "Machine Learning for Electronic Design Automation: A Survey", https://arxiv.org/abs/2101.09557, ACM Transactions on Design Automation of Electronic Systems
Bor-Sung Liang, MediaTek 29


正在檢視 Bs.Liang MTK 的螢幕
若無法改善運算的能源效益 現今的計算成長難以持續
Compute Energy J/year
1.E+22
1.E+20
1.E+18
1.E+16
1.E+14
1.E+12
1.E+10
1.E+08
World's energy production
Scaling limit
Current
Scaling
Compute principles
New trajectory
2010
2030
2050
Source: Decadal Plan for Semiconductors, SRC (Semiconductor Research Corporation), SIA (Semiconductor Industry Association), DoE (Department of Energy)
Bor-Sung Liang, MediaTek | 32