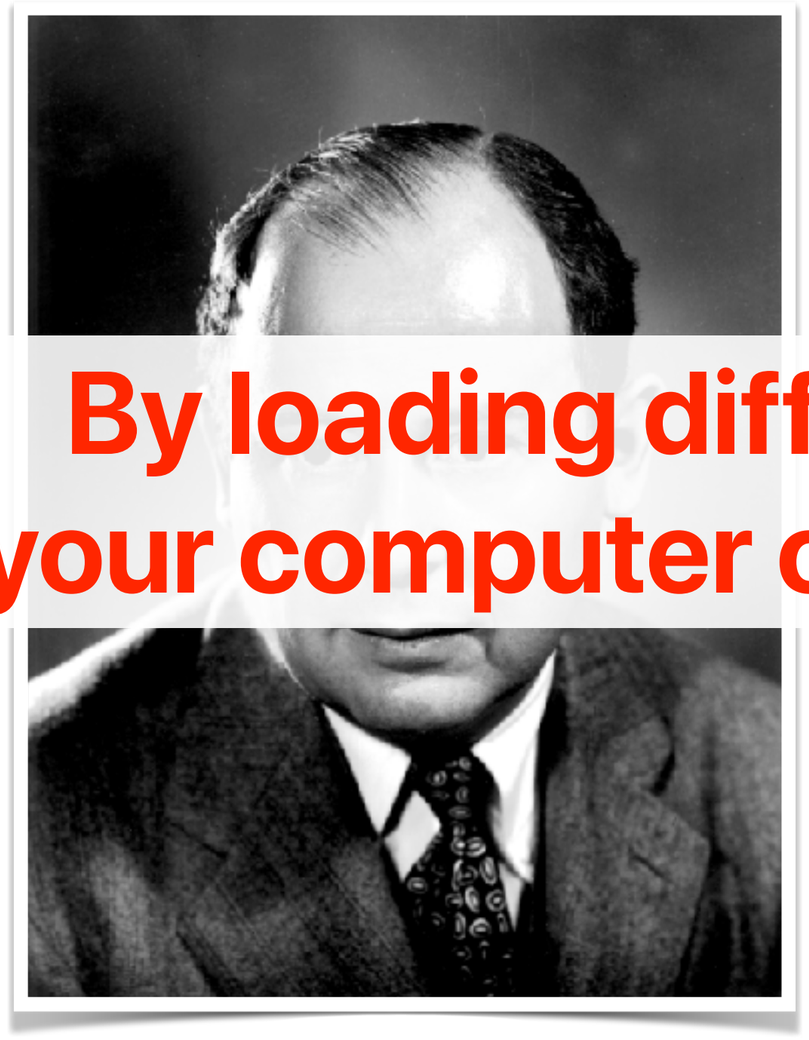
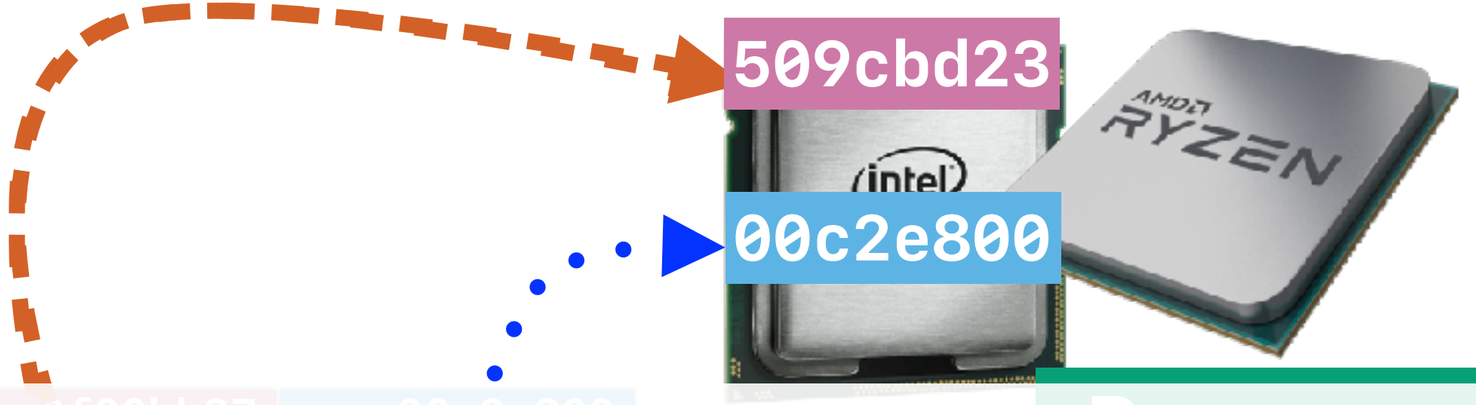
Performance (1): What does "perfect" mean?

Hung-Wei Tseng

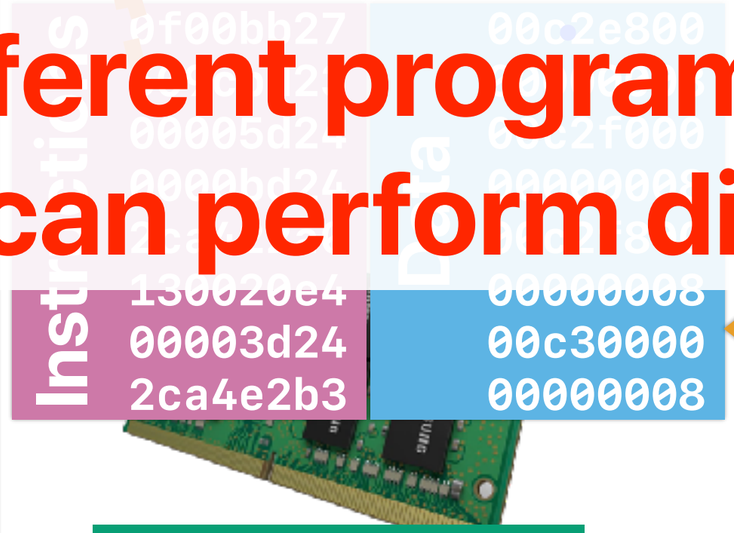
Recap: von Neumman Architecture





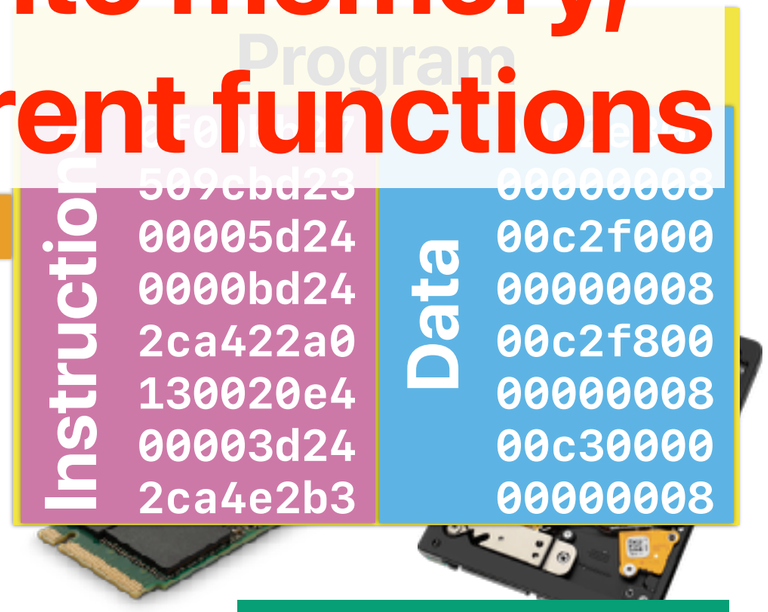
Processor

By loading different programs into memory, your computer can perform different functions



Memory

Program



Storage

Recap: Start with this simple program in C

```
int A[] =
{1,2,3,4,5,6,7,8,9,10,1,2,3,4
,5,6,7,8,9,10};
```

Compiler

```
Contents of section .data:
 0000 01000000 02000000 03000000 04000000
 0010 05000000 06000000 07000000 08000000
 0020 09000000 0a000000 01000000 02000000
 0030 03000000 04000000 05000000 06000000
 0040 07000000 08000000 09000000 0a000000
```

Data

```
int main()
{
    int i=0, sum=0;
    for(i = 0; i < 20; i++)
    {
        sum += A[i];
    }
    return 0;
}
```

control flow operations

logical operations

memory access

arithmetic operations

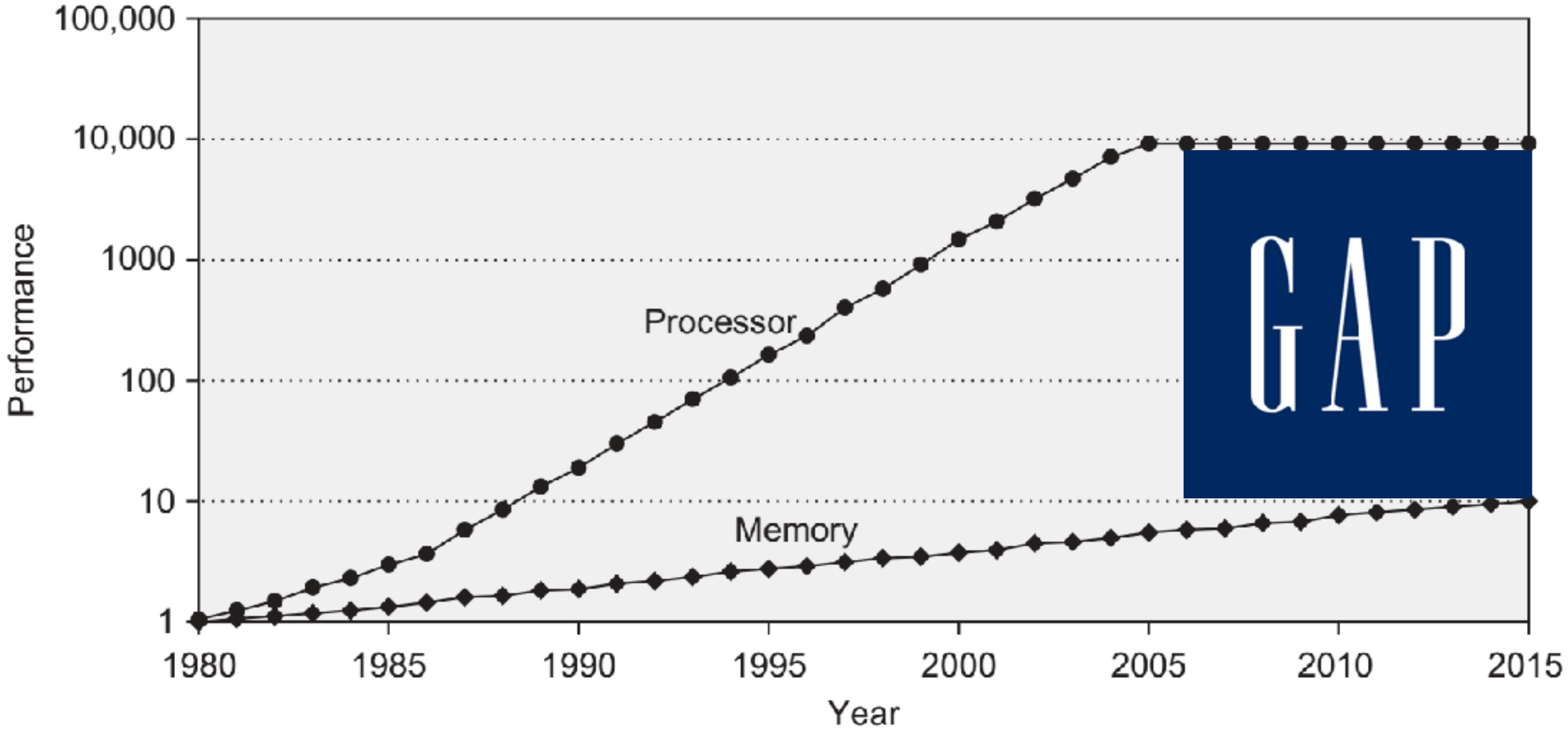
```
main:
.LFB0:
    endbr64
    pushq   %rbp
    movq    %rsp, %rbp
    movl    $0, -8(%rbp)
    movl    $0, -4(%rbp)
    movl    $0, -8(%rbp)
    jmp .L2
.L3:
    movl    -8(%rbp), %eax
    cltq
    leaq    0(,%rax,4), %rdx
    leaq    A(%rip), %rax
    movl    (%rdx,%rax),
%eax
    addl    %eax, -4(%rbp)
    addl    $1, -8(%rbp)
.L2:
    cmpl    $19, -8(%rbp)
    jle .L3
    movl    $0, %eax
    popq    %rbp
    ret
```

Compiler

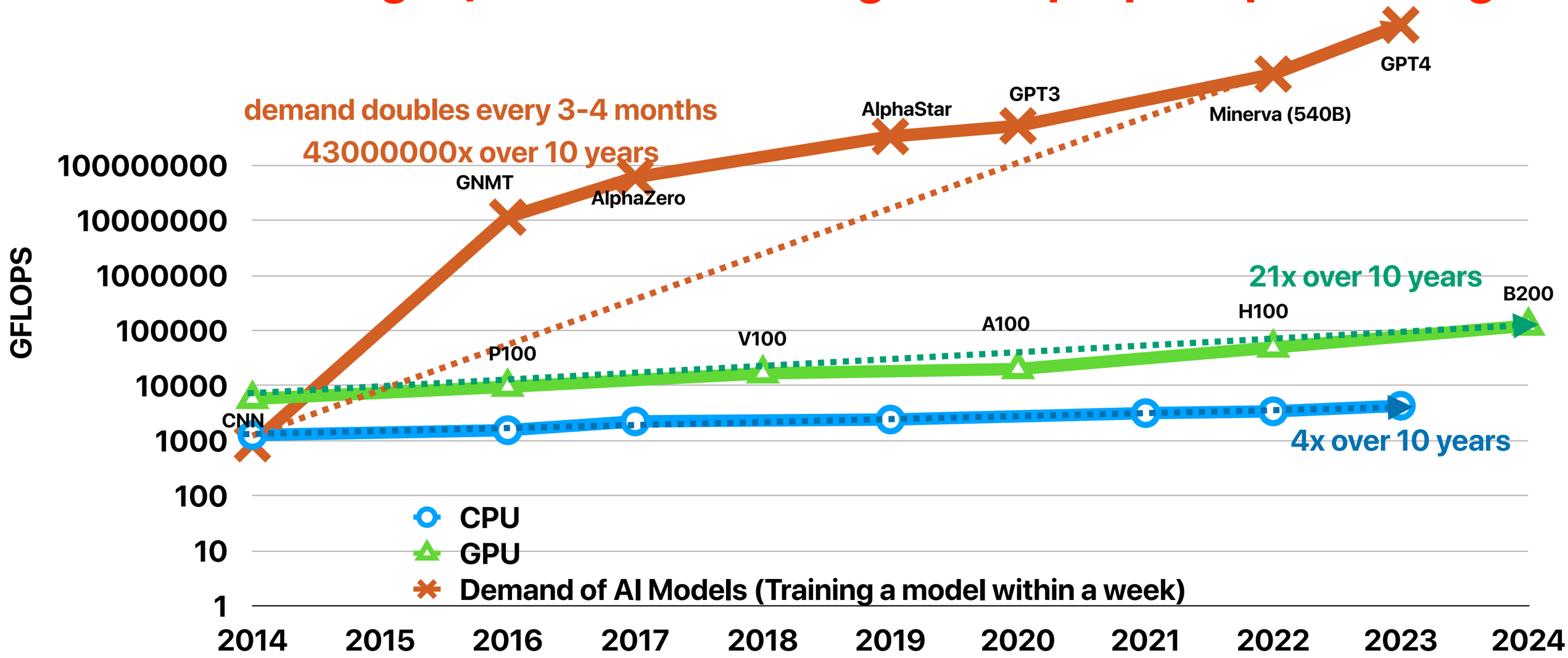
```
Contents of section .text:
 0000 f30f1efa 554889e5 c745f800 000000c7
 0010 45fc0000 0000c745 f8000000 00eb1e8b
 0020 45f84898 488d1425 00000000 488d0500
 0030 0000008b 04020145 fc8345f8 01837df8
 0040 137edcb8 00000000 5dc3
```

Instructions

Recap: Performance gap between Processor/Memory



Mis-matching AI/ML demand and general-purpose processing



https://ourworldindata.org/grapher/artificial-intelligence-training-computation

3:00

Is UCSD the best university? Why or why not?

Outline

- Definition of “Performance”
- The performance equation
- Other metrics and are they good?

Best Nation

Schools in the National Univers
a full range of undergraduate r
producing groundbreaking res

To unlock full rankings, SAT/A

SUMMARY

443

Scho

Loca

Rank

Nat

Best Compute

Best Computer Science Schools

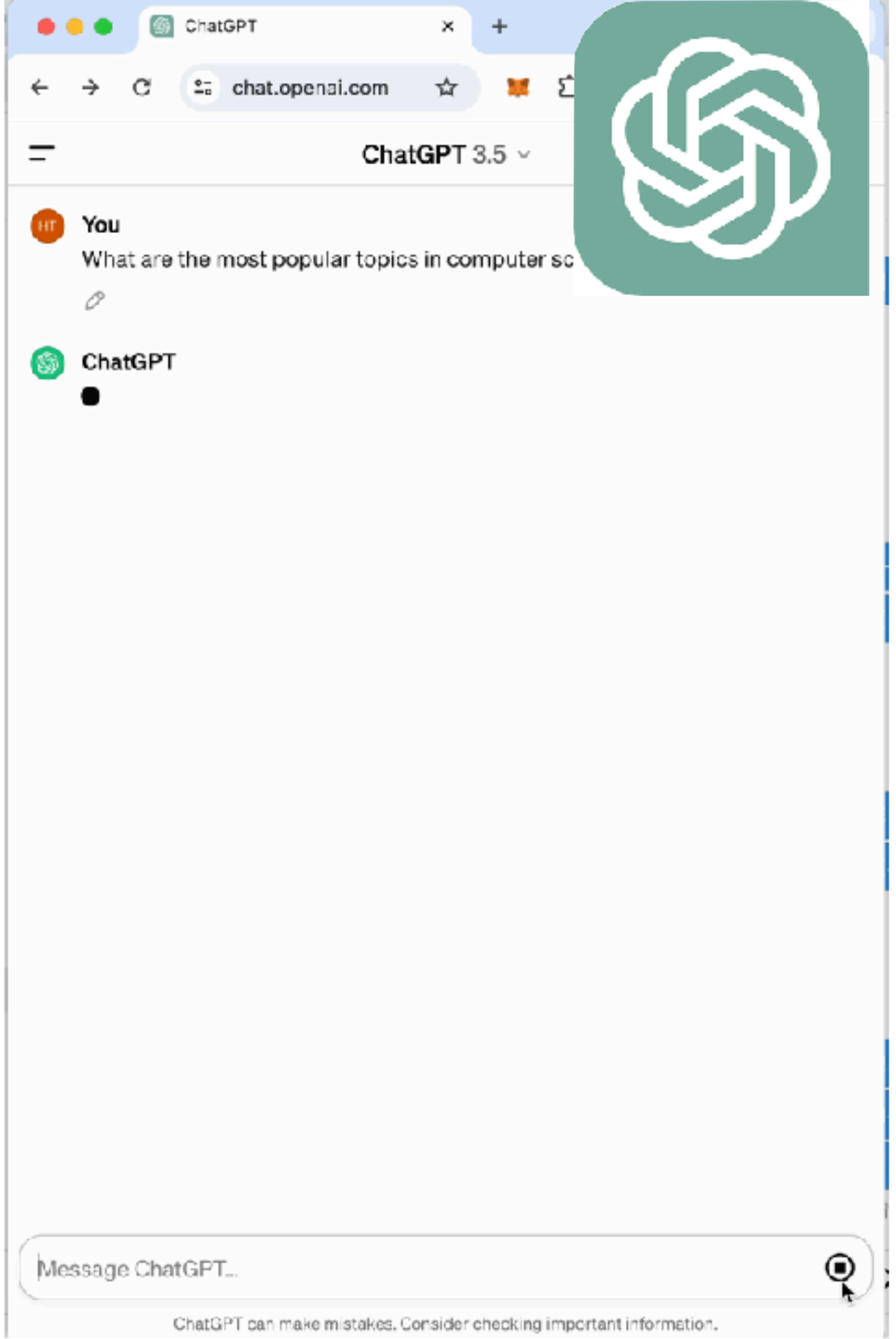
Ranked in 2022, part of Best Science S

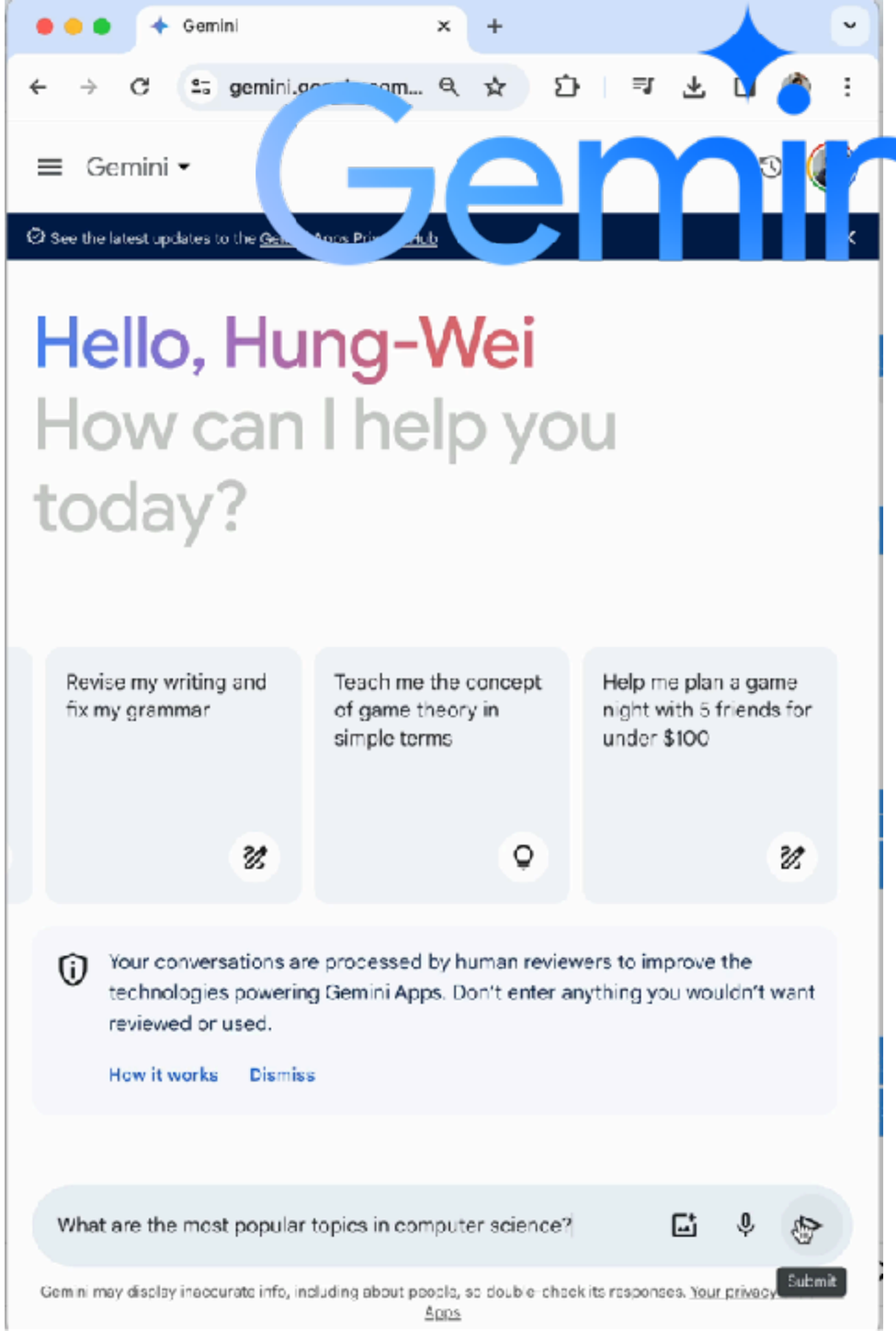
Earning a graduate degree in computer
technology companies and colleges ar
reflects its average rating on a scale fr
institutions. Read the methodology »

UC San Diego Ranked No. 1 Public University by Washington Monthly

Campus celebrated as a leader in social mobility, research and public

What does it really mean by “better” performance


ChatGPT
chat.openai.com
ChatGPT 3.5
You
What are the most popular topics in computer sc
ChatGPT
Message ChatGPT...
ChatGPT can make mistakes. Consider checking important information.


Gemini
Gemini
Hello, Hung-Wei
How can I help you today?
Revise my writing and fix my grammar
Teach me the concept of game theory in simple terms
Help me plan a game night with 5 friends for under $100
Your conversations are processed by human reviewers to improve the technologies powering Gemini Apps. Don't enter anything you wouldn't want reviewed or used.
How it works
Dismiss
What are the most popular topics in computer science?
Submit

Gemini v.s. ChatGPT

- Comparing the experiments we have done with Gemini and ChatGPT, how many of the following metrics does Gemini outperforms ChatGPT?
  - ① Response time
  - ② Throughput
  - ③ End-to-end latency (i.e., total execution time)
  - ④ Quality of results
  - A. 0
  - B. 1
  - C. 2
  - D. 3
  - E. 4

Important performance metrics

- End-to-end latency — how much time the program/operation takes from the beginning to the end
- Response time — how much time the user starts to feel the program is running/finishing
- Throughput/bandwidth — the average amount of work/data can the program/system deliver within the execution time
- Energy consumption — the aggregated power during the execution time
- Cost of operation — the amount of money necessary for finishing an operation
- Quality of results — the human perception of the execution result
- Power consumption — the heat generation produced by the circuit

Important performance metrics

- End-to-end latency — how much **time** the program/operation takes from the beginning to the end
- Response time — how much **time** the user starts to feel the program is running/finishing
- Throughput/bandwidth — the average amount of work/data can the program/system deliver within the execution **time**
- Energy consumption — the aggregated power during the execution **time**
- Cost of operation — the amount of money necessary for finishing an operation (related to **time**)
- Quality of results — the human perception of the execution result
- Power consumption — the heat generation produced by the circuit

Takeaways: What does “perfect” mean?

- Latency is the most fundamental performance metric

Let’s start with “end-to-end latency” as the default metric — how long it takes to execute a program?

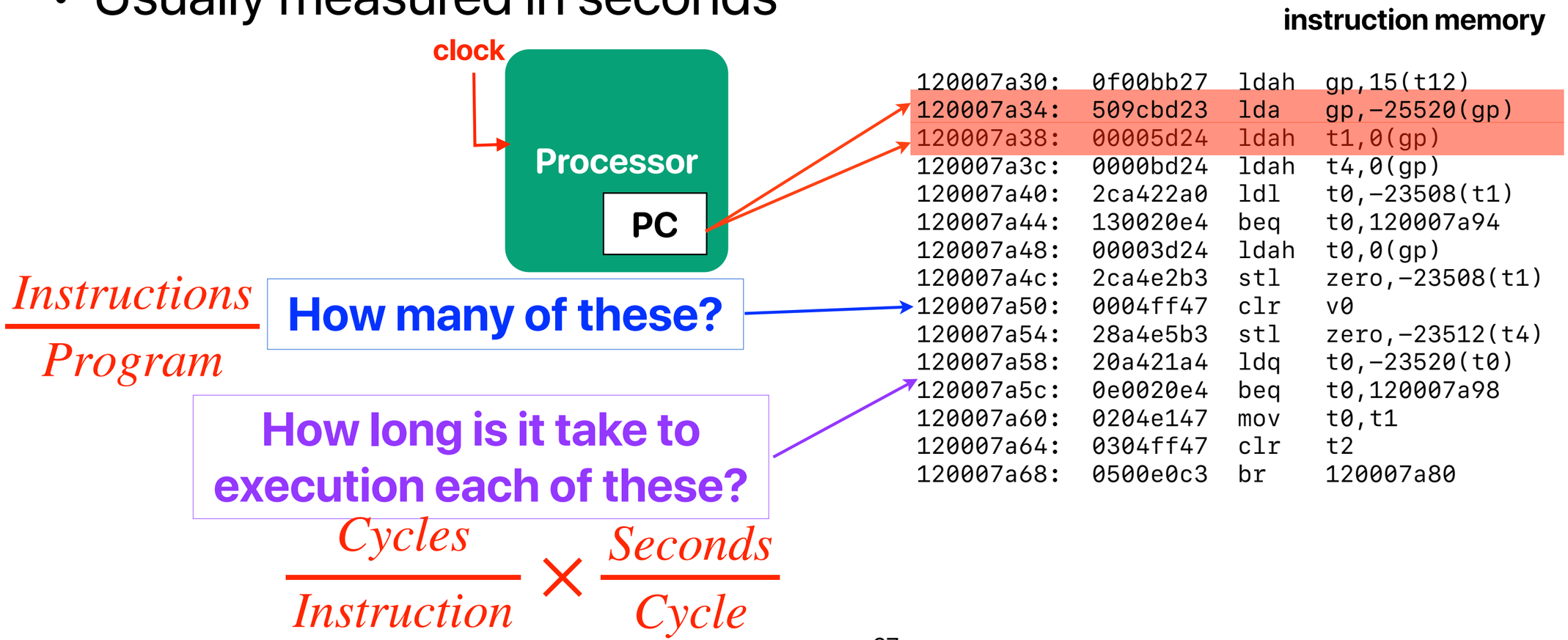
https://www.pollev.com/hungweitseng close in 1:30

CPU Performance Equation (X)

- Assume that we have an application composed with a total of **5,000,000,000** instructions, in which **20%** of them are "Type-A" instructions with an average **CPI of 4** cycles, **20%** of them are "Type-B" instructions with an average **CPI of 3** cycles and the rest instructions are "Type-C" instructions with average **CPI of 1** cycle. If the processor runs at **4 GHz**, how long is the execution time?
  A. 1.25 sec
  B. 2.5 sec
  C. 3.75 sec
  D. 7.5 sec
  E. 40 sec

Execution Time

- The simplest kind of performance
- Shorter execution time means better performance
- Usually measured in seconds



CPU Performance Equation

$$Performance = \frac{1}{Execution\ Time}$$

$$Execution\ Time = \frac{Instructions}{Program} \times \frac{Cycles}{Instruction} \times \frac{Seconds}{Cycle}$$

$$ET = IC \times CPI \times CT$$

$$1GHz = 10^9 Hz = \frac{1}{10^9} sec\ per\ cycle = 1\ ns\ per\ cycle$$

$$\frac{1}{Frequency(i.e., clock\ rate)}$$

Performance Equation (X)

- Assume that we have an application composed with a total of **5,000,000,000** instructions, in which **20%** of them are "Type-A" instructions with an average **CPI of 4** cycles, **20%** of them are "Type-B" instructions with an average **CPI of 3** cycles and the rest instructions are "Type-C" instructions with average **CPI of 1** cycle. If the processor runs at **4 GHz**, how long is the execution time?

  A. 1.25 sec
  B. 2.5 sec
  C. 3.75 sec
  D. 7.5 sec
  E. 40 sec

$$ET = IC \times CPI \times CT$$

$$ET = (5 \times 10^9) \times \underbrace{(20\% \times 4 + 20\% \times 3 + 60\% \times 1)}_{\text{average CPI}} \times \frac{1}{4 \times 10^9}\ sec = 2.5\ sec$$

Performance equation (round 2)

- Consider the following c code snippet and x86 instructions implement the code snippet

C	x86

`for(i = 0; i < count; i++) {`
`  s += a[i];`
`}`
	`.L3:`
	`movslq (%rdi), %rdx`
	`addq $4, %rdi`
	`addq %rdx, %tax`
	`cmpq %rcx, %rdi`
	`jne .L3`

If (1) count is set to 1,000,000,000, (2) a memory instruction takes 4 cycles, (3) a branch/jump instruction takes 3 cycles, (4) other instructions takes 1 cycle on average, and (5) the processor runs at 4 GHz, how much time is it take to finish executing the code snippet?

A. 0.5 sec
B. 1 sec
C. 2.5 sec
D. 3.75 sec
E. 4 sec

Performance equation (round 2)

- Consider the following c code snippet and x86 instructions implement the code snippet

C	x86

`for(i = 0; i < count; i++) {`
`  s += a[i];`
`}`
	`.L3:`
	`movslq (%rdi), %rdx`
	`addq $4, %rdi`
	`addq %rdx, %tax`
	`cmpq %rcx, %rdi`
	`jne .L3`

If (1) count is set to 1,000,000,000, (2) a memory instruction takes 4 cycles, (3) a branch/jump instruction takes 3 cycles, (4) other instructions takes 1 cycle on average, and (5) the processor runs at 4 GHz, how much time is it take to finish executing the code snippet?

A. 0.5 sec
B. 1 sec
C. 2.5 sec
D. 3.75 sec
E. 4 sec

$$ET = IC \times CPI \times CT$$

$$ET = \underbrace{(5 \times 10^9)}_{\text{total \# of dynamic instructions}} \times \underbrace{(20\% \times 4 + 20\% \times 3 + 60\% \times 1)}_{\text{average CPI}} \times \frac{1}{4 \times 10^9}\ sec = 2.5\ sec$$

Takeaways: What does “perfect” mean?

- Latency is the most fundamental performance metric
- Instruction count, cycles per instruction, cycle time define the latency of execution on CPUs

Choose the right metric — Latency v.s. Throughput/Bandwidth

nvidia.com

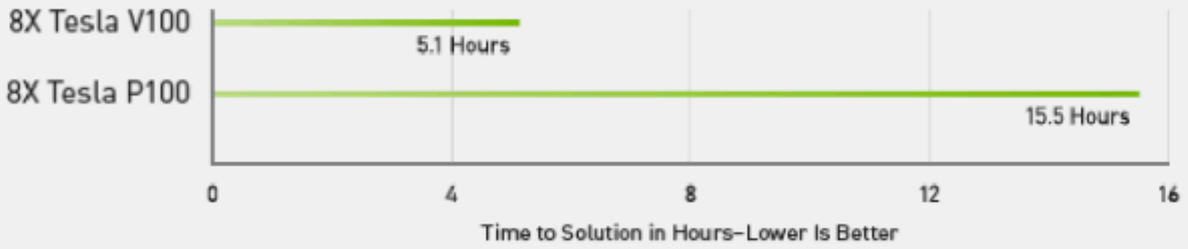
NVIDIA

Artificial Intelligence Computing Leadership from NVIDIA

CLOUD & DATA CENTER | PRODUCTS | SOLUTIONS | APPS | FOR DEVELOPERS | TECHNOLOGIES

Tesla V100 | AI TRAINING | AI INFERENCE | HPC | DATA CENTER GPUs | SPECIFICATIONS

Deep Learning Training in Less Than a Workday



Server Config: Dual Xeon E5-2699 v4 2.6 GHz | 8X NVIDIA® Tesla® P100 or V100 | ResNet-50 Training on MXNet for 90 Epochs with 1.28M ImageNet Dataset.

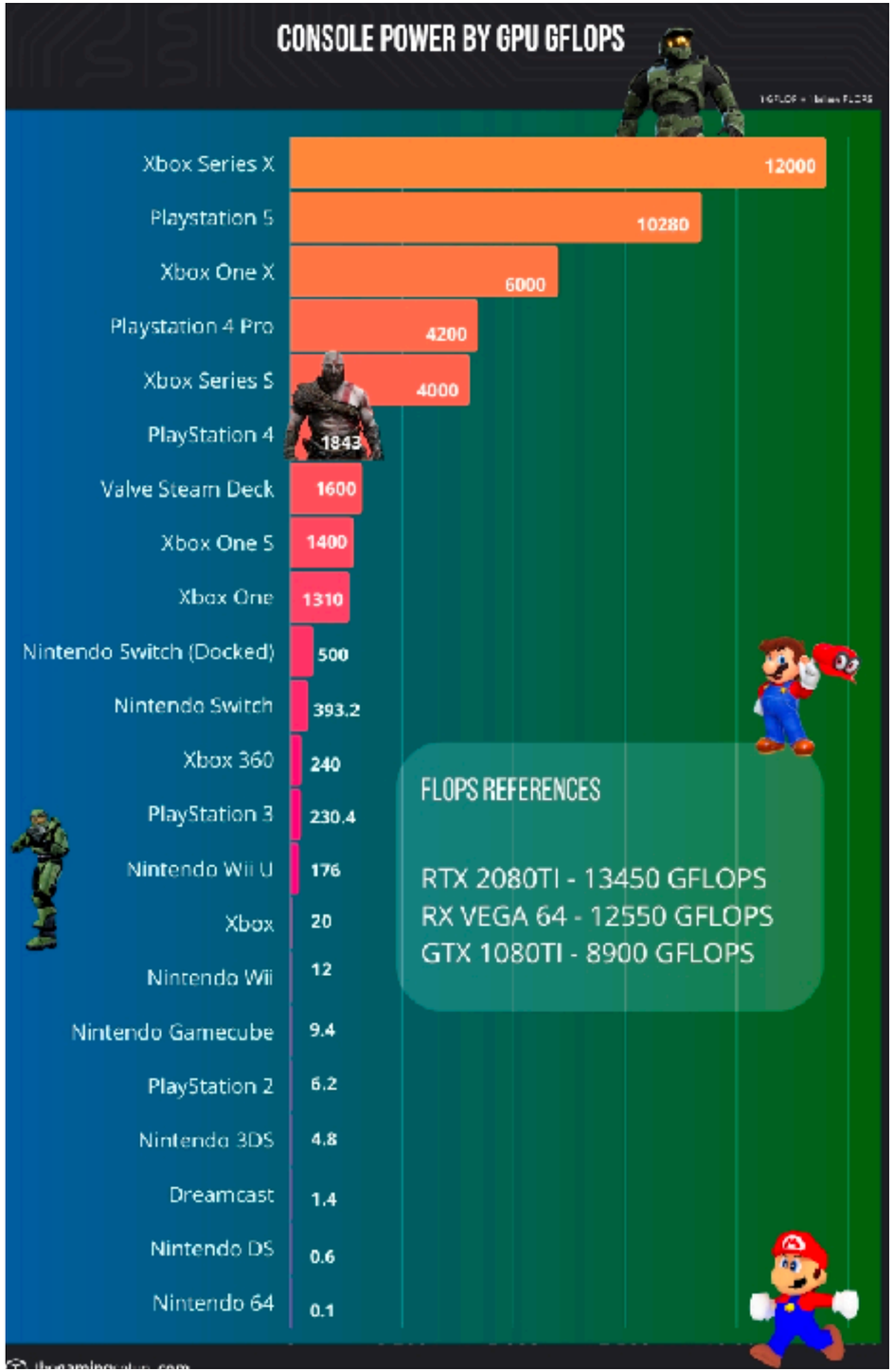
AI TRAINING

From recognizing speech to training virtual personal assistants and teaching autonomous cars to drive, data scientists are taking on increasingly complex challenges with AI. Solving these kinds of problems requires training deep learning models that are exponentially growing in complexity, in a practical amount of time.

With 640 Tensor Cores, Tesla V100 is the world's first GPU to break the 100 teraFLOPS (TFLOPS) barrier of deep learning performance. The next generation of NVIDIA NVLink™ connects multiple V100 GPUs at up to 300 GB/s to create the world's most powerful computing servers. AI models that would consume weeks of computing resources on previous systems can now be trained in a few days. With this dramatic reduction in training time, a whole new world of problems will now be solvable with AI.

TFLOPS (Tera FLoating-point Operations Per Second)

	TFLOPS	clock rate
Switch	1	921 MHz
PS5	10.28	2.23 GHz
XBox Series X	12	1.825 GHz
GeForce RTX 3090	40	1.395 GHz



Let's measure the FLOPS of matrix multiplications

```
for(i = 0; i < ARRAY_SIZE; i++) {
  for(j = 0; j < ARRAY_SIZE; j++) {
    for(k = 0; k < ARRAY_SIZE; k++) {
      c[i][j] += a[i][k]*b[k][j];
    }
  }
}
```

Floating point operations (FLOP"s"):

$$i \times j \times k \times 2$$

Given $i = j = k = 2048$

$2^{3\times11} \times 2 = 2^{34}$ **FLOPs in total**

Floating point operations per second (FLOP"S"):

$$FLOPS = \frac{2^{34}}{ET_{seconds}}$$

https://www.pollev.com/hungweitseng close in 1:30

How reflective is FLOPS?

- If you're given the FLOPS of an underlying GPU, how many situations below can the FLOPS be representative to the real performance?
  - ① The FLOPS remains the same on the same GPU even if we change the data size
  - ② The FLOPS remains the same on the same GPU even if we change the data type to double
  - ③ The FLOPS remains the same on the same GPU if we change the algorithm implementation
  - ④ The ratio of FLOPS on two different GPUs reflects the ratio execution on these two GPUs when executing floating point applications

A. 0
B. 1
C. 2
D. 3
E. 4

How reflective is FLOPS?

- If you're given the FLOPS of an underlying GPU, how many situations below can the FLOPS be representative to the real performance?
  - ① The FLOPS remains the same on the same GPU even if we change the data size
  - ② The FLOPS remains the same on the same GPU even if we change the data type to double
  - ③ The FLOPS remains the same on the same GPU if we change the algorithm implementation
  - ④ The ratio of FLOPS on two different GPUs reflects the ratio execution on these two GPUs when executing floating point applications

A. 0
B. 1
C. 2
D. 3
E. 4

TFLOPS (Tera FLoating-point Operations Per Second)

$$TFLOPS = \frac{\#\ of\ floating\ point\ instructions \times 10^{-12}}{Exection\ Time}$$

Demo: matmul on GPU

Size	Latency	Relative Latency	Throughput (Output Numbers Per Second)	Relative Throughput
16x16x16	~ 0.09ms	1	0.09ms/256	1
32x32x32	~ 0.09ms	1	0.09ms/1024	4
64x64x64	~ 0.09ms	1	0.09ms/4096	16

Larger throughput doesn't mean shorter latency!

Is TFLOPS (Tera FLoating-point Operations Per Second) a good metric?

$$TFLOPS = \frac{\#\ of\ floating\ point\ instructions \times 10^{-12}}{Exection\ Time}$$

$$= \frac{IC \times \%\ of\ floating\ point\ instructions \times 10^{-12}}{IC \times CPI \times CT}$$

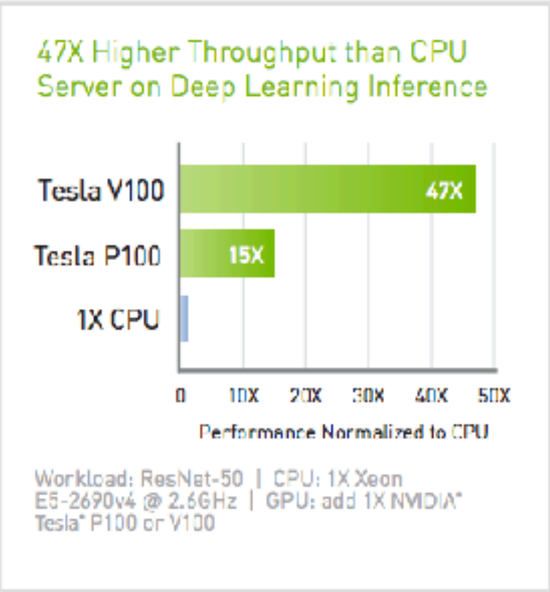
$$= \frac{\%\ of\ floating\ point\ instructions \times 10^{-12}}{CPI \times CT}$$

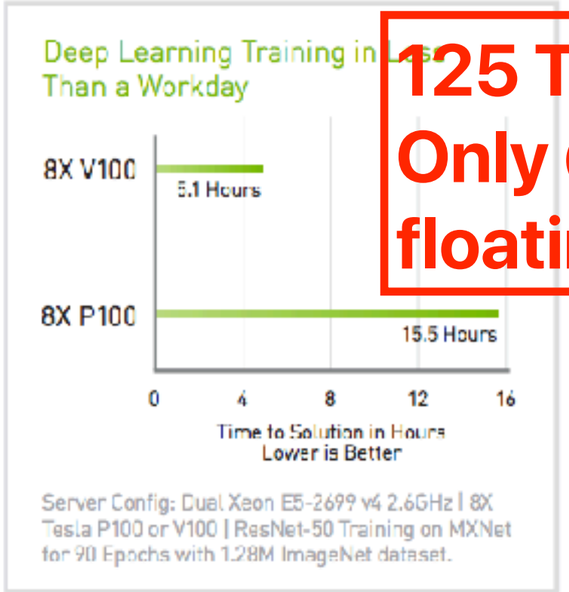
IC is gone!

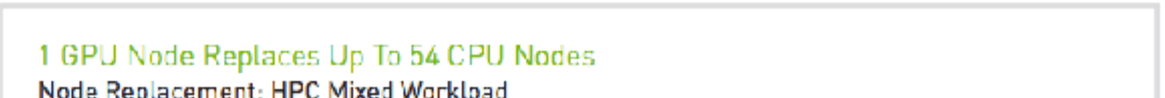
- Cannot compare different ISA/compiler
  - What if the compiler can generate code with fewer instructions?
  - What if new architecture has more IC but also lower CPI?
- Does not make sense if the application is not floating point intensive

The Most Advanced Data Center GPU Ever Built.

NVIDIA® Tesla® V100 is the world's most advanced data center GPU ever built to accelerate AI, HPC, and graphics. Powered by NVIDIA Volta, the latest GPU architecture, Tesla V100 offers the performance of up to 100 CPUs in a single GPU—enabling data scientists, researchers, and engineers to tackle challenges that were once thought impossible.

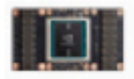




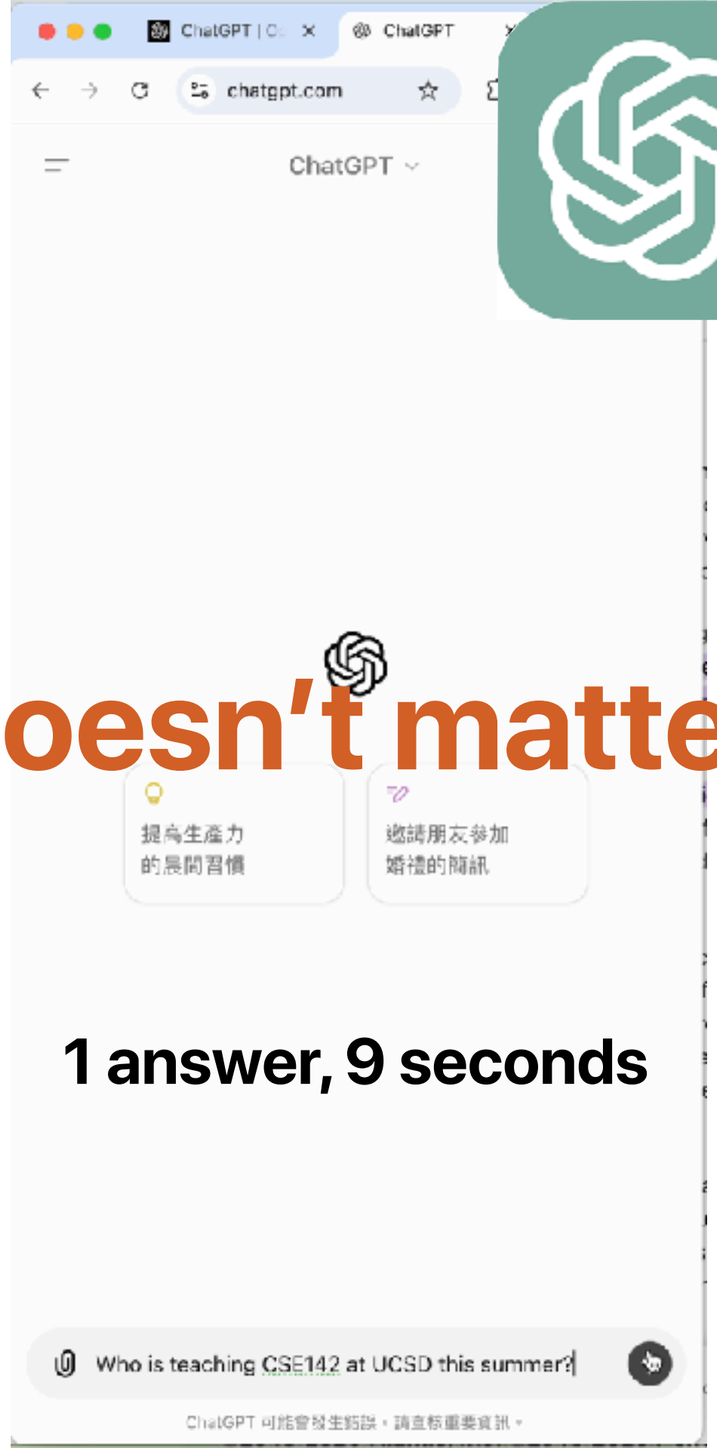


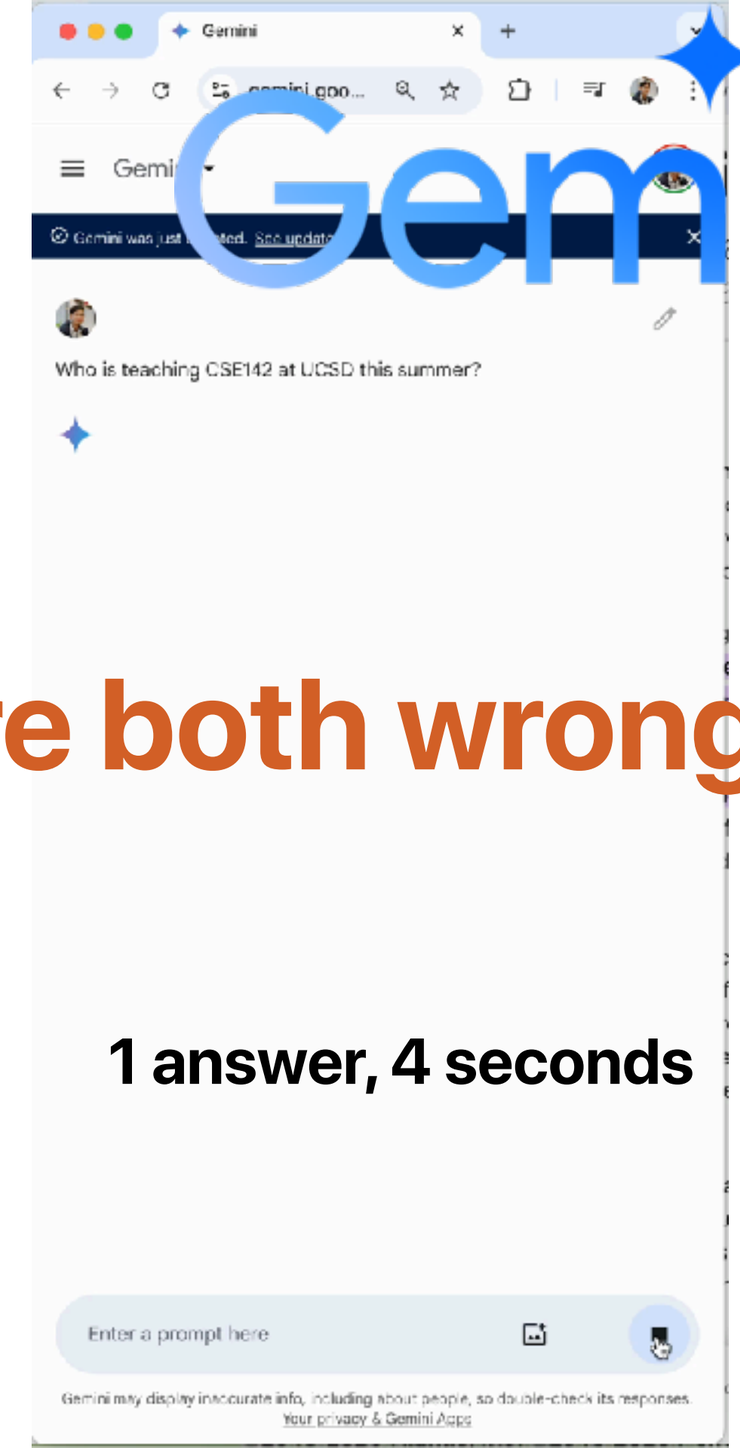
125 TFLOPS Only @ 16-bit floating point

SPECIFICATIONS



	Tesla V100 PCIe	Tesla V100 SXM2
GPU Architecture	NVIDIA Volta	
NVIDIA Tensor Cores	640	
NVIDIA CUDA® Cores	5,120	
Double-Precision Performance	7 TFLOPS	7.8 TFLOPS
Single-Precision Performance	14 TFLOPS	15.7 TFLOPS
Tensor Performance	112 TFLOPS	125 TFLOPS
GPU Memory	32GB /16GB HBM2	
Memory Bandwidth	900GB/sec	
ECC	Yes	
Interconnect Bandwidth	32GB/sec	300GB/sec
System Interface	PCIe Gen3	NVIDIA NVLink
Form Factor	PCIe Full Height/Length	SXM2
Max Power		





Doesn’t matter — they’re both wrong :)

1 answer, 9 seconds

1 answer, 4 seconds

gamespot.com/articles/nvidia-accused-of-cheating-in-3dmark-03/1100-6028694/#:~:text="We%20have%20now%20established%20that,d

Nvidia accused of cheating in 3DMar

Futuremark, the maker of 3DMark 03, releases a patch for the g to correct the way Nvidia drivers manipulate results in the popu

By David Becker on May 23, 2003 at 4:27PM PDT

Saratoga, California-based Futuremark on Friday said in a statement that Nvidia tweaked the software needed to run its new GeForce FX 5900 processor to distort performance in Futuremark's 3DMark 03 testing application. The company said drivers for the new Nvidia chip were altered to detect activity characteristic of a benchmark and adjust performance accordingly.

Recently, there have been questions and some confusion regarding 3DMark 03 results obtained with certain Nvidia products, Futuremark said in the statement. "We have now established that Nvidia's Detonator FX drivers contain certain detection mechanisms that cause an artificially high score when using 3DMark 03," the statement read.

Futuremark has released the version 330 patch for 3DMark 03, which prevents the Nvidia drivers from identifying the benchmark. By Futuremark's measure, the performance of the Nvidia GeForce FX 5900 Ultra drops 24 percent with the patch, compared with a drop of less than 2 percent with ATI's Radeon 9800 Pro with the latest ATI drivers.

A representative at Nvidia questioned the validity of Futuremark's conclusions. "Since Nvidia is not part of the Futuremark beta program (a program which costs of hundreds of thousands of dollars to participate in), we do not get a chance to work with Futuremark on writing the shaders like we would with a real applications developer," the representative said. "We don't know what they did, but it looks like they have intentionally tried to create a scenario that makes our products look bad."

Nvidia is amid a hard-fought battle with rival ATI Technologies to claim the performance lead in PC graphics processors. After years of Nvidia dominating both in market share and performance, ATI

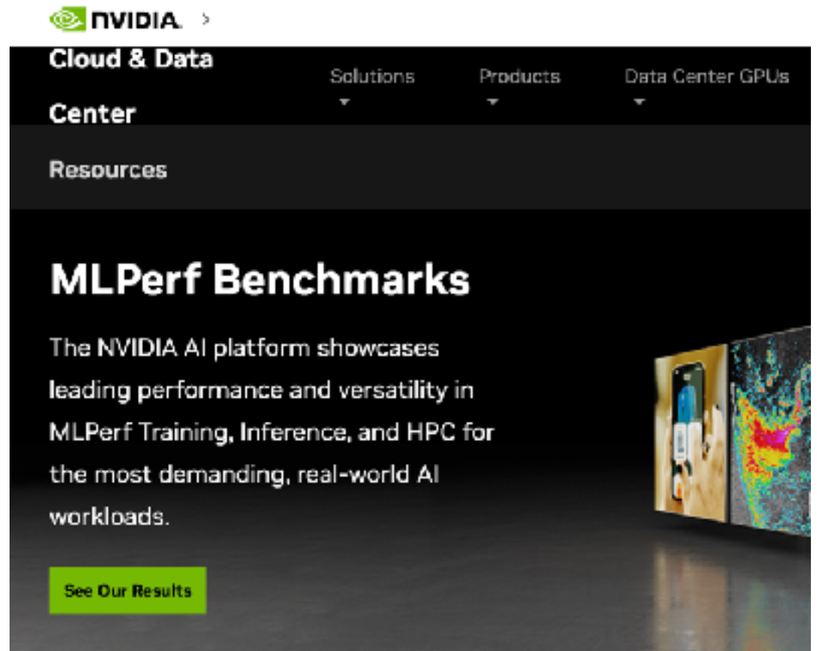
https://www.gamespot.com/articles/nvidia-accused-of-cheating-in-3dmark-03/1100-6028894/#:~:text=%22We%20have%20now%20established%20that,drivers%20from%20identifying%20the%20benchmark.

Latency v.s. Bandwidth/Throughput

- Latency — the amount of time to finish an operation
  - End-to-end execution time of “something”
  - Access time
  - Response time
- Throughput — the amount of work can be done within a given period of time (typically “something” per “timeframe” or the other way around)
  - Bandwidth (MB/Sec, GB/Sec, Mbps, Gbps)
  - IOPs (I/O operations per second)
  - FLOPs (Floating-point operations per second)
  - IPS (Inferences per second)
  - FPS (Frames per second)

The reason of “Benchmark Suites”

- Allowing people evaluate systems with exactly the same program and the same inputs and validate results from different machines
- Popular benchmark suites
  - SPEC — CPU benchmark
  - MLPerf — ML systems



Takeaways: What does “perfect” mean?

- Latency is the most fundamental performance metric
- Classic CPU performance equation —Instruction count (IC), cycles per instruction (CPI), cycle time (CT) define the latency of execution on CPUs
- Performance metrics without considering all three factors in the classic performance equation can mislead — anything throughput typically miss one of them

Announcement

- Reading quiz due this Thursday before the lecture — we will drop two of your least performing reading quizzes
- Check our website for slides, Gradescope for assignments, piazza for discussions
- Youtube channel for lecture recordings: https://www.youtube.com/c/ProfUsagi/playlists
- Don’t forget to check your access to escalab.org/datahub and piazza
- CSE142L lecture today 5p @ WLH 2205

Computer
Science &
Engineering

142

つづく